THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES
et de L'UNIVERSITÀ *DEGLI STUDI DI MILANO-BICOCCA*

École doctorale : LLSH - Langues, Littératures et Sciences Humaines
Spécialité : Sciences du langage Spécialité Informatique et sciences du langage
PhD in Education in the Contemporary Society
Unité de recherche : Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles (Université Grenoble-Alpes) et Department of Human Sciences for Education in the Contemporary Society (Università degli studi di Milano-Bicocca)

Outiller la description linguistique de la continuité référentielle dans un corpus d'écrits scolaires français et italien avec une perspective didactique

Developing tools for the linguistic description of referential continuity in a corpus of French and Italian school writing: a didactic perspective

Présentée par :

Martina BARLETTA

Direction de thèse :

Catherine BRISSAUD — **Directrice de thèse**
PROFESSEURE DES UNIVERSITES EMERITE, Université Grenoble Alpes
Federica DA MILANO — **Co-directrice de thèse**
PROFESSEURE, Università degli studi di Milano-Bicocca
Claude PONTON — **Co-encadrant de thèse**
MAITRE DE CONFERENCES, Université Grenoble Alpes

Rapporteurs :

Matteo VIALE
PROFESSEUR, Università di Bologna
Frédéric LANDRAGIN
DIRECTEUR DE RECHERCHE, CNRS

Thèse soutenue publiquement le 19 juin 2026, devant le jury composé de :

Olivier KRAIF, — **Président**
PROFESSEUR DES UNIVERSITES, Université Grenoble Alpes.
Catherine BRISSAUD, — **Directrice de thèse**
PROFESSEURE DES UNIVERSITES EMERITE, Université Grenoble Alpes
Federica DA MILANO, — **Co-directrice de thèse**
PROFESSEURE, Università degli studi di Milano-Bicocca
Matteo VIALE, — **Rapporteur**
PROFESSEUR, Università di Bologna
Frédéric LANDRAGIN, — **Rapporteur**
DIRECTEUR DE RECHECHE, CNRS
Luisa REVELLI, — **Examinatrice**
CHARGEE DE RECHERCHE, Università della Valle d'Aosta.

Invités :

Claude PONTON
MAITRE DE CONFERENCES, Université Grenoble Alpes

Titre : Outiller la description linguistique de la continuité référentielle dans un corpus d'écrits scolaires français et italien avec une perspective didactique

Mots clés : TAL, corpus d'écrits scolaires, coréférence, continuité référentielle, linguistique de corpus

Résumé :

Quelles sont les qualités qui permettent de distinguer un texte cohérent d'un texte peu ou pas cohérent ? Comment peut-on décrire ces caractéristiques de manière objectivable ? La compréhension et l'écriture d'un texte reposent sur plusieurs plans langagiers. Parmi les phénomènes contribuant à la cohésion et à la cohérence textuelle, les chaines de référence occupent une place centrale. Une chaine de référence est constituée d'au moins trois expressions linguistiques renvoyant à un même référent dans un texte donné. Si de nombreux travaux portent sur la coréférence et les anaphores dans des corpus d'écrits standard, peu de ressources ont été développées pour analyser ces phénomènes dans les écrits d'apprenants.

De plus, l'étude des écrits scolaires à l'aide de la linguistique outillée demeure complexe. Cette difficulté tient à l'éloignement des productions de la norme, qui complique l'annotation, ainsi qu'aux contraintes de constitution d'un tel corpus. Leur élaboration implique un processus long et coûteux, expliquant leur rareté. Ces ressources sont pourtant essentielles pour la description linguistique et les recherches en didactique de l'écriture. Dans ce contexte, plusieurs questions de recherche se posent : comment annoter des écrits scolaires afin de décrire la continuité référentielle ? Peut-on outiller cette description dans la continuité des travaux sur la coréférence ? Une fois cette description établie, quel est l'impact de la langue ou du niveau scolaire sur la réalisation des chaines de référence ?

Pour répondre à ces questions, cette thèse s'appuie sur le corpus *Scolinter*, un corpus longitudinal et comparable d'écrits scolaires recueillis en France, en Italie et en Espagne. Il regroupe environ 4 000 productions d'élèves du primaire, recueillies à partir des mêmes consignes, puis numérisées et traitées selon des protocoles communs. L'étude proprement dite a permis d'annoter un sous-corpus français et italien de *Scolinter* (148 textes en français, 150 en italien) produit par des élèves de 2e et 3e années des écoles primaires. L'annotation, fondée sur un guide inspiré des travaux existants, a été validée par une procédure d'adjudication permettant la constitution d'un *corpus gold* fiable. Les analyses combinent approches quantitative et qualitative, selon une perspective comparative et didactique.

Les résultats mettent en évidence des régularités dans l'organisation des chaines de continuité référentielle. Les textes s'articulent autour d'un nombre restreint de référents principaux, dont le maintien repose sur des stratégies situées sur un continuum entre gestion pronominale et gestion nominale. Les élèves privilégient des stratégies explicites afin de limiter les ambiguïtés dans les cas où plusieurs référents coexistent.

Les syntagmes nominaux introduisent le plus souvent les référents, tandis que les formes moins informatives (pronoms, anaphores zéro) assurent leur maintien. Les mentions courtes dominent et leur type grammatical varie selon la position dans la chaine. Si des différences liées à la langue apparaissent dans la distribution des types de mentions, les écarts entre niveaux scolaires restent

limités. Les variations entre langues tiennent en partie à la longueur et à la densité des productions, les textes italiens étant plus développés, ainsi qu'à des propriétés morphosyntaxiques propres aux deux langues.

Enfin, cette recherche souligne l'intérêt des chaines de référence pour analyser la cohésion des écrits d'élèves, en mettant en évidence des régularités, mais aussi des difficultés (ambiguïtés référentielles, ruptures de continuité), pouvant affecter l'interprétation, notamment en cas de compétition entre référents saillants. Ces résultats ouvrent des perspectives pour une didactique de l'écriture intégrant la continuité référentielle dès les premiers niveaux scolaires, ainsi que pour le développement d'une ressource annotée à plus grande échelle.

Title: Developing tools for the linguistic description of referential continuity in a corpus of French and Italian school writing: a didactic perspective

Keywords: NLP, school writings' corpora, coreference, referential continuity, corpus linguistics

Abstract:

What distinguishes a coherent text from a less coherent or incoherent one? How can these characteristics be described objectively? Text comprehension and production rely on several linguistic levels. Among the phenomena contributing to textual cohesion and coherence, reference chains play a central role. A reference chain consists of at least three linguistic expressions that denote the same referent within a given text. While much research has focused on coreference and anaphora in corpora of standard written texts, few resources have been developed to analyse these phenomena in learners' writing.

Furthermore, the study of school writing using computational linguistics remains challenging. This is due both to such texts deviating from the standard norm, which complicates annotation, and to the constraints involved in compiling such corpora. The development of such corpora is time-consuming and costly, which explains their scarcity. Nevertheless, these resources are essential for linguistic description and research in writing didactics. In this context, the following research questions arise: How can school writings be annotated to account for referential continuity? Can this description be achieved using tools based on existing literature on coreference? And what impact do language and educational level have on the construction of reference chains?

To address these questions, this thesis makes use of the *Scolinter* corpus, a longitudinal, comparable collection of school writing from France, Italy and Spain. It comprises around 4,000 pieces of writing by primary school pupils, collected using the same instructions, then digitised and processed according to shared protocols. The study involved annotating a French and Italian sub-corpus from *Scolinter* comprising 148 French and 150 Italian texts produced by pupils in Years 2 and 3 of primary school. The annotation, based on a guide inspired by previous works on coreference, underwent a validation procedure, resulting in a reliable gold-standard corpus. The analyses combine quantitative and qualitative approaches from a comparative and didactic perspective.

The results highlight recurring patterns in the structure of reference chains. The texts are structured around a limited number of referents, whose maintenance relies on strategies situated along a continuum ranging from pronominal to nominal expressions. Pupils tend to favour explicit strategies to reduce ambiguity when multiple referents coexist.

Referents are primarily indicated using noun phrases, whereas less informative forms (pronouns and zero anaphora) are used to maintain those referents. Short mentions predominate, and their grammatical type varies according to their position in the chain. While cross-linguistic differences emerge in the distribution of grammatical types, variation across school levels remains limited.

These differences are partly explained by text length and density, with Italian texts being more elaborate, as well as by language-specific morphosyntactic properties.

This research also highlights the relevance of reference chains for understanding cohesion in pupils' writing by identifying both regular patterns and difficulties, such as referential ambiguities and breaks in continuity, that could affect interpretation, especially in cases of competition between salient referents. These findings open new possibilities for teaching writing by integrating referential continuity from the earliest school years, as well as for developing larger-scale annotated resources.

Titolo: Sviluppo di strumenti per la descrizione linguistica computazionale della continuità referenziale in un corpus di produzioni scritte scolastiche in francese e italiano, in una prospettiva didattica

Parole chiave: Trattamento automatico del linguaggio, corpus scolastici, coreferenza, continuità referenziale, linguistica dei corpora

Abstract:

Quali sono le caratteristiche che consentono di distinguere un testo coerente da un testo poco coerente? È possibile descrivere tali caratteristiche in modo oggettivo? La comprensione e la scrittura di un testo si basano su diversi livelli linguistici. Tra i fenomeni che contribuiscono alla coesione e alla coerenza testuale, le catene di coreferenza hanno un posto centrale. Una catena di coreferenza è costituita da almeno tre espressioni linguistiche che rimandano a un unico referente in un testo. Sebbene numerosi studi abbiano approfondito la coreferenza e le anafore nei corpora di scritti standard, poche risorse sono state sviluppate per analizzare questi fenomeni su testi scritti da studenti di scuola primaria.

In effetti, lo studio della scrittura scolastica tramite la linguistica computazionale rimane complesso. Questa difficoltà è dovuta alla distanza delle produzioni dalla norma, che complica l'annotazione, nonché ai limiti legati alla costituzione di questi corpora. La loro elaborazione implica un processo lungo e costoso, che ne giustifica la rarità. Tuttavia, tali risorse sono essenziali per la descrizione linguistica e la ricerca in didattica della scrittura. In questo contesto, sorgono diverse questioni di ricerca: come annotare gli scritti scolastici per descrivere la continuità referenziale? È possibile fornire strumenti per questa descrizione in continuità con i lavori esistenti sulla coreferenza? Una volta effettuata questa descrizione, qual è l'impatto della lingua o del livello scolastico sulla realizzazione delle catene di coreferenza?

Per rispondere a queste domande, la presente tesi si avvale del corpus *Scolinter*, un corpus longitudinale e comparabile di scritti scolastici raccolti in Francia, Italia e Spagna. Questo corpus raggruppa circa 4000 produzioni di alunni di scuola primaria, raccolte sulla base delle stesse consegne, poi digitalizzate ed elaborate secondo protocolli di trattamento comuni (trascrizione e normalizzazione). Il nostro studio vero e proprio ha permesso di annotare un sotto-corpus francese e italiano di *Scolinter* (148 testi in francese, 150 in italiano) prodotto da alunni di classi seconde e terze di scuola primaria. L'annotazione, basata su una guida ispirata ai lavori esistenti, è stata convalidata da una procedura di aggiudicazione che ha permesso la costituzione di un corpus *gold* affidabile. Le analisi combinano approcci quantitativi e qualitativi, secondo una prospettiva comparativa e didattica.

I risultati evidenziano alcune regolarità nell'organizzazione delle catene di continuità referenziale. I testi si articolano attorno a un numero limitato di referenti principali, il cui mantenimento si basa su strategie che si collocano su un continuum tra gestione pronominale e gestione

nominale. Gli studenti privilegiano strategie esplicite per limitare le ambiguità nei casi in cui più referenti coesistono nel testo.

I sintagmi nominali introducono più spesso i referenti, mentre le forme meno informative (pronomi, anafore zero) ne assicurano il mantenimento. Prevalgono le menzioni brevi e il loro tipo grammaticale varia a seconda della posizione nella catena. Se nella distribuzione dei tipi di menzione emergono differenze legate alla lingua, le differenze tra i livelli scolastici rimangono limitate. Le variazioni tra le lingue dipendono in parte dalla lunghezza e dalla densità delle produzioni: i testi italiani, infatti, sono più articolati. Altre differenze derivano, in parte, da proprietà morfosintattiche proprie delle due lingue.

In conclusione, questa ricerca sottolinea l'importanza delle catene di coreferenza per analizzare la coesione dei testi scritti dagli studenti, mettendo in evidenza sia le regolarità che le difficoltà (ambiguità referenziali, interruzioni di continuità) che possono influire sull'interpretazione, in particolare in caso di competizione tra referenti salienti. Questi risultati aprono nuove prospettive per una didattica della scrittura che integri la continuità referenziale fin dai primi livelli scolastici, nonché per lo sviluppo di una risorsa annotata su scala più ampia.

Remerciements

Tant de thèses s'achèvent de manière plus ou moins tumultueuse. Cette thèse semble s'achever de manière plutôt pacifique, et ce n'est pas totalement grâce à moi. Dans mon cas, j'ai eu la chance d'avoir trois encadrants de thèse exceptionnels qui m'ont accompagnée tout au long de ce parcours. Ils m'ont non seulement enseigné le métier de chercheuse, mais ils m'ont également donné suffisamment confiance en moi pour que je puisse profiter du chemin. Je souhaite remercier chaleureusement Claude Ponton, Catherine Brissaud et Federica Da Milano pour tout le travail et l'attention qu'ils ont consacrés à mon parcours de recherche et à ce manuscrit, pour leur bienveillance et leur pédagogie.

Je pense que tous les directeurs de thèse devraient s'inspirer de votre exemple : vous m'avez permis de développer mes idées et mes intuitions de manière indépendante, sans jamais me laisser seule pour les exprimer. Vous m'avez enseigné à écrire un article (et écrire un bon article est difficile !), à mettre en place une méthodologie solide, à définir mon projet de recherche tel que je l'avais imaginé, à devenir une chercheuse en TAL comme en linguistique, et peut-être aussi un peu en didactique. Vous avez toujours pris le temps de discuter avec moi lorsque c'était nécessaire et vous avez su me rassurer sur le fait que mon travail avait encore un sens dans un monde qui devient de plus en plus difficile.

Ces remerciements ne seront jamais à la hauteur de vos efforts et de l'énorme travail que vous avez consacré à mon parcours pendant ces dernières années, depuis mon Master jusqu'aux derniers jours de finalisation de ce manuscrit. Je suis chanceuse de vous avoir eus comme directrices et directeur de thèse.

Je souhaite remercier les rapporteurs Monsieur Frédéric Landragin et Monsieur Matteo Viale, qui m'ont fait l'honneur d'accepter de relire ce travail, ainsi que Luisa Revelli et Olivier Kraif qui ont accepté de faire partie de mon jury en tant qu'examinateurs. Merci d'avoir accepté d'évaluer mon travail et de participer à cette étape importante de mon parcours académique et d'avoir contribué à la poursuite de mon travail avec vos précieuses contributions.

Merci également à Lilia Teruggi et Elisa Farina de l'Université Milano-Bicocca pour leur immense travail de recueil des textes du corpus Scolinter, sans lequel cette thèse n'existerait pas. Merci aussi de m'avoir aidée à entrer dans les données de l'italien avec les yeux des enseignants et d'avoir joué le jeu de l'annotation avec moi. Merci d'avoir rendu mes séjours à Milan si agréables et joyeux. Un remerciement également à Francesca Antonacci, ma tutrice italienne, pour sa bienveillance et son aide précieuse, ainsi qu'à mes collègues du Doctorat en Education dans la Société Contemporaine pour les échanges enrichissants. Merci en particulier à Angela pour m'avoir fait comprendre quel est le véritable apport du TAL aux sciences de l'éducation et pour m'avoir montré que tout est possible dans la recherche, tant qu'on a le courage de le faire. Un grand merci également à Rafaela Gutierrez Cáceres de l'Université d'Almería pour avoir géré le recueil et le traitement de la partie espagnole de ce corpus, ainsi que pour sa contribution aux réflexions sur l'adaptation de notre annotation à cette langue.

Merci à Paola Zancanaro, Léa Wichroff et Baptiste Delpech pour avoir contribué à l'annotation et aux discussions autour de mon corpus. Merci aussi aux étudiantes et étudiants du Master Sciences du Langage qui ont contribué à faire évoluer mon guide d'annotation au cours des années. Et encore une fois merci à Claire Wolfarth, dont le travail fondateur sur le corpus Scoledit m'a permis de poser le jalon d'une analyse outillée de ce corpus, et aux étudiantes et étudiants, aux vacataires et enseignants qui ont contribué dans les années au recueil et au traitement des textes de Scolinter.

Je tiens aussi à remercier les membres de mon Comité de Suivi Individuel, Saskia Mugnier et Thomas Lebarbé, qui ont pris le temps d'écouter mes craintes, les résultats de mes recherches, mes accomplissements et même mes échecs (car il y en a eu, bien sûr). Merci pour votre soutien tout au long de ce parcours et pour vos conseils bienveillants et éclairants. Vous avez su reconnaître mon évolution au cours de ces trois années ; merci de m'avoir aidée à la voir moi-même lorsque c'était difficile d'avancer.

Je souhaite également remercier tous les membres du laboratoire LIDILEM pour ces années magnifiques, pour leur bienveillance, leur disponibilité et leur présence tout au long de mon parcours. Je remercie notamment Olivier Kraif pour m'avoir suggéré de réaliser une des expériences présentées dans cette thèse, Aurélie Nardy et Isabelle Rousset pour leur aide en statistiques et en méthodologie de représentation des données, Rachel Gaubil pour avoir pris soin de mes données. Un énorme merci à tous les membres des axes 1 et 2 du labo. Merci aussi à Claire pour m'avoir permis de m'embarquer dans une nouvelle et incroyable aventure dans le projet MOCONDI, aventure qui démarrait lors que cette thèse s'achevait. Et si j'ai oublié quelqu'un, je vous dois un verre.

Un remerciement spécial aux magnifiques doctorants (ou jeunes docteurs) du laboratoire LIDILEM, entre autres Iris, Manon, Sonia, Ping Ping, Laurence, Timothée, Stella, Jonas, Solenne, Roxanne, Luca, Clémence, Véronique, et toutes celles avec qui j'ai pu collaborer, organiser des séminaires ou juste rigoler à la pause de midi. Merci de tous ces beaux moments, la thèse a été vraiment agréable aussi grâce à vous. Je souhaite remercier en particulier Triscia, qui a été là depuis mon tout premier jour à Grenoble. J'ai appris à travailler avec toi, et surtout, tu as été mon premier modèle de chercheuse ici. Merci pour ton support pendant ces 9 ans. Merci aussi à Monica Masperi pour m'avoir donné l'opportunité de travailler dans le projet Innovalangues depuis 2017, ce qui m'a permis d'arriver et de rester à Grenoble et de prendre goût au métier de la recherche, et à toute l'équipe d'Innovalangues que j'ai pu côtoyer, mes premiers collègues.

Je remercie les enseignants-chercheurs du CLLE pour leur invitation à Toulouse et pour les nombreux échanges autour de ma thèse, en particulier Mai Ho-Dac pour l'accueil chaleureux, les enseignements lors de l'école d'été AnnoDemo et les conseils précieux sur ma thèse.

Je remercie également mes amis de thèse, géographiquement proches ou plus éloignés, rencontrés lors d'écoles d'été ou de colloques. Merci à ceux qui ont relu mes écrits, à ceux qui m'ont apporté du café pendant les sessions posters, ou simplement à ceux qui ont été là avec moi, un moment ou tout au long de ce magnifique chemin : Julien, présent depuis le début, avec

qui nous avons affronté un sanglier à Marseille ; les amis montpelliérains Jean et Lara ; les amis du GETALP, en particulier Nils et Fabien pour leurs relectures de mon manuscrit et pour tant de soirées à décompresser ; Elena et Soline à Toulouse ; William pour les photos des poules ; Elisa Gugliotta pour les heures de discussion sur ma thèse et sur nos recherches, et tous les chercheurs et chercheuses qui ont pris le temps d'écouter ce que je fais, et parfois aussi de me répondre. Un remerciement particulier à Silvia Federzoni, copine de coréférence, avec qui j'ai tant échangé sur les indices de calcul des chaines. Certaines des réflexions que l'on retrouve dans ce manuscrit sont issues de nos discussions.

Un immense merci également à toutes les doctorantes et tous les doctorants, chercheurs et chercheuses rencontrés au fil de ces années, avec qui j'ai pu échanger sur des théories linguistiques, sur des choses étranges en informatique, sur des livres, pester contre les LLM, boire des verres, faire semblant de comprendre des trucs, angoisser et rire pendant ces magnifiques années de thèse.

Surtout, merci à Jesús, mon compagnon : je suis dans tes remerciements de thèse, maintenant c'est à toi d'être dans les miens. Je suis fière du chercheur que tu es devenu, et c'est aussi grâce à toi que cette thèse arrive à son terme. Merci d'avoir été là.

Un ringraziamento speciale alla mia famiglia tutta, soprattutto a papà, mamma e Serena e alle mie nonne, che mi sostengono da lontano da tanti anni e che sono comunque sempre vicini a me. Se sono riuscita a fare tutto questo, è perché ci siete sempre stati. Vi voglio bene. Grazie.

Table des matières

Remerciements ..... **xi**

Table des matières ..... **xv**

Glossaire ..... **xvii**

Exemples d'annotation en coréférence et en continuité référentielle ..... **xxi**

Introduction **1**

1. Le corpus Scolinter 3
2. Problématique et méthodologie 3
3. Plan de la thèse 5

Partie 1 – État de l'art **7**

Chapitre 1. *La référence dans les textes* *9*

1. Cohérence textuelle et référence 11
2. Types d'anaphore 12
3. Coréférence 22
4. Chaine de (co-)référence : définition formelle et indices de description 27
5. Continuité référentielle 41
6. Typologie des mentions dans les chaines de référence 48
7. Synthèse de chapitre : anaphore, coréférence, chaine de référence et continuité référentielle 55

Chapitre 2. *Corpus scolaires constitués en France et en Italie* *59*

1. Constitution et exploitation des corpus d'écrits scolaires 59
2. Bref historique des corpus d'écrits scolaires en France 61
3. Le projet E-Calm 63
4. Corpus scolaires constitués en Italie 72
5. Synthèse de chapitre : corpus scolaires en France et en Italie 83

Chapitre 3. *La coréférence en TAL et les corpus annotés en coréférence pour le TAL* *85*

1. La coréférence en Traitement Automatique des Langues 85
2. Corpus annotés en coréférence 91
3. Corpus annotés en coréférence en France 92
4. Synthèse de chapitre : coréférence en TAL et corpus annotés 98

Conclusion *101*

Partie 2 – Description du corpus et méthodologie d'annotation **107**

Introduction *109*

Chapitre 4. *Recueil et constitution d'un corpus longitudinal trilingue : contexte, difficultés et enjeux* *111*

1. Lire et écrire au CP 111
2. Corpus Scoledit 115
3. Corpus Scolinter 116
4. Protocole de recueil 117
5. Recueil en Italie et Espagne 118
6. Métadonnées associées aux productions 120
7. Traitement des données : du manuscrit à sa version numérique 121
8. Description du corpus recueilli 139
9. Corpus longitudinaux français et italien : caractérisation 142

Chapitre 5. *Annotation de corpus* *145*
1. Objectifs et description de la chaine de prétraitement 146
2. Méthodologie d'annotation et de sélection du corpus à annoter 155
3. Guide d'annotation 159
4. Le rôle de l'adjudication dans la création d'un corpus *gold* 163

Chapitre 6. *Campagnes d'annotation* *165*
1. Pré-campagnes : plusieurs allers-retours vers le guide et échanges entre enseignantes-chercheuses 167
2. Rodages : campagnes d'annotation en TP, ou le guide à l'épreuve des groupes d'annotateurs de Master 169
3. Première campagne d'annotation : validation méthodologique de la constitution d'un corpus *gold* 171
4. Deuxième campagne d'annotation : constitution d'un corpus *gold* par adjudication 177

Conclusion *183*

Partie 3 – Résultats **187**

Chapitre 7. *Approche quantitative de l'analyse des chaines de continuité référentielle* *189*
1. Description des indicateurs retenus et analyses quantitatives 191
2. Interdistance 200
3. Densité référentielle 204
4. Indice de stabilité ou d'instabilité 215
5. Distribution des catégories grammaticales de mentions 222
6. Conclusion 239

Chapitre 8. *Approche qualitative de l'analyse des chaines de continuité référentielle* *242*
1. Longueur des mentions 244
2. Composition des chaines 251
3. Nombre de personnages et modes de cohabitation des chaines 261
4. Différences de longueur des textes scolaires : entre choix pédagogiques et contraintes linguistiques. 289
5. Conclusion 296

Chapitre 9. *Conclusions* *299*
1. Quelles perspectives pour un travail multidisciplinaire ? 302
2. Perspectives en linguistique de corpus et en TAL 303
3. Perspectives en didactique 304

Références bibliographiques **309**

Liste des tableaux **323**

Liste des figures **325**

Annexes **329**

Annexe A. Consigne corpus Univers-ITA 329
Annexe B. Lettre aux directeurs d'écoles italiens 330
Annexe C. Description du protocole de recueil adopté par l'équipe italienne 331
Annexe D. Guide de transcription – corpus Scoledit et Scolinter 333
Annexe E. Guide de normalisation – corpus Scoledit et Scolinter 337
Annexe F. Test de transcription automatique par OCR 346
Annexe G. Guide d'annotation en continuité référentielle du corpus Scolinter 348
Annexe H. Statistiques 359
Annexe I. Séquences des trois premiers maillons 361
Annexe J. Scan issu du corpus Scolinter 363
Annexe K. Corpus des mentions longues 363
Annexe L. Corpus français annoté 365
Annexe M. Corpus italien annoté 403

Glossaire

Annotation

- Accord inter-annotateur (*inter-annotator agreement*) : Mesure du degré de cohérence entre plusieurs annotateurs appliquant le même schéma d'annotation à un corpus.
- Adjudication : Phase de résolution des divergences d'annotation entre annotateurs afin d'obtenir une version finale consensuelle de l'annotation.
- Annotation en partie du discours (POS tagging) : Processus consistant à attribuer à chaque token une catégorie grammaticale (nom, verbe, adjectif, etc.).
- Annotation syntaxique : Analyse de la structure grammaticale d'une phrase permettant d'identifier les relations syntaxiques entre les mots.
- Schéma d'annotation : Ensemble de catégories, règles et conventions définissant comment les données doivent être annotées dans un corpus.

Corpus : Ensemble structuré de données langagières, collectées ou rassemblées selon des critères définis, ou dans le cadre d'une situation de communication donnée. Ces textes (écrits ou oraux) sont généralement stockés dans un format exploitable informatiquement. Ils peuvent être enrichis de données supplémentaires, tels que des métadonnées ou des annotations linguistiques.

- Corpus annoté : Dans ce travail de thèse, nous faisons référence à travers ce terme à l'échantillon de 298 textes issus du corpus *Scolinter* que nous avons annoté en continuité référentielle.
- Corpus d'écrits scolaires : Il est constitué de productions écrites d'enfants en âge scolaire, habituellement recueillies dans le contexte d'une école donné et représentant des genres de textes scolaires. Il est utilisé pour analyser les pratiques d'écriture et leur évolution.
- Corpus longitudinal : Corpus dans lequel les données sont collectées à plusieurs moments dans le temps auprès des mêmes locuteurs ou apprenants, afin d'étudier l'évolution de leurs productions dans le temps.
- Corpus *Scoledit* : corpus français d'écrits scolaires longitudinal, recueillis du CP au CM1.
- Corpus *Scolinter* : corpus d'écrits scolaires longitudinal en français, italien et espagnol, comparable entre les trois langues. La partie longitudinale de *Scoledit* constitue la partie française de ce corpus.
- Métadonnées : Informations descriptives associées aux données linguistiques d'un corpus (par exemple, âge des scripteurs, niveau scolaire, langue, date de production ou contexte de collecte).
- Sous-corpus : Sous-ensemble d'un corpus plus large, constitué selon certains critères (par exemple langue, niveau scolaire ou type de tâche).

Format de données

- CoNLL-2012 : Format standard de représentation de corpus annotés en coréférence.
- CoNLL-U : Format standard de représentation de corpus annotés, notamment utilisé dans le cadre des *Universal Dependencies*, permettant d'encoder des informations morphologiques et syntaxiques.
- XML : Langage de balisage permettant de structurer et représenter des données annotées.

Interopérabilité : Capacité de différents systèmes, outils ou ressources numériques à échanger, utiliser et traiter des données selon des formats et des conventions compatibles, permettant leur exploitation conjointe.

Mot et token

- Mot : Unité lexicale abstraite du lexique d'une langue.
- Token : En linguistique de corpus et en TAL, un token correspond généralement à une unité séparée lors de la tokenisation. Elle peut correspondre au mot et à la ponctuation ou à des sous-unités de mot selon le type de tokenisation adopté. Dans le contexte de notre travail, elle correspond aux mots et à la ponctuation.
- Tokenisation **:** Processus consistant à segmenter un texte en unités élémentaires appelées *tokens* (mots, ponctuation ou autres segments selon les conventions adoptées).

Référence : Rapport entre une expression linguistique et un objet, entité ou notion qui se situent en dehors du langage, à savoir, dans le monde extralinguistique. De manière plus large, la référence permet de « fixer une entité dans la représentation mentale des états de choses dénotés par l'énoncé » (Charolles, 2002, p. 36).

- Anaphore : Relation référentielle dans laquelle une expression donnée, *l'anaphore*, renvoie à un référent déjà introduit dans le discours, qui constitue son *antécédent*. Elle se configure comme une relation asymétrique : l'expression anaphorique ne peut pas être interprétée sans son antécédent.
- Chaine de (co-)référence : Trois expressions ou plus qui renvoient au même référent dans un texte donné.
- Continuité référentielle : L'ensemble des procédés linguistiques qui assurent la *cohésion textuelle*. Parmi ces procédés, nous retrouvons aussi les *expressions référentielles*.
- Coréférence : Relation entre plusieurs expressions linguistiques qui renvoient au même *référent*. Elle se configure comme une relation symétrique : toutes les expressions peuvent être interprétées de manière indépendante les unes des autres.
- Entité (référent) : Objet, personne, événement ou concept auquel une *expression* linguistique (ou *référentielle*) renvoie dans le discours.
- Expression référentielle, maillon ou mention référentielle : Expression linguistique (mot ou syntagme) utilisée pour désigner un référent donné dans un texte.
- Stratégie de gestion référentielle : Ensemble des procédés linguistiques permettant d'introduire, maintenir ou réactiver un référent dans le discours.

Texte : Discours produit dans un contexte de communication donné. Un texte peut être oral ou écrit et constitue une unité linguistique cohérente analysable à différents niveaux (lexical, syntaxique, discursif, pragmatique).

- Contexte : Ensemble des éléments situationnels, sociaux, culturels et discursifs dans lesquels un texte est produit et interprété.
- Cotexte : Environnement linguistique immédiat d'un élément dans un texte, c'est-à-dire les mots, syntagmes ou phrases qui l'entourent.

Traitement automatique des langues (TAL) : Domaine interdisciplinaire visant à développer des méthodes informatiques pour analyser, comprendre ou générer des langues naturelles.

- Affinement (*fine tuning*) : Processus consistant à adapter un modèle préentraîné à une tâche spécifique en poursuivant l'entraînement sur un jeu de données plus ciblé.
- Campagne d'évaluation : Évaluation collective dans laquelle plusieurs systèmes sont testés sur un même jeu de données et selon un protocole commun.
- Entraînement (*training*) : Phase d'apprentissage automatique durant laquelle un modèle apprend à généraliser après observation d'une grande quantité de données. L'apprentissage peut être supervisé (les données sont annotées) ou non supervisé (les données ne sont pas annotées).
- Évaluation : Processus visant à mesurer la qualité d'un système de TAL sur une tâche donnée.
- LLM (*Large Language Model*) : Modèle de langage de grande taille entraîné sur de vastes quantités de données textuelles et capable d'effectuer différentes tâches linguistiques.
- Métrique : Indicateur calculé pour mesurer la performance d'un système (par exemple précision, rappel ou F-mesure). Différentes tâches en TAL prévoient des métriques ciblées à l'évaluation de la tâche.
- Tâche (*task*) : Objectif spécifique d'un système ou module de TAL (par exemple analyse syntaxique, classification de texte ou résolution de coréférence).

Transcription et normalisation

- Transcription : La transcription d'un texte manuscrit a comme objectif sa représentation dactylographique, qui le rend manipulable à l'aide d'outils informatiques et que, dans le cas des corpus scolaires, et selon les projets, peut représenter au plus près le texte d'origine dans ces caractéristiques génétiques et visuelles.
- Normalisation : Dans notre travail, la normalisation constitue une forme d'annotation déportée du texte d'origine. Il s'agit ici de fournir une version du texte qui respecte autant que possible l'original tout en se rapprochant des normes de la langue cible.

Niveaux scolaires en France et en Italie

Afin de permettre au lecteur de se repérer dans les différentes descriptions de corpus issus de niveaux scolaires différents et de deux pays, nous décrivons dans ce tableau le parallélisme entre les systèmes scolaires français et italien. Cette correspondance permet de comparer les productions écrites d'élèves à des stades similaires de leur scolarité.

Cycle école française	**Classe française**	**Cycle école italienne**	**Classe italienne**	**Âge moyen des élèves en fin d'année**
École primaire Cycle 2	CP	Scuola primaria	1e année	7
	CE1		2e année	8
	CE2		3e année	9
École primaire Cycle 3	CM1		4e année	10
	CM2		5e année	11
Collège	6e	Scuola secondaria di primo grado	*1° media*	12
	5e		*2° media*	13
	4e		*3° media*	14
	3e	Scuola secondaria di secondo grado	*1° liceo*	15
Lycée	Seconde		*2° liceo*	16
	Première		*3° liceo*	17
	Terminale		*4° liceo*	18
-	-		*5° liceo*	19

Cycle éducation supérieure française	**Classe française**	**Cycle éducation supérieure italienne**	**Classe italienne**
Licence	L1	Laurea Triennale	1 année
	L2		2 année
	L3		3 année
Master	M1	Laurea Magistrale	1 année
	M2		2 année

Exemples d'annotation en coréférence et en continuité référentielle

Ce manuscrit contient trois types différents d'exemples d'annotation : des exemples issus du corpus *Scolinter*, que nous avons annotés dans le cadre des campagnes d'annotation décrites ou annotés a posteriori pour illustrer des phénomènes particuliers ; des exemples issus de corpus annotés en coréférence ou d'autres sources (articles en linguistique ou en informatique sur les thématiques abordées dans cette thèse) ; et des exemples inventés, mais fonctionnels pour décrire un phénomène défini. Les exemples issus du corpus *Scolinter* proviennent des textes normalisés, sauf indication contraire.

Nous avons utilisé une représentation commune, déjà attestée dans la bibliographie en coréférence, où chaque mention dans le texte est délimitée à l'aide de crochets. Dans le cas d'exemples issus de notre corpus annoté, les entités représentées par les mentions sont indiquées en exposant, comme dans l'exemple suivant. Si plusieurs référents sont représentés dans une mention, ils sont tous indiqués en exposant, séparés par des virgules (surligné dans le texte).

> Il était une fois **[une sorcière]**^{witch} **[qui]**^{witch} était dans **[sa]**^{witch} maison. Une fois **[elle]**^{witch} a décidé d' aller mettre un sort au village donc **[elle]**^{witch} a pris **[son]**^{witch} balai magique et **[alla]**^{witch} mettre un sort au village et même sur le château de **[la princesse]**^{ext1}. Après **[elle]**^{witch} est rentrée chez **[elle]**^{witch} pour préparer de la potion magique pour en donner au village et au château pour qu' **[ils]**^{ext1,ext2} meurent **[tous]**^{ext1,ext2}. *(corpus français, CE1, 2035)*

En ce qui concerne les exemples tirés d'autres ressources, les chaines de coréférence ont été annotées de la même manière, mais les référents ont été indiqués à l'aide de lettres différentes si nécessaire, sinon seul le marquage des expressions référentielles est présent. Si la provenance d'un exemple n'est pas indiquée, celui-ci est un exemple inventé pour illustrer un phénomène déterminé.

> Il était une fois **[une petite fille de village, la plus éveillée qu'on eût su voir]**^{a} : **[[sa]**^{a} **mère]**^{b} en était folle, et **[[sa]**^{a} **mère-grand]**^{c} plus folle encore. **[Cette bonne femme]**^{c} **[lui]**^{a} fit faire un petit chaperon rouge, qui **[lui]**^{a} seyait si bien, que partout on **[l']**^{a}appelait **[le petit Chaperon rouge]**^{a}.

Dans le cas des textes issus du corpus *Scolinter*, certaines caractéristiques issues de la normalisation des textes sont également indiquées. L'indication de titre, entourée d'accolades, a notamment été conservée, ainsi que les mots omis et intégrés par les chercheurs, surlignés dans le texte. Les paragraphes sont également reconstitués par des retours à la ligne.

La nomenclature des exemples est systématiquement composée de trois éléments : la langue du corpus de provenance, le niveau scolaire et le numéro d'identification de l'élève scripteur du texte.

Introduction

Combien de fois, à l'école, les annotations inscrites en rouge sur une copie ont-elles rappelé l'importance de la cohérence d'un texte, parfois de manière péjorative : « Texte peu cohérent », « Attention à la cohérence ici ! », « Ce texte n'a ni queue ni tête » ? Pourtant, si la cohérence constitue un critère récurrent d'évaluation des productions écrites, les enseignements explicites portant sur la cohérence et la cohésion textuelles demeurent relativement rares.

Les notions de cohésion et de cohérence sont pourtant étroitement liées entre elles et à la réussite d'un texte écrit. Ces deux propriétés caractérisent un texte proprement dit et le distinguent d'une simple séquence de phrases : « comme tout tas de mots ne donne pas une phrase, tout tas de phrases ne forme pas un texte. À l'échelle du texte ainsi que sur le plan de la phrase, il existe donc des critères efficients de bonne formation instituant une **norme minimale de composition textuelle** » (Charolles, 1978, p. 8, en gras dans le texte original). Ces normes minimales, implicitement reconnues par les membres d'une communauté linguistique, permettent ainsi au lecteur ou lectrice de comprendre un texte conformément à l'intention de son auteur ou autrice. Si la cohérence désigne la propriété interprétative globale d'un texte, la cohésion, quant à elle, se réalise à travers l'utilisation de marques linguistiques explicites, telles que les connecteurs ou les reprises anaphoriques. Ces marques contribuent ainsi à la cohérence textuelle.

Cependant, tout objet didactique doit être décrit dans sa réalité avant de pouvoir être enseigné. La description d'un phénomène dans son contexte réel de production constitue un enjeu essentiel non seulement pour la linguistique, mais dans le cas des corpus d'écrits représentant des phases d'apprentissage, aussi pour la didactique de la langue écrite à l'école.

Si Charolles (1978) décrivait déjà les manières d'enfreindre les normes de cohésion et de cohérence ainsi que les réactions des enseignants face à ces écarts à la norme, Garcia-Debanc et son équipe ont ensuite proposé d'effectuer une cartographie des formes langagières qui contribuent à la cohérence textuelle, à travers l'annotation et l'étude des chaines de continuité référentielle dans un corpus d'écrits scolaires (Garcia-Debanc et al., 2021). L'annotation puis l'analyse de la manière de convoquer et maintenir les référents tout au long d'un texte deviennent ainsi le champ privilégié pour l'étude du développement de la cohérence et de la cohésion textuelle lors de l'apprentissage de l'écriture.

C'est dans cette perspective que se positionne notre travail, à savoir, contribuer à la cartographie des formes qui assurent la continuité référentielle et éclairer la manière dont les expressions qui la constituent se configurent dans des productions écrites d'élèves d'école primaire. Notre recherche interroge plus précisément la gestion des expressions référentielles désignant des personnages animés dans des textes narratifs : comment ces expressions s'articulent-elles au sein du texte, en interaction avec les autres éléments langagiers et entre elles ? Dans quelle mesure les choix linguistiques opérés garantissent-ils l'interprétabilité du récit et l'alignement entre intention narrative et compréhension du lecteur ? Enfin, quel lien peut-on établir entre la gestion

référentielle et la capacité à produire un texte narratif conforme à la norme de la langue d'une part et au genre narratif de l'autre ?

Ces interrogations ouvrent un ensemble de questions complémentaires. Elles supposent d'identifier, à partir d'une description linguistique, les régularités et les variations observables dans les productions des élèves. Une telle description ne constitue toutefois pas seulement une fin en soi, mais vise également à nourrir une réflexion à portée didactique. En effet, analyser la gestion des chaines référentielles devrait permettre de déterminer les compétences déjà maîtrisées par les élèves, les stratégies qu'ils mobilisent pour gérer la présence de différents référents dans le contexte de la production d'un texte narratif, mais aussi les zones de fragilité susceptibles de faire obstacle à l'écriture et à la compréhension du texte. La description linguistique devient ainsi le point d'appui d'une analyse des acquis en termes de construction de la cohérence à une étape donnée, mais aussi des besoins d'apprentissage, permettant d'identifier ce qui doit faire l'objet d'activités didactiques spécifiques, explicitement centrées sur la gestion de la continuité référentielle.

Répondre à ces questionnements implique d'abord de choisir un cadre théorique apte à rendre compte des phénomènes observés. Dans le cas de notre travail, cela demande aussi de l'opérationnaliser dans une démarche d'annotation efficace et comparable entre deux langues, en l'occurrence le français et l'italien. Ces annotations devraient ainsi permettre d'articuler des analyses quantitatives issues des méthodes de la linguistique de corpus et du Traitement Automatique des Langues (TAL) avec une analyse qualitative qui porte sur les dimensions narratives et didactiques des textes.

Le travail mené dans cette thèse tente d'apporter des éléments de réponse à ces différentes questions. Il montre que les cadres théoriques existants pour l'étude de la continuité référentielle sont applicables aux écrits d'élèves et que les indicateurs proposés dans la littérature linguistique et en TAL permettent, pour l'essentiel, de décrire ces productions. Certaines catégories d'indicateurs nécessitent toutefois des ajustements ou des précisions lorsqu'elles sont confrontées à des textes produits par des élèves en cours d'apprentissage, ou dans une langue autre que le français. Par ailleurs, si les analyses quantitatives fournissent des tendances générales, elles ne suffisent pas à rendre compte de la complexité des stratégies textuelles mises en œuvre ; celles-ci exigent un examen qualitatif, ce que nous avons réalisé à travers des analyses qui portent à la fois sur le contenu des chaines et sur leur articulation dans la linéarité du texte et avec son contenu narratif.

Les descriptions fondées sur de tels corpus permettent d'éclairer ainsi le développement des compétences rédactionnelles des élèves et d'ajuster les attentes didactiques selon les niveaux scolaires (Garcia-Debanc et al., 2021). Elles peuvent également contribuer à identifier des leviers pédagogiques pour l'enseignement de l'écriture.

Cette recherche repose ainsi sur une approche fondée sur les corpus : dans cette perspective, et en continuité avec des travaux antérieurs menés sur les écrits scolaires, nous abordons l'étude des phénomènes de cohésion et de cohérence textuelles sous un angle à la fois longitudinal et

contrastif, à travers l'annotation de la continuité référentielle dans le corpus *Scolinter* (Farina & Teruggi, 2021; Ponton et al., 2021).

1. Le corpus Scolinter

Le corpus *Scolinter* représente l'objet d'annotation et d'étude de ce travail. Il a été constitué à partir d'un même protocole de recueil et de constitution, déjà expérimenté dans le contexte scolaire français avec la création du corpus *Scoledit* (Ponton et al., 2019; Wolfarth, Brissaud, & Ponton, 2018; Wolfarth, 2019). À partir du constat du manque de corpus scolaires longitudinaux et accessibles en Italie et en Espagne, des équipes de chercheurs des trois pays se sont fédérés pour recueillir un grand corpus longitudinal et comparable en français, italien et espagnol.

Ce corpus a été constitué dans le cadre d'une collaboration entre l'Université Grenoble Alpes (laboratoire LIDILEM), l'Université de Milano-Bicocca et l'Université d'Almería. Constitué selon un protocole commun dans les trois pays, le corpus rassemble des textes narratifs produits à partir d'une même consigne intégrant quatre images représentant quatre personnages différents (un loup, une sorcière, un chat et un robot), et invite les élèves à mobiliser un ou deux de ces personnages dans leur production. La présence de cette consigne commune permet d'assurer la comparabilité des productions entre langues et de mener des études longitudinales. Le recueil, initié en France dans le cadre du projet *Scoledit* puis étendu à l'Italie et à l'Espagne, vise à combler le manque de ressources longitudinales numérisées dans ces contextes éducatifs. Les textes ont été collectés auprès d'élèves du CP au CM2, selon des protocoles de recueil standardisés, puis numérisés, transcrits, normalisés et anonymisés pour en autoriser la diffusion et l'exploitation à travers des outils de TAL et de linguistique outillée. Le corpus est accompagné de métadonnées permettant la caractérisation des productions et le suivi longitudinal des élèves. Malgré les perturbations liées à la pandémie de Covid-19, qui ont affecté partiellement la continuité du recueil en Italie et en Espagne, le corpus *Scolinter* constitue aujourd'hui une ressource de grande ampleur, qui offre un cadre particulièrement adapté à l'étude contrastive du développement des compétences d'écriture à l'école primaire dans les trois pays européens.

Si des travaux ont déjà été consacrés à des niveaux tels que la segmentation en mot (Ponton et al., 2021), l'orthographe (Ponton, Brissaud, et al., 2022) et la morphographie verbale (Wolfarth, Ponton, et al., 2018), aucun travail n'a étudié à ce jour les niveaux de cohérence et cohésion textuelles sur ce corpus, ni proposé d'analyse comparable sur deux langues différentes. Par ailleurs, en France, peu de travaux existent qui décrivent de manière véritablement empirique les formes qui manifestent ces niveaux textuels, à l'exception des travaux menés sur le corpus *RésolCo* (Garcia-Debanc et al., 2017, 2021). En Italie, les recherches sur ce phénomène sont encore plus limitées et reposent le plus souvent sur des échantillons restreints (Perugini & Zucchini, 2025). C'est à ce manque de description empirique des phénomènes de cohésion et cohérence textuelles que notre travail tente de répondre.

2. Problématique et méthodologie

Si la cohérence et la cohésion textuelles constituent des propriétés primordiales de la textualité et font l'objet d'évaluations scolaires fréquentes, les études systématiques de ces niveaux fondées

sur des corpus longitudinaux annotés demeurent encore limitées, en particulier dans une perspective contrastive. Par ailleurs, les outils de TAL, bien que prometteurs dans leur utilisation en linguistique outillée, présentent encore des limites lorsqu'ils sont appliqués à des corpus caractérisés par une forte variabilité et un éloignement de la norme.

Dans la continuité des travaux de Wolfarth (2019), qui interrogent l'apport du TAL à la constitution et à l'exploitation de corpus d'écrits scolaires, cette thèse se concentre plus spécifiquement sur l'analyse de la continuité référentielle dans ce type de données. La continuité référentielle est définie ici comme la manière d'établir et ensuite maintenir l'existence des référents donnés au fil du texte. Elle renvoie ainsi à la capacité du scripteur à assurer l'accessibilité des entités évoquées et à garantir la cohérence interprétative du texte pour le lecteur.

Dans cette perspective, plusieurs questions se posent : comment décrire la continuité référentielle de manière systématique dans des écrits d'élèves ? Peut-on appliquer les mêmes critères et méthodes déjà utilisés pour décrire le même phénomène dans la continuité des travaux en coréférence pour apprécier leurs caractéristiques ? Dans quelle mesure les méthodes et les outils issus du TAL peuvent-ils être mobilisés pour permettre l'annotation et l'analyse de ce phénomène dans un corpus à forte variabilité ?

Dans cette mesure, il nous a semblé nécessaire de contribuer à la cartographie de la continuité référentielle dans des écrits d'élèves en constituant une ressource annotée permettant de la décrire à partir d'annotations.

Pour étudier ce phénomène, nous avons choisi d'exploiter les possibilités offertes par l'annotation de corpus et par l'outillage issu du TAL. Les analyses qui en découlent s'appuient sur l'état de l'art existant en matière de corpus annotés en coréférence en TAL ainsi que sur des travaux consacrés aux chaines de continuité référentielle dans les textes narratifs d'élèves, notamment ceux menés dans le cadre du corpus *RésolCo*.

Afin d'atteindre cet objectif, nous avons conçu un guide d'annotation applicable aux deux langues étudiées, le français et l'italien. Ce guide a été élaboré de manière itérative à travers plusieurs phases d'annotation portant sur des sous-ensembles du corpus. Sa conception et sa mise à l'épreuve ont impliqué plusieurs enseignants-chercheurs appartenant aux trois groupes de recherche qui ont participé à la constitution du corpus ainsi que quatre annotateurs experts (étudiants en linguistique et TAL de niveau Master 1 et 2 et une doctorante). Les ajustements successifs opérés à partir des premières campagnes ont conduit à la stabilisation progressive du guide.

Plusieurs campagnes d'annotation ont ainsi été menées, aboutissant à une dernière phase fondée sur une version stabilisée du guide. Les annotations produites ont ensuite fait l'objet d'une étape d'adjudication, permettant de constituer une annotation de référence (*gold standard*) sur 298 textes (148 français et 150 italiens) de niveaux CE1 et CE2. Ces deux niveaux scolaires ont été retenus pour analyse, car les textes étaient déjà disponibles pour une exploitation automatique, particulièrement dans le cas du corpus italien, dont le traitement était encore en cours lors du début de ce travail de thèse.

Les analyses menées à partir de cette ressource se déclinent en deux volets complémentaires. D'une part, une analyse quantitative des chaines de continuité référentielle, qui s'appuie sur des critères descriptifs issus des travaux précédents de description des chaines de référence. D'autre part, une analyse qualitative menée sur une partie restreinte des textes permet d'approfondir l'interprétation de certains indicateurs quantitatifs afin d'éclairer les stratégies de construction des chaines effectivement mises en œuvre et de les recontextualiser dans le cadre de la linéarité du texte d'origine. L'articulation entre analyse quantitative et analyse qualitative vise ainsi à rendre compte à la fois des tendances générales et de la complexité des configurations textuelles observées.

3. Plan de la thèse

La thèse s'organise en trois parties : la première est consacrée à l'élaboration du cadre théorique, la deuxième à la description de la méthodologique d'annotation adoptée, et la troisième à l'analyse des annotations du point de vue quantitatif puis qualitatif.

La première partie est dédiée à la description du cadre théorique de notre travail, à la croisée de la linguistique, de la linguistique de corpus, en particulier des corpus d'écrits scolaires, et du TAL. Le premier chapitre pose les fondements conceptuels nécessaires à l'analyse des phénomènes référentiels dans les textes. Il revient notamment sur les notions d'*anaphore*, *coréférence*, *chaines de référence* et *continuité référentielle*, en mettant en évidence leurs différences et similarités, ainsi que certains enjeux théoriques qui leur sont associés. Ces éléments permettent d'aboutir, entre autres, à une taxonomie des types de mentions, qui servira de base aux analyses ultérieures, et d'assurer la continuité entre le cadre théorique retenu en linguistique et son opérationnalisation méthodologique. Le deuxième chapitre propose un état des lieux sur les travaux en constitution de corpus d'écrits scolaires, en particulier dans les contextes français et italien. Il met en évidence l'apport des corpus longitudinaux à l'étude du développement des compétences d'écriture et situe le corpus *Scolinter* dans cette tradition de recherche empirique. Ce positionnement permet de souligner à la fois la continuité avec des travaux antérieurs et la nouveauté que représente l'apport de cette recherche dans le domaine. Le troisième chapitre s'inscrit au croisement entre linguistique et Traitement Automatique des Langues. Il examine les principales modélisations de la tâche de résolution de la coréférence en TAL ainsi que les corpus annotés existants, notamment en français. Cette mise en perspective permet de situer notre travail dans un espace interdisciplinaire, où les exigences de la description linguistique rencontrent celles de la modélisation computationnelle.

La deuxième partie de la thèse est consacrée à la méthodologie adoptée et à la constitution de la ressource annotée. Le quatrième chapitre décrit le contexte de la recherche ainsi que les étapes de constitution du corpus *Scolinter*. Il définit les protocoles de recueil des données, en particulier dans le cas des recueils italien et espagnol, qui sont présentés ici pour la première fois. En outre, il détaille les traitements appliqués afin de rendre ce corpus d'écrits d'élèves exploitable à travers des outils de TAL et un état de lieux des textes recueillis est proposé. Le cinquième chapitre présente les choix méthodologiques relatifs à l'annotation. Il décrit d'autres étapes de traitement automatique que nous avons dû appliquer pour autoriser l'annotation de nos textes à l'aide de

l'outillage retenu. Il expose aussi les résultats obtenus grâce à une expérimentation portant sur le traitement des mots omis par les élèves à l'aide des LLM (*large language models*). Après avoir explicité les caractéristiques de la méthodologie d'annotation agile que nous avons adoptée dans notre travail, il retrace la démarche adoptée pour élaborer un guide applicable aux deux langues étudiées, en explicitant les principes théoriques retenus ainsi que les ajustements successifs ayant conduit à la stabilisation du protocole. Le sixième chapitre prolonge cette réflexion en détaillant les campagnes d'annotation que nous avons menées, les procédures d'adjudication et l'analyse des désaccords inter-annotateurs effectuées. Il met en lumière les critères qui ont permis d'aboutir à une annotation de référence et discute les enjeux de fiabilité et de reproductibilité associés à la constitution d'un tel corpus.

La troisième partie présente les résultats issus de l'exploitation du corpus annoté. Le septième chapitre est consacré à l'analyse quantitative des chaines de continuité référentielle : il définit les indicateurs retenus, leurs opérationnalisations, il compare les données selon les niveaux scolaires et les langues étudiées, et met en évidence les principales tendances observées. Le huitième chapitre propose une analyse qualitative qui met en relation les enchainements référentiels et l'organisation narrative des récits. Il vise à identifier les stratégies privilégiées par les élèves dans la gestion des personnages et à dégager, à partir de ces observations, des pistes de réflexion pour la didactique de l'écriture.

Enfin, le neuvième chapitre présente les conclusions générales de la recherche et ouvre des perspectives scientifiques et méthodologiques. Il souligne les prolongements possibles du corpus annoté et les apports du travail à l'articulation entre linguistique, didactique et traitement automatique des langues, en inscrivant la recherche dans une dynamique interdisciplinaire appelée à se développer au-delà du cadre doctoral.

Partie 1 – État de l'art

Chapitre 1. *La référence dans les textes* *9*
1. Cohérence textuelle et référence 11
2. Types d'anaphore 12
3. Coréférence 22
4. Chaine de (co-)référence : définition formelle et indices de description 27
5. Continuité référentielle 41
6. Typologie des mentions dans les chaines de référence 48
7. Synthèse de chapitre : anaphore, coréférence, chaine de référence et continuité référentielle 55

Chapitre 2. *Corpus scolaires constitués en France et en Italie* *59*
1. Constitution et exploitation des corpus d'écrits scolaires 59
2. Bref historique des corpus d'écrits scolaires en France 61
3. Le projet E-Calm 63
4. Corpus scolaires constitués en Italie 72
5. Synthèse de chapitre : corpus scolaires en France et en Italie 83

Chapitre 3. *La coréférence en TAL et les corpus annotés en coréférence pour le TAL* *85*
1. La coréférence en Traitement Automatique des Langues 85
2. Corpus annotés en coréférence 91
3. Corpus annotés en coréférence en France 92
4. Synthèse de chapitre : coréférence en TAL et corpus annotés 98

Conclusion *101*

Chapitre 1. La référence dans les textes

Dans ce chapitre, nous définissons la notion clé de ce travail de thèse, à savoir *la référence dans le texte*. Elle est capitale pour comprendre nos choix d'annotation en continuité référentielle, les enjeux qui en découlent dans le cadre de l'annotation d'un corpus d'écrits scolaires, ainsi que les difficultés rencontrées au cours de ce processus d'annotation.

Selon Schnedecker (1997) (qui, dans son ouvrage, cite à son tour une célèbre vignette de la BD Snoopy[1]), pour avoir une bonne narration, a minima compréhensible, il faut relier les différents fils ouverts dans un discours, où chaque fil représente, par exemple, un personnage donné. « Il ne suffit pas de les multiplier [les personnages ou les situations] et de les additionner pour que le texte « prenne ». Comme le dit Snoopy, *in fine*, il faut relier les différents fils. » (Schnedecker, 2021, p. 1).

Mais pour introduire ces personnages, pour établir les protagonistes d'une intrigue ou tout simplement poser le décor d'une narration, il faut tout d'abord introduire les entités qui constituent ces personnages et ce décor, à travers des expressions susceptibles de les indiquer. Par la suite, ces personnages vont probablement contribuer à faire avancer l'intrigue : il est donc utile d'avoir la capacité à les désigner plusieurs fois dans le même texte. C'est intéressant de noter ici que l'on peut pointer le même référent à travers des expressions différentes, par exemple, pour ne pas avoir à répéter que le Petit Chaperon Rouge est une petite fille qui porte un bonnet rouge à chaque fois qu'on a besoin de la mentionner. Cela se fait pour une question d'économie dans le récit, pour éviter des répétitions qui risquent d'alourdir le récit, ennuyer le lecteur ou, pire, lui faire perdre le fil de la narration.

Dans l'extrait du conte « Le Petit Chaperon Rouge » de l'exemple (1), nous avons annoté toutes les expressions faisant référence aux personnages animés présents dans le conte, autrement dit, aux personnages principaux de l'histoire.

(1) Il était une fois **[une petite fille de village]**^{a}, la plus éveillée qu'on eût su voir : **[[sa]**^{a} **mère]**^{b} en était folle, et **[[sa]**^{a} **mère-grand]**^{c} plus folle encore. **[Cette bonne femme]**^{c} **[lui]**^{a} fit faire un petit chaperon rouge, qui **[lui]**^{a} seyait si bien, que partout on **[l']**^{a}appelait [**le petit Chaperon rouge]**^{a}. Un jour, **[[sa]**^{a} **mère]**^{b}, ayant fait des galettes, **[lui]**^{a} dit : « **[Va]**^{a} voir comment se porte**[[ta]**^{a} **mère-grand]**^{c} ; car on **[m']**^{b}a dit qu'**[elle]**^{c} était malade : **[porte]**^{a}**-[lui]**^{c} une galette et ce petit pot de beurre. » **[Le petit Chaperon rouge]**^{a} partit aussitôt pour aller chez **[[sa]**^{a} **mère-grand]**^{c}, **[qui]**^{c} demeurait dans un autre village. En passant dans un bois, **[elle]**^{a} rencontra [**compère le Loup]**^{d}, **[qui]**^{d} eut bien envie de **[la]**^{a} manger ; mais **[il]**^{d} n'osa, à cause de **[quelques bûcherons]**^{e} **[qui]**^{e} étaient dans la forêt. *(Perrault, 1826)*

[1] Dans cette vignette, le célèbre chien (et écrivain) écrit les phrases suivantes dans des vignettes successives : « C'était une sombre nuit d'orage », « Soudain, il y eut une détonation. Une porte claqua. La bonne hurla », « Soudain, un navire de pirates surgit à l'horizon ! », « Tandis que des millions de gens mouraient de faim, le roi vivait dans le luxe », « Cependant, dans une petite ferme du Kansas vivait un jeune garçon ». La bande se conclut avec Snoopy annonçant que « Dans le chapitre deux, je relie les différents fils » (Schulz, 2010).

Une fois un personnage nommé pour la première fois dans un texte, d'autres syntagmes ou mots peuvent être utilisés pour le désigner, qui peuvent reprendre de manière identique la première expression utilisée, fournir plus de détails par rapport au personnage ou encore renvoyer vers lui sans pour autant utiliser son nom de manière explicite. Par exemple, dans l'incipit de l'histoire du Petit Chaperon Rouge, dès les premières lignes du conte, nous pouvons voir comment les différents personnages qui vont ensuite participer à l'intrigue de l'histoire sont introduits à l'aide de syntagmes nominaux. Dans le cas de la protagoniste, elle est introduite dans le texte au moyen d'un syntagme nominal indéfini, « une petite fille de village », complété par une apposition « la plus éveillée qu'on eût su voir » ; elle est identifiée comme « le petit Chaperon rouge » seulement lors de sa deuxième parution dans le texte, suite à la description du couvre-chef qu'elle porte. D'autres personnages sont introduits à travers des syntagmes qui déterminent le lien entre eux et la protagoniste, « *sa mère* » et « *sa mère-grand* » ; ou encore le personnage du méchant loup qui est introduit à l'aide d'un syntagme, à cheval entre syntagme nominal déterminé et nom propre, « *compère le Loup* ».

Ces expressions peuvent servir à rajouter davantage de détails dans l'image mentale qu'on se fait d'un référent au fur et à mesure qu'on avance dans la lecture d'un texte, ou simplement rappeler sa présence dans le texte pour permettre à l'intrigue d'avancer. Dans ce travail, nous nous intéressons à la description de ces *expressions* que l'on peut qualifier de *référentielles*, aux liens qu'elles entretiennent entre elles pour désigner une même entité, pour contribuer à la cohérence et à la cohésion d'un texte. Ces expressions sont aussi appelées *mentions* ou, par le biais d'une expression métaphorique, *maillons* d'une chaine de (co-)référence. Par la suite, nous allons décrire plus en détail quelles sont les catégories syntaxiques qui entrent dans la notion d'expression référentielle, et nous verrons également dans quelle mesure la catégorie syntaxique est influencée par des caractéristiques textuelles ou par l'articulation de la chaine dans le texte. Nous utilisons dans ce travail les trois termes *expression référentielle*, *mention* et *maillon* comme synonymes désignant cette même notion.

Ces ensembles d'expressions, quand elles pointent vers le même référent, peuvent être ainsi définis à l'aide des notions d'*anaphore*, de *coréférence* ou encore de *continuité référentielle*. Nous allons par la suite décrire ces trois notions, les distinctions épistémiques entre elles, ainsi que la manière dont elles ont été abordées en linguistique. Nous nous positionnerons par rapport à cet état de l'art et expliquerons les choix effectués dans notre projet d'annotation de la continuité référentielle dans un corpus d'écrits scolaires.

À la suite de la description de l'état de l'art, nous illustrons ces différentes notions à l'aide d'exemples inventés, d'exemples issus de corpus variés ainsi que des exemples tirés du corpus objet de notre recherche. Nous n'avons toutefois pas pu nous limiter à l'utilisation exclusive de notre corpus annoté, dans la mesure où celui-ci ne permet pas d'illustrer l'ensemble des typologies de phénomènes décrites dans la littérature. Notre corpus étant représentatif du genre narratif et écrit par des élèves d'école primaire, certains types d'anaphores ne sont pas présents dans ces écrits, ou elles demeurent très rares. Si les exemples utilisés sont tirés d'un corpus différent de *Scolinter*, cela a été indiqué. Tous les exemples tirés du corpus *Scolinter* sont issus des versions orthographiquement normalisées des textes (cf. Chapitre 4, §7.3 pour la définition

de la normalisation dans le cadre de notre travail). Ils sont également accompagnés de la mention du corpus d'origine (français ou italien), le niveau scolaire et le numéro identifiant de l'élève auteur du texte.

1. Cohérence textuelle et référence

Avant d'aborder la notion de référence, il est nécessaire de définir la cohérence textuelle. Cette idée de relier les différents fils dans un discours, que nous avons évoquée dans l'introduction de ce chapitre, est étroitement liée à la notion de cohérence textuelle. Comme le définit Charolles, « la cohérence n'est pas une qualité que pourraient avoir ou non les discours, elle est constitutive de l'idée même de discours » (Charolles, 2011, p. 2). Également appelée *pertinence* par certains auteurs, comme Charolles, elle constitue ainsi une des conditions *sine qua non* de la textualité. Elle renvoie, ensemble avec la notion de *cohésion*, à la capacité d'un discours de former un ensemble interprétable, où les énoncés s'organisent en une progression significative (Charolles, 1988b).

Selon Charolles, quatre types de dispositifs concourent à la mise en place de la cohérence textuelle : les connecteurs (par exemple *et*, *mais*, *donc*…) ; les expressions introductrices de cadres de discours délimitant des domaines ou cadres temporels ou spatiaux, (il était une fois *dans un bois sombre*) ; les marques configurationnelles qui délimitent le texte en unités selon un certain critère dispositionnel (les alinéas ou les organisateurs métadiscursifs) ; enfin, les différentes formes d'anaphore et en particulier celles qui constituent les chaines de référence (Charolles, 1995, p. 124). C'est à cette dernière catégorie que nous nous intéressons dans le cadre de notre travail.

Les expressions repérées et analysées dans le texte (1) servent à conférer à une entité le statut de *référent* dans l'univers d'un texte donné. Autrement dit, elles permettent qu'une entité soit désignée, rendue présente et identifiable dans le monde discursif qui est représenté dans le texte en question. La référence peut ainsi être définie comme « le rapport entre une expression linguistique et un objet qui se situe en dehors du langage, à savoir, dans le monde extralinguistique » (Kim, 2023, p. 14). Toutefois, même si cette définition est opérationnelle, elle simplifie quelque peu la portée réelle des expressions référentielles. En effet, ces expressions linguistiques peuvent renvoyer aussi bien à des entités du monde dit « réel » qu'à des entités fictives, conceptuelles ou imaginaires, telles que des personnages de fiction, des représentations mentales, des idées ou des notions abstraites. Il est donc nécessaire d'« étendre l'accès de la référence à des mondes ou univers autres que le seul monde réel et donc [d']élargir la définition initiale pour y intégrer ces mondes possibles » (Kleiber, 1997, p. 11). Nous reprenons ainsi la définition proposée par Charolles, qui permet de dépasser cette difficulté : « La seule condition requise pour qu'une expression réfère est donc qu'elle permette de fixer une entité dans la représentation mentale des états de choses dénotés par l'énoncé » (Charolles, 2002, p. 36). En d'autres termes, la référence se mesure à la capacité du discours à construire un espace mental cohérent où ces objets, ou référents, prennent une consistance suffisante pour être mobilisés

dans le discours. Cette approche permet ainsi d'intégrer pleinement les référents « imaginaires », tels que les personnages de fiction, au champ de l'analyse linguistique.

La construction de la référence repose ainsi sur des mécanismes discursifs, tels que l'anaphore et la coréférence, dont l'analyse permettra de préciser comment les référents sont introduits et maintenus dans le discours. Nous examinerons dans la suite de ce chapitre les différentes formes que peuvent prendre ces mécanismes et leurs modalités de fonctionnement.

2. Types d'anaphore

Parmi les éléments de langage qui contribuent à la réalisation de la cohérence d'un texte, nous avons déjà mentionné les *anaphores*. Celles-ci représentent certains des éléments qui permettent d'indiquer un référent donné dans un texte (cf. supra). Nous allons en fournir une définition, avant d'illustrer différents types d'anaphores décrites en littérature, à travers des exemples issus de la littérature ou de corpus variés.

L'origine du mot *anaphore* se trouve dans la langue grecque. Composé du préfixe *ana-* (en arrière) et du verbe -*pherein* (porter), ou bien du verbe αναφέρω, qui veut dire « porter en haut » (Merlin, 2016), il était déjà utilisé dans la Grèce antique. Dans son sens rhétorique, le terme désigne la figure consistant en la répétition de mots ou de syntagmes en début de vers ou de phrases contiguës. Cette figure rhétorique a pour but de mettre l'accent sur un élément du discours et de le renforcer dans l'esprit du lecteur. Un exemple d'anaphore dans la littérature italienne se trouve dans une partie de la « Divine Comédie » de Dante Alighieri, « L'Enfer ». Au début du Chant III, Dante lit l'inscription présente sur la porte de l'Enfer.

(2) **Per me si va** nella città dolente, **Per me si va** nell'eterno dolore, **Per me si va** tra la perduta gente (…)	Par moi l'on va tout droit dans la cité des pleurs, Par moi l'on va souffrir d'éternelles douleurs, Par moi l'on va parmi les maudits, race abjecte. (…)
(Dante Alighieri, Enfer, Chant III, 1-3)	(Traduction de Villain-Lami, (1867))

La répétition de « per me si va » (par moi l'on va) au début de chaque vers renforce le sens d'inexorabilité de ce qui attend Dante une fois rentré dans l'Enfer. Ce procédé littéraire partage avec l'anaphore en linguistique, dite aussi *anaphore grammaticale*, la construction de sens ou de cohésion à travers la répétition : des mots ou syntagmes dans le cas du procédé littéraire, des expressions liées à un référent donné dans le cas de l'anaphore grammaticale.

Le terme d'anaphore ou anaphorique en linguistique est probablement dérivé du nom de la figure rhétorique. Il est souvent utilisé pour indiquer « toute reprise d'un élément antérieur dans un texte » (Riegel et al., 2004, p. 610), ou encore des « expressions dont l'interprétation dépend des objets introduits dans l'univers du discours U, étant mentionnés ou seulement inférés »[2]

[2] Dans le texte d'origine, « (…) expressions whose interpretation depends on objects introduced in universe of discours U either by virtue of being explicitly mentioned or by being inferred ».

(Poesio, 2016, p. 38, notre traduction). L'anaphore linguistique reprend elle aussi un élément précédent du texte, qui, dans ce cas-là, est une entité et non plus une portion plus ou moins réduite de texte.

Plus précisément, l'anaphore se définit comme une expression linguistique dont l'interprétation dépend crucialement d'une autre expression présente dans le contexte textuel antérieur, qualifiée d'*antécédent* ou de *source sémantique* (Kleiber et al., 1993). Riegel et al. (2018, p. 1029) précisent également qu'« une expression est anaphorique si son interprétation référentielle dépend nécessairement d'une autre expression qui figure dans le contexte antérieur ».

Les anaphores sont habituellement prises en considération en tant que paires antécédent-anaphore[3]. Ce critère numérique (deux expressions reliées entre elles), avec d'autres critères qui seront approfondis par la suite, le distingue des *chaines de coréférence* (auxquelles nous avons dédié la partie 4 de ce chapitre).

L'extrait de texte (3) représente un exemple d'anaphore issu du corpus français. Ici sont annotées seulement les expressions qui font référence au référent *chat*.

(3) (…) La sorcière a **[un chat]**^{cat} très gentil. Mais quand **[il]**^{cat} trouve quelqu'un **[il]**^{cat} le griffe. (…) *(corpus français, CE1, 1518)*.

Dans cet exemple, le syntagme « un chat » présent dans la première phrase du texte, est repris dans la deuxième phrase à travers l'utilisation du pronom personnel « il », ce qui permet de remettre le chat en question au centre de l'action. Ici, le syntagme nominal « un chat » constitue l'*antécédent* du pronom personnel, ainsi appelé puisqu'il le précède dans l'ordre de la phrase. Sa présence est nécessaire pour donner une signification au pronom « il » dans le contexte de la phrase qui le contient. En outre, le deuxième pronom « il » peut être également relié de manière non ambigüe à l'antécédent *chat*, grâce à la saillance[4] de l'antécédent dans la phrase et à l'absence de référents en compétition. Sans pouvoir la relier à un antécédent, une expression de type anaphorique reste vide d'interprétation ou de sens. Son interprétation dépend de la disponibilité d'une autre expression dans le contexte, expression qui possède un sens autonome du point de vue de la référence. Pour cette raison, dans l'interprétation de Milner (1982, p. 33), la relation anaphorique entre deux expressions linguistiques est définie comme *asymétrique* et non autonome car elle « suppose la mise en relation d'une expression non autonome du point de vue de la référence et d'une expression référentielle susceptible de la « saturer » (…) » (Corblin, 1995; Schnedecker, 2019, p. 11).

L'anaphore est aussi caractérisée par sa *non-transitivité* : par exemple, si un référent A est anaphorique d'une expression B et B est anaphorique d'expression C, A et C ne seront pas nécessairement anaphoriques à cause de la relation qu'ils ont avec B. Selon ce principe, « Un

[3] Plusieurs termes ont été adoptés en littérature pour indiquer la paire : le premier élément de la paire est appelé *interprétant* (Ducrot & Todorov, 1979), *référent* (Dubois, 1965), *contrôleur de l'anaphorique*, *source* (Corblin, 1995), *terme anaphorique* (Savin, 2025) et le deuxième élément *anaphorique* ou *anaphorisé* (Savin, 2025). Nous préférons ici nous tenir à la terminologie qui fait le plus de consensus, *antécédent* et *anaphore*.

[4] Pour la notion de saillance, voir §5.2 de ce chapitre.

pronom anaphorique ne peut jamais fonctionner comme le premier terme d'une relation d'anaphore. » (Milner, 1982, p. 33 ; Schnedecker, 2019). Par exemple, dans le cas de l'exemple suivant, nous pouvons voir comment le deuxième pronom « *il* » est nécessairement anaphorique avec l'expression « *le chat* », et le pronom qui le précède ne peut pas être son antécédent.



Figure 1. Illustration des liens anaphoriques dans l'exemple de texte (3).

En outre, l'anaphore est aussi *non-réflexive* : une expression ne peut pas être l'antécédent d'une anaphore qui le reprend elle-même, c'est-à-dire c'est inconcevable qu'« un terme soit anaphorisant ou anaphorisé de lui-même » (Milner, 1982, p. 33; Perdicoyianni-Paleologou, 2001).

Dans le cas où l'expression autonome du point de vue référentiel est précédée par une expression dépendante, c'est-à-dire dans les cas où l'anaphore précède son antécédent, cela est définie comme une *cataphore.* De manière similaire à *anaphore*, l'origine de ce mot vient de la langue grecque : formé par le linguiste Karl Bühler à partir du préfixe *cata-* (en bas) et du verbe -*pherein* (porter). Dans les exemples (4) et (5), le nouveau personnage est introduit à travers un pronom indéfini suivi par un syntagme nominal défini en (4) et par un syntagme nominal indéfini et descriptif qui introduit le personnage avant de le nommer de manière explicite dans l'exemple (5).

(4) (…) Un jour **[quelqu'un]**^{witch} vint, c'est **[la sorcière]**^{witch}. (…) *(corpus français, CE1, 114)*

(5) (…) Un jour Flique rentre à la décharge, BOUM ! Il fonça dans **[quelque chose solide et résonnante]**^{robot}. C'était **[un autre robot]**^{robot}. (…) *(corpus français, CE1, 2886)*

Il existe différents types d'anaphore, classées selon plusieurs critères et de manière différente selon les auteurs[5]. Certains relèvent de la forme de surface, comme les anaphores fidèles (reprise lexicale directe) et infidèles (reformulation). D'autres dépendent de critères sémantiques, comme les anaphores conceptuelles, associatives, génériques ou collectives, qui impliquent un lien inférentiel ou métonymique avec l'antécédent. Enfin, des critères syntaxiques permettent d'identifier des formes comme les anaphores pronominales, adverbiales, verbales ou les anaphores zéro, dont le lien avec l'antécédent repose sur des structures grammaticales spécifiques. Nous adoptons pour décrire cette taxonomie les points de vue de plusieurs auteurs et autrices qui nous semblent décrire de manière cohérente entre eux les possibles variétés

[5] Voir par exemple la distinction entre anaphores par le genre lexical et anaphores par le genre discursif de Lerat reportés par Schnedecker (1997, p. 26-27).

d'anaphore existantes. Entre autres, nous allons nous inspirer des taxonomies proposées par Schnedecker (2021)[6].

2.1. Anaphore pronominale

Le pronom représente la catégorie la plus fréquente au sein des anaphores, ou « l'expression référentielle préférée pour le maintien de la référence » (Bonnemaison, 2018, p. 100). Des pronoms tels que les pronoms personnels, relatifs, et les démonstratifs peuvent être anaphoriques, et renvoyer à un antécédent donné.

Selon la théorie de l'accessibilité d'Ariel (1990), le pronom personnel représente un marqueur de haute accessibilité du référent. Il indique souvent un référent qui est fortement activé dans la représentation mentale du texte, par saillance ou proximité dans le texte. C'est-à-dire que nous retrouvons le pronom comme élément anaphorique dans ces cas où l'antécédent permet de récupérer facilement le référent indiqué, soit parce qu'il a été nommé peu avant dans le texte, soit parce qu'il s'agit d'un référent principal et donc saillant dans le texte. Cet aspect est illustré dans l'exemple (6), où le pronom personnel reprend le groupe nominal de la phrase principale, situé à quatre mots de distance.

(6) (…) Mais un jour **[une vilaine sorcière]**$^{\textbf{witch}}$ voulait un chat, et **[elle]**$^{\textbf{witch}}$ savait où en trouver un. (…) *(corpus français, CE1, 841)*

Le pronom représente l'exemple le plus emblématique de l'expression anaphorique car il a forcément besoin d'un antécédent « plein » pour identifier son antécédent. En revanche, certains pronoms personnels (première et deuxième personne du singulier en particulier) sont considérés comme déictiques de manière inhérente : leur interprétation est intrinsèquement liée à la situation d'énonciation, et ils ne peuvent pas posséder d'interprétation figée en dehors de ce contexte d'énonciation. En particulier, les pronoms « je » et « tu » changent de référent à chaque changement de locuteur. Les similitudes et différences entre anaphores et déictiques seront approfondies dans la partie 2.10 de ce chapitre.

2.2. Anaphore nominale (fidèle ou infidèle)

L'anaphore nominale fidèle, ou anaphore par répétition (Schnedecker, 1997, 2021, p. 15), est caractérisée par la répétition du nom contenu dans l'antécédent quasiment à l'identique, ce qui permet d'établir le lien entre antécédent et anaphorique de manière non ambigüe. Cette reprise est souvent accompagnée par la variation du déterminant qui introduit le syntagme[7].

(7) Il était une fois [**un chat**]^{cat} et [**une sorcière**]^{witch}. [**La sorcière**]^{witch} voulait transformer [**le chat**]^{cat} en cheval (…). *(corpus français, CE1, 1830)*

[6] Schnedecker fait référence à plusieurs auteurs en tant qu'inspiration pour cette taxonomie dans son ouvrage (2021, p. 14), dont Cornish, Lerat, Apothéloz, Charolles, Kleiber et Reichler-Béguelin.

[7] Nous reviendrons sur le rôle des déterminants dans la classification des types d'anaphore et d'expressions coréférentielles dans la section 6 de ce chapitre.

Dans les cas des anaphores relatives aux entités *chat* et *sorcière* de l'exemple (7), l'antécédent est constitué par le nom précédé d'un déterminant indéfini, alors que la deuxième expression est introduite par un déterminant défini.

L'anaphore infidèle ou anaphore par le genre lexical (Schnedecker, 1997, 2021, p. 15), quant à elle, présente de la variation lexicale entre antécédent et anaphorique, qui conservent la relation référentielle. Celle-là peut être établie à travers des relations comme la synonymie, l'hyperonymie ou l'hyponymie.

(8) Ma sœur a trouvé [**un chat**] dans la rue. [**Ce félin**] est très mignon.

Dans l'exemple (8), un lien anaphorique est créé entre *chat* et *félin* car le deuxième terme et hyperonyme de *chat* (représente la catégorie à laquelle il appartient).

2.3. Anaphore présuppositionnelle, pragmatique ou anaphore par le genre discursif

Cette anaphore permet d'apporter des informations nouvelles par rapport à leur source, notamment en exprimant « un point de vue sur le référent » (Schnedecker, 2021, p. 15). Ces nouvelles informations peuvent être apportées à l'aide d'une métaphore ou de noms de qualité, comme illustré dans l'exemple (9).

(9) Un jour **[un robot]**^{robot} se promenait dans la forêt, il rencontra un chat. **[Le chaton]**^{cat} était abandonné alors Alexe dit » *je vais l'appeler Croquette »* et là, **[la boule de poil]**^{cat} sourit. Alors **[le robot]**^{robot} ramena Croquette à sa maison. Le lendemain ils s'amusèrent, c'était devenu les meilleurs amis. En vrai le chat savait parler, mais les deux amis ne le disaient à personne sinon c'était la fin. Mais **[le tas de ferrailles]**^{robot} avait une chose qu'il ne disait à personne, c'était qu'il avait une base secrète avec des vaisseaux, des gardes et des ninjas. (…) *(corpus français, CE2, 2973*)[8]

Dans cet exemple, le référent *chat* est par la suite identifié par la description métaphorique « boule de poil », et le robot par le syntagme nominal « le tas de ferrailles ». Ces descriptions permettent d'enrichir ainsi la représentation mentale des référents dans l'univers narratif du texte en question.

Comme nous le verrons dans les définitions suivantes, bien que toutes les expressions de coréférence soient anaphoriques, toutes les relations anaphoriques ne sont pas nécessairement coréférentes. En effet, il n'existe pas toujours de relations d'identité entre référents au sein de certains types d'anaphore. Les types d'anaphore qui suivent sont ceux traditionnellement considérés comme non-coréférentielles.

[8] Dans ce texte nous avons seulement annoté les expressions fonctionnelles à l'explication du type d'anaphore analysée, bien que des chaines de références complètes soient présentes déjà dans ce court extrait.

2.4. Anaphore conceptuelle ou résomptive

L'anaphore résomptive, également appelé *anadeixis* de discours (Cornish, 1999; Gardelle et al., 2019) permet de résumer ou de recatégoriser un contenu propositionnel ou un ensemble d'informations présenté antérieurement dans le texte. Son antécédent n'est pas un syntagme ou une expression référentielle précise, mais une portion plus large du texte, qu'elle résume et reprend (Schnedecker, 2021, p. 25), d'où son nom d'*anaphore résomptive* (Maillard, 1974). Autrement dit, « les anaphores résomptives (…) sont un moyen de reprendre et d'encapsuler l'information déjà connue tout en indiquant que le référent est nouveau » (Schnedecker, 2021, p. 26). Elles opèrent ainsi une recatégorisation discursive : l'expression anaphorique construit un nouveau référent en s'appuyant sur le contenu du co-texte précédent. C'est le cas dans l'exemple (10), où le syntagme « cette tranquille obstination » crée un nouveau référent à partir du contenu du cotexte, et fournit un nouveau point de vue interprétatif sur le cinéaste japonais protagoniste de l'extrait en question.

(10) Kurosawa, 84 ans cette année, n'est toujours pas décidé à céder la place: sur le mode de l'ironie, Madadayo (1993), son dernier film en date, matérialise **[cette tranquille obstination]** sur la personne d'un autre créateur; l'écrivain Hyakken Uchida. *(Apothéloz, 1995, son ex. 40, tiré de Schnedecker, 2021, p. 26)*

Cette anaphore peut s'exprimer, entre autres, à travers l'utilisation d'un syntagme nominal souvent accompagné d'un déterminant démonstratif ; elle peut aussi s'exprimer à l'aide d'un démonstratif seul, comme montré dans l'exemple (11), où « cela » indique l'idée de la sorcière de transformer les loups en chat, décrite dans la phrase précédente et liée par une conjonction de coordination à la phrase contenant le démonstratif.

(11) Il était un fois une sorcière qui vivait au fin fond des bois et qui n'aimait pas les loups et un jour elle eut *une idée de les transformer, mais en chat* et **[cela]** n'avait pas grand succès (…). *(corpus français, CM1, 3071)*

L'anaphore résomptive et la deixis textuelle sont deux phénomènes linguistiques qui présentent des similitudes formelles et fonctionnelles, au point où le terme d'anaphore est souvent utilisé aussi pour indiquer des usages qui relèvent plutôt du champ des déictiques. Nous reviendrons de manière approfondie sur les critères permettant de différencier ces deux phénomènes dans la section 2.10, dédiée à la description des déictiques.

2.5. Anaphore associative

L'anaphore associative constitue une anaphore indirecte, où une expression anaphorique introduit un nouveau référent via le référent de l'expression antécédent (Kleiber, 1999), c'est-à-dire, dans ces anaphores « l'antécédent active un élément du cadre ou du scénario (au sens cognitif) qui lui est typiquement associé, et permet donc une référence à cet élément par un anaphorique. » (Gardelle et al., 2019, paragr. 14). Par exemple, dans l'exemple (12) proposé par Kleiber et Vassiliadou (2007), « l'église » et « le village » qui constitue l'antécédent de « village » sont considérés comme anaphoriques en vertu de ce que Schnedecker appelle une « relation de localisation stéréotypique fonctionnelle » (2021, p. 17). Le discours suggère implicitement, via le

contexte, qu'un village comprend probablement une église et que l'église mentionnée se trouve dans ce village.

(12) Nous entrâmes dans **[un village]. [L'église]** était située sur une butte.
(Kleiber & Vassiliadou, 2007, p. 155)

Une autre type d'anaphore qui fait partie des anaphores associatives est celle de type membre/collection (Schnedecker, 2021, p. 17). Dans l'exemple (13), le rapport établi entre l'antécédent « famille » et l'anaphorique « maman » se base sur le rapport implicite d'appartenance de l'anaphorique au groupe indiqué par l'antécédent.

(13) Il était une fois **[une famille heureuse]**. **[La maman]** était très sympathique.

2.6. Anaphore lexicale ou nominale

L'anaphore lexicale fait partie des anaphores indirectes, tout comme les anaphores associatives et les anaphores « avec *ils* grégaire ou collectif, ainsi *Paul a été à l'hôpital ; ils l'ont soigné énergiquement* (cité par Kleiber, 2001 : 49) ; les pluriels génériques textuels indirects, ainsi *Paul a acheté une Toyota parce qu'elles sont robustes* (cité par Kleiber, 2001 : 49) ; et des anaphores indirectes autres » (Gardelle et al., 2019, paragr. 14).

(14) Nous avons loué **[un kayak]** et nos amis ont amené **[le leur]**.

Dans l'exemple (14), il est possible d'interpréter l'anaphore « le leur » en lien avec l'antécédent « un kayak », bien qu'il n'y ait pas de relation d'identité entre les deux kayaks en question. Ces types de pronoms peuvent aussi désigner les phrases de type *paycheck*. Cette terminologie provient du célèbre exemple de Kartunnen « L'homme qui donne **son salaire** à sa femme est plus sage que celui qui **le** donne à sa maitresse »[9], cité par Gardelle et al. (2019). Dans le cas de ces phrases, antécédent et anaphore sont bien anaphoriques, mais ne sont pas pourtant coréférentiels, car il est « question de scénarios parallèles » (Gardelle et al., 2019, paragr. 13) ou, tout simplement, question de référence à deux objets différents dans la réalité qui sont pourtant indiqués par le même token.

2.7. Anaphore générique

L'anaphore générique « mets aux prises un référent particulier dont la mention antérieure est à l'origine d'une référence à la classe à laquelle appartient l'antécédent » (Kleiber, 1990, p. 253). Ce type d'anaphore peut souvent impliquer une relation de non-identité (et donc une relation anaphorique et non pas coréférentielle) entre antécédent et anaphorique, comme dans le cas de l'exemple (15), illustré dans Kleiber (1990, p. 253) :

(15) J'ai acheté **[une Toyota]**, parce que **[ces voitures]** sont robustes.
(Kleiber, 1990, p. 253)

Ce type d'anaphore met en œuvre un passage entre un référent déterminé et sa classe d'appartenance d'origine, comme dans le cas de l'exemple (16), tiré du corpus annoté.

[9] Dans le texte d'origine: « The man who gives his paycheck to his wife is wiser than the man who gives it to his mistress. »

(16) Mais un jour une vilaine sorcière voulait **[un chat]**^{cat}, et elle savait où en trouver **[un]**^{cat}, elle allait au magasin des animaux, mais il n'y avait plus **[de chat noir]**^{cat} (…).*(corpus français, CE1, 841)*[10]

La première occurrence, « un chat » introduit un référent spécifique non encore identifié dans le discours. Les occurrences suivantes, « un » et « chat noir », ne renvoient pas à un individu déjà établi mais réactivent la classe ou le type d'entité introduit par l'antécédent. Cette configuration confirme l'observation de Kleiber selon laquelle certaines anaphores reposent sur une relation de non-identité entre antécédent et anaphorique, la relation étant médiatisée par la classe d'appartenance du référent initial.

2.8. Anaphore possessive

Les déterminants possessifs sont considérés comme anaphoriques et représentatifs de la présence du possesseur. Comme remarqué par Kleiber (2011), l'anaphore indirecte définie de type associatif (« le tronc » dans l'exemple (17)) présent des parallélismes avec l'anaphore possessive, car les deux types d'anaphore indiquent un rapport d'appartenance entre antécédent-possesseur et anaphorique-possédé. Dans les deux exemples parallèles (17) et (18), les entités E1 qui indique l'arbre et E2 qui indique le tronc peuvent donner lieu à un syntagme binominal de la forme *Le N2 de SN1* (« le tronc | de l'arbre »), et cela peut être exprimé de manière similaire à travers les deux types d'anaphore.

(17) Il s'abrita sous [un vieux tilleul]. **[Le tronc]** était tout craquelé.
(18) Il s'abrita sous [un vieux tilleul]. **[Son tronc]** était tout craquelé.
(Kleiber, 2011, p. 3)

La différence entre les deux types d'anaphore réside, selon Kleiber, dans la modalité de donation différente d'un référent, c'est-à-dire la modalité d'accès au référent par le lecteur. Dans le cas de l'anaphore associative en (17), c'est l'anaphorique qui est mis au premier plan et donc son antécédent passe en deuxième plan. Dans le cas d'expression de la même relation à travers le possessif comme dans l'exemple (18), celui-ci détermine un effet de saillance majeure sur le possesseur, et donc un retour de l'antécédent au premier plan au niveau topical, « le possessif par l'anaphore coréférentielle réalisée active explicitement ce lien topical. » (Kleiber, 2011, p. 11). Ce type d'anaphore est très présent dans notre corpus annoté, surtout dans la représentation de possession des référents souvent « possédés » type *chat* ou *robot*, souvent associés à des propriétaires comme dans les exemples (19) et (20) (le binôme sorcière/chat étant un de plus fréquent au sein de notre corpus, comme nous le verrons dans le chapitre dédié aux résultats, cf. Chapitre 7, §1.3).

(19) Dans une forêt il y avait une maison où habitaient **[une sorcière]**^{witch} et **[son]**^{witch} chat. **[Son]** ^{witch} chat était noir avec des yeux perçants qui faisaient peur à un chien. *(corpus français, CE2,855)*

[10] Cet exemple, bien qu'utile à illustrer un exemple d'anaphore générique, ne corresponds pas à l'annotation que nous avons effectuée sur ce même texte. Les textes annotés sont disponibles en annexe.

(20) Il était une fois **[un loup]**[wolf] et **[son]**[wolf] fidèle serviteur le robot. *(corpus français, CE2, 563)*

2.9. Anaphore zéro

Dans les langues dites *à sujet zéro*, comme c'est le cas pour l'italien ou l'espagnol, la présence d'un référent donné peut se réaliser à l'aide d'un verbe qui n'est pas accompagné d'un sujet explicite au sein de la même phrase. Ce verbe renvoie néanmoins à un référent représenté précédemment dans le texte. Ce phénomène, nommé *zero anaphora* en anglais, désigne la reprise d'un référent sans employer de groupe nominal ou de pronom précis. La référence au sujet est alors implicite, souvent explicitée par le contexte physique ou discursif. Selon la langue, la morphologie du verbe dans la même phrase contenant ce « gap » peut également jouer un rôle pour indiquer ce référent absent. Dans l'échelle d'accessibilité d'Ariel (cf. ce chapitre, §5.1), l'anaphore zéro fait partie des expressions à *haute accessibilité*, c'est-à-dire que le référent qu'elle reprend doit être extrêmement saillant et accessible dans le texte. Bien que ce phénomène soit courant en italien, on le retrouve également en français, mais seulement dans certains cas précis. Dans le manuel d'annotation du corpus *DEMOCRAT*, certaines constructions sont ainsi analysées comme comportant une anaphore zéro. C'est le cas notamment dans les structures coordonnées où le sujet de la seconde proposition n'est pas réalisé (21), ainsi que dans les phrases impératives où le sujet est implicite (22).

(21) Pierre boit et **[fume]**.

(22) **[Ferme]** la porte.

(Exemples tirés du manuel d'annotation du corpus DEMOCRAT)

Dans le cas de l'italien, la question de l'annotation de l'anaphore zéro demeure plus complexe. Comme illustré dans les exemples suivants tirés du corpus annoté, l'ellipse du sujet peut apparaitre dans des phrases coordonnées, mais également dans des propositions relatives, y compris lorsqu'elles sont fautives, comme dans le cas de « che lo chiamò Sregino » de l'exemple (23) (**qu'il l'appellait Sregino*). Elle peut aussi se rencontrer dans d'autres types de phrases complexes, telles que les propositions coordonnées ou subordonnées, comme la proposition temporelle subordonnée « mentre volava » (pendant qu'elle volait) dans l'exemple (24).

(23) C'era una volta **[una strega]**[witch] **[che]**[witch] si chiama [Marina]witch, **[era]**[witch] sempre sola e **[andò]**[witch] a prendere un gatto che lo **[chiamò]**[witch] Sregino e lo **[portò]**[witch] a casa **[sua]**[witch] **[della strega Marina]**[witch] e gli **[dà]**[witch] da mangiare il pesce e un po' d'acqua (…). *(corpus italien,CE1, 1495)*

(24) C'era una volta **[una strega]**[witch] **[che]**[witch] viveva in una casetta di montagna e **[viveva]**[witch] sola. Un giorno mentre **[volava]**[witch] sulla **[sua]**[witch] scopa **[vide]**[witch] un gattino (…). *(corpus italien, CE1, 15)*[11]

[11] Dans les exemples 23 et 24, seules les chaines qui font référence à la *sorcière* sont annotées dans le texte.

2.10. Anaphore et deixis

Parmi les phénomènes anaphoriques, rentrent aussi d'autres modalités d'introduction d'une entité dans le texte, comme la deixis. Ces expressions demandent à être reliées soit à des éléments du contexte d'énonciation, soit au contenu même du texte identifié en tant que contexte d'énonciation, afin de pouvoir identifier leur référent (Conte, 1999; Hirst, 1981, p. 7). Comme pour l'anaphore, plusieurs types de deixis existent : la deixis spatiale, la deixis temporelle, la deixis textuelle et la deixis personnelle (Conte, 1999; De Cesare, 2010).

La deixis spatiale fait référence à la localisation dans l'espace, tandis que la deixis temporelle concerne la position dans le temps. Par exemple, dans « viens *ici* », l'adverbe *ici* illustre la deixis spatiale ; dans « *demain* elle va discuter sa thèse », c'est l'adverbe *demain* qui indique une deixis temporelle.

La deixis textuelle, quant à elle, se manifeste lorsque certaines expressions renvoient à une partie du discours lui-même pendant que la production de celui-ci est en cours, ou à une portion du texte déjà produite ou à venir (Conte, 1999; Da Milano, 2025, p. 235). Ce type de deixis ne porte pas sur le contexte externe au texte, mais sur le co-texte lui-même (Da Milano, 2025). Ainsi, dans l'énoncé: « Nous approfondirons *ceci* dans le paragraphe suivant », le pronom démonstratif « ceci » renvoie à une idée déjà développée dans le texte, tandis que le syntagme « le paragraphe suivant » anticipe une section successive du texte. De cette manière, la deixis textuelle joue donc un rôle d'orientation du lecteur dans le texte et contribue à structurer la progression discursive.

L'exemple (25), cité par Da Milano, illustre une des co-occurrences fréquemment observées dans le corpus Univers-ITA avec le déictique italien « sopra » (ci-dessus). Ce marqueur apparait souvent en association avec des formes de participe passé telles que *cité, dit, écrit, vu* etc. (« citati, detto, esposti, scritto, trattati, elencati, descritte, riportato, menzionati, indicata, visto » (Da Milano, 2025, p. 235)), comme « detto sopra » (dit ci-dessus) dans le texte en question. Ce marqueur a ici fonction de logodéictique (*logodeittico*) : ce n'est pas une entité précise dans le texte qui est indiquée ici mais une portion de ce dernier.

(25) Detto ciò però è anche giusto osservare come sia stato a dir poco *sorpendente* il breve tempo in cui tutto ciò è passato dall'essere ideato alla sua effettiva messa in pratica, sorprendendo me per quanto la rapidità non sia andata ad influire sull'efficienza, anzi tutt' altro, e ciò merita soltanto un ringraziamento perché come già detto **sopra**, ci ha permesso di proseguire in questo splendido viaggio che è l'istruzione *(Corpus Univers-ITA, texte n. 37, tiré de Da Milano, (2025, p. 236))*[12]

La fonction de la deixis textuelle peut être considérée comme similaire à celle de l'anaphore résomptive : dans les deux cas, ces expressions renvoient à une portion du texte précédente ou à venir. Autrement dit, ces deux expressions établissent un lien avec le co-texte (Conte, 1999). Toutefois, Conte souligne que les deux procédés se distinguent par la nature de l'entité à laquelle s'effectue la référence. Dans le cas de l'anaphore, elle renvoie à une entité extra-textuelle mais

[12] Le texte est ici présenté dans sa forme d'origine, y compris les erreurs d'orthographe (voir *sorpendente*).

qui a déjà été introduite dans le discours. Le référent indiqué existe déjà dans la représentation du monde extra-textuelle, bien que sa représentation soit « diluée » dans une portion de texte. Dans le cas de la deixis textuelle, l'expression renvoie directement à une partie du texte lui-même, donc à une entité intra-textuelle, et a fonction connective (Conte, 1999, p. 19), qualifiant donc toute deixis textuelle non plus comme procédé anaphorique[13]. Ainsi, si l'anaphore établit une relation avec un référent du monde du discours, la deixis textuelle établit une relation avec le texte lui-même. Pour cette raison, Conte (1999, p. 19) propose de considérer la deixis textuelle comme un procédé distinct de l'anaphore, dans la mesure où la référence ne porte plus sur une entité extra-textuelle mais sur une portion du discours en tant que telle.

Comme décrit auparavant (cf. §2.1, §2.4), certaines expressions sont considérées intrinsèquement déictiques, par exemple les pronoms personnels de première et deuxième personne singuliers et pluriels : leur interprétation est toujours liée au contexte d'énonciation ou au co-texte d'un discours donné. Cette utilisation déictique de première et deuxième personne peut également être exprimée à travers la morphologie verbale dans le cas des langues à sujet omis (Bonomi et al., 2010). Ces pronoms déictiques sont souvent présents dans les discours directs reportés dans les textes narratifs. En tant que partie intégrante des dialogues, en suivant le choix déjà effectué dans d'autres projets d'annotation de la coréférence, nous pouvons considérer la deixis personnelle comme faisant partie des chaines de référence des personnages qu'elle indique (cf. §4 pour la notion de chaine de référence) et donc l'annoter au sein des chaines.

3. Coréférence

Dans cette partie, nous allons définir la notion de *coréférence* avant d'aborder quelques-unes des questions débattues dans la définition et l'annotation de ce phénomène et qui sont liées à des questions abordées par la suite, en particulier nos choix en termes d'annotation des référents évolutifs dans la partie méthodologique (cf. Chapitre 5, §3.1) et les cas de mise en œuvre des stratégies référentielles ciblées pour éviter les ambiguïtés référentielles dans le chapitre dédié à nos résultats (cf. Chapitre 8, §3). En nous inspirant aussi de la présentation de cas d'études faite par Schnedecker (2021, p. 34-49), nous allons en particulier décrire la question des référents évolutifs ainsi que celle de l'ambigüité référentielle. Nous allons revenir dans la partie relative à l'état de l'art en annotation de la coréférence en TAL (cf. Chapitre 3) sur les différentes théories qui existent dans cette discipline pour le traitement de ce phénomène.

Pour introduire la notion de coréférence, il convient de décrire les similitudes et les différences qu'elle entretient avec la notion d'*anaphore*, que nous avons décrite précédemment. Comme le souligne Milner (1982, p. 32), les deux notions sont fréquemment confondues. « Elles sont pourtant distinctes. Il y a relation de coréférence entre deux unités référentielles A et B quand

[13] Dans le texte d'origine « Ciò che differenzia anafora e deissi testuale è che nell'anafora il riferimento è fatto ad un'entità extratestuale alla quale si è già fatto riferimento con un'espressione antecedente che si trova nel co-testo; nella deissi testuale, invece, il riferimento è fatto direttamente ad una parte del co-testo stesso, ad un'entità intratestuale. » (Conte, 1999, p. 19)

elles se trouvent avoir la même référence - ce qui peut arriver sans que l'interprétation de l'une soit affectée par l'interprétation de l'autre. » (Milner, 1982, p. 32) La coréférence suppose, en fait, « une forme d'identité référentielle entre les référents évoqués » (Schnedecker, 2019, p. 13) sans pour autant reposer sur une interdépendance entre expressions qui indiquent le même référent afin de pouvoir les interpréter et les relier au référent évoqué, ce qui configure la relation de coréférence comme *symétrique*, au contraire de celle d'anaphore qui est, quant-à-elle, caractérisées comme *asymétrique* (cf. ce chapitre, §2). Autrement dit, dans le cas de l'anaphore une des deux expressions, l'antécédent, est indispensable pour codifier l'autre, l'anaphorique, alors que dans le cas des expressions coréférentielles l'interprétation de l'une est indépendante de l'autre.

Dans l'exemple (26), « Barack Obama » et « Le président des États-Unis » représentent deux mentions coréférentes : les deux expressions apportent assez d'information sur le référent pour permettre de l'identifier de manière indépendante l'une de l'autre. L'interprétation de l'une est indépendante de l'interprétation de l'autre, elles indiquent de manière égale le même référent, peut-être en apportant des informations ou un point de vue différent vis-à-vis de celui-ci.

(26) **[Barack Obama]** a nommé Hillary Rodham Clinton secrétaire d'État lundi. **[Le président des États-Unis]** l'a choisie parce qu'elle avait l'expérience des affaires étrangères en tant qu'ancienne première dame.

Dans le texte (27), le personnage de Pierre du roman « Le Capitaine Fracasse » est introduit à l'aide de l'expression « un bonhomme, moitié paysan, moitié domestique », puis repris plus tard dans le texte par le SN nominal défini « le vieillard », qui apporte cette nouvelle information sur le personnage et qui confirme son lien avec l'expression précédente le début de la chaine, « une personne âgée ».

(27) La marmite bouillait toujours, et le chat restait immobile à son poste, comme une sentinelle qu'on a oublié de relever. Enfin un pas se fit entendre, pas lourd et pesant, celui d'**une personne âgée** ; une petite toux préalable résonna, le loquet de la porte grinça, et **[un bonhomme, moitié paysan, moitié domestique]**, fit son entrée dans la cuisine. À l'apparition **[du nouveau venu]**, le chat noir, qui semblait lié de longue date avec **[lui]**, quitta les cendres de l'âtre et se vint frotter amicalement contre **[ses]** jambes, arquant le dos, ouvrant et refermant ses griffes, en faisant sortir de sa gorge ce murmure enroué qui est le plus haut signe de satisfaction chez la race féline. «Bien, bien, Béelzébuth, dit **[le vieillard]** en se courbant pour passer à deux ou trois reprises **[sa]** main calleuse sur le dos pelé du chat, afin de n'être pas en reste de politesse avec un animal; (…).
(Capitaine Fracasse, extrait annoté issu du corpus Democrat)

Delaborde (2020, p. 18) rappelle d'ailleurs que si ces deux notions de coréférence et d'anaphore peuvent parfois se recouper, l'une n'implique pas automatiquement l'autre. Ainsi, l'anaphore peut être coréférentielle, mais toute relation de coréférence n'est pas nécessairement anaphorique. Cette distinction est cruciale pour comprendre comment les chaines de référence

se construisent dans un texte : alors que l'anaphore structure la progression discursive par des rappels au contexte discursif antérieur et au co-texte, la coréférence devrait permettre d'identifier plusieurs expressions comme renvoyant à une même et seule entité. Dit autrement, dans la coréférence les expressions référentielles devraient permettre d'identifier toujours un seul et même référent, qu'il appartienne à la réalité extra-textuelle ou à l'univers d'un texte donné.

A priori, l'interprétation de ces expressions linguistiques construit donc une relation d'*identité référentielle* (Corblin, 1985). C'est cette notion d'identité qui est parfois discutée, surtout en TAL, à cause de la notion de *quasi-identité* (Recasens et al., 2010) qui peut parfois remettre en question un schéma d'annotation donné ainsi que la qualité des annotations obtenues. La notion de quasi-identité peut être définie comme la « zone grise » entre l'identité et la non-identité : lorsque deux expressions font référence au même référent en partageant la plupart de ses caractéristiques, mais sont au moins différentes dans une dimension cruciale pour son identification, elles se situent dans la zone de quasi-identité (Recasens et al., 2010, p. 151). Cette conception de la coréférence implique de formaliser les traits de coréférent ou non coréférent d'une expression donnée dans un continuum où ces deux dimensions constituent les extrêmes, entre lesquelles peuvent se retrouver des relations plus ou moins strictes de quasi-identité (Recasens et al., 2011). Dans les catégories faisant partie de la quasi-identité selon Recasens et al. (2010), on retrouve la métonymie, la méronymie, la classe et la fonction spatio-temporelle.

C'est cette même notion d'identité qui est remise en cause dans l'annotation des *référents évolutifs*. Cette question, qui a donné lieu à de longs débats sur la nature même de la notion de référent (Charolles, 2001), se pose encore plus dans le cas de textes narratifs qui s'inspirent largement des contes de fées, surtout là où la présence d'un personnage magique comme la sorcière peut provoquer entre autres des métamorphoses des personnages présents dans le texte narratif.

Si dans le cas des référents évolutifs nous sommes parvenue à fournir une réponse qui nous semble justifier les choix effectués dans le cadre de notre annotation, les cas d'ambiguïtés référentielles nous semblent encore plus compliqués à trancher en donnant une solution opérationnelle immédiate et claire, à cause de la nature peu normée des textes que nous analysons (mais aussi à cause de la nature ambigüe du langage naturel).

3.1. Référents évolutifs

Les référents évolutifs peuvent être définis, selon Charolles et Schnedecker, comme ces entités qui subissent des changements, au fur et à mesure qu'on avance dans un texte, ou encore comme une entité qui développe « divers avatars, portant gravement atteinte à son état, au point que l'on peut se demander si, au terme de ces avatars, on a toujours affaire à la même entité et si donc il est encore possible de parler de coréférence » (Charolles & Schnedecker, 1993, p. 106).

Un des exemples les plus utilisés en littérature pour introduire l'analyse des référents évolutifs est celui de la transformation à laquelle on soumet un ingrédient dans une recette de cuisine[14]. Dans l'exemple (28), qui décrit le processus de réalisation de la confiture de fraises, le référent initial est soumis à une série de changements physiques.

(28) Lavez et équeutez **[les fraises**], coupez-**[les]** en 2, et déposez **[les moitiés]** dans un grand saladier en alternant les couches de sucre et **[de fraises]**. À l'aide d'une passoire, filtrez **[le sirop [de fraise]] [qui]** sera apparu. Mettez **[la pulpe obtenue]** dans une casserole et portez **[le tout]** à feu vif.

À cet égard, le paradoxe du bateau de Thésée[15] illustre de manière exemplaire la problématique liée à l'évolution des référents et à l'influence des métamorphoses sur les relations d'identité entre étapes de transformation, ou encore liés aux changements des caractéristiques inhérentes aux référents. Si dans un bateau on change de manière progressive toutes ses parties d'origine avec des nouvelles parties, au fur et à mesure des années, le bateau ne va plus contenir aucune de ses parties d'origine. Ce bateau est encore le même qu'il était au début ou s'agit-il d'un nouveau bateau que l'on a créé ? Ce même paradoxe explique bien les doutes que l'on peut rencontrer lors qu'on doit annoter un texte comme celui de l'exemple (28) : les fraises mentionnées en début de texte constituent-elles une chaine référentielle distincte de celles associées au sirop de fraise et à la pulpe, ou bien s'agit-il d'un unique référent dont les caractéristiques externes évoluent au fil du texte ? Ce questionnement peut s'appliquer à tout référent sujet à des métamorphoses dans un texte : si l'on change toutes les caractéristiques d'un référent, est-ce que le référent reste le même tout au long du texte, ou encore, est-ce que ce référent est indiqué par une seule chaine ou par une chaine différente pour chacune de ses étapes de transformation ?

Cette notion de référent évolutif a été longuement source de débat entre différents linguistes (voir Achard-Bayle, 2016; Apothéloz & Reichler-Béguelin, 1995; Charolles, 2001; Charolles & Schnedecker, 1993 entre autres). La notion de référent évolutif constitue un objet d'analyse particulièrement délicat, car elle se situe à l'intersection de plusieurs domaines théoriques : la philosophie du langage, la linguistique, la sémantique textuelle et la théorie de la valence des prédicats. Sa compréhension approfondie implique par ailleurs le recours aux apports des sciences cognitives et de la psycholinguistique, dans la mesure où elle engage des mécanismes mentaux complexes et fortement dépendants du contexte discursif. Même lorsqu'on limite l'analyse à une approche strictement linguistique, il demeure inévitable de prendre position sur la notion de référent que l'on retient dans notre approche. Charolles et François (1998) ont recensé et classé « les verbes à même d'exprimer des changements » (Charolles, 2001, p. 4) dans une catégorisation qui prévoit trois typologies : les verbes de création, les verbes d'annihilation et un troisième groupe qui comprend les verbes de changement qualitatif qui n'affectent pas la forme de l'entité, ou qui l'affectent par un passage de l'ensemble à la partie ou *viceversa* ou par

[14] L'exemple plus célèbre étant celui de Brown et Yule (1983) qui décrivent une recette de poulet rôti pour justifier la non fonctionnalité du concept de remplacement dans l'analyse de l'anaphore pronominale.
[15] Pour le paradoxe du bateau de Thésée, voir Engel et Nef (1988), cité par Charolles et Schnedecker (1993).

modification de la structure originaire. Dans cette taxonomie, les constructions du type *transformer X en Y* signalent des opérations de reclassification du référent, c'est-à-dire indiquent un changement radical entre identité originaire et *identité sortale* (Charolles & François, 1998, p. 18), changement qui n'est pas pris en compte dans la taxonomie précédemment décrite en littérature. Nous pouvons illustrer les difficultés rencontrées à cet égard à travers l'exemple (29) issu du corpus annoté.

(29) Il était une fois **[une vieille sorcière méchante]**^{witch} et plus loin habitait **[une chatte blanche]**^{cat}. Et **[cette chatte]**^{cat} alla voir **[la sorcière]**^{witch} et puis **[la sorcière]**^{witch} **[la]**^{cat} transforma en **[crapaud]**^{cat}... En **[souris]**^{cat}, en **[poisson baveux]**^{cat} mais sans faire exprès **[la sorcière]**^{witch} donna les pouvoirs à **[la chatte]**^{cat} et **[la chatte]**^{cat} fit ABRACADABRI ! en rigolant... (…). *(corpus français, CE1, 493)*

Dans ce texte, la sorcière transforme le personnage *chat* plusieurs fois. Et pourtant, nonobstant plusieurs métamorphoses, l'élève fait le choix en fin de texte (dernière phrase) de se référer au personnage à travers le syntagme nominal qui identifie sa forme première de chat. De même, faire le choix de considérer les chaines relatives aux personnages dès leur première parution et toujours en utilisant une seule étiquette nous permet de suivre le même personnage tout au long du texte de manière indépendante de ses possibles transformations. Cela nous semble correspondre donc à nos objectifs d'annotation, sur lesquels nous allons revenir dans le chapitre dédié à la méthodologie d'annotation (cf. Chapitre 5).

3.2. Ambiguïté référentielle

Comme affirmé par Schnedecker (2021, p. 35), « un premier obstacle à l'établissement de la coréférence tient aux cas dits d'« ambiguïté référentielle » où une expression référentielle donnée se trouve avoir potentiellement deux sources » ou plus. Cela peut arriver pour plusieurs raisons, dont la compétition entre référents saillants dans la sélection de l'antécédent d'une expression anaphorique. Cela peut se produire dans le cas où les deux sources s'accordent en genre et en nombre avec l'expression référentielle à rattacher. Cela est le cas dans le texte (30), où les expressions surlignées dans le texte se trouvent dans une « zone floue » quant à leur rattachement aux deux chaines annotées dans le texte.

(30) Il était une fois **[un chat]**^{cat} **[qui]**^{cat} se promenait dans la forêt, mais soudain **[un loup]**^{wolf} a mangé des cailloux. C'est pour reconnaitre **[son]**^{wolf} chemin **[du chat]**^{cat}. **[Il]**^{cat} se retourne. **[Le chat]**^{cat} a griffé **[le loup]**^{wolf} et **[tombe]**^{wolf} par terre boum. Donc **[le chat]**^{cat} est perdu mais **[le loup]**^{wolf} a fait semblant de tomber alors **[il]**^{wolf} se déguise en **[chat]**^{cat}. **[Le chat]**^{cat} se retourne et **[voit]**^{cat} **[un énorme chat]**^{wolf} donc **[le loup]**^{wolf} et **[le chat]**^{cat} sont devenus amis. Fin de l'histoire. *(corpus français, CE1, 950)*

Dans la troisième phrase, la mention anaphorique « il » devrait en théorie être liée au *loup* : bien que sa dernière mention nominale remonte à la première phrase, la présence de « son » dans la deuxième phrase devrait rendre le *loup* le dernier personnage saillant et donc le candidat le plus probable pour sélectionner l'expression anaphorique « il ». Cependant, le fait que la mention

nominale la plus proche, à la fois en contexte gauche et en contexte droit, fasse référence au *chat* peut également influencer l'interprétation de cette anaphore pronominale. En outre, le fait que les deux référents ont les mêmes caractéristiques morphosyntaxiques de genre masculin et nombre singulier, complique davantage l'opération de désambiguïsation nécessaire pour rattacher de manière certaine l'anaphore à un personnage plutôt qu'à l'autre.

Si certains éléments peuvent contribuer aux opérations de désambiguïsation de rattachement dans le cas d'anaphores pronominales, comme nous avons vu, le genre et le nombre grammaticaux, mais aussi d'autres paramètres cotextuels comme la sémantique des verbes ou encore les effets de saillance donnés par l'emploi d'une expression référentielle « pleine » plutôt qu'une expression référentielle « vide », dans le cas de textes peu normés, ces éléments ne sont pas toujours suffisants pour reconstruire correctement le rattachement d'une expression référentielle à un référent donné (autrement dit, l'appartenance d'un maillon à une *chaine de référence* donnée).

4. Chaine de (co-)référence : définition formelle et indices de description

4.1. Définition de chaine de référence

La notion de *chaine de référence* (parfois aussi appelée *chaine de coréférence*) est une notion plus récente par rapport à celle d'anaphore. De ce fait, et parce que ses mécanismes d'interprétation reposent sur des notions plus « anciennes » et établies comme celles d'anaphore et de coréférence, la notion de chaine de référence reste peu présente dans les grammaires de la langue française. Elle est en effet mentionnée dans la *Grande grammaire du français* (Abeillé et al., 2021), où la notion de chaine est définie ainsi : « Des énoncés successifs peuvent partager le même référent de discours, désigné par des expressions différentes, qui forment alors une chaine de référence (…) » (Abeillé et al., 2021, p. 102).

Cette notion est reprise et précisée plus loin dans le même ouvrage. Cette deuxième définition l'illustre à travers un exemple concret :

> « Dans les récits comme **4**, qui comporte plusieurs expressions anaphoriques, toutes les expressions renvoyant à un même personnage forment ce qu'on appelle une *chaine référentielle* ou *chaine de référence*. (…) Une expression anaphorique intégrée dans une chaine de référence n'a pas pour antécédent l'autre expression anaphorique qui la précède : la proforme *l'* n'a pas pour antécédent *il*, mais tous deux ont pour antécédent *le lieutenant hussard* en **4.**
>
> **(4)** *En 1812, [le lieutenant hussard André Duvalier] se trouve perdu sur la côte baltique. [Il] est secouru par [une ravissante jeune fille] qui [l]'entraine au château de l'horrible baron von Leppe, où [elle] disparait. [André] est alors saisi par l'incroyable*

> *ressemblance qu'[elle] présente avec le portrait de la femme du baron.* (*CRF, L'Halluciné*) » (Abeillé et al., 2021, p. 2027)[16]

Comme observé par Schnedecker (2019, paragr. 1), cette notion est définie en linguistique pour la première fois par Chastain (1975), puis apparait en linguistique française (Charolles, 1978, 1988c; Corblin, 1985, 1995). Charolles (1988c, p. 8) la définit de la manière suivante :

> « Les chaines sont constituées par des suites d'expressions coréférentielles [...]. Seules peuvent appartenir (donner lieu à) une chaine des expressions employées référentiellement, c'est-à-dire toutes et rien que les expressions nominales (ou pronominales) permettant d'identifier un individu (un objet de discours) quelle que soit sa forme d'existence (personne humaine, événement, entité abstraite). »

Ces expressions qui composent les chaines peuvent être appelées, par métaphore, *maillons* de la chaine de référence. D'autres auteurs, spécialement anglophones, appellent ces expressions *mentions*.

Cette première définition, quelque peu contraignante, a été ensuite revisitée et redéfinie pour comprendre toutes ces expressions qui indiquent un même référent, en plus des expressions nominales et pronominales, telles que les noms propres ou les « maillons invisibles » comme les anaphores zéro. La définition de *chaine* que nous avons adoptée est formellement celle donnée par Schnedecker, que nous approfondissons ici de suite.

La notion de chaine peut être distinguée de celle d'anaphore et de coréférence sur la base d'un critère purement numérique (Schnedecker, 2021, p. 50) : pour construire une chaine de coréférence, il faut au moins trois expressions faisant référence à la même entité, appartenant à l'univers du texte. Si on a moins de trois composantes, les notions d'anaphore et de coréférence sont suffisantes pour une description pertinente des paires d'enchainements référentiels (Schnedecker, 1997, 2019, 2021). Ce critère essentiellement numérique sera retenu dans la suite de notre travail pour distinguer les instances de chaine de celles d'anaphore (deux mentions liées au même référent dans notre travail, qu'ils soient coréférents ou seulement anaphoriques) ou de *singleton* (les cas où un référent est évoqué dans un texte par le biais d'une seule mention).

Selon Schnedecker (2021, p. 53-55), deux approches existent dans les études des chaines de référence : l'approche *paradigmatisante* et l'approche *syntagmatisante*. Selon l'approche *paradigmatisante*, les chaines sont considérées comme des lots d'expressions qui peuvent remplacer le référent de manière interchangeable, ou encore elles sont toutes considérées équivalentes, « abstraction faite de leur environnement discursif » (Schnedecker, 2021, p. 86). Cette approche ne semble pas prendre en considération la linéarité du texte et la place que chaque type d'expression peut avoir à un moment déterminé du déroulement textuel ; elle néglige aussi les effets déterminés par la cohabitation entre chaines différentes qui constitue la réalité textuelle. Cependant, la conception définie par Schnedecker comme *syntagmatisante* prend en considération la linéarité textuelle et les contraintes liées à la construction de la référence qui proviennent de la textualité (et des caractéristiques typiques des genres discursifs

[16] Italique et gras dans le texte original.

représentés dans les textes). C'est à cette deuxième conception que nous faisons référence dans la suite de notre travail, et en particulier dans le cadre de nos analyses qualitatives sur les textes annotés de notre corpus (cf. Chapitre 8). En effet, une description plus approfondie des textes annotés ayant comme objectif ultime d'observer les pratiques d'écriture à but didactique rend nécessaire d'adopter une telle approche, qui prend en compte le déroulement des chaines et les stratégies de cohabitation mises en œuvre dans le texte.

4.2. Indices de description des chaines de coréférence

Avant d'introduire les différents indices utilisés pour décrire les chaines de référence, il faut tout d'abord évoquer le travail mené par Landragin et son équipe sur le corpus *DEMOCRAT* (Landragin, 2016a)[17]. Ce corpus représente le premier grand corpus français annoté en coréférence stricte, et sa réalisation a aussi eu le mérite de regrouper les différents critères de description des chaines existant en littérature autour de la caractérisation d'un seul corpus, représentant des genres de textes et des époques variées. Le travail d'annotation effectué sur ce corpus se base également sur le cadre théorique décrit par Ariel et Givón que nous allons décrire par la suite (cf. §5.1 et §5.2).

Une *chaine de coréférence* ou *chaine de référence* peut être décrite par plusieurs caractéristiques quantitatives ou numériques ainsi que qualitatives. Au-delà du travail réalisé sur le corpus *DEMOCRAT*, plusieurs auteurs ont proposé divers descripteurs (pour une sélection, Boudreau & Kittredge, 2005; Federzoni et al., 2020; Landragin et al., 2024; Oberlé et al., 2018; Obry et al., 2017; Perret, 2000; Rousier-Vercruyssen & Landragin, 2019; Schnedecker, 1997, 2019, 2021; Schnedecker & Landragin, 2014). Ces contributions se fondent sur des approches théoriques et méthodologiques variées, allant de l'analyse linguistique fine à des traitements plus computationnels. Plusieurs auteurs ont proposé des définitions variées de ces critères permettant de décrire les chaines de coréférence, dont l'interprétation peut varier selon les auteurs. Avant d'en proposer une synthèse, il convient donc de présenter les principales définitions existantes, afin de mieux cerner les divergences opérationnelles que l'on retrouve chez ces différents auteurs.

Ces indicateurs sont, entre autres :

- Le nombre de référents du texte ;
- Le nombre d'expressions référentielles ;
- La longueur de la chaine, c'est-à-dire le nombre de maillons qui composent une chaine de référence. Ce critère sert aussi à distinguer les chaines des paires d'expressions nominales liées au même référent et des singletons ;
- La portée de la chaine : elle peut avoir une portée soit locale (elle se déroule sur un ou deux paragraphes), soit globale (elle est persistante tout au long du texte ou elle recouvre une bonne partie de celui-ci) ;

[17] La constitution de ce corpus est approfondie en Chapitre 3, §3.4.

- La distance moyenne entre maillons de la chaine (combien de mots ou tokens il y a entre un maillon et l'autre. Cette mesure peut être aussi représentée en nombre de phrases dans le cas de textes longs) ;
- La longueur des mentions en termes de tokens ;
- La catégorie et la fonction grammaticale des maillons, notamment celle du premier et deuxième maillon de la chaine ;
- Le patron de la chaine : les séquences de types de maillons les plus fréquents en termes de catégories grammaticales ;
- Le mode de cohabitation des chaines dans le texte : succession, entrecroisement, dérivation, partition, fusion, déroulement en parallèle (voir Schnedecker, 1997, p. 13-15).

Ces critères sont hautement variables selon le niveau de maîtrise du scripteur, la longueur et le genre du texte analysé, comme démontré dans plusieurs travaux, entre autres, Longo & Todirascu (2014); Schnedecker (2005, 2014, 2021); Schnedecker & Longo (2012). Des critères supplémentaires existent, qui peuvent être calculés à partir de certains des critères déjà définis. Parmi eux, nous pouvons retrouver la *densité référentielle* d'un texte et le *coefficient de stabilité* d'une chaine.

Densité référentielle

La densité référentielle est définie comme le ratio entre le nombre d'expressions référentielles et le nombre total de tokens dans un texte (Boudreau & Kittredge, 2005; Obry et al., 2017; Schnedecker, 2017). Elle varie selon le genre de texte ou par opposition diamésique (oral vs. écrit) (Kim, 2023).

Pour définir de manière plus précise la manière de calculer la densité référentielle, nous pouvons nous appuyer sur un exemple de texte annoté tiré de notre corpus. Dans l'exemple (31), nous pouvons compter 11 expressions référentielles (6 dans la chaine de *cat*, 5 dans la chaine de *wolf*) dans un texte composé de 53 mots. La densité référentielle sera égale à 11/53, soit 20,8%.

(31) Il était une fois **[un petit chat blanc]^{cat} [qui]^{cat}** vivait dans la forêt. Un jour que **[le petit chat]^{cat}** se promenait et **[il]^{cat}** vit **[un loup]^{wolf} [qui]^{wolf}** était noir. **[Le loup]^{wolf}** était gentil, **[il]^{wolf}** voulait un ami et **[le petit chat blanc]^{cat}** voulait un ami**.** **[Le petit chat]^{cat}** et **[le loup]^{wolf}** sont ami. *(corpus français, CE1, 212)*

Néanmoins, dans le cas de la présence de mentions référentielles composées par plusieurs tokens (ce qui est le cas dans les SN, les syntagmes nominaux), cet indice peut être biaisé par le manque de prise en compte du nombre total de tokens inclus dans les mentions dans ce calcul de densité. Nous allons revenir sur ces questions dans la partie dédiée à l'opérationnalisation des indices de description et à leur application pratique (cf. Chapitre 7, §3).

Le Tableau 1 présente les valeurs de densité référentielle rapportées dans la littérature. Ces valeurs ne sont pas directement comparables entre eux à cause des différences entre pratiques d'annotation et entre type de référents annotées (ce sont que les référents animés pour certains auteurs, tous les référents pour d'autres). Dans ces résultats, on peut observer une grande variation de cet indice, qui est déterminée par la période historique de production du corpus

ainsi que par le genre de texte représenté dans chaque corpus. Cette variation, située entre 4% pour les portraits journalistiques et 40% pour les textes encyclopédiques du 12e – 13e siècle, témoigne de la haute densité référentielle que l'on peut rencontrer dans les textes narratifs, par rapport à une densité qui semble être plus basse dans le cas de textes appartenant à des genres divers.

Publication corpus	**Genre textuel**	**Période de production du corpus**	**Densité référentielle (texte)**
(Tutin, 2002)[18]	*Sciences humaines, textes scientifiques*	*21e siècle*	*2,53%*
	Presse	*21e siècle*	*3,04%*
	Manuel technique	*21e siècle*	*0,05%*
	Roman	*19e siècle*	*4,32%*
(Schnedecker, 2005)	Portraits journalistiques (Le Monde)	21e siècle	4%
(Schnedecker, 2014)	Recettes de cuisine	21e siècle	34%
	Intro de textes encyclopédiques	21e siècle	29%
(Obry et al., 2017) DEMOCRAT	Textes narratifs (récits brefs)	8e – 16e siècle	19,6% (Des Périers) 21,6% (Gombert) 23,6% (Vilain) 15,4% (CNN)
(Schnedecker, 2017)	Faits divers	21e siècle	9%
	Contes de fées	21e siècle	30%
(Oberlé et al., 2018) DEMOCRAT	Textes encyclopédiques	12e – 13e siècle	40%
	Textes encyclopédiques	14e – 20e siècle	30%
(Kim, 2023)	Roman "Le Fantôme de l'Opéra"	20e siècle	19%
(Landragin, 2020b; Landragin et al., 2024) DEMOCRAT	Poème épique	11e siècle	25,20%
	Roman	12e siècle	22,86%
	Hagiographie, chartes	13e siècle	21,58%
	Registre, didactique, chronique historique	14e siècle	17,34%
	Mémoires, roman, manuel d'éducation	15e siècle	18,54%
	Traité argumentatif, roman, pamphlet, récit de voyage	16e siècle	16,05%
	Roman, traités historiques, récits brefs, traité en agriculture, texte philosophique	17e siècle	17,25%
	Romans, traités argumentatifs, traité didactique, fables	18e siècle	14,61%
	Romans, nouvelles, traité argumentatif, textes juridiques, biographie	19e siècle	13,93%
	Romans, articles de presse, textes juridiques, pamphlet, biographie	20e siècle	14,73%
	Texte juridique, articles encyclopédiques (Wikipedia)	21e siècle	13,81%

Tableau 1. Synthèse des taux de densité référentielle observés dans divers corpus.

[18] Dans le cas de cette publication, il faut préciser que le phénomène annoté est plutôt la densité anaphorique.

Stabilité référentielle

Le *coefficient de stabilité référentielle* dit aussi *coefficient de Perret* (2000), a été initialement conçu pour mesurer la pauvreté lexicale dans les œuvres narratives. Il s'est par la suite ensuite avéré utile pour rendre compte des différentes caractéristiques des textes modernes en termes de variété dans la dénomination d'un référent donné (Obry et al., 2017).

Ce coefficient sert à décrire la « stabilité désignationnelle » des reprises nominales et il a été utilisé pour la première fois pour comparer l'utilisation de la désignation nominale dans des textes du XIVe et du XVe siècle. Dans sa première définition, ce coefficient de stabilité est « obtenu en divisant, pour un référent donné (ie. un personnage), le nombre total *d'anaphores nominales* par le nombre de désignations différentes (…). Plus le coefficient de stabilité est élevé, moins il y a de désignations différentes par rapport au nombre d'anaphores et donc plus la stabilité référentielle est grande. » (Perret, 2000, p. 17, notre italique). Toute désignation formellement différente est considérée comme une variation, y compris dans le cas de changement du déterminant qui introduit le syntagme (comme montré par les syntagmes *sa mère*, *leur mère*, en italique dans la liste ci-dessous), car ces syntagmes « peuvent avoir des fonctions narratives différentes », et enfin les désignations simples du type *Mélusine* sont considérées comme différentes par rapport aux désignations expansées comme « Mélusine la mainsnee, Mélusine la dame belle, Mélusine, la douce, courtoise et bénigne (Coudrette) » (Perret, 2000, p. 18). Les syntagmes contenus dans les titres de chapitres, dans les légendes, ainsi que les syntagmes collectifs et les premiers syntagmes indéfinis des chaines « qui ser[ven]t de pivot référentiel » ne sont pas considérés dans le calcul afin de n'étudier que les désignations « de même niveau » (Perret, 2000, p. 18). Dans sa taxonomie des désignations neutres et marquées, une liste des mentions considérées dans les textes est proposée, et si on prend l'exemple des désignations du personnage de Mélusine, nous pouvons remarquer que le nombre total de désignations prises en compte par Perret correspond à la liste suivante, qui comprend les noms propres dans les désignations du calcul.

« Mélusine : Désignations neutres :
86 Melusigne,
1 celle Melusigne,
31 la dame,
9 sa dame,
7 sa mere,
7 leur mere,
2 sa femme,
1 Melusigne, leur mere,
1 Melusigne, sa mere,
1 sa moillier.

Désignations marquées :
7 la serpente,
6 l'espousee,
1 celle qui si bon confort lui avoit donné,
1 la dame, en guise de serpente,
1 la plus seignourie,
1 Melusigne en guise de serpente,
1 Melusigne, la bonne dame. »

(Perret, 2000, p. 19)

Cependant, tel qu'il est décrit dans la littérature, cet indice présente plusieurs difficultés, et parfois même des incohérences mineures dans sa définition opératoire selon les différents auteurs. Si on regarde le calcul de la stabilité proposé par Perret (2000), le chiffre des mentions

prises en compte pour Mélusine correspond tout à fait à la somme des mentions de la liste ci-dessus qui comprend les noms propres : « Mélusine (J. d'A) : Mélusine 9,64 (164/17) ; Raymondin 32 (352/11) » (Perret, 2000, p. 18). Cependant, selon la définition donnée par Schnedecker « ce coefficient pose un certain nombre de questions : limité aux anaphores nominales, *il laisse de côté le nom propre (et ses sous-types)* qui n'est pas anaphorique, les pronoms personnels (et autres). » (Schnedecker, 2021, p. 97, notre italique). Cette dernière position nous semble, d'une part, reposer sur une interprétation stricte de l'anaphore nominale. En effet, les noms propres sont traditionnellement considérés comme des mentions autonomes du point de vue de la référence, ils participent ainsi à des relations de coréférence et ne sont pas, à proprement parler, anaphoriques. D'autre part, cette définition ne semble pas prendre en considération les exemples concrets utilisés par Perret pour expliquer l'usage de son coefficient, dans lesquels les syntagmes nominaux qui contiennent des noms propres et les noms propres seuls sont effectivement inclus dans le décompte des mentions utilisées pour le calcul de l'indice. Ainsi, on observe un décalage entre la définition théorique et la mise en œuvre empirique de l'indice, ou les noms propres semblent être intégrés au calcul.

Ainsi, en plus des incohérences que nous avons identifiées, ce coefficient tel que décrit par Perret, présente à notre avis les limites suivantes :

- le fait de ne pas prendre en compte les possibles démarrages de chaine peut être problématique lors qu'on analyse des textes qui n'appartiennent pas au genre narratif ;
- le calcul de l'indice ne prend pas en considération les pronoms personnels, ni d'ailleurs les anaphores zéro, car cette analyse est ciblée sur la variété dans la dénomination : c'est donc un très bon outil pour gérer la variation lexicale à l'intérieur des chaines, mais il ne peut pas rendre compte de manière complète de la variété tout court à l'intérieur des chaines en termes de différences entre types de mentions utilisées ;
- cet indice est facilement influencé par la longueur des chaines : à parité de variation, une chaine plus longue sera forcément plus stable qu'une chaine plus courte.

Une autre définition opératoire légèrement différente de celle de Perret est fournie par Oberlé (Oberlé, 2019; Oberlé et al., 2018). Dans les deux articles, la définition reste globalement identique et reléguée à une note en bas de page :

> Le coefficient de stabilité (adapté de Perret, 2000) est calculé en ne prenant en compte que la tête syntaxique des groupes nominaux (GN) (les pronoms et déterminants ne sont pas inclus dans le calcul). Le coefficient est compris entre 0 et 1 : plus il est élevé, moins il y a de variations (s'il est à 1, toutes les reprises nominales sont les mêmes, s'il est de 0, elles sont toutes différentes). Nous avons utilisé la formule suivante :
> 1-(x-1) / (n-1), où x est « le nombre de désignations différentes » et n « le nombre total d'anaphores nominales », pour reprendre les termes de Perret. (Oberlé, 2019, p. 9)

L'idée d'utiliser seulement les têtes syntaxiques des groupes nominaux représente une nouveauté par rapport à la définition originale de Perret, en plus de la normalisation de la métrique à l'aide de la modification de son calcul original. La définition opératoire donnée par Oberlé a comme but de créer « un indicateur normalisé, c'est-à-dire dont les valeurs sont comprises entre 0 (ou 0 % : aucune stabilité, toutes les expressions sont différentes) et 1 (ou 100 % : stabilité maximale, toutes les expressions sont identiques) » (Oberlé, 2016, p. 83) Ce coefficient, appelé *coefficient normalisé de stabilité lexicale* (CNSL) est donc défini à l'aide de la formule suivante :

$$\mathrm{CNSL} = 1 - \frac{(x-1)}{(n-1)}$$

où x est le nombre de désignations différentes et n est le nombre total d'anaphores nominales.

Cependant, si cette formule du coefficient tente de résoudre le problème de la comparaison entre chaines différentes, grâce à la normalisation entre 0 et 1 du coefficient obtenu, elle ne parvient pas à combler d'autres lacunes de cet indice.

Dans la définition de Obry et al. (2017), cet indice est calculé «(...) en divisant, pour un référent donné, le nombre total d'ER nominales (G[roupes] N[ominaux] et noms propres) par le nombre de ses désignations nominales différentes » (Obry et al., 2017, p. 99). Il inclut de manière explicite les noms propres dans le calcul en plus des syntagmes nominaux. Rousier-Vercruyssen & Landragin (2019) redéfinissent cet indice comme instabilité référentielle, et proposent de le calculer comme le ratio entre les dénominations uniques et le nombre total de syntagmes nominaux, multiplié par 100 pour rendre les chaines comparables entre elles et entre genres de textes différents. Leur indice d'instabilité sera donc calculé selon la formule suivante, où x est le nombre de désignations différentes et *n* le nombre d'anaphores nominales dans la chaine analysée :

$$Instabilité = \frac{x}{n} * 100$$

Cependant, tout en étant utilisée en littérature pour évaluer la richesse lexicale à l'intérieur des chaines, cette mesure présente plusieurs difficultés. Tout d'abord, elle tend à augmenter avec la longueur des chaines, malgré le tentative de la « normaliser (…) en la rapportant à la longueur de la CR » (Obry et al., 2017, p. 99). Outre ce biais lié à la longueur des chaines, cet indice ne semble pas prendre en compte la réelle composition des chaines, car il exclut tout autre type de mention non nominale comme les pronoms ainsi que les anaphores zéro. Dans la partie dédiée aux résultats (cf. Chapitre 7, §4), nous allons proposer un calcul alternatif de cette mesure pour la rendre opérationnelle.

4.3. Incidence du genre de texte sur la répartition des types de maillons

Comme anticipé dans la définition des descripteurs des chaines, celles-ci sont habituellement constituées de maillons hautement variables, tant sur le plan lexical que sur le plan morphosyntaxique. Les chaines peuvent ainsi être catégorisées et décrite à travers les types d'expression qui les composent. Nous allons approfondir la catégorisation des expressions adoptée dans notre travail dans la section 6 de ce chapitre.

De nombreux chercheurs se sont intéressés à la relation entre la nature du matériau composant les chaines de référence ou les phénomènes qui relèvent de la représentation de la référence et le genre textuel. Le genre textuel peut être défini comme un ensemble de caractéristiques partagées par différentes productions langagières (orales ou écrites), mobilisant plusieurs niveaux du langage, et permettant aux locuteurs d'une langue donnée d'identifier un texte en le rattachant à un contexte culturel spécifique ainsi qu'à des objectifs communicatifs particuliers, malgré les variations observables d'un texte à l'autre (Adam, 1992; Bakhtine, 1984; Bronckart, 1997; Schneuwly, 1995).

L'un des premiers travaux à avoir exploré cette question est celui de Tutin (2002), qui décrit la distribution des différents types d'expressions anaphoriques dans un corpus de textes appartenant à divers genres : quatre articles scientifiques en sciences humaines, un article du journal « Le Monde », le manuel technique d'une imprimante et un extrait du roman de Jules Verne *De la Terre à la Lune*, pour un total de 74 969 mots (Tutin, 2002). En plus d'étudier la distance entre une anaphore et sa source en nombre de phrases, pour découvrir qu'elle est moindre dans les textes techniques et scientifiques et plus importante pour les textes de presse et de littérature, cette analyse a permis d'identifier les catégories d'anaphores les plus productives selon les genres analysés. Celles-ci sont, à savoir, les pronoms personnels (entre 46% et 68,5%), auxquels s'ajoutent les déterminants possessifs, particulièrement fréquents dans l'ensemble des genres étudiés (entre 26% et 36,5% sur trois des quatre genres représentés dans le corpus), à l'exception du manuel technique (4%). Les autres types d'expressions anaphoriques, tels que les pronoms démonstratifs, les pronoms disjonctifs ou les pronoms indéfinis, apparaissent quant à eux dans des proportions nettement plus réduites. Le Tableau 2 contient les valeurs correspondant au pourcentage de chaque type d'expression anaphorique dans les textes étudiés, d'après Tutin (2002).

Textes	Pronoms clitiques	Déterminants possessifs	Pronoms démonstratifs	Pronoms disjonctifs	Ellipses nominales	Pronoms numéraux	Pronoms indéfinis
Articles SH	46	26	10	6	7	0,5	3
Article journal	52	31	6	4	3	2	1,5
Manuel technique	68,5	4	18,5	0	4	0	5,5
Texte littéraire	54	36,5	3	3	2	2	2

Tableau 2. Distribution (%) des types d'expressions anaphoriques selon le genre textuel dans le corpus étudié par Tutin (2002).

Schnedecker reprend par la suite la question du lien entre genre textuel et distribution des types de maillons référentiels. Cette corrélation est notamment examinée dans des travaux portant sur des corpus homogènes, tels que les faits divers (Schnedecker & Longo, 2012) ou sur le portrait journalistique (Schnedecker, 2005). Cette question a également été explorée dans des études comparant plusieurs genres textuels (Boudreau & Kittredge, 2005; Longo, 2013; Schnedecker,

2014). Ces études, réalisées par Schnedecker sur des corpus appartenant à des genres variés, confirment dans une certaine mesure la variabilité de la distribution des types de mentions et suggèrent une possible influence du genre textuel. Toutefois, cette corrélation doit être interprétée avec prudence. En effet, nous pourrions faire cette hypothèse pour des genres fortement codifiés, tels que les recettes de cuisine ou les faits divers, dont les contraintes discursives et pragmatiques favorisent l'emploi récurrent de certaines formes référentielles. En revanche, dans le cas de genres narratifs plus ouverts, la distribution des types de maillons pourrait dépendre davantage de facteurs stylistiques propres à l'auteur ou au texte. Cependant, ce travail étant ciblé sur l'analyse de textes narratifs, nous ne pourrions pas trouver ici une réponse à une telle hypothèse.

Dans cette perspective, certains aspects de l'organisation référentielle, tels que les modalités de démarrage des chaines référentielles, pourraient constituer des indicateurs plus pertinents pour comparer les genres. L'étude de l'incidence du genre textuel sur la distribution des types de maillons permet néanmoins de mieux comprendre la manière dont les différentes pratiques discursives façonnent la cohésion référentielle. Le Tableau 3 présente, à titre illustratif, la distribution des différents types de mentions observée dans plusieurs genres textuels étudiés par Schnedecker et ses collaborateurs.

En particulier, ici nous pouvons observer les distributions obtenues dans le cas de l'annotation et analyse des *incipit* des contes de fées (Schnedecker, 2017). Ce sous-genre présente en effet certaines affinités avec celui représenté dans notre corpus annoté. Les données ont été obtenues à partir de l'annotation des incipit de sept contes de Charles Perrault. Comme le précise Schnedecker : « Il s'agit des incipit de sept contes de Perrault : T1, Le Petit Chaperon rouge (121 mots) ; T2, Cendrillon (114 mots) ; T3, Le Petit Poucet (155 mots) ; T4, Barbe Bleue (258 mots) ; T5, Le Chat botté (221 mots) ; T6, Les Fées (277 mots) ; T7, Riquet à la houppe (159 mots). Le corpus totalise 1 305 mots, 393 expressions référentielles et 23 CR » (Schnedecker, 2017, p. 57).

Ces données offrent ainsi un point de comparaison particulièrement pertinent pour l'analyse de la distribution des types de mentions dans les textes narratifs étudiés dans le présent travail. Dans le cas des incipit de contes de fées, trois catégories apparaissent comme particulièrement représentées. Les syntagmes nominaux définis constituent la catégorie la plus fréquente (16,79 %), ce qui s'explique par leur rôle de désignation à basse accessibilité et par leur utilisation fréquente dans le maintien référentiel. Les pronoms personnels représentent également une proportion importante des mentions (16,5 %). Ils contribuent en effet à assurer la continuité référentielle tout en évitant la répétition des formes nominales et peuvent être utilisés lorsque le référent (généralement un personnage) est hautement accessible dans le texte. Enfin, les syntagmes nominaux indéfinis occupent eux aussi une place significative dans la distribution des mentions (13,23 %). Cette catégorie joue un rôle central dans les incipit narratifs, où elle sert souvent à introduire de nouveaux référents dans l'univers du récit.

Dans le chapitre consacré à nos analyses quantitatives (cf. Chapitre 7, §5), nous examinerons la distribution observée dans notre corpus et évaluerons dans quelle mesure elle correspond à celle décrite par Schnedecker pour les incipit de contes de fées.

	Noms propres	SN Ø	SN indéfinis	SN quantité	SN définis	SN possessifs	Pronoms personnels	Pronoms réfléchis	Pronoms relatifs	Pronoms démonstratifs	Anaphore zéro	SN démonstratifs	Déterminants possessifs	Divers
(Schnedecker, 2005, 2014) portrait journalistique[19]	20,8	/	/	/	7,6	/	38			/	4,3	1,7	21 (son)	6,6
(Schnedecker & Longo, 2012) faits divers*	3,2	1	17	/	32,5	2,5	19	7	3	0,5	1,5	1,3	11,5	/
(Schnedecker, 2014) recettes de cuisine	0,5	29	4	24	28,5	0,5	3,5	0,5	0,5	/	6,5	/	/	2
(Schnedecker, 2014) textes encyclopédiques	21	8,5	8,5	6	38,5	3,5	3	2	4	/	1	/	/	4
(Schnedecker, 2017) incipit conte de fées	0,5	6,6	13,2	6,6	16,8	9,1	16,5	1,7	6,6	/	1	3,3	/	17

Tableau 3. Distribution des mentions par type dans différents genres textuels.

[19] Les chiffres reportés pour le portrait journalistique sont tirés de Schnedecker (2014).

4.4. Modalités de cohabitation des chaines

Comme nous l'avons déjà observé dans les exemples déjà introduits, plusieurs référents sont habituellement présents dans un texte, ce qui se traduit par l'existence de plusieurs chaines au sein du même texte. Selon Schnedecker (1997, p. 14, 2021, p. 100), cinq modalités de cohabitation existent : succession, entrecroisement, dérivation, partition ou fusion et parallélisme. Voyons plus en détail leur définition au travers d'exemples tirés de notre corpus annoté.

Succession : une chaine commence seulement une fois la chaine précédente terminée. Dans le texte de l'exemple (32), la chaine du référent *chat* se termine avant le début de la chaine référée à la *sorcière*.

(32) **[Ce chat]**^{cat} veut des caresses, **[il]**^{cat} est mignon. On voit qu' **[il]**^{cat} veut plein de libertés. **[Cette sorcière]**^{witch} a l' air gentille plus ou moins. **[Elle]**^{witch} peut voler sur **[son]**^{witch} balai magique et **[elle]**^{witch} peut faire des potions : sort de gel, potion de lenteur, de trou noir, de téléportation ou de double. **[Elle]**^{witch} peut faire bonne cuisinière dans les restaurants. *(corpus français, CE2, 1959)*

Entrecroisement : les maillons de deux ou plusieurs chaines différentes alternent dans le texte. C'est le cas dans l'exemple (33), où les chaines du *loup* (ou *loup-garou*) et du *chat* sont instanciées de manière simultanée et leurs mentions alternent au début du texte.

(33) Il était une fois **[un loup-garou]**^{wolf} **[qui]**^{wolf} voulait attaquer **[un chat]**^{cat}. **[Le chat]**^{cat} vivait dans une maison. **[Le chat]**^{cat} était tout seul . Un jour **[le loup - garou]**^{wolf} essayait de casser la maison **[du chat]**^{cat}. Il faisait nuit et **[le chat]**^{cat} **[il]**^{cat} se réveilla. **[Il]**^{cat} sortit pour voir ce qui se passait . **[Il]**^{cat} ne voit rien, **[il]**^{cat} va se recoucher. Le lendemain matin **[il]**^{cat} posa un piège . Et un soir **[le loup - garou]**^{wolf} tombe dans le piège. Le lendemain matin **[le chat]**^{cat} dit « enfin » et **[il]**^{cat} va dans **[sa]**^{cat} maison. *(corpus français, CE1, 2983)*

Dérivation : cette modalité « s'appuie sur une relation de type anaphorique » (Kim, 2023, p. 41). Par exemple, dans le cas de la famille de la sorcière, « la famiglia della sua amica » (la famille de son amie), ce syntagme nominal entretien des relations de type anaphoriques associatives (cf. ce chapitre, §2.5) avec les expressions nominales référées aux membres de cette famille : « la mamma della strega », « la sorella maggiore » et « il papà » (la mère de la sorcière, la sœur ainée, le père), qui apparaissent successivement.

(34) **[Il robot]**^{robot} e **[la strega]**^{witch} C'era una volta **[una strega]**^{witch} **[che]**^{witch} faceva il compleanno, ma **[pensava]**^{witch} che **[il [suo]$^{\mathbf{witch}}$ amico robot]**^{robot} l'avesse dimenticato. **[Il robot]**^{robot} invece non se l'è dimenticato, **[voleva]**^{robot} fare una sorpresa **[alla [sua]$^{\mathbf{robot}}$ amica]**^{witch} allora **[si mise]**^{robot} d' accordo con **[la famiglia della [sua]$^{\mathbf{witch}}$ amica]**^{ext1,ext2,ext3,ext4} e allora **[la mamma della [strega]$^{\mathbf{witch}}$]**^{ext2} "si va bene allora" dice **[la mamma]**^{ext2}. "Dico **[io]**^{ext2} penso che la torta la farà **[la sorella**

maggiore][ext3] e i preparativi li farà **[il papà]**[ext4] invece **[io]**[ext2] farò la cena e **[tu]**[robot] **[robot]**[robot] prendi i palloncini neri" (…). *(corpus italien, CE2, 1437)*

Partition/fusion : une chaine peut se diviser en plusieurs entités ou se fusionner à nouveau dans un référent de type groupe/pluriel. Dans l'exemple (35), les rencontres du chat avec plusieurs animaux différents donnent lieu à un groupe auquel les personnages se rajoutent à fur et à mesure qu'ils avancent dans la narration, donnant lieu à des expressions au pluriel telles que « les deux amis », « tous les animaux » ou à des pronoms au pluriel, « ils », qui regroupent ces référents (la girafe, le lapin, le lion etc.).

(35) **[Un chat tout seul]**[cat] mais un jour **[il]**[cat] rencontra **[un loup]**[wolf] alors **[il]**[cat] courut puis **[se reperdit]**[cat]. **[Il]**[cat] trouva **[une girafe]**[ext1]. **[Il]**[cat] était plus soulagé. Puis **[les deux amis]**[cat,ext1] repartirent puis **[ils]**[cat,ext1] trouvent **[un lapin]**[ext2]. Puis **[ils]**[cat,ext1,ext2] repartirent, **[ils]**[cat,ext1,ext2] trouvèrent **[un lion]**[ext3]. **[Tous les animaux]**[cat,ext1,ext2] disent « au secours, **[le tigre]**[ext4] va **[nous]**[cat,ext1,ext2] manger » puis **[ils]**[cat,ext1,ext2] ont réussi à s'échapper. **[Ils]**[cat,ext1,ext2] retrouvèrent **[le loup]**[wolf] au même endroit puis la faim arrive. *(corpus français, CE1, 2984)*

Parallélisme : si deux référents sont « traités sur un pied d'égalité et présentés comme aussi importants l'un que l'autre » (Schnedecker, 2021, p. 102), il s'agit de chaines parallèles. Cela peut se réaliser par le biais des pronoms corrélés (*l'un… l'autre ; le premier… le second*) (Schnedecker, 2006) ou par des maillons ayant la même fonction syntaxique dans la phrase, par exemple, des syntagmes nominaux avec fonction de sujet (Schnedecker, 2021). L'exemple (36) illustre l'utilisation des pronoms corrélés, référés aux deux personnages de l'extrait, Saccard et Gunderman.

(36) Autrefois, dans les spéculations sur les terrains de la plaine Monceau, Saccard avait eu des discussions, toute une brouille même avec Gundermann. **[Ils]** ne pouvaient s'entendre. **[L'un]**, passionné et jouisseur, [l'autre], sobre et d'une froide logique. Aussi **[le premier]**, dans sa colère, exaspéré encore par cette entrée triomphale, s'en allait-il lorsque **[l'autre]** l'appela. *(Zola, L'argent, extrait et annotation tirés de Schnedecker (2021, p. 102))*

L'exemple (37), issu de notre corpus, illustre un cas de parallélisme entre chaines mis en œuvre à travers l'utilisation de chaines majoritairement nominales : les deux référents sont pour la plupart indiqués à travers de noms propres dans le texte.

(37) C'era una volta **[un robot chiamato Mattia]**[robot] e **[il robot]**[robot] va nel bosco e sull'albero c'era **[un gatto]**[cat] e **[*il* Mattia]**[robot] disse "**[vieni]**[cat] giù" **[al gatto]**[cat] e **[il gatto]**[cat] veniva giù. **[Mattia]**[robot] disse "**[ti]**[cat] **[chiamerò]**[robot] **[Silvestro]**[cat]". **[Mattia]**[robot] e **[Silvestro]**[cat] vanno a casa. **[Mattia]**[robot] dava il latte a **[Silvestro]**[cat] e dopo **[giocavano]**[cat,robot]. *(corpus italien, CE1, 325)*

5. Continuité référentielle

Avant d'aborder plus en détail les trois notions d'*anaphore*, *coréférence* et *chaine de coréférence*, il nous semble important de résumer deux cadres théoriques qui orientent notre approche de ces phénomènes. Dans l'analyse qualitative que nous proposons (cf. Chapitre 8), nous adoptons en effet une perspective fonctionnelle de la référence inspirée notamment de la théorie de l'accessibilité d'Ariel (1990) et de la théorie de continuité référentielle de Givón (1983).

Ces approches considèrent les expressions référentielles non seulement comme des relations formelles entre unités linguistiques, mais également comme des dispositifs permettant de gérer l'introduction, le maintien et la réactivation des référents dans le discours. Dans cette perspective, le choix d'une forme référentielle dépend notamment du degré d'accessibilité cognitive du référent pour l'interlocuteur et de son statut dans le discours.

En nous inspirant également des travaux réalisés dans le cadre du corpus *RésolCo* (cf. Chapitre 2, §3.1), nous mobilisons plus particulièrement la notion de *continuité référentielle*. Cette notion permet d'appréhender la dynamique discursive de la référence au-delà de la seule relation de coréférence au sens strict. En outre, dans la perspective de l'opérationnalisation de ces notions pour l'annotation de notre corpus, l'adoption de cette perspective facilite la comparaison ultérieure de nos annotations avec celles réalisées sur le corpus *RésolCo* (Garcia-Debanc et al., 2021).

5.1. Théorie de l'accessibilité de Ariel

La théorie de l'accessibilité d'Ariel (1990) trouve sa place au sein des théories développées dans le périmètre des approches fonctionnelles de la grammaire.

Cette théorie a comme objectif de décrire la manière dont les locuteurs et les scripteurs choisissent différents types d'expressions référentielles en fonction du degré d'accessibilité (ou d'activation) mentale d'un référent donné dans la mémoire de l'interlocuteur. Selon cette théorie, l'accessibilité mesure à quel point un référent est facilement récupérable par l'interlocuteur, en fonction de facteurs divers tels que la distance depuis sa dernière mention, sa saillance, la compétition avec d'autres référents ou l'unité du discours dans laquelle elle se trouve.

Ariel développe sa théorie (1990, 1991) à partir de la notion d'accessibilité telle que décrite par la *Relevance Theory* de Sperber et Wilson (1996) (théorie de la pertinence), inspirée à son tour des maximes de Grice (1975). Ces auteurs décrivent la variation dans la disponibilité d'un contexte donné à des moments différents d'un discours, et de l'effort correspondant demandé aux locuteurs pour introduire, réintroduire ou maintenir un référent selon son degré d'accessibilité dans le discours précédent. Par exemple, un référent qui vient d'être nommé dans l'énoncé précédent sera largement plus accessible qu'un référent qui a été introduit plusieurs phrases auparavant dans le texte.

Ariel enrichit cette approche en développant une théorie complète de l'accessibilité référentielle qui vise à décrire le rôle joué par la notion d'accessibilité dans la sélection d'un marqueur donné,

c'est-à-dire le type d'expression référentielle utilisée dans le discours. Ce choix est déterminé par « l'état d'accessibilité que la représentation mentale du référent est supposée avoir pour le destinataire à l'étape actuelle du discours »[20] (Ariel, 2014, p. 69, notre traduction). L'accessibilité d'un référent dépend ainsi de plusieurs facteurs, parmi lesquels Ariel identifie notamment la *distance*, la *compétition*, la *saillance* et *l'unité*.

La *distance* indique l'écart entre une anaphore et son antécédent, ou entre deux expressions référentielles. Elle peut être mesurée en termes de tokens ou de phrases. Plus cette distance augmente, moins le référent est accessible dans le discours, puisqu'il est moins facile à récupérer dans la représentation mentale du texte par l'interlocuteur. À l'inverse, lorsque l'antécédent est proche dans le discours, le référent peut être récupéré plus facilement. Ce critère est aussi le seul véritablement analysé dans l'ouvrage de Ariel (1990), car il représente le seul critère que l'on peut quantifier dans l'analyse empirique.

Dans l'exemple (38), le référent *chat* peut être repris à l'aide de pronoms (d'abord un pronom relatif, puis un pronom personnel) en raison de la proximité de l'antécédent dans la phrase. De même, la chaine du *loup* démarre par un syntagme nominal immédiatement repris par un pronom relatif. Dans les deux cas, la distance entre ces expressions étant minimale, voire nulle, les reprises pronominales sont non seulement possibles, mais également préférentielles. La faible distance favorise ainsi un haut degré d'accessibilité du référent dans la représentation discursive, conformément aux prédictions de la théorie de l'accessibilité.

(38) Il était une fois **[un petit chat blanc]**^{cat} **[qui]**^{cat} vivait dans la forêt. Un jour que **[le petit chat]**^{cat} se promenait et **[il]**^{cat} vit **[un loup]**^{wolf} **[qui]**^{wolf} était noir. *(corpus français, CE1, 212)*

La *compétition* entre antécédents est définie comme « le nombre de compétiteurs pour le rôle d'antécédent » (Ariel, 1990, p. 28). Cela renvoie au nombre de référents susceptibles d'être rattachés à une expression anaphorique dans un contexte donné. Si des traits morphosyntaxiques tels que le genre et le nombre peuvent contribuer à la désambiguïsation du référent, la présence de plusieurs candidats partageant les mêmes caractéristiques diminue leur accessibilité et peut introduire une interférence dans l'identification du référent, rendant alors nécessaire l'emploi d'expressions référentielles plus informatives pour permettre de désambiguïser correctement.

Si dans un texte nous retrouvons plusieurs référents du même genre, cela peut donc provoquer des interférences dans le rattachement à une chaine donnée si l'on utilise des expressions peu informatives. Au contraire, l'ambiguïté de rattachement peut être évitée en utilisant des formes plus informatives telles que les syntagmes nominaux, comme c'est le cas de l'exemple (39).

Dans ce texte, deux personnages masculins sont introduits dans le récit : un *loup-garou* et un *chat*. Ces deux référents étant compatibles avec la forme pronominale *il*, l'emploi d'un pronom

[20] Dans le texte d'origine, « (…) In fact, the choice of a referring expression is dependent on the Accessibility status the mental representation of the referent is assumed to have for the addressee at the current stage of the discourse. »

personnel pourrait potentiellement créer une ambiguïté dans l'identification du référent. On observe ainsi que l'élève auteur du texte recourt régulièrement à des syntagmes nominaux définis tels que « le chat » ou « le loup-garou ». L'utilisation de ces formes plus informatives permet de maintenir une distinction claire entre les deux chaines référentielles et d'éviter toute confusion dans l'interprétation du texte.

(39) Il était une fois **[un loup-garou]**^{wolf} **[qui]**^{wolf} voulait attaquer **[un chat]**^{cat}. **[Le chat]**^{cat} vivait dans une maison. **[Le chat]**^{cat} était tout seul. Un jour **[le loup-garou]**^{wolf} essayait de casser la maison **[du chat]**^{cat}. Il faisait nuit et **[le chat]**^{cat} **[il]**^{cat} se réveilla. **[Il]**^{cat} sortit pour voir ce qui se passait. **[Il]**^{cat} ne voit rien, **[il]**^{cat} va se recoucher. Le lendemain matin **[il]**^{cat} posa un piège. Et un soir **[le loup-garou]**^{wolf} tombe dans le piège. Le lendemain matin **[le chat]**^{cat} dit « enfin » et **[il]**^{cat} va dans **[sa]**^{cat} maison. *(corpus français, CE1, 2983)*

La *saillance* représente la proéminence cognitive d'un référent dans la représentation mentale du discours partagée entre locuteur et destinataire. « Être saillant, c'est ressortir particulièrement, au point de capter l'attention et de donner une accroche, un point de départ à la prise de connaissance et à la comprehénsion », selon Landragin (2020a, paragr. 25).

Différentes théories existent par rapport à la manière dont un référent devient saillant dans nos représentations mentales. Entre ces théories, l'hypothèse de la saillance relative (ou graduée) de Giora (2003) définit cette dernière comme le degré d'accessibilité cognitive d'un référent pour le locuteur ou récepteur d'un discours. Un référent devient saillant dès qu'il est enregistré dans la mémoire d'un locuteur. Cette théorie permet aussi de « rendre compte du mécanisme de choix du bon sens lexical : les différents sens (…) sont hiérarchisés dans la mémoire (…) plus un sens a été utilisé pour un mot considéré, plus ce sens est saillant. L'intérêt est que la saillance permet d'ordonner des hypothèses, de même que la saillance référentielle ordonnait par ordre de plausibilité les référents potentiels. » (Landragin, 2020a, paragr. 32) Des référents familiers, fréquents, prototypiques et conventionnels seront plus accessibles que des référents qui ne possèdent pas ces qualités. Plusieurs facteurs peuvent influencer la saillance, entre autres le statut de référent animé peut rendre un référent plus saillant qu'un non animé ou abstrait. En outre, la fréquence de mention et la *persistance topicale* augmentent le degré de saillance d'un référent. Givón (1983) définit la *persistence* dans son ouvrage comme un indice corrélé à la distance intermaillonnaire. Cet indice permet d'évaluer l'importance d'un référent dans un texte. Un référent est dit *continu* lorsque la distance entre deux mentions n'excède pas les 20 phrases ; il est défini *persistant* s'il est immédiatement réinstancié, alors que dans le cas contraire il sera dans une condition de persistance *minimale*, aussi dite « désintégration immédiate ».

L'*unité* représente la cohérence du cadre discursif dans lequel les locuteurs emploient les expressions référentielles. Ce concept englobe la continuité spatio-temporelle, énonciative, textuelle et conceptuelle du discours. Plus l'unité est forte (le discours représente la même scène, on est situé dans le même paragraphe ou dans la même phrase), plus l'accessibilité du référent est élevée. Les ruptures d'unité (changements de cadre, de paragraphe ou de monde possible) nécessitent des expressions référentielles plus informatives pour indiquer le même référent.

La différence d'accessibilité marquée par les différents types d'expressions est représentée par Ariel comme une échelle (cf. Figure 2), un continuum qui va des référents à basse accessibilité, comme les noms propres et les expressions définies (où le référent est censé être moins présent dans l'esprit du lecteur/récepteur du discours et une expressions plus précise est nécessaire pour l'indiquer), jusqu'aux marqueurs à plus haute accessibilité, comme les pronoms personnels (le référent doit être très saillant, très proche ou avec peu de compétition avec d'autres référents pour être indiqué par un pronom ou une expression référentielle peu marquée).

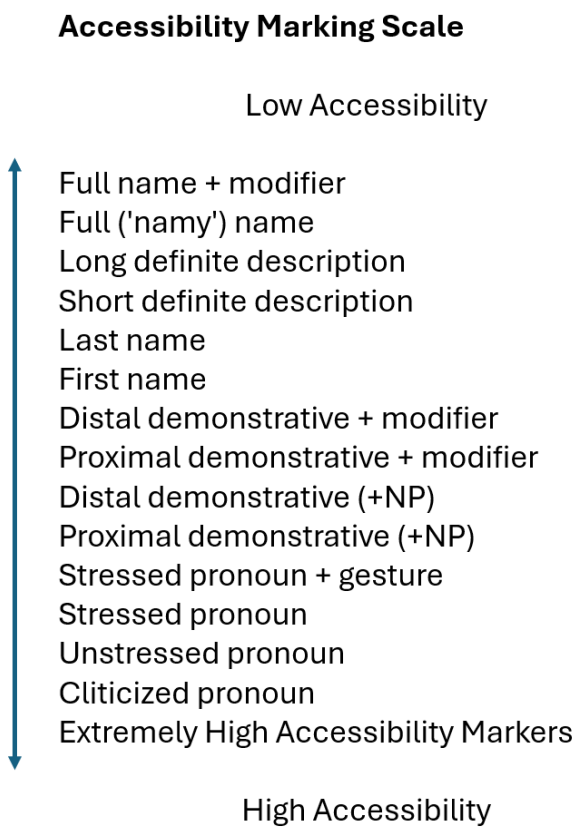


Figure 2. Échelle de l'accessibilité de Ariel (1991)

Trois principes fondamentaux guident ainsi le choix des marqueurs d'accessibilité (autrement dit des expressions référentielles) selon Ariel ; il s'agit de l'*informativité*, la *rigidité* et l'*atténuation*.

L'*informativité* (*Informativity*) est définie comme le degré d'information sémantique contenue dans une expression référentielle. Dans cette perspective, les expressions associées à une faible accessibilité du référent sont celles qui contiennent généralement le plus d'informations, car elles jouent le rôle de sources sémantiques permettant d'identifier le référent. Comme le souligne Ariel, « ce sont des dispositifs de récupération qui renvoient le destinataire vers des sources moins accessibles, dans des cas où il y a également des nombreux compétiteurs »[21] (Ariel,

[21] Dans le texte d'origine, « They are retrieval devices which refer addressees to less accessible sources, where there are many competitors too. »

2014, p. 80, notre traduction). Ainsi, les marqueurs tels que les syntagmes nominaux complets, éventuellement accompagnés de modificateurs, se situent du côté de la faible accessibilité du référent. Plus un référent est difficile à récupérer dans le discours, plus il devient nécessaire de recourir à une expression riche en contenu sémantique afin de permettre son identification.

On peut observer cela dans l'exemple (40), déjà cité, tiré du conte *Chaperon Rouge*. Dans ce texte, le personnage du Chaperon Rouge est repris à l'aide d'un syntagme nominal défini à la 4e phrase du texte : cela parce que le dernier syntagme nominal l'indiquant se trouve deux phrases auparavant (ce qui nous renvoie au critère de distance que nous avons déjà analysé), et qu'entretemps deux autres chaines de référence du même genre (féminin) ont cohabité avec la chaine du Chaperon Rouge dans le texte, à travers la présence de maillons nominaux, pronominaux et verbaux.

(40) Il était une fois **[une petite fille de village, la plus éveillée qu'on eût su voir]**[a] :
[[sa][a] **mère]**[b] en était folle, et **[[sa]**[a] **mère-grand]**[c] plus folle encore.
[Cette bonne femme][c] **[lui]**[a] fit faire un petit chaperon rouge, qui **[lui]**[a] seyait si bien, que partout on **[l']**[a] appelait **[le petit Chaperon rouge]**[a].
Un jour, **[[sa]**[a] **mère]**[b], ayant fait des galettes, **[lui]**[a] dit : « **[Va]**[a] voir comment se porte **[[ta]**[a] **mère-grand]**[c] ; car on **[m']**[b]**a dit qu'[elle]**[c] était malade : **[porte]**[a]**-[lui]**[c] une galette et ce petit pot de beurre. »
[Le petit Chaperon rouge][a] partit aussitôt pour aller chez **[[sa]**[a] **mère-grand]**[c], **[qui]**[c] demeurait dans un autre village.
En passant dans un bois, **[elle]**[a] rencontra **[compère le Loup]**[d], **[qui]**[d] eut bien envie de **[la]**[a] manger ; mais **[il]**[d] n'osa, à cause de **[quelques bûcherons]**[e] **[qui]**[e] étaient dans la forêt. *(Perrault, 1826)*

D'un autre côté, des formes telles que les pronoms représentent les formes de plus haute accessibilité ainsi que les moins informatives. Ces formes permettent de récupérer un référent seulement si sa dernière mention est assez proche et assez saillante dans la représentation mentale du lecteur. Dans le texte de l'exemple (40), le Petit Chaperon Rouge est le sujet d'une reprise pronominale dans la phrase 5 pour plusieurs raisons : pour saillance narrative, car on sait que entre tous les personnages féminins du conte c'est elle qui se met en chemin pour rejoindre sa grand-mère, donc il n'y a pas de risque de penser que ce pronom puisse reprendre les autres personnages féminins du conte ; pour saillance grammaticale, car c'est elle le sujet de la dernière phrase principale.

La rigidité (*Rigidity*) représente la capacité d'une expression à indiquer de manière non ambigüe un référent donné dans un contexte donné (Ariel, 1990, p. 81), c'est-à-dire à quel point une expression peut indiquer un référent donné sans l'aide du contexte textuel. Une expression avec une rigidité haute permet d'identifier un référent avec un support minimal du contexte alors qu'une expression avec une faible rigidité a besoin du contexte pour être désambiguïsée. Dans l'échelle d'accessibilité de Ariel, les formes les plus rigides permettent donc d'indiquer ces référents qui présentent une accessibilité mineure. Ce critère se superpose partiellement avec la richesse lexicale et donc avec le critère d'informativité que nous avons décrit auparavant.

Le dernière critère, l'atténuation (*Attenuation*) correspond à la quantité d'information contenue dans une expression. Ce contenu est directement proportionnel au contenu phonique de l'expression, c'est-à-dire à la longueur de sa forme de surface, ou autrement dit « Le plus informative est l'expression, le plus elle nécessite de mots pour être exprimée, toutes choses égales par ailleurs »[22] (Ariel, 2014, p. 81, notre traduction). Cette notion est similaire à celle de *phonological size* de Givón (1983).

La théorie que nous avons présentée ici nous semble constituer un cadre théorique opérationnel pour établir certains des critères de description quantitative et qualitative que l'on peut appliquer à nos analyses. En effet, les principes de distance, compétition et saillance peuvent être traduits en indicateurs quantifiables, ce qui permet leur mise en œuvre dans une analyse empirique. Cette approche nous permet non seulement d'analyser les distances entre maillons des chaines et le lien avec le type de mention utilisée, mais il va nous permettre aussi d'identifier les sources potentielles d'ambiguïté ou de rupture dans la cohérence référentielle, et également de quantifier les relations entre les différents types d'expressions utilisées et leur degré d'accessibilité dans la chaine à l'aide de l'analyse des types de mentions. L'application de ce modèle théorique à notre corpus d'étude, est approfondie dans la partie 3 de ce travail, dédiée aux résultats ; cela pourrait ainsi permettre de décrire des patterns significatifs dans l'usage des expressions référentielles et contribuer à une meilleure description des stratégies utilisées pour assurer la cohésion textuelle dans la production de l'écrit narratif à l'école primaire.

5.2. Théorie de la continuité référentielle de Givón

L'adoption de la notion de continuité référentielle comme objectif de notre annotation nous a été fortement inspirée des travaux effectués sur le corpus *RésolCo* (Garcia-Debanc et al., 2017, 2019, 2021). Ce corpus, composé d'écrits d'élèves à des niveaux scolaires différents, est annoté selon les chaines de continuité référentielle limitées aux personnages de la consigne[23]. Etant donné la similarité des objectifs de notre annotation avec celle effectuée sur le corpus *RésolCo*, nous avons fait le choix d'adopter un cadre théorique similaire. Mais si l'adoption de cette notion nous a été inspirée par ce projet, nous allons voir par la suite comment nous avons intégré notre interprétation de cette notion dans le cadre de ce travail et comment elle a été reprise dans nos schémas et guide d'annotation (cf. Chapitre 5, §3).

La continuité référentielle selon Givón peut être définie comme l'ensemble des procédés linguistiques qui assurent la cohésion textuelle en permettant l'introduction, le maintien et la reprise des référents dans le discours. Parmi ces procédés figurent notamment les différentes expressions référentielles utilisées pour maintenir ou réactiver un référent au fil du texte. Givón (1983) théorise l'existence des sujets d'un discours sur une échelle de prévisibilité ou de continuité dans le texte, continuité qui peut être mesurée à travers des indices grammaticaux et discursifs. La continuité référentielle (*referential continuity*) fait donc référence au degré de maintien d'un référent dans un discours ou dans un texte donné. Givón a aussi proposé une

[22] Dans le texte d'origine, « The more informative the expression the more wording it takes, other things being equal. »
[23] La constitution et la méthodologie d'annotation du corpus *RésolCo* seront traitées de manière plus approfondie dans le Chapitre 2, §3.1.

échelle de topicalité (Givón, 1983), qui décrit le degré d'accessibilité du référent selon le type d'expression utilisé pour l'indiquer. Cette échelle propose un mélange entre critères « aussi bien catégoriels que structurels » (Schnedecker, 1997, p. 64).

La Figure 3 présente cette échelle de topicalité, dans laquelle les expressions référentielles sont ordonnées selon leur degré d'accessibilité. Les formes situées en haut de l'échelle correspondent aux référents les plus continus ou les plus accessibles dans le discours, tandis que celles situées vers le bas signalent des référents moins accessibles ou nouvellement introduits. Ainsi, les formes les plus réduites (l'anaphore zéro ou les pronoms clitiques) sont généralement utilisées lorsque le référent est hautement accessible et déjà fortement établi dans le discours. Par exemple, dans une séquence narrative comme « Le chat entra dans la maison et [se coucha] près du feu », l'anaphore zéro renvoie au référent *chat*, dont la continuité dans le discours est maximale. À l'inverse, lorsque l'accessibilité du référent diminue ou que plusieurs référents concurrents sont présents dans le contexte, des formes plus explicites tendent à être employées, comme les syntagmes nominaux définis (le chat, la sorcière) ou les constructions de focalisation. Enfin, les syntagmes nominaux indéfinis apparaissent généralement dans les positions les plus basses de l'échelle, puisqu'ils servent souvent à introduire de nouveaux référents dans le discours, comme dans « Un loup apparut dans la forêt ». Cette échelle permet ainsi de rendre compte de la relation entre la forme linguistique utilisée et le degré de continuité du référent dans le discours.

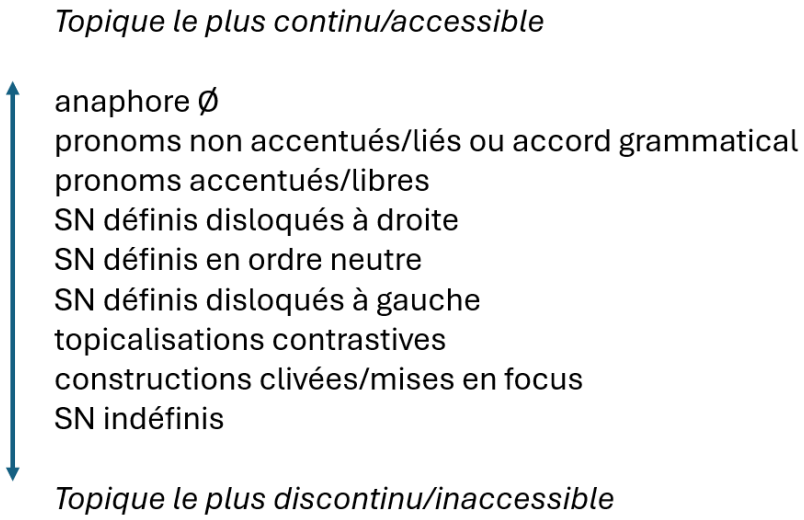


Figure 3. Échelle de topicalité de Givón (1983). Traduction de Schnedecker (1997).

Les approches de Givón et d'Ariel s'inscrivent dans une même tradition fonctionnelle de l'analyse de la référence en discours. Toutes deux reposent sur l'idée selon laquelle la forme des expressions référentielles est étroitement liée au statut discursif du référent : plus celui-ci est accessible ou continu dans le discours, plus les formes utilisées pour le désigner tendent à être réduites (par exemple des pronoms ou des anaphores zéro), tandis que les référents moins accessibles nécessitent des expressions plus explicites, comme des syntagmes nominaux complets.

Chez Givón, l'un des apports majeurs consiste à introduire le rôle de la distance discursive par rapport à la mention précédente pour expliquer le choix des formes référentielles dans le discours. Plus la reprise est éloignée de la mention antérieure, plus la forme utilisée tend à être informative ; à l'inverse, lorsque la reprise est proche, les formes employées sont généralement moins informatives. Cette relation s'explique également par le fait que l'augmentation de la distance discursive accroît la probabilité que d'autres référents aient été introduits dans l'intervalle, ce qui peut engendrer une situation de compétition référentielle.

Voyons maintenant comment nous avons catégorisé les différentes typologies de mentions existantes selon le critère syntaxique, qui relève des distinctions utilisées par Ariel et Givón dans leurs théories ainsi que des différentes catégorisations existantes en littérature.

6. Typologie des mentions dans les chaines de référence

Dans une chaine de référence, nous retrouvons des types de mentions différentes. Celles-ci peuvent être catégorisées selon plusieurs critères : soit selon la ou les parties du discours contenue dans la mention (nom commun, nom propre, pronom etc.) soit par la représentation de surface de l'entité mentionnée (par exemple le lemme choisi pour l'identifier) ou par la nature sémantique de la mention choisie (hyperonyme, synonyme etc.) (Žitkus et al., 2024). Ce premier critère de catégorie grammaticale a été utilisé par Federzoni et al. (2021) dans leur travail sur la catégorisation des chaines de coréférence pour décrire les différentes catégories de mentions que l'on peut détecter dans le corpus *Annodis* (Péry-Woodley et al., 2011). Dans ce travail de recherche, huit catégories de mentions ont été reconnues, catégories qui se recoupent avec celles que nous allons présenter par la suite, en faisant référence en particulier aux données issues du corpus objet de notre étude. Cette catégorisation présente également des points de convergence avec celle décrite par Ariel dans sa théorie de l'accessibilité (cf. §5.1), qui établit une hiérarchie des expressions selon le degré d'accessibilité des référents dans le discours. Toutefois, la définition de ce qui est effectivement une mention a fait l'objet de plusieurs évolutions dans la littérature. Selon une définition donnée initialement par Charolles (1988c), seules les expressions nominales ou pronominales susceptibles d'identifier une entité du texte peuvent être considérées comme des mentions. Une description ultérieure (Charolles, 2002) étend cette caractéristique aux noms propres, tout en incluant les pronoms et en divisant les mentions nominales en SN définis, indéfinis et démonstratifs.

Cette description demeure néanmoins plus restreinte que celle proposée par Ariel (1990), dont l'approche, plus fonctionnelle, prend en compte un éventail plus large d'expressions référentielles. Dans le cadre du présent travail, nous adoptons une taxonomie intermédiaire, qui regroupe sous une même catégorie syntaxique des expressions présentant des propriétés similaires dans les langues étudiées.

Dans les exemples présentés dans les sections suivantes, seules les expressions appartenant à la catégorie analysée seront indiquées à l'aide de crochets **[ ]**.

6.1. Les noms propres

Les noms propres représentent des marqueurs très informatifs, très rigides et de basse accessibilité selon Ariel (1990, p. 36-41). Ils sont facilement identifiés à l'aide de leur premier caractère toujours en majuscule. Ils permettent d'identifier un référent spécifique et unique, « un particulier dans tous ses états » (Charolles, 2002, p. 55). Ce type de mention permet aussi de doter un référent de saillance à l'intérieur d'un discours donné (Schnedecker, 1997). Plusieurs types de noms propres existent, catégorisés selon le référent identifié (anthroponymes pour les noms de personnages, toponymes pour les noms de lieu, etc.). Dans le cadre de notre travail, nous nous intéressons plus particulièrement aux anthroponymes qui indiquent des personnages dans les textes de notre corpus. Ce type de noms propres caractérise souvent les personnages principaux d'un texte narratif ou ces personnages récurrents ou saillants.

Les noms propres de personnes s'emploient habituellement sans déterminant[24] (Charolles, 2002, p. 53). Le nom propre peut constituer une mention en lui-même, comme le nom « Bobis » dans l'exemple (42). On peut également le retrouver dans une expression nominale, définie ou indéfinie, formée par une juxtaposition du type de personnage au nom propre, comme dans les exemples (41) et (43), où l'exemple (43) contient le nom propre à l'intérieur d'un SN défini plus élargi. L'exemple (43) est d'ailleurs représentatif d'une structure de type *SN indéfini descriptif + nommé + nom propre* (où l'on considère que « di nome » est l'équivalent de « nommé » en français), que l'on peut retrouver dans les deux langues du corpus. Les noms propres peuvent également être introduits à travers une formule de présentation : dans les exemples (42) et (44), cette formule (en italique dans les exemples) est composée d'un syntagme nominal (défini comme « le robot » ou indéfini comme « un robot ») qui fait référence à un personnage de la narration suivi par un verbe réfléchi, *chiamarsi* (s'appeler en français).

(41) Il était une fois une sorcière . elle avait un chat magique, le chat avait le pouvoir de changer de couleur. La maison de **[la sorcière Gigi]** était noire donc vu que **[le chat Victor]** pouvait changer de couleur, Gigi tomba, mais un jour elle a dit « j'en ai marre oust. » (…) *(corpus français, CE2, 212)*

(42) Maitre chat a un robot, il nettoie tout, même le grenier. *Le robot s'appela* **[Bobis]**. (…) *(corpus français, CE2, 214)*

(43) La strega felice. Un giorno che non era ancora giorno **[una vecchia strega di nome Riunione]** si svegliò alle 7:00 per vestirsi, fare colazione e lavare i denti. (…) *(corpus italien, CE1, 11)*

(44) C'era una volta *un robot che si chiamava* **[Arbo]**. Poi c'era *una strega che si chiamava* **[Eva]**. (…) *(corpus italien, CE1, 1376)*

[24] Sauf dans certains cas particuliers, dont Charolles (2002) donne des exemples : les emplois dénominatifs, de fractionnement, exemplaires, métaphoriques, quantifiés et métonymiques. Dans certaines variétés diatopiques de l'italien (italien du Nord de l'Italie), les noms propres de personne peuvent être introduits par un déterminant défini (ex. **la Carlotta*.

6.2. Les expressions nominales définies

Ce sont des syntagmes nominaux introduits par des déterminants définis. Dans le cas de ce type de mention, nous pouvons relier l'expression à son référent sur la base du contexte d'énonciation. Ces syntagmes peuvent être divisés en deux catégories, *complètes* ou *incomplètes*. Dans le cas des expressions complètes, elles « ne sont valides que pour un seul référent ou un ensemble de référents » déterminé (Charolles, 2002, p. 75). Ce sont des expressions habituellement hautement descriptives, qui peuvent soit fournir plus d'informations sur un référent donné, soit introduire un nouveau référent à l'aide d'une définition de ses caractéristiques, comme dans le cas de la mention « la strega cattiva » (la méchante sorcière) de l'exemple (45).

Les expressions définies incomplètes, en revanche, indiquent de manière moins stricte un référent donné : leur utilisation implique que le référent soit déjà d'une certaine manière « activé » (Salles, 2015), elles seront moins souvent utilisées en première mention de la chaine. C'est le cas pour la mention « la strega » que l'on retrouve dans l'exemple (45).

(45) Una sera **[la strega cattiva]** trovò un gatto bianco molto vivace, **[la strega]** se lo porta via in una palude lontana dove aveva una casa di legno con tante pozioni di tutti i tipi. (…) *(corpus italien, CM2, 32)*

6.3. Les expressions nominales indéfinies

Les expressions nominales indéfinies sont toujours introduites par un article indéfini. Ces expressions sont autonomes du point de vue de la référence, et peuvent introduire un référent perçu comme entièrement nouveau pour l'interlocuteur (Manuélian & Schnedecker, 2021). Corblin (1994) définit comme « la condition de nouveauté » celle qui consiste « à présenter un syntagme nominal comme un terme de première mention, présumé non introduit pour l'interlocuteur » (Manuélian & Schnedecker, 2021; Muller, 2019, p. 16). Ces expressions se retrouvent donc souvent en position de premier maillon de la chaine, pour introduire un nouveau personnage au sein du texte narratif, comme dans le cas des exemples (46) et (47) ou bien de la mention « una vecchia strega di nome Riunione » dans l'exemple (43) vu précédemment.

(46) C'est l'histoire d'**[un chat]** très malin qui vit dans un château (…). *(corpus français, CE2, 584)*

(47) Un jour une navette spatiale s'écrase à Chatville, le pays des chats quand descendit **[un drôle de bonhomme fait en un métal et en clous]** (…). *(corpus français, CE2, 2000)*

6.4. Les expressions nominales démonstratives

L'utilisation du syntagme démonstratif fait toujours référence à une entité déjà introduite dans le discours et il est perçu comme expression *focalisante*, donc qui détermine une entité en particulier et qui remet sa présence au centre de l'attention du lecteur. Dans l'exemple (48), le référent *petite fille* est introduit dans le discours à travers une expression nominale, introduite par une préposition ; celui-ci est ensuite repris sous forme d'expression nominale démonstrative « cette petite fille ». Dans l'exemple (49), le référent *chat* nommé en début de texte est refocalisé

à travers une expression démonstrative qui précède une description dans la dernière phrase du texte.

(48) Un jour la sorcière **[le]** transforma en **[petite fille]**, et la sorcière trouva **[cette petite fille]** très gentille et très mignonne. (…) *(corpus français, CE2, 1931)*

(49) C'era una volta **[un gatto carino] [che]** si chiamava **[Ghiacciolina]** (…). E **[questo gatto]** è bianco con pochissimi pallini marroni. (…) *(corpus italien, CE1, 627)*

6.5. Les déterminants possessifs

Comme déjà décrit auparavant (cf. ce chapitre, §2.8 sur les Anaphore possessives), les déterminants possessifs indiquent la présence d'un possesseur/agent et sont donc indicatifs d'anaphore, comme dans les exemples (50) et (51). Ces déterminants introduisent souvent un syntagme nominal indiquant un autre référent.

(50) Dans une forêt il y avait une maison où habitaient [une sorcière]^{witch} et **[son]**^{witch} chat. (…) *(corpus français, CE2, 855)*

(51) Il y avait une fois [une sorcière appelée Camille]^{witch} [qui]^{witch} vivait heureuse avec **[son]**^{witch} chat mignon, qui s'ennuyait. (…) *(corpus français, CE2, 2978)*

Dans cet exemple, la mention « son chat mignon » introduit un nouveau référent dans le discours, le *chat*, mais contient aussi une référence au référent précédemment mentionné, « la *sorcière* Camille ». Cela est lié à la notion d'ancrage définie par Prince (1981) : celle-ci distingue les entités *brand-new*, nouvellement introduites dans le texte, les entités *brand-new anchored*, c'est-à-dire introduites en s'appuyant sur un ancre, déjà accessible dans le texte, et les entités *brand-new unanchored*, introduites de manière autonome. Dans sa définition, en fait, une entité du discours est ancrée si un SN qui la représente est lié à un autre SN déjà présent dans le discours, qui constitue son ancre. Cette ancre doit toujours être déjà accessible dans le texte.

En français, le déterminant possessif introduit généralement directement un syntagme nominal. En italien, en revanche, il est le plus souvent précédé par un autre déterminant, généralement un article défini ou indéfini, comme dans l'exemple (52)[25]. Dans cet extrait, les nouveaux référents *Nella* et *Gimi* sont introduits au moyen de syntagmes nominaux possessifs tels que « la loro amica Nella » (« leur amie Nella ») et « il suo amico gattino Gimi » (« son ami le chaton Gimi »). Cette structure reflète la règle grammaticale de l'italien selon laquelle le déterminant possessif apparaît généralement en cooccurrence avec un déterminant.

(52) C'era una volta **[una strega]**^{witch} **[che]**^{witch} si chiama **[Marina]**^{witch} , **[era]**^{witch} sempre da sola e **[andò]**^{witch} a prendere **[un gatto]**^{cat} che **[lo]**^{cat} chiamò **[Sregino]**^{cat} e **[lo]**^{cat} **[portò]**^{witch} a casa **[sua]**^{witch} **[della strega Marina]**^{witch} e **[gli]**^{cat} **[dà]**^{witch} da mangiare il pesce e un po' di acqua e poi **[andavano]**^{cat,witch} a fare una passeggiata

[25] En italien, le déterminant possessif est normalement précédé d'un article défini ou indéfini (il mio libro, mon livre ; un suo amico, un de ses amis). Une exception concerne les noms de membres de la famille au singulier, qui peuvent apparaître sans article, par exemple suo padre, sua madre ou suo fratello (son père, sa mère, son frère).

e **[videro]**[cat,witch] **[la [loro]**[cat,witch] **amica Nella]**[ext1] con **[il [suo]**[ext1] **amico gattino Gimi]**[cat1] (…). *(corpus italien, CE1, 1495)*

6.6. Les pronoms personnels

Les pronoms jouent un rôle syntaxique bien défini, car ils sont liés au contexte d'énonciation et nécessitent la présence d'un référent explicite pour qu'on puisse leur attribuer un rôle sémantique. Ils représentent donc des expressions anaphoriques (cf. ce chapitre, §2.1) ou ils peuvent également avoir fonction de déictique (cf. ce chapitre, §2.10), comme dans l'exemple (53), où les pronoms personnels utilisés dans le discours direct ont fonction de déictique dans leur référence au chat ou à la sorcière (mentions « tu », « je » et « moi », en gras et en italique dans l'exemple).

(53) Il était une fois **[une sorcière]**[witch] **[qui]**[witch] vivait dans une maison hantée. Et dehors **[elle]**[witch] trouve **[un petit chat noir]**[cat] avec **[une souris]**[ext1] dans la bouche et **[elle]**[witch] **[lui]**[cat] dit « ***[tu]***[cat] vas cracher **[cette souris]**[ext1]. Pauvre ***[tu]***[cat] veux des croquettes et du lait ? » « ***[j']***[cat]aime le lait » « viens dans **[ma]**[witch] maison hantée. ***[Tu]***[cat] seras bien sage, ***[tu]***[cat] viens manger avec ***[moi]***[witch], ***[tu]***[cat] prendras un goûter » (…). *(corpus français, CE1, 2986)*

Dans le cas où les pronoms indiquent un référent à la troisième personne, ils ont habituellement un rôle anaphorique, comme dans les exemples (54) et (55). Néanmoins, les pronoms étant entre les marqueurs à plus haute accessibilité selon la hiérarchie de l'accessibilité d'Ariel (cf. ce chapitre, §5.1), ils doivent se référer à des référents très saillants (ou très proches) dans le texte. Ils représentent également les expressions plus compliquées à désambigüiser dans le cas de compétition entre référents.

(54) La sorcière et sa chatte font une potion magique avec des plantes et des serpents et des myrtilles tellement [**elles]** étaient les plus gentilles et des fois étaient méchantes. (…) *(corpus français, CE2, 576)*

Cependant, en français la présence des pronoms impersonnels dans un texte, vides de sens ou du moins ne renvoyant pas à un référent déterminé, peut être problématique pour l'annotation de chaines de référence. Ce type de pronom est très présent dans le corpus analysé dans la formule d'incipit des productions écrites comme dans le cas du pronom « il » dans l'exemple (55).

(55) **[Il]** était une fois un chat qui se promenait dans la forêt et qui ramassait des fleurs. (…) *(corpus français, CE2, 2531)*

6.7. Les pronoms réfléchis, les pronoms relatifs et les anaphores zéro

Ces trois types de mentions, à des degrés différents, représentent des mentions moins informatives et moins marquées que celles décrites précédemment. Elles sont donc utiles pour la reprise de référents plus accessibles dans le texte, avec l'anaphore zéro localisée à l'extrémité de l'échelle d'accessibilité d'Ariel (cf. §5.1) comme forme qui indique la plus haute accessibilité possible d'un référent.

En ce qui concerne les pronoms réfléchis, ceux-ci ne sont pas considérés comme coréférents dans le cas de constructions où ils sont coréférents avec le sujet de la phrase où ils sont contenus. Cela est le cas aussi dans le guide d'annotation du corpus *DEMOCRAT*, qui spécifie que les pronoms réfléchis clitiques ne sont pas annotés, comme dans l'exemple (56), où « se » n'est pas à annoter comme coréférent.

(56) Il **se** lave.

En revanche, les pronoms clitiques qui signalent un référent autre que celui indiqué par le sujet du verbe peuvent représenter la présence d'un référent au sein de la construction linguistique analysée, et cela nous semble particulièrement fréquent en italien, où ces pronoms sont habituellement utilisés pour indiquer des référents à haute accessibilité. Dans l'exemple (57), le pronom « la », faisant partie de la forme verbale, sert à confirmer que le référent qui va subir potentiellement l'action de la faire devenir méchante est le petit chaton introduit auparavant.

(57) C' era una volta **[Valentina una gattina]**^{cat}. E un giorno **[una strega di nome Emma]**^{witch} **[gli]**^{cat} voleva fare un incantesimo per far**[la]**^{cat} diventare cattiva. (…) *(corpus italien, CE1, 1380)*

Le pronom relatif peut introduire des phrases de type relatif explicatif (dites « proposizioni appositive » en italien, voir (Ferrari, 2014, p. 190)). Ces pronoms peuvent aussi modifier un syntagme nominal qui suit. Dans ces cas, ils peuvent constituer une nouvelle mention pronominale à part entière (en italique, la phrase relative présente dans l'exemple (58)).

(58) (…) Il était une fois un robot **[qui]** *vivait dans sa maisonnette, au pays des robots.* (…) *(corpus français, CE2 ; 816)*

Ces pronoms sont aussi fréquents dans la structure du type *il était une fois +* SN nominal *+ qui +* phrase relative de description du personnage ou de la situation dans laquelle se trouve le personnage en question.

Comme déjà décrit précédemment (cf. ce chapitre, §2.9), l'anaphore zéro ou verbe sans sujet explicite est un type de reprise anaphorique très fréquent dans les langues à sujet zéro comme l'italien et l'espagnol : le fait que le sujet n'est pas exprimé de manière obligatoire, car déjà encodé dans la morphologie verbale rend possible et fréquente l'utilisation de ce type de maillon. En français moderne ce type de mention est moins présent mais existe encore (cf. ce chapitre, §2.9).

Pour cela, dans le corpus DEMOCRAT, le choix était d'indiquer l'anaphore zéro par le biais de la forme verbale concernée. En français nous avons des cas d'anaphore zéro souvent dans des phrases coordonnées et dans le cas de verbes à l'impératif (ce dernier illustré dans l'exemple (59)), alors qu'en italien ce type de mention est beaucoup plus fréquent, à cause de la nature de la langue elle-même (cf. ce chapitre, §2.9). Dans l'exemple (60), nous pouvons remarquer que la majorité des expressions référentielles utilisées pour indiquer le référent *loup* sont des anaphores zéro : le loup étant considéré comme un référent hautement accessible, par proximité et, car il représente le seul personnage présent au début du texte, il est indiqué de manière quasiment

exclusive à l'aide d'anaphores zéro, fonctionnelles aussi à la description des actions qu'il est en train d'effectuer.

(59) (…) Au bout de quelques minutes de marche, il vit un loup féroce qui était prêt à le manger. Courageusement, il dit « Ne me **[mangez]** pas ! Je ne suis qu'un tas de ferraille ! »(…) *(corpus français, CE2, 816)*

(60) Nel bosco lontano dalla città viveva un lupo. [**Amava]** correre, saltare, curiosare…. Un giorno [**vide]** uno stagno e [**provò]** a entrarci dentro. [**Si buttò]**, [**stava]** per annegare quando all'improvviso un robot di metallo gli salvò la vita. (…) *(corpus italien, CE1, 162)*

7. Synthèse de chapitre : anaphore, coréférence, chaine de référence et continuité référentielle

Dans ce chapitre, nous avons introduit et discuté la notion de référence ainsi que les phénomènes linguistiques qui permettent sa réalisation dans les textes : l'anaphore, la coréférence et la chaine de coréférence, trois phénomènes qui s'inscrivent plus largement dans les mécanismes qui assurent la cohérence textuelle. Bien que ces trois notions soient étroitement liées, elles présentent des distinctions, notamment en termes de représentation de l'identité du référent dans le cas de l'anaphore et de la coréférence, et en termes numériques et de rapport avec la textualité dans le cas de la chaine de référence. L'anaphore met en jeu une relation d'interprétation asymétrique entre deux expressions, une dépendante de l'autre, alors que la coréférence met en œuvre une relation de type symétrique, où les deux expressions renvoient de manière indépendante à la même entité. Lorsque trois expressions ou plus renvoient au même référent, celles-ci peuvent être considérées comme faisant partie d'une chaine de référence.

Nous avons ainsi décrit les principaux types d'anaphore issues de la littérature, en les illustrant à travers d'exemples provenant à la fois de notre corpus et de travaux antérieurs. Lorsque cela s'avère pertinent, nous avons également examiné certaines différences observées entre les deux langues analysées. Cette comparaison a permis de souligner certaines spécificités propres à chaque système linguistique, notamment en ce qui concerne les anaphores zéro, plus fréquemment utilisées en italien qu'en français. Nous avons aussi identifié certaines des problématiques abordées en littérature vis-à-vis de la notion de coréférence, notamment les questions liées aux référents évolutifs et à l'ambiguïté référentielle, qui constituent des enjeux majeurs pour la modélisation et l'annotation des chaines de coréférence. Nous avons également présenté une typologie des expressions référentielles, ou mentions, qui constituent les chaines de référence, ainsi que différents indices permettant de décrire ces chaines dans les textes.

Nous avons ensuite introduit les approches théoriques d'Ariel et de Givón, dont les critères définitoires serviront de base à l'analyse des chaines de référence dans notre corpus annoté, telle qu'elle sera développée dans la partie 3 (cf. Chapitre 7, Chapitre 8). L'approche des chaines de référence adoptée dans cette recherche peut être caractérisée par deux propriétés principales. Elle repose d'une part sur une perspective syntagmatisante (Schnedecker, 2021), qui considère les chaines dans leur déroulement textuel. D'autre part, elle s'inscrit dans une approche de type *entity-based*, et non *pairwise*, où l'unité centrale de l'analyse est l'entité elle-même plutôt que les relations binaires entre expressions contiguës. Autrement dit, l'annotation porte sur la présence et l'évolution d'une entité dans le texte, plutôt que sur la succession de liens entre mentions isolées. Dans cette perspective, la notion de continuité référentielle constitue le cadre conceptuel principal adopté dans ce travail pour analyser la gestion des référents dans les textes.

Cette notion présente cependant plusieurs facettes, qui ont été analysées sous le point de vue de la linguistique fonctionnelle, de la psycholinguistique ou acquisitionnelle. Plusieurs travaux ont abordé les chaines de coréférence ou les phénomènes de continuité référentielle selon

différentes perspectives. Si l'on se limite ici à quelques thèses soutenues au cours des dix dernières années, on peut citer notamment Dupret (2022) qui a abordé cette notion dans une perspective acquisitionnelle et fonctionnaliste, proche de la nôtre, en analysant des récits oraux produits par des enfants âgés de 4 à 8 ans dans le cadre d'une approche multifactorielle. Federzoni (2025) propose quant à elle une typologie des chaines de référence fondée sur l'analyse systématique de corpus annotés et sur l'étude de l'enchainement des expressions référentielles dans différents types de textes. Dans une perspective contrastive, Guo (2023) examine la structure et la composition morphosyntaxique des chaines de référence en français et en chinois, en s'intéressant à leur organisation dans différents genres textuels. De manière complémentaire, Hou (2021) analyse la saillance référentielle dans ces deux langues à partir d'une approche multifactorielle, en étudiant les facteurs syntaxiques, sémantiques et discursifs qui influencent le degré de saillance des entités dans le discours. Enfin, Kim (2023) s'intéresse aux mécanismes de cohésion référentielle dans les textes narratifs dans une perspective discursive, en analysant notamment le rôle des chaines de référence dans la structuration du genre du roman en français et en coréen. Cependant, notre travail prenant en compte l'existence de ces différents points de vue, nous avons décidé de sélectionner une perspective qui tient à la description linguistique et fonctionnelle des phénomènes de coréférence et continuité référentielle pour supporter la réflexion en didactique.

En plus de ces fondements théoriques, notre approche tient également compte des contraintes liées à l'âge des auteurs des textes analysés ainsi qu'aux caractéristiques du type de discours et du sous-genre textuel représenté dans notre corpus. En raison de la nature de la consigne proposée aux élèves pour la production des textes, ceux-ci relèvent majoritairement du discours narratif et s'apparentent souvent, d'après une première observation empirique, au sous-genre des contes de fées. Or, comme l'ont montré plusieurs travaux (Longo & Todirascu, 2009 ; Schnedecker, 2005, 2014 ; Schnedecker & Longo, 2012), le genre du texte conditionne « le matériau linguistique contenu dans les chaines de référence » (Schnedecker & Longo, 2012). Le genre textuel et le type de discours peuvent ainsi faire varier sensiblement les indices de description des chaines référentielles et conditionner leur déroulement dans le texte.

Toutefois, au-delà des effets liés au genre et au type de discours, il est également nécessaire de prendre en compte les compétences scripturales des auteurs des textes analysés, à savoir des élèves d'école primaire âgés d'environ huit à neuf ans (CE1-CE2). Les compétences de gestion de la référence se développent en effet différemment à l'oral et à l'écrit. Si les enfants sont capables, à l'oral, de mobiliser des formes déictiques étroitement liées au contexte immédiat de l'interaction, le passage à l'écrit implique de nouvelles contraintes cognitives et discursives (Dupret, 2022). Les productions écrites d'élèves se caractérisent ainsi souvent par des structures référentielles plus simples que celles observées dans les écrits de scripteurs experts, ainsi que par une variété lexicale plus limitée dans les mentions utilisées.

L'ensemble de ces éléments théoriques et contextuels constitue le cadre dans lequel s'inscrit l'étude de la continuité référentielle que nous proposons dans ce travail. Ils orientent les choix

méthodologiques présentés dans la partie suivante, notamment en ce qui concerne la conception du guide d'annotation et l'analyse des chaines de référence dans le corpus *Scolinter*.

Chapitre 2. Corpus scolaires constitués en France et en Italie

Ce chapitre a pour objectif de définir la notion de *corpus scolaire*, puis de présenter certains corpus de ce type constitués en France et en Italie.

Le choix de centrer ce chapitre sur les corpus écrits français et italiens s'explique par le cadre empirique de notre recherche : le corpus analysé dans ce travail a en effet été recueilli dans ces deux pays. Il apparait donc nécessaire de situer notre étude par rapport aux ressources existantes dans ces deux contextes nationaux. Au cours des dernières années, la France et l'Italie ont par ailleurs connu un intérêt croissant pour l'étude des productions écrites ou orales d'enfants et d'élèves de différents niveaux scolaires, ce qui a conduit à la constitution et à la diffusion de plusieurs corpus. Sans prétendre à l'exhaustivité, ce chapitre vise à rassembler les principaux corpus que nous avons identifiés au cours de notre travail de description d'un état de l'art comparatif des corpus d'écrits scolaires ou d'apprenants en langue première (L1) dans les deux pays.

1. Constitution et exploitation des corpus d'écrits scolaires

L'appellation *corpus d'écrits scolaires* (parfois raccourcie en *corpus scolaire*) est issue des travaux d'Elalouf et Boré (2007). Ces corpus se caractérisent d'abord par le profil de leurs scripteurs : des élèves généralement âgés de 3 à 18 ans, à différents stades de l'apprentissage de l'écriture, ce qui fait de l'école le lieu privilégié pour leur constitution[26]. Les textes collectés relèvent le plus souvent des genres spécifiques aux pratiques d'enseignement et d'apprentissage, pouvant être définis comme *genres scolaires* au sens de Schneuwly et Dolz (1997), c'est-à-dire comme des formes stabilisées servant de médiation aux pratiques langagières des apprenants.

Ces corpus ont ainsi le mérite de représenter de manière la plus fidèle possible les conditions de production écrite des élèves à l'école, et de tracer les « trajets développementaux de l'acquisition de la langue écrite et de ses usages » (Doquet & Ponton, 2021, p. 11). Ils permettent notamment de mettre en évidence les régularités dans l'acquisition de l'écriture ainsi que dans les types d'erreurs produites par les élèves, en plus d'examiner l'influence de différents facteurs de variation sur l'écriture, tels que l'âge des élèves, le genre de texte ou le type de tâche à accomplir. En outre, comme déjà mentionné, ces corpus peuvent également refléter les pratiques scolaires d'enseignement et apprentissage de la lecture et de l'écriture, ainsi que des éléments sociolinguistiques relevant de la culture scolaire ou nationale, tels que certaines normes implicites, routines pédagogiques ou attentes institutionnelles. En linguistique, les corpus scolaires constituent des ressources essentielles pour l'étude des productions écrites et orales des élèves en situation d'apprentissage. En didactique, ces mêmes corpus peuvent servir comme

[26] Les corpus d'écrits de niveau universitaire ou assimilés ne sont généralement pas considérés comme des corpus d'écrits scolaires. Ces corpus sont habituellement caractérisés par des productions réalisées par des élèves encore soumis à l'obligation scolaire.

outils de formation à destination des futurs enseignants, permettant d'observer la réalité de la production langagière selon une tâche donnée.

Le recueil peut s'effectuer selon deux méthodologies principales : de manière écologique, à partir de productions réalisées dans des situations ordinaires d'enseignement et d'apprentissage, ou bien à partir de consignes élaborées spécifiquement par les chercheurs dans le cadre d'un protocole expérimental. Dans le premier cas, les textes sont issus des pratiques habituelles de classe ; dans le second, les élèves répondent à des tâches d'écriture comparables à celles proposées dans le cadre d'un devoir ou d'une évaluation scolaire. Les productions sont le plus souvent collectées sous forme manuscrite, leur exploitation numérique suppose ainsi une étape préalable de transcription, souvent nécessairement manuelle. Pendant longtemps, les recherches reposaient le plus souvent sur des corpus de taille restreinte, recueillis par des enseignants-chercheurs, souvent impliqués du point de vue didactique dans la ou les classes concernées, et réalisés pour des projets d'études spécifiques pour lesquels ils sont consultés. Ils étaient rarement diffusés sous forme numérique. En effet, la constitution et la diffusion de tels corpus représentent un travail particulièrement exigeant, mobilisant d'importantes ressources humaines et nécessitant un investissement temporel considérable (Elalouf & Boré, 2007; Wolfarth, 2019).

Comme le souligne Wolfarth, « les scripteurs sont généralement en cours d'apprentissage de l'écrit. Il en résulte des corpus souvent assez éloignés de la norme attribuée aux locuteurs/scripteurs experts » (Wolfarth, 2019, p. 45). En effet, les productions d'élèves présentent fréquemment des variations orthographiques, morphosyntaxiques ou graphiques qui compliquent leur traitement automatique. « Les écrits scolaires mettent à l'épreuve les approches et les ressources classiques du traitement automatique des langues » (Doquet & Ponton, 2021, p. 11-12), ce qui rend nécessaire la mise en place d'opérations de normalisation des textes, destinée à rapprocher ces productions de la norme de la langue cible et à les rendre interrogeables à l'aide d'outils de traitement automatique des langues. L'ensemble de ces étapes, particulièrement coûteuses en temps et en ressources humaines, a longtemps constitué un frein à la constitution et à la diffusion de grands corpus scolaires. Les évolutions récentes de l'outillage TAL pourraient toutefois contribuer à alléger partiellement ces contraintes, en facilitant certaines phases de constitution de corpus et d'annotation des textes.

Nous commencerons par proposer un bref historique des projets pionniers de constitution de grand corpus d'écrits scolaires. Parmi ceux-ci figure le corpus *Éma* (Boré et al., 2018), l'un des premiers corpus français publiés et exploitable à travers la linguistique outillée. Nous présenterons ensuite le projet *E-CALM* (Ponton, Doquet, et al., 2022), qui constitue une ressource d'envergure pour l'analyse des pratiques d'écriture à l'école en France. Ce projet fédère plusieurs corpus, dont *Scoledit*, qui constitue la partie française de notre corpus d'analyse, et le corpus *RésolCo* (Garcia-Debanc et al., 2019). Ce dernier représentait, avant notre travail, le seul corpus d'écrits d'élèves annoté en continuité référentielle en France, et il constitue l'une des principales sources d'inspiration de notre démarche. Enfin, cette partie se conclura par une présentation de certains corpus scolaires italiens, en précisant leurs caractéristiques et les objectifs de recherche

qui les ont motivés : les corpus *ISACCO* (Brunato & Dell'Orletta, 2015), *CItA* (Barbagli et al., 2015), *CoDiSSc* (Revelli, 2024), *Univers-ITA* (Grandi et al., 2023), et un corpus objet de test d'annotation en continuité référentielle élaboré par deux chercheurs de l'Université de Bologna (Perugini & Zucchini, 2025).

2. Bref historique des corpus d'écrits scolaires en France

Nous présentons à présent les principales étapes historiques de la constitution des corpus d'écrits scolaires en France, en mettant l'accent sur les projets qui ont directement influencé notre démarche Après avoir évoqué certains travaux pionniers du domaine, nous nous concentrons plus particulièrement sur les ressources diffusées entre 2019 et 2025, qui viennent compléter l'état des lieux proposé par Wolfarth (2019).

Parmi ces initiatives, le corpus publié par Elalouf (2005) constitue une étape pionnière dans l'histoire de la constitution de corpus d'écrits scolaires en France. Il est composé de plus de cinq cents textes d'élèves, produis selon des consignes variées. Ces textes ont été numérisés, transcrits et normalisés, diffusés sous forme de CD rom. Le projet visait explicitement une exploitation des textes recueillis en linguistique outillée et accompagnait un ouvrage destiné à la formation des enseignants. Il représente ainsi « le premier travail de ce type pour le français langue de scolarisation » (Wolfarth, 2019, p. 53).

Cette initiative a ouvert la voie à des projets de plus grande ampleur, notamment le corpus *ÉMA* (Boré et al., 2018), disponible sur Ortolang[27]. Ce projet, qui s'inscrit dans la continuité du travail initié en 2005 par la même autrice, avait comme but la constitution d'un grand corpus longitudinal « destiné à la connaissance de la langue écrite des élèves de l'enseignement primaire et secondaire et des dispositifs dans lesquels elle s'élabore » (Elalouf et al., 2025). Conçu comme un corpus longitudinal écologique, il rassemble des productions d'élèves couvrant une diversité de genres scolaires. Il comprend les scans anonymisés, les transcriptions, les annotations et les métadonnées contextuelles. L'adoption de normes communes de transcription pour ce projet, partagées avec d'autres projets tels qu'*Écriscol* (Doquet, 2020) et *E-Calm* (Ponton, Doquet, et al., 2022) et mises en place dans les dernières mises à jour du corpus, a contribué à renforcer l'interopérabilité des données, c'est-à-dire la capacité à utiliser, échanger et traiter des données selon des formats et des conventions compatibles, et à faciliter leur exploitation outillée. Ces travaux pionniers ont mis en évidence la nécessité de disposer de données véritablement empiriques pour analyser les écrits d'élèves. Ils ont également montré que l'adoption de normes communes de numérisation favorise l'interopérabilité des données et des outils d'analyse, condition essentielle à leur diffusion et à la comparabilité des résultats d'analyse obtenus avec ceux issus d'autres corpus. En parallèle de ces projets collectifs, plusieurs recherches doctorales récentes ont porté sur la constitution et/ou l'analyse de ces corpus, soit par la description du développement langagier des élèves, soit par l'analyse des pratiques pédagogiques associées à

[27] Le corpus ÉMA est disponible au lien suivant : https://www.ortolang.fr/market/corpora/ema-ecrits-scolaires-1

la production et à la correction des écrits. Nous présentons ci-dessous certains de ces travaux afin de situer plus précisément notre recherche dans cet ensemble.

Comme le soulignent Doquet et Ponton (2021), plusieurs thèses récentes ont porté sur la constitution et/ou l'analyse des corpus scolaires et universitaires (Mazziotti, 2021 ; Miquet-Savin, 2024 ; Moysan, 2024 ; Wolfarth, 2019). Ces recherches s'intéressent soit au développement des compétences d'écritures des élèves et à leur analyse à travers l'utilisation du TAL ou de la linguistique outillée, comme c'est le cas des travaux de Wolfarth et Savin, soit aux pratiques d'enseignement et pédagogiques telles que les interventions des enseignants, comme dans le cas des recherches de Mazziotti et Moysan[28].

La thèse de Wolfarth (2019) constitue une référence importante et fait figure de préambule méthodologique à notre propre travail. Ce travail, menée au laboratoire LIDILEM, explore l'apport du TAL à la constitution et à l'exploitation linguistique de corpus d'écrits scolaires, en particulier pour analyser le développement de l'écriture chez des élèves du primaire en France. Il s'inscrit dans le projet *Scoledit*, qui a abouti à la constitution d'un corpus longitudinal d'écrits d'élèves du CP au CM2, intégrant dictées et productions narratives, et qui constitue la partie française du corpus mobilisé dans notre étude. L'une des contributions majeures de ce travail, outre la constitution d'un corpus, réside dans la conception d'une *approche par comparaison* entre transcriptions et normalisations des textes du corpus en vue de les analyser. Cette comparaison systématique, réalisée à l'aide d'outils de TAL conçus pour cet objectif, permet d'identifier les écarts à la norme et de caractériser notamment les phénomènes orthographiques et morphosyntaxiques aux différentes étapes du développement de l'écriture.

Une approche plus directement textuelle est développée dans la thèse de Miquet-Savin (2024), également menée au laboratoire LIDILEM et consacrée à l'étude des phénomènes anaphoriques dans les écrits d'élèves du secondaire en France. Le corpus objet de ce travail a été constitué de manière écologique dans plusieurs classes de seconde de l'agglomération grenobloise, la chercheuse étant également enseignante dans ces classes. Il comprend 223 textes produits par 48 élèves, représentant 42 476 tokens au total. Les copies manuscrites ont été transcrites à l'aide d'outils de reconnaissance vocale puis corrigées manuellement, selon un protocole de normalisation partielle conservant au format d'origine seules les formes susceptibles de jouer un rôle anaphorique. Les chaines de référence ont ensuite été annotées à l'aide du logiciel SACR (Oberlé, 2018), permettant d'identifier et de classifier les différentes catégories d'anaphores mobilisées par les élèves. Les résultats mettent en évidence la diversité des stratégies référentielles participant à la cohérence textuelle. Cependant, dans ce travail, certains choix de transcription adoptés, notamment la correction systématique d'une partie des formes non normées, peuvent limiter la réutilisation du corpus pour d'autres objectifs linguistiques.

[28] Nous sommes en outre au courant de plusieurs thèses en cours qui ont comme objectif la constitution ou l'annotation d'un corpus d'écrits scolaire ou universitaire. Entre autres, nous citerons ici les thèses actuellement en cours de Sonia Tesson sur l'annotation d'anaphoriques sur un corpus issu du projet Ecri+, et Gabriella De Luca qui a constitué un corpus de copies d'examens de niveau universitaire.

D'autres travaux se sont davantage concentrés sur les pratiques pédagogiques entourant la production et la correction des écrits. La thèse de Mazziotti (2021) analyse ainsi les interactions entre les interventions des enseignants et les réécritures d'élèves à partir d'un corpus de copies annotées, recueillies en contexte scolaire authentique en France et en Italie à partir d'une consigne proposée par la chercheuse. Le corpus recueilli compte au total 512 copies, 202 françaises et 310 italiennes des niveaux CE1 et CM2, rédigées respectivement par 101 et 155 élèves. Les textes manuscrits ont été transcrits en adoptant l'approche semi-diplomatique et génétique décrite par le protocole de transcription du corpus *Écriscol* (Doquet, 2020). Deux axes d'analyse différents ont été développés sur ce corpus : un axe portant sur l'évaluation de l'influence du système linguistique (et probablement scolaire) sur les annotations effectuées par les enseignants, avec comme objectif de créer une description des interventions les plus répandues ; le deuxième axe vise à évaluer le degré de prise en compte des interventions de l'enseignant par l'élève lors de la deuxième « réécriture » du texte (Mazziotti, 2020). En étudiant les commentaires, corrections et reformulations proposés par les enseignants, cette recherche met en évidence le rôle structurant des gestes didactiques dans la transformation des textes et dans l'appropriation progressive des normes d'écriture. Les résultats montrent que les modalités de correction influencent directement les stratégies de révision adoptées par les élèves, révélant une dimension fortement médiée de l'apprentissage scriptural.

Dans la même veine, la thèse de Moysan (2024) s'intéresse aux pratiques d'évaluation et aux annotations portées par les enseignants sur les copies d'élèves du secondaire. Son travail se base sur un corpus de 657 copies d'élèves de 6e et de 3e, recueillies de manière écologique et corrigées par les enseignants, complété par des entretiens menés avec sept enseignants de français. Ce travail examine les critères mobilisés pour juger la qualité des textes et les représentations de l'écriture qui sous-tendent ces pratiques, et en particulier la place de l'orthographe au sein des pratiques de correction. L'étude met en lumière une tension persistante entre exigences normatives et accompagnement des processus rédactionnels, montrant que les dispositifs d'évaluation participent eux-mêmes à façonner les formes d'écriture scolaire.

Ces travaux illustrent la complémentarité des approches actuelles : certains mettent l'accent sur la constitution de ressources outillées et sur l'analyse linguistique des textes d'élèves, tandis que d'autres interrogent les dimensions didactiques et institutionnelles qui encadrent leur production. Ils confirment ainsi l'intérêt de corpus scolaires pour décrire à la fois les caractéristiques linguistiques des écrits d'apprenants et les conditions pédagogiques de leur élaboration, perspective dans laquelle s'inscrit également notre propre démarche.

3. Le projet E-Calm

Initié en 2016 et financé par l'Agence Nationale de la Recherche, le projet *E-Calm* ambitionnait le développement d'un large corpus scolaire et universitaire en langue française pour mener différentes études (orthographe, cohérence/cohésion des textes, interventions des enseignants). Un des objectifs principaux de ce projet était de mettre à disposition, en accès libre, « un grand nombre d'écrits produits dans des contextes d'apprentissage variés afin d'en analyser les

caractéristiques linguistiques et discursives » (Doquet et al., 2023, p. 1). La constitution de cette ressource répondait à un manque évident en France de corpus d'écrits scolaires accessibles et à disposition de la communauté des chercheurs au moment du début du projet. Cet ensemble de corpus, déposé sur Ortolang[29] (Ponton, Doquet, et al., 2022), est « aujourd'hui repéré par l'Inspection Générale des Lettres comme une ressource importante pour les enseignants et utilisé en formation » (Mazziotti et al., 2024, p. 1). Ce corpus avait comme objectif d'une part de fournir aux linguistes et aux didacticiens des ressources suffisantes pour décrire l'acquisition de l'écriture à des étapes diverses de l'éducation en France, d'une autre part cette ressource servait aussi pour observer les pratiques d'interventions des enseignants sur les copies et étudier comment ces interventions orientent l'écriture (Mazziotti et al., 2024).

Parmi les quatre sous-corpus composant le corpus **E-Calm**, deux sont constitués de productions écologiques, c'est-à-dire de textes recueillis dans les classes lors du déroulement normal des cours, à partir de tâches « conçues et mises en œuvre par l'enseignant de la classe au moment où elles doivent prendre place dans la programmation prévue » (Boré & Elalouf, 2017, paragr. 8) :

- Le corpus *Écriscol* (Doquet, 2020).

Ce corpus rassemble des textes scolaires du CE1 jusqu'à l'entrée à l'université (première année de Licence et de BTS), rédigés selon des consignes diverses, ainsi que les notes et brouillons relatifs à leur genèse. Ce corpus a comme objectif l'étude de la génétique textuelle des écrits scolaires, c'est-à-dire « les processus de réécriture d'un texte avec ou sans intervention de l'enseignant » (Ponton, Doquet, et al., 2022). Comme nous verrons par la suite, le type de transcription adopté dans ce projet permet de visualiser les opérations d'écriture fondamentales pour étudier l'évolution d'un texte de son « brouillon » à la version finale sans ou avec les interventions des enseignants.

- Le corpus *Littératie avancée* (Jacques & Rinck, 2017).

Il rassemble des productions d'étudiants universitaires, allant de la première année de licence à la deuxième année de master. Les étudiants concernés appartenaient aux domaines des sciences du langage, des sciences de l'éducation et de la formation des enseignants de français. Il rassemble une grande diversité de productions tapuscrites : mémoires, fiches de lecture, lettres de motivation, comptes-rendus.

Les deux corpus restants contiennent des productions suscitées par la recherche :

- Le corpus *Scoledit* (Ponton, 2018).

Ce corpus, que nous avons déjà mentionné précédemment et que nous allons décrire plus en détail par la suite (cf. Chapitre 4, §2), trouve son origine dans les recueils de textes narratifs effectués au sein du projet *Lire et écrire au CP* (Goigoux, 2016), recueils poursuivis de manière longitudinale du CP jusqu'au CM2 par les chercheurs du laboratoire LIDILEM de l'Université Grenoble-Alpes. La méthodologie de recueil et de traitement des textes de ce corpus a été

[29] E-Calm est disponible au lien suivant https://www.ortolang.fr/market/corpora/e-calm/v2

appliquée par la suite à la collecte et traitement de données comparables dans des écoles en Italie et Espagne, dans le cadre de la constitution du corpus *Scolinter*[30] (LIDILEM, 2025; Ponton et al., 2021). C'est ce corpus qui fait l'objet de notre travail de recherche (cf. Chapitre 4).

- Le corpus *RésolCo* (Garcia-Debanc & Bonnemaison, 2014; Garcia-Debanc et al., 2017, 2021).

Le corpus *RésolCo* rassemble des textes collectés à l'école primaire (CE2, CM1, CM2), au collège (tous les niveaux) et à l'université (2e année de Master). Les textes produits répondent tous à une même consigne, qu'impose aux élèves la résolution de problèmes de cohésion textuelle dans la création de leur texte. Comme déjà mentionné, il représente le seul corpus annoté en continuité référentielle en France avant la publication du nôtre. Pour cette raison, nous allons décrire ci-dessous plus en détail ses caractéristiques et en particulier celles que nous avons retenues pour la suite de notre travail analogue d'annotation d'écrits scolaires en continuité référentielle.

3.1. Le corpus RésolCo

Le corpus *RésolCo* (Garcia-Debanc et al., 2017, 2021) a abordé l'annotation de la continuité référentielle sur des écrits de niveaux scolaires variés. Composé de 400 textes, il a été récolté dans des classes de niveaux différents, du CE2 à l'université (Garcia-Debanc et al., 2021). La distribution par niveau scolaire et la taille du corpus sont illustrées dans le Tableau 4. En s'appuyant sur l'expérience d'annotation faite lors de la conception d'*Annodis* (cf. Chapitre 3, §3.2) et sur le guide d'annotation conçu pour le corpus *DEMOCRAT* (cf. Chapitre 3, §3.4) , les chercheurs responsables de la constitution de ce corpus se posent le double objectif de réaliser une ressource annotée en continuité référentielle et d'élaborer une cartographie des formes linguistiques qui manifestent les compétences textuelles et discursives en cours de développement (Garcia-Debanc et al., 2021), à travers la description « (…) exhaustive et systématique des formes linguistiques utilisées pour construire la référence dans les textes d'élèves (…) » (Garcia-Debanc et al., 2021 : 100). Ces données « tracent des trajets développementaux de l'acquisition de la langue écrite et de ses usages » (Doquet & Ponton, 2021, p. 11), en permettant de mettre au premier plan les obstacles persistants dans l'apprentissage et d'effectuer un véritable travail scientifique de modélisation des acquisitions.

[30] Le corpus est disponible sur Ortolang au lien suivant https://www.ortolang.fr/market/corpora/scolinter/v1.

Niveaux scolaires	Nombre de textes	Nombre de mots
CE2	53	5 334
CM1	50	5 447
CM2	57	9 365
6e	106	18 260
5e	4	1 061
4e	45	11 689
3e	57	16 553
M2	13	5 164
Tot.	385	72 873

Tableau 4. Composition du corpus RésolCo. Données tirées de la page internet http://redac.univ-tlse2.fr/corpus/resolco/telechargement.html, consultée le 01/02/2026.

Les productions sont issues d'une tâche imposant aux élèves la résolution de problèmes de cohésion textuelle (Garcia-Debanc & Bonnemaison, 2014; Garcia-Debanc & Bras, 2016). La même consigne était proposée pour tous les niveaux scolaires. Trois phrases contenant des pronoms personnels (« ils » et « elle ») et des syntagmes nominaux introduits par un déterminant démonstratif (« cette maison », « ce grand bruit », « cette aventure ») sont proposées aux étudiants. La tâche impose d'insérer trois phrases dans un texte narratif fictionnel.

Voici la consigne proposée :

> « Raconte une histoire dans laquelle tu insèreras séparément et dans l'ordre donné les trois phrases suivantes :
> P1 - **Elle** habitait dans cette maison depuis longtemps.
> P2 - **Il** se retourna en entendant ce grand bruit.
> P3 - Depuis cette aventure, **les enfants** ne sortent plus la nuit. »

Les trois phrases sont données aux élèves sous forme de bandelettes de papier qu'ils peuvent coller à l'endroit de leur choix dans leur texte manuscrit. Ce dispositif a été mis en place pour éviter la copie des trois phrases et pour laisser aux scripteurs la liberté d'écrire autant qu'ils souhaitent entre une phrase et une autre. Les trois phrases et la consigne proposée forcent le scripteur à insérer dans le récit des personnages « à partir d'une expression anaphorique imposée » (Garcia-Debanc et al., 2021 : 101), à la différence des consignes habituellement proposées à l'école, dans lesquelles « le choix de l'identité référentielle du personnage est réalisé antérieurement à la mise en texte » (Garcia-Debanc et al., 2021 : 101).

Un site en ligne[31] contient le corpus qui est disponible au téléchargement sous différents formats d'exploitation ainsi qu'en exploration libre à travers deux pages différentes : une page contenant la version transcrite du texte (cf. Figure 4) et l'autre une version normalisée d'un point de vue orthographique et annotée selon les référents principaux de la tâche, à savoir les référents reliés par le scripteur aux pronoms personnels et au syntagme nominal pluriel explicités dans les trois phrases qui constituent la consigne : « elle », « il » et « les enfants » (cf. Figure 6).

	Ile de France	2017-04-26	école publique en zone mixte hors REP. Entre 300 et 400 élèves	Rédaction

Il était une fois, une ~~/~~fille qui habitait dans une petite mai~~x~~ |Son en forêts. |Elle habitait dans cette maison depuis longtemps.
Cette fille s'appellait Mégane. Un jour, dans la forêt d'à côté, un garçon se ballada. Le garçon entendit un éternuement. |Il se retourna en entendant ce grand bruit, et vu une petite maison. La nuit commença à tombé, alors, il entra. Léon, le garçon, vu Mégane, et il lui dit :

- « Bonjour, puis-je dormir ici ? demanda Léon

- oui, bien sûr, je m'appelle Mégane

- Moi je m'appelle Léon.

- Veut-tu faire une ballade dehors ?

- OK ! »

Ils sortirent dehors et partîrent se ballader. Les enfants rencontrèrent un ours affamé.
L'ours les poursuivaient en courant, ~~x~~ heureusement il y avait une grotte... mais c'était la grotte de l'ours ! Puis l'ours trouva du miel. Les deux enfants ~~e~~n profitèrent pour rentrer. |Depuis cette aventure, les enfants ne sortent plus la nuit.

Lien vers la version normalisée.

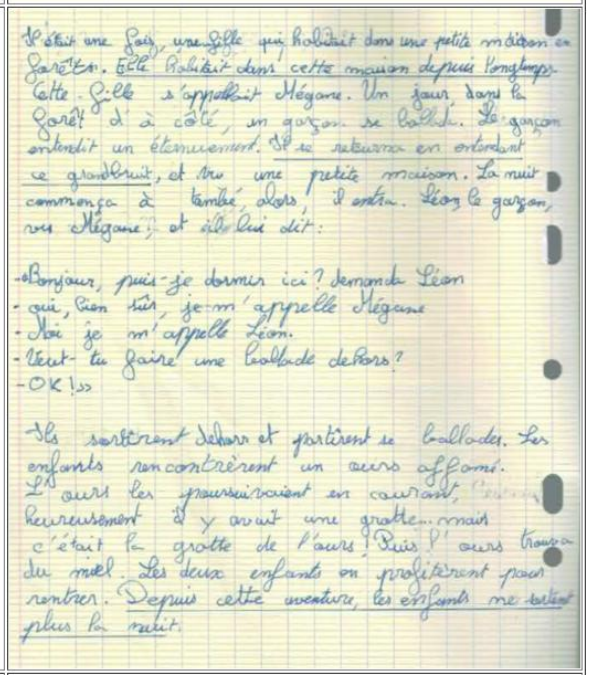


Figure 4. Texte du corpus RésolCo. Transcription (à gauche) accompagnée du scan (à gauche). Texte EC-CM2-2016-SGLEA-D1-R11-V1.

Les deux pages contiennent également le scan de la copie d'origine présenté à droite de la transcription (cf. Figure 4) ou de la version normalisée (cf. Figure 5).

[31] Le corpus est disponible en ligne au lien suivant : http://redac.univ-tlse2.fr/corpus/resolco/

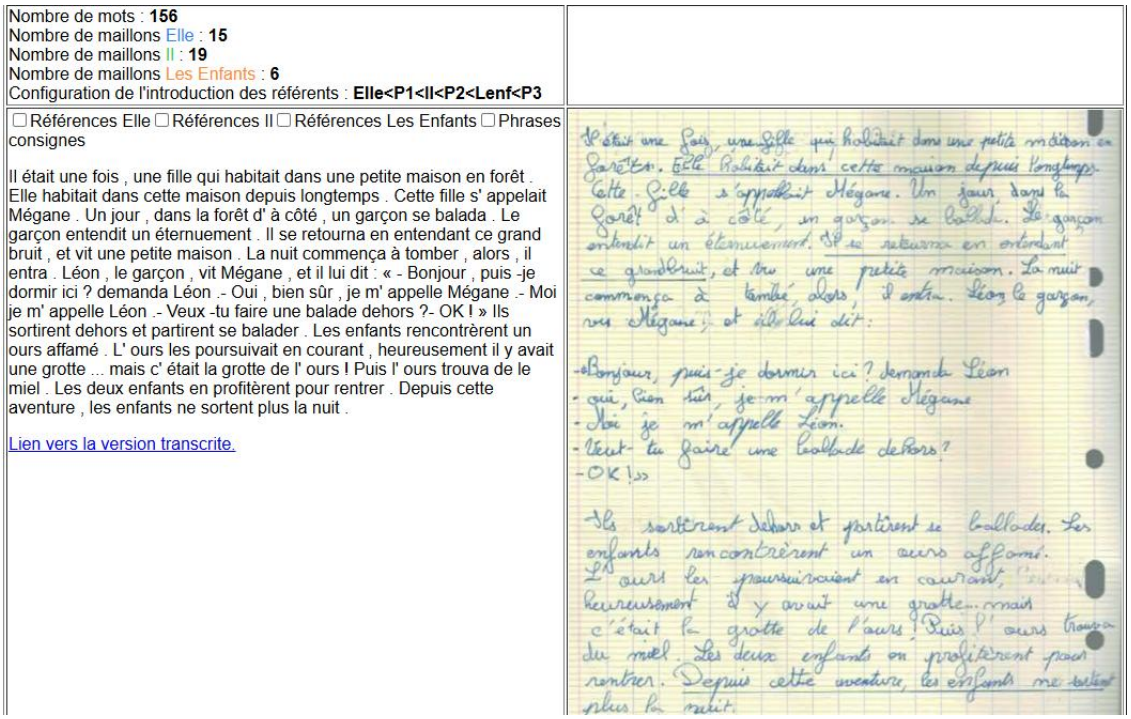


Figure 5. Texte du corpus RésolCo. Version normalisée (à gauche) accompagné du scan (à droite). Texte EC-CM2-2016-SGLEA-D1-R11-V1.

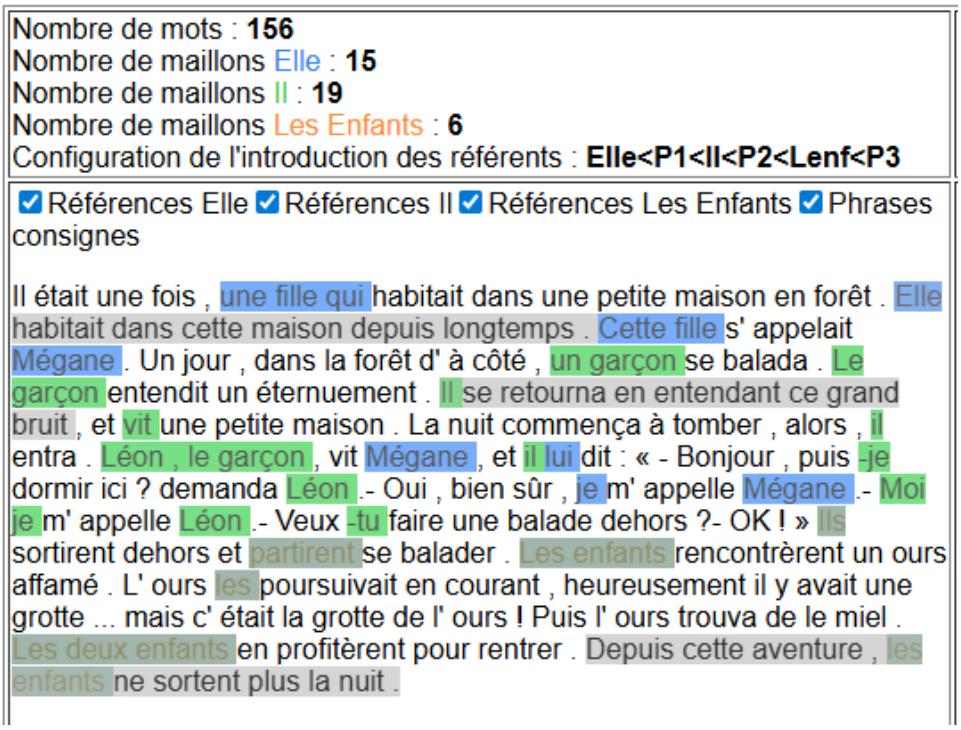


Figure 6. Version normalisée et annotée d'un texte aves les filtres possibles sélectionnés. Texte EC-CM2-2016-SGLEA-D1-R11-V1.

La méthodologie de transcription adoptée pour le corpus *RésolCo* s'appuie sur les principes de la génétique textuelle afin de rendre compte du processus d'écriture. Les ajouts, suppressions ou substitutions sont ainsi pris en considération dans la phase de transcription et font l'objet d'une annotation spécifique. L'interface de consultation des textes du corpus permet d'afficher, en regard du scan de la copie originale, une version normalisée et annotée du texte (cf. Figure 6); des

filtres interactifs facilitent par ailleurs la visualisation des mentions impliquées dans les chaines de continuité référentielle, ainsi que des segments directement liés aux consignes d'écriture.

Comme spécifié dans la page de présentation du site du corpus[32], chaque texte est enrichi de plusieurs couches d'annotation. Outre les traces du processus d'écriture déjà mentionnées, on y retrouve la normalisation orthographique, l'étiquetage morphosyntaxique, l'analyse syntaxique en dépendances et l'annotation de structures discursives (Garcia-Debanc et al., 2017) comprenant notamment la continuité référentielle, la segmentation en unités de discours élémentaires, les relations de discours et les problèmes de cohérence. L'étape d'analyse syntaxique a été réalisée à l'aide de l'outil *Stanza* (Qi et al., 2020) et l'annotation de la continuité référentielle a été effectuée manuellement grâce à l'interface d'annotation *Glozz* (Widlöcher & Mathet, 2012). Dans sa version actuelle, le corpus contient au total 385 textes[33].

Les données sont mises à disposition sous forme de ressources téléchargeables distinctes : d'une part, le corpus *RésolCo* proprement dit, qui rassemble les transcriptions encodées en XML-TEI, les scans des copies au format image et les versions textuelles normalisées ; d'autre part, la ressource *RésolCo Continuité Référentielle*, contenant les textes annotés au format *Glozz*[34].

La méthodologie d'annotation adoptée dans le corpus *RésolCo* s'inspire des procédures issues des projets *Annodis* et *DEMOCRAT*, tout en l'adaptant aux spécificités des productions d'élèves. Le phénomène annoté est ici la continuité coréférentielle. Les mentions annotées sont celles en lien avec les personnages introduits par la consigne ; elles sont donc circonscrites « aux seuls référents présents dans l'ensemble des textes du corpus ; autrement dit, aux référents provoqués par la consigne *RésolCo* », c'est-à-dire les trois référents humains de la consigne : « elle », « il » et « les enfants ». Il n'y a donc pas d'identité référentielle stricte, ce qui est un des prérequis pour l'annotation de la coréférence (Garcia-Debanc et al., 2021, paragr. 17). À part permettre à l'annotateur de se concentrer sur les référents qui jouent un rôle de premier plan dans le récit, ce choix d'annotation limité aux personnages probablement de premier plan dans le récit permet aussi de créer des annotations comparables entre les textes narratifs du corpus, très variables à cause de la nature libre de la consigne proposée.

Un prémarquage automatique d'assistance à l'annotateur humain a été appliqué sur certains éléments des textes susceptibles dits « maillons-candidats »: les éléments ciblés sont les noms propres, les pronoms personnels de troisième personne, les déterminants possessifs et les séquences composées de "les" suivies de "enfants".

La démarche d'annotation employée par le projet, comme décrite dans le guide d'annotation (Federzoni, Rebeyrolle, et al., 2021), est composée de quatre phases : 1) la sélection de la chaine référentielle à annoter (seulement trois chaines par texte) ; 2) l'identification des maillons qui

[32] http://redac.univ-tlse2.fr/corpus/resolco/presentation.html (consulté le 29/08/2025).

[33] Données tirées de la page de consultation du corpus http://redac.univ-tlse2.fr/corpus/resolco/telechargement.html (consulté le 13/03/2026).

[34] Informations contenues dans la page « Téléchargements » du site web du projet http://redac.univ-tlse2.fr/corpus/resolco/telechargement.html (consulté le 13/03/2026). Une troisième ressource, consacrée à la cohérence discursive et incluant la segmentation en unités de discours et les relations de discours, est annoncée comme à paraître.

signalent le référent de la CR choisie ; 3) la délimitation de la portion de texte qui constitue les maillons ; 4) la mise en relation des maillons appartenant à la même chaine référentielle. Une des particularités, qui comme on le verra, constitue une des difficultés dans l'annotation des corpus pour le TAL, est l'inclusion des référents multiples dans l'annotation (du type pronom pluriel, SN pluriel ou SN collectif). Si le référent multiple inclut un des référents humains de la consigne, il est annoté comme maillon de la chaine de ce référent et le trait « groupe » lui est attribué. Comme dans le corpus *DEMOCRAT*, les pronoms réfléchis ne sont pas annotés.

Les analyses du corpus *RésolCo* ont permis de mettre en évidence plusieurs résultats significatifs concernant les stratégies de gestion de la continuité référentielle par les élèves dont l'identification de deux stratégies de gestion référentielle. Cela a permis aussi d'effectuer une comparaison quantitative entre les chaines de continuité référentielle des niveaux CE2, 6e et 3e , et d'étudier le type de première mention des référents comme indice de la stratégie rédactionnelle dominante (Garcia-Debanc et al., 2021).

L'analyse des configurations d'introduction des référents a permis d'identifier deux principales stratégies rédactionnelles : la stratégie pas à pas et la stratégie de planification. La stratégie « pas à pas », préférée par les élèves de 6e, consiste dans la gestion des chaines de référence par succession (cf. Chapitre 1, §4.4). L'introduction de chaque référent est faite juste avant la phrase imposée qui le contient. La stratégie de « planification », plus fréquente chez les élèves de 3e, implique l'introduction des référents « elle » et « il » dès le début du texte, permettant une gestion par entrecroisement (cf. Chapitre 1, §4.4) des continuités référentielles. Cette dernière stratégie peut se réaliser de manière plus ou moins complexe.

Des indices qualitatifs, tels que la position de la première mention du ou des référents dans le texte par rapport à la position des bandelettes papier de la consigne, permettent de décrire « les stratégies utilisées par les rédacteurs pour introduire, maintenir et faire cohabiter les différents référents dans le discours » (Garcia-Debanc et al., 2021, p. 105). Les chercheuses identifient ainsi deux stratégies principales : la stratégie « pas à pas » et la stratégie dite « de planification ».

Lorsque le texte présente l'introduction d'un personnage immédiatement suivie par le collage de la bandelette papier contenant le pronom susceptible de constituer l'anaphore du référent introduit, puis l'introduction d'un nouveau suivie par le collage de la bandelette correspondante, on peut inférer l'utilisation de la stratégie « pas à pas ». Cette stratégie était préférée par les élèves de 6e. Dans ce cas, l'élève gère les chaines en mettant en place des stratégies de succession. Le texte tiré de Garcia-Debanc et al. (2021), présentée en (61), illustre un exemple d'utilisation minimale de cette stratégie. Le personnage de la *vieille dame* est introduit dans la phrase précédant celle de la consigne (« Elle vivait dans cette maison depuis longtemps », surlignée dans le texte). L'introduction d'un nouveau personnage, *Luke le petit*, entraîne ensuite la disparition de la chaine référentielle associée à la vieille dame.

(61) Un jour il y a **[une vieille dame]^{elle} [qui]^{elle}** voulait déménager. **[Elle]**^{elle} habitait dans cette maison depuis longtemps.
[Elle]^{elle} mit **[sa]^{elle}** maison en vente. Un jeune couple passa et l'aperçut ils décident d'emménager. La première journée tout se passa bien mais la première nuit des bruits retentissent **[Luke le petit]**^{il} se leva et **[alla]**^{il} dans le salon puis
[il]^{il} se retourna en entendant ce grand bruit.
Et **[il]**^{il} retourna dans **[son]**^{il} lit en courant. Les autres nuits allaient mieux et maintenant.
Depuis cette aventure, **[les enfants]^{ils}** ne sortent plus la nuit.
(texte normalisé issu du corpus RésolCo, niveau 6e, tiré de Garcia -Debanc et al., (2011))

La stratégie dite de « planification », plus fréquente chez les élèves de 3e, consiste en l'anticipation de l'organisation du texte par l'intégration, dès son début, des trois référents « il », « elle » et « les enfants ». Cette stratégie implique donc la gestion simultanée de plusieurs chaines de référence, et la mise en place de stratégies de cohabitation dans le discours. « Cette stratégie peut se réaliser de manière plus ou moins complexe dans les textes des élèves » (Garcia-Debanc et al., 2021, p. 107). Dans le texte (62), les deux référents sont introduits conjointement (« Camille et Manu ») dès le début du texte et leur présence est maintenue jusqu'à la fin du texte, principalement à travers l'utilisation du pronom personnel pluriel « ils ».

(62) **[Camille]**^{elle} **et [Manu]**^{il}**]**^{ils} jouaient la nuit dans la forêt. Tout à coup **[elle]**^{elle} eut une idée : « **[Viens]**^{il} **[Manu]**^{il} **[je]**^{elle} vais **[te]**^{il} montrer **[ma]**^{elle} cabane ».
[Elle]^{elle} habitait dans cette maison depuis longtemps. Et dans la maison **[il]**^{il} se retourna en entendant ce grand bruit. **[Ils]**^{ils} montèrent dans le grenier et **[ils]**^{ils} entendirent un autre bruit. Et **[ils]**^{ils} virent une ombre de rat géant. Et **[ils]**^{ils} retournèrent chez **[eux]**^{ils} en courant.
Depuis cette aventure, **[les enfants]**^{ils} ne sortent plus la nuit.
(texte normalisé issu du corpus RésolCo, niveau CM1, tiré de Garcia -Debanc et al., (2011))

Les analyses quantitatives conduites sur le nombre de maillons dans le corpus indiquent une grande hétérogénéité des profils au niveau CE2 et une augmentation du nombre des maillons selon le niveau scolaire. Cependant, comme nous pouvons observer grâce à la comparaison entre nombre de maillons et longueur des textes en termes de tokens, cela pourrait être lié à la longueur des textes elle-même. Ces indicateurs et les observations qu'on peut en tirer permettent de mieux comprendre comment les élèves construisent la cohérence de leur récit et comment ils établissent les liens entre les différents personnages.

Le corpus *RésolCo* constitue ainsi une ressource précieuse pour l'étude de la continuité référentielle dans les productions écrites d'élèves français. Les résultats obtenus mettent en évidence une évolution significative dans la gestion de la continuité référentielle avec l'âge et le niveau scolaire. On observe notamment une augmentation du nombre de maillons dans les chaines

référentielles et une transition d'une stratégie "pas à pas" vers une stratégie de "planification" plus élaborée.

Cette première partie a permis de retracer les principales étapes de la constitution de corpus d'écrits scolaires en France. Les premiers travaux pionniers ont posé les bases de la numérisation et de l'exploitation linguistique des écrits d'élèves, ouvrant la voie à des projets de plus grande envergure tels que *ÉMA* ou le projet *E-Calm*. Ces initiatives ont contribué à structurer un environnement de recherche où les corpus scolaires deviennent des ressources partagées, diffusées en ligne et enrichies de multiples niveaux d'annotation.

L'examen plus détaillé du corpus *RésolCo* a montré l'intérêt particulier de ces ressources pour l'étude de la continuité référentielle dans les textes d'élèves. En proposant une tâche d'écriture visant la résolution de problèmes de cohérence et un protocole d'annotation spécifique, ce corpus permet d'observer les stratégies mises en œuvre par les scripteurs pour introduire et maintenir les référents dans un récit. Les résultats obtenus mettent notamment en évidence l'évolution des stratégies rédactionnelles au fil de la scolarité, depuis une gestion « pas à pas » des référents vers des formes plus planifiées d'organisation textuelle. Cette ressource constitue ainsi un point de référence important pour les recherches portant sur la cohérence et la cohésion des textes d'élèves et une inspiration importante de notre travail de recherche.

4. Corpus scolaires constitués en Italie

Comme indiqué précédemment, si la constitution du corpus *Scolinter* répondait en partie à l'absence de corpus longitudinaux d'école primaire directement exploitables à l'aide d'outils de traitement automatique des langues, le paysage italien ne se caractérise pas pour autant par une absence de ressources. Plusieurs projets ont en effet permis de constituer des corpus scolaires couvrant différents niveaux de scolarité, modalités de recueil et types d'annotation, qu'il s'agisse de productions écrites ou orales, annotées manuellement ou automatiquement.

Cette section propose ainsi un aperçu des principales ressources italiennes susceptibles d'éclairer le contexte scientifique dans lequel s'inscrit notre travail. Quatre corpus apparaissent particulièrement représentatifs des orientations méthodologiques et des objets d'étude développés dans le pays: les corpus *ISACCO* (Brunato & Dell'Orletta, 2015), *CItA* (Barbagli et al., 2016), le corpus d'écrits de niveau universitaire *Univers-ITA* (Grandi et al., 2023) ainsi que *CoDiSSc* (Revelli, 2024), qui offre une perspective diachronique sur l'écriture scolaire italienne de l'Unité d'Italie à nos jours. Il convient toutefois de souligner que, malgré leur diversité, aucun de ces corpus ne repose sur un suivi longitudinal des mêmes élèves au fil de leur scolarité. À ces ressources s'ajoute une étude exploratoire d'annotation des chaines de référence menée par Perugini et Zucchini (Perugini & Zucchini, 2025) sur un corpus de textes narratifs d'élèves de collège. Cette étude présente des similitudes avec notre travail concernant le choix du cadre théorique et certains aspects des annotations, mais diffère quant à la conception du schéma d'annotation lui-même. Nous avons néanmoins tiré inspiration de cette étude, qui nous semble présenter des pistes d'analyse qualitative intéressantes.

4.1. ISACCO (Italian School-Age Children Corpus)

Le corpus *ISACCO* (*Italian School-Age Children Corpus,* Brunato & Dell'Orletta, 2015) est composé de 181 récits de restitution d'un texte narratif, recueillis à l'oral et à l'écrit auprès de 56 élèves de 2e et de 4e années d'école primaire italienne (correspondant aux niveaux CE1 et CM1). Les textes ont été collectés en deux temps auprès des mêmes élèves : après une lecture collective du texte en classe, chaque élève réécoutait individuellement le texte puis produisait immédiatement un récit de restitution oral. Après un intervalle de deux semaines, une deuxième lecture collective du texte était proposée, à l'issue de laquelle les élèves devaient rédiger individuellement une version écrite du même récit, sans limite de temps ni contrainte typographique particulière (le choix entre écriture cursive ou script restant libre). Les productions ont ensuite été numérisées (transcription des enregistrements oraux et des manuscrits) puis annotées à plusieurs niveaux : enrichissement des transcriptions avec des indications de phénomènes de l'oral, annotation manuelle de certains types d'erreurs, étiquetage morphosyntaxique, analyse en dépendance et extraction terminologique automatique pour l'analyse de contenus.

Des études préliminaires sur ce corpus (Brunato & Dell'Orletta, 2015) ont permis d'effectuer des comparaisons diamésiques entre oral et écrit (c'est-à-dire des comparaisons entre différentes modalités de réalisation du langage selon le médium de production), ainsi que des analyses de contenu évaluant la correspondance entre l'histoire d'origine et la restitution effectuée par l'élève, afin «[d']identifier les patterns de développement typiques à utiliser pour des comparaisons, par exemple dans les contextes cliniques » (Brunato & Dell'Orletta, 2015, p. 63, notre traduction).

Différentes analyses quantitatives ont été menées sur ce corpus, avec l'objectif de comparer les textes oraux aux textes écrits. En ce qui concerne la description des textes bruts, les indices disponibles sont la longueur des textes, la longueur moyenne des mots, et des mesures de la complexité morphosyntaxique comme la densité lexicale et la distribution des mots par *POS tag*.

Ces comparaisons quantitatives donnent un aperçu de l'influence de la modalité de communication dans les productions des enfants. Les productions orales demeurent plus longues que les productions écrites, laissant penser que le coût cognitif plus haut de l'écriture pourrait influencer la mémoire et provoquer de la perte d'information ; des mots plus courts sont utilisés en général dans les textes oraux, dont un pourcentage important est représenté par des « mots outils », ce qui mène aussi à une densité lexicale plus importante dans les textes écrits. Enfin, les textes écrits semblent présenter une cohésion textuelle plus faible que les textes oraux.

4.2. CItA (Corpus Italiano di Apprendenti L1)

Le *CItA* ou *Corpus Italiano di Apprendenti L1* (*Corpus Italien d'apprenants L1*, Barbagli et al., 2015, 2016) a été constitué avec l'objectif de mettre à disposition le premier corpus numérisé et librement accessible d'écrits d'élèves italiens du collège[35]. Il s'agit d'un corpus diachronique, collecté auprès de 156 élèves de 1ère et 2ème années de collège italiens, correspondant aux classes de 6ème et 5ème du système scolaire français. Le corpus regroupe 1 352 textes, pour un total de

[35] Le corpus est disponible pour téléchargement au lien suivant http://www.italianlp.it/resources/cita-corpus-italiano-di-apprendenti-l1/.

366 335 tokens. Il comprend deux textes par élève, recueillis par les chercheurs selon une consigne commune à toutes les classes, ainsi que toutes les productions écrites réalisées durant l'année scolaire, recueillies de manière écologique par les enseignants. En moyenne, le corpus comprend ainsi 8,66 textes par élève. Les textes sont accompagnés de métadonnées issues des questionnaires sociobiographiques soumis aux élèves et qui concernaient le statut socioéconomique de la famille, la provenance géographique ainsi que des questions relatives à leur rapport avec la lecture et d'autres informations.

Les textes du corpus, en plus d'être ordonnés chronologiquement pour chaque élève, ont été annotés pour définir les catégories d'erreurs les plus fréquentes et proposer des corrections possibles de chaque erreur au niveau du mot. Ces annotations, qui portent sur les erreurs de grammaire, d'orthographe et de lexique, ont permis d'effectuer des études sur les variations en fréquence et en typologie des erreurs entre niveaux scolaires contigus ou au même niveau scolaire. Chaque type d'erreur fréquent a été associé à un code (par exemple, l'utilisation du temps verbal, une omission, une utilisation erronée d'un pronom relatif, d'un article etc.). L'annotation se réalise sous forme d'une balise XML contenant le code de l'erreur ainsi qu'une proposition de correction possible du token concerné. Dans la Figure 7 nous avons un exemple d'annotation : la phrase « io avevo salire fin lassù ma mi sono fata coraggio […]" contient deux erreurs : la première porte sur le verbe « avevo », il est marqué comme erreur lexicale et propose le verbe « dovevo » comme correction ; la deuxième erreur est une erreur orthographique, portant sur une utilisation erronée des doubles consonnes dans le cas du participe *« fata », « fatta » dans la forme attendue.

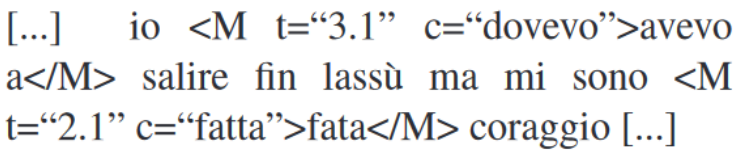


Figure 7. Exemple d'annotation des erreurs, tiré de Barbagli et al., 2015 (p. 32).

En ce qui concerne la variation du nombre d'erreurs entre années scolaires, ce corpus met en évidence une diminution globale entre niveaux (en passant de 1929 à 1525). Toutefois cela ne concerne pas toutes les catégories d'erreurs de manière homogène. Notamment dans le cas d'erreurs d'utilisation des temps verbaux, celles-ci passent de 150 à 239 entre la 1e et la 2e année. Les auteurs expliquent ce phénomène par le changement du type de consigne proposée. En 1e année les élèves produisent davantage des textes de type narratif, typologie textuelle qui prévoit l'utilisation de séquences temporelles, plus simples et mieux connues par les élèves. En 2e année, les textes à rédiger étaient plutôt de type argumentatif, avec une structure plus complexe nécessitant l'usage de formes verbales plus difficiles à maitriser. L'hypothèse est donc que cette complexité conduit les élèves à mobiliser des temps plus élaborés, ce qui entraîne une hausse du taux d'erreur, comme le confirment les analyses automatiques réalisées sur les textes concernés (Barbagli et al., 2015, p. 34).

4.3. CoDiSSc (Corpus Digitale delle Scritture Scolastiche)

Le *Corpus Numérique des Écritures Scolaires* (*CoDiSSc, Corpus Digitale delle Scritture Scolastiche*), initialement *CoDiSV* (*Corpus Digitale delle Scritture scolastiche d'ambito Valdostano*), peut être vu comme « une archive numérique » (Revelli, 2011, p. 10) des productions scolaires en diachronie. Il comprend des reproductions numériques de cahiers d'école produits en Vallée d'Aoste en Italie entre l'Unité d'Italie (1863) et la fin du XXe siècle (2014), sur une période d'environ 150 ans, puis enrichie de documents variés, administratifs (certificats, diplômes…) ou pertinents pour l'activité de l'enseignant (registres de classe ou carnet de l'enseignant) provenant de différentes régions d'Italie du Nord, en plus des différentes productions d'élèves telles que des dessins, ou des copies d'examen. Le socle du corpus est constitué par 1000 cahiers d'école. La distribution des types de documents dans le corpus est visible dans la Figure 8, issue du site internet dédié à l'exploration du corpus[36].

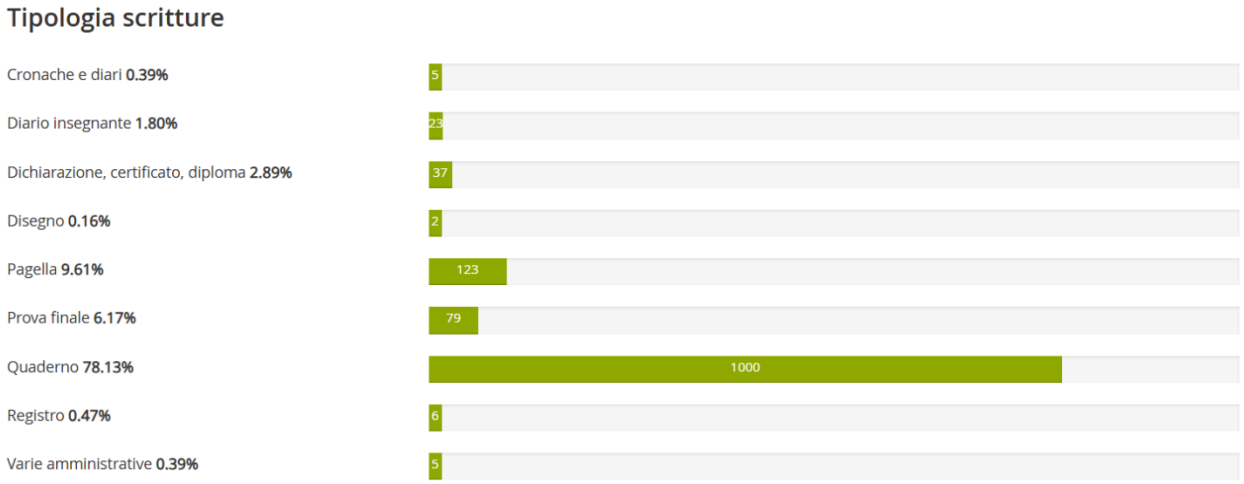


Figure 8. Représentation des types d'écrits présents dans le CoDiSSc issus du site web du projet (consulté le 20/08/2025).

La description faite par Luisa Revelli, enseignante-chercheuse responsable du projet, nous semble résumer parfaitement les objectifs du projet et la raison pour laquelle il est centré sur la collecte et la numérisation des cahiers d'école. Cette description nous rappelle aussi l'importance des travaux sur la patrimonialisation de telles ressources :

> « Les cahiers sont, en effet, des objets ordinaires en raison de leur banalité et de leur signification restreinte et contingente, qu'ils revêtent individuellement ; ils peuvent toutefois devenir des outils extraordinaires s'ils sont examinés en série et à distance dans le temps, car de nombreuses informations qu'ils contiennent ne sont pas récupérables autrement : une fois rassemblés, les témoignages individuels contenus dans les collections de cahiers perdent en effet leur dimension subjective pour prendre le rôle de pièces d'une mosaïque qui, une fois composée de manière

[36] Le corpus est disponible au lien suivant https://codissc.it/. (Consulté le 01/02/2026).

> appropriée, restitue une image plus large et globale de phénomènes qui ne sont plus individuels, mais sociaux et collectifs. »[37] (Revelli, 2011, p. 5)

Dans la mesure du possible, chaque document est accompagné d'informations biographiques sur l'auteur ou autrice, quand celles-ci sont disponibles, cela pour identifier « le contexte sociolinguistique auquel les écrits font référence » (Revelli, 2011, p. 12).

Une partie des textes avait déjà été transcrite en 2011 : le type de transcription choisi tenait surtout compte de l'écriture des élèves, en laissant de côté les interventions des enseignants, sauf dans des cas de transcriptions faites pour des études sur les interventions de correction (Revelli, 2011). Dans cette optique, le choix est tombé sur une transcription diplomatique[38], qui tient compte de la « physionomie globale de la page écrite » (Doquet et al., 2021, p. 32).

Les différents cahiers et documents recueillis sont actuellement disponibles sur la plateforme du projet, qui permet de « feuilleter » de manière virtuelle les documents de l'archive (dont un exemple de cahier produit en 1938 est illustré dans la Figure 9). Les filtres disponibles sur la plateforme permettent de sélectionner les documents par région, typologie et période de production. Ce corpus, qui a été fait objet de nombreuses publications, surtout en didactique et en histoire de la didactique[39], représente une ressource d'une richesse inédite dans le panorama de la recherche italienne. Sa double finalité, de conservation d'un patrimoine matériel qui est habituellement négligé ou totalement perdu, et de numérisation de ce patrimoine afin d'étudier le développement des pratiques d'enseignement tout au long d'un siècle, rend ce projet unique en son genre.

[37] Dans l'original « I quaderni sono, in effetti, oggetti ordinari per la loro dozzinalità e per il significato, ristretto e contingente, che singolarmente rivestono; possono però diventare strumenti straordinari se esaminati in serie e a distanza di tempo, perché molte informazioni in essi rinvenibili non sono recuperabili altrimenti: quando raccolte, le singole testimonianze contenute nelle collezioni di quaderni perdono, infatti, la loro dimensione soggettiva, per assumere il ruolo di tessere di un mosaico che — opportunamente composto — restituisce un'immagine più ampia e globale di fenomeni non più individuali ma sociali e collettivi. » (Revelli, 2011, p. 5).

[38] Nous allons décrire plus en détail les différents types de transcriptions existants par la suite (cf. Chapitre 4, 7.2).

[39] Sur le site du projet est disponible la liste des contributions écrites sur ce corpus, au lien suivant https://codissc.it/pubblicazioni-codissc-codisv-prodotti-della-ricerca.

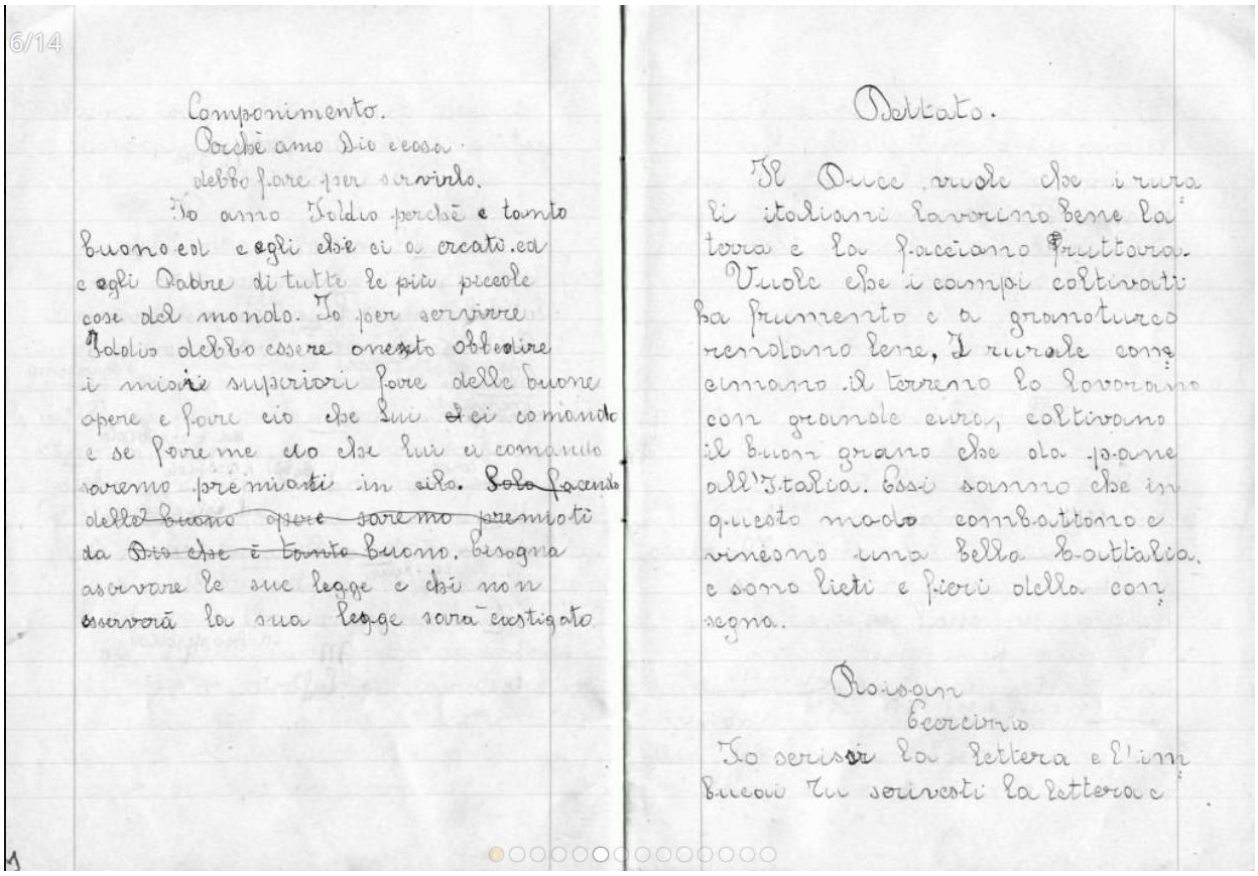


Figure 9. Page 6/14 du cahier 0486 du corpus CoDiSsC, produit en 1938.

4.4. Univers-ITA

Ce projet et les corpus qui en découlent avaient comme objectif de représenter de la manière la plus fidèle possible l'état de la langue italienne telle qu'elle est utilisée à l'université. Il est composé de trois sous-parties, recueillies selon des méthodologies différentes :

- Le corpus *Univers-ITA*, que nous allons décrire plus en détail par la suite.
- Le corpus *Univers-ITA-ProUniv*, composé de mémoires, rapports, et écrits universitaires variés dans leur version d'origine, c'est-à-dire avant les corrections apportées par les enseignants. Ce corpus est composé de 773 textes pour un total de 6 267 765 tokens. À partir de ce corpus, un sous-corpus a été sélectionné selon des critères de représentativité géographique ainsi que de représentativité des types de domaine disciplinaire d'appartenance de l'étudiant.
- Le corpus *Univers-ITA-ProGior* : il est composé de 1 630 textes (1 692 846 tokens) de type journalistique rédigés entre 2012 et 2021 et extraits des blogs, périodiques universitaires en ligne et sites web d'information gérées par étudiants universitaires de différents établissements italiens.

Nous allons présenter de manière détaillée le corpus *Univers-ITA*, qui constitue à ce jour le corpus le plus étudié parmi les trois présentés. Ce corpus a été recueilli pour « tracer un profil exhaustif de l'italien écrit des étudiants et étudiantes universitaires en analysant la production écrite formelle d'un échantillon représentatif de la population universitaire par discipline et zone

géographique »[40] (Grandi, 2025, p. 12, notre traduction). À ce stade, il s'agit de la première étude systématique consacrée à ce public en Italie.

Le recueil de données s'est déroulé au cours de l'année académique 2020/2021 auprès d'étudiants universitaires inscrits dans différents domaines d'études (lettres, santé, sciences humaines et sociales et sciences) et provenant de zones géographiques différentes (Nord, Centre et Sud, y compris la Sicile et la Sardaigne) répartis dans 44 universités. Parmi les productions recueillies, 2 137 textes (810 715 tokens) ont été sélectionnées selon des critères de représentativité géographique et disciplinaire. Ces textes ont fait l'objet de traitements automatiques (tokenisation, lemmatisation, POS-tagging et traits morphosyntaxiques), d'analyses automatiques de complexité[41], puis d'une annotation manuelle des phénomènes linguistiques « déviants » par rapport à la norme attendue de l'italien standard. Tout en étant un corpus de taille relativement réduite, le corpus est enrichi de métadonnées détaillées, et constitue ainsi une ressource précieuse pour étudier le corpus selon différentes perspectives (Grandi, 2025).

La tâche d'écriture comprenait deux étapes, réalisées à travers une plateforme dédiée : la rédaction d'un texte de longueur comprise entre 250 et 500 mots, suivie d'un questionnaire sociobiographique comportant plus de cinquante questions. Le temps total alloué à la tâche était entre 50 et 60 minutes. La consigne de rédaction du texte[42] précisait le registre de langue à adopter, formel, et la type de texte attendu, argumentatif. Les participants devaient écrire un court texte exprimant leur point de vue sur le fonctionnement de l'enseignement à distance pendant la pandémie de Covid-19. Le questionnaire sociobiographique « visait à recueillir des données pertinentes sur les dimensions sociologiques et sociolinguistiques, afin d'enrichir les perspectives d'analyse des textes et de faire ressortir à la fois les corrélations systématiques entre les paramètres sociaux et les traits linguistiques, et les perceptions et les besoins de la population cible en matière d'écriture dans le contexte universitaire »[43] (Grandi, 2025, p. 13, notre traduction).

Ce corpus a été fait l'objet de plusieurs études sur des phénomènes variés, tels que l'utilisation des formes de déictiques textuelles (Da Milano, 2025), ou les écarts à la norme liés la ponctuation (Pecorari, 2025). En particulier, deux tendances majeures sont soulignées par Grandi : d'une part « la prévalence écrasante, parmi les annotations [des formes déviantes de la norme *ndr*], de celles qui concernent la ponctuation, qui couvrent environ 50 % des annotations globales » (Grandi, 2025, p. 15). La deuxième observation, qui pourrait être considérée en lien avec la première, concerne la corrélation entre déviations liées à la syntaxe et à la cohésion textuelle avec celles liées à la cohérence textuelle. C'est-à-dire, même à ce niveau de l'éducation, que les difficultés

[40] Dans le texte d'origine, « Il progetto di cui qui presentiamo i risultati, finanziato nel bando PRIN 201711, ha invece lo scopo di tracciare un profilo esaustivo dell'italiano scritto di studenti e studentesse universitari analizzando la produzione scritta formale di un campione della popolazione universitaria rappresentativo per aree disciplinari e geografiche. »

[41] Les analyses ont été effectuées par le biais du logiciel READ-IT (Dell'Orletta et al., 2011).

[42] La consigne dans son intégralité traduite est disponible en Annexe A.

[43] Dans le texte d'origine, « Il questionario sociobiografico di oltre 50 domande (…) era finalizzato a rilevare dati pertinenti a dimensioni sociologiche, oltre che sociolinguistiche, in grado di arricchire le prospettive di analisi dei testi e di fare emergere sia correlazioni sistematiche tra parametri di ordine sociale e tratti di tipo linguistico, sia percezioni e bisogni della popolazione di riferimento in relazione al tema della scrittura in ambito universitario. »

majeures dans l'écriture restent liées aux questions de « planification, organisation et structuration des textes eux-mêmes »[44] (Grandi, 2025, p. 15, notre traduction).

Ces observations nous semblent faire écho aux préoccupations que nous avons déjà rencontrées dans l'analyse des grands corpus scolaires en France et, plus particulièrement, aux objectifs de recherche de ce travail de thèse, qui porte sur l'étude des « zones de résistance » dans la maîtrise de la cohérence textuelle au cours des premières années d'apprentissage de l'écriture. Les résultats présentés par Grandi apparaissent particulièrement pertinents, dans la mesure où ils mettent en évidence l'importance des questions de planification, d'organisation et de structuration des textes, qui demeurent centrales même dans un contexte d'apprentissage avancé de l'écriture, tel que l'université. La persistance de ces difficultés à des niveaux aussi élevés souligne la nécessité de poursuivre les recherches sur les compétences rédactionnelles et sur les pratiques d'enseignement de l'écriture tout au long du parcours scolaire, dès l'école primaire jusqu'à l'enseignement supérieur, en particulier en ce qui concerne les enjeux de cohérence et de cohésion textuelles.

Dans cette perspective, l'analyse des mécanismes assurant la continuité référentielle (notamment les chaines de référence, qui participent directement à l'articulation locale et globale du texte) constitue un point d'entrée privilégié pour étudier ces dimensions discursives. C'est précisément à ce niveau que se situe l'étude présentée dans la section suivante.

4.5. Annotation des chaines de référence dans un corpus scolaire de textes narratifs : présentation d'une étude exploratoire

Dans une contribution récente, Perugini et Zucchini (2025), présentent une étude exploratoire consacrée à l'annotation des chaines de référence dans un corpus scolaire du collège italien. Cette étude vise à analyser la distribution des types de mentions (syntagmes nominaux, pronoms, anaphores zéro) et leur degré d'*informativité* (cf. Chapitre 1, §5.1 pour la notion d'informativité chez Ariel), en comparant textes produits par élèves monolingues et plurilingues.

Les textes du corpus annoté ont été collectés dans le cadre d'un travail doctoral (Zucchini, 2023) auprès des classes de 3e année de *scuola secondaria di primo grado* ou *scuola media* (dernier niveau du collège en Italie, équivalent à la 4e française) dans des établissements des provinces de Modène et Bologne, entre 2020 et 2023. Recueillis de manière écologique, les textes proviennent d'activités d'écriture réalisées en classe (*tema di italiano*), et appartenaient à des genres de textes différents, dont des récits et des fiches de lecture[45]. Parmi ces productions, 40 textes ont été retenus pour l'annotation des chaines de référence : 20 rédigés par des élèves italophones et vingt par des élèves issus de familles alloglottes. Les élèves participants ont également rempli un

[44] Dans le texte d'origine, « Il primo riguarda la prevalenza schiacciante, tra le annotazioni, di quelle che riguardano la punteggiatura, che coprono circa il 50% delle annotazioni globali e sulle quali tornerò a breve. Il secondo concerne invece la correlazione tra diverse classi di annotazioni, che rivela un forte legame tra devianze relative alla sintassi (e, più specificamente, alla coesione testuale) e devianze relative alla coerenza testuale. »

[45] Les fiches de lecture écrites par les élèves ont été retenues en tant que textes narratifs vu que la plupart du texte est constitué par le résumé de l'œuvre narrative analysée, et présente donc des caractères qui rendent le texte similaire à un texte narratif, par exemple la présence de personnages, objets de l'annotation.

questionnaire sociobiographique visant à reconstruire leur répertoire langagier et les années scolaires de permanence dans des pays autres que l'Italie. Parmi les données recueillies, nous trouvons le ou les pays d'origine de la famille, ainsi que les langues parlées en famille.

De manière similaire à notre projet et à d'autres projets existants (Pallotti et al., 2021, cité par Perugini & Zucchini, 2025), l'annotation est limitée aux personnages de la narration, et donc aux « référents à sémantique animée » (Perugini & Zucchini, 2025, p. 95, notre traduction). Le schéma d'annotation adopté est composé de trois niveaux différents :

- L'annotation de la fonction discursive, doté de deux traits possibles, introduction et maintien ;
- La typologie du maillon annoté (SN, pronom, etc.) ;
- L'informativité du maillon, dotée de deux traits : *sur-spécifié* et *sous-spécifié (*notre traduction des termes italiens *sovraspecificato* et *sottospecificato*).

Le niveau d'informativité, encodé par deux étiquettes, correspond au niveau d'adéquation entre type de maillon et accessibilité du référent. La *sur-spécification* consiste dans l'utilisation d'un type de maillon trop informatif ou trop marqué par rapport au statut d'accessibilité du référent donné ; la *sous-spécification*, au contraire, consiste dans l'utilisation d'un type de maillon peu informatif dans des cas de faible accessibilité du référent et donc dans l'utilisation de maillons insuffisants à récupérer le référent correct (surtout dans des cas de compétition entre référents). En pratique, cette annotation pourrait être envisagée comme un système à trois valeurs, incluant également une valeur « neutre » du maillon.

Cette étude confirme certaines observations déjà formulées dans Barletta et Ponton (2025a), qui formulaient des hypothèses concernant les types de maillons les plus fréquents en italien selon leur position dans la chaine référentielle. Il montrait notamment qu'en italien comme en français, les chaines sont majoritairement introduites par des syntagmes nominaux, tandis qu'en italien, les positions de maintien sont principalement réalisées au moyen d'anaphores zéro.

Les résultats de Perugini et Zucchini (2025) vont dans le même sens. L'analyse des récits d'élèves italiens, monolingues et plurilingues, montre que les syntagmes nominaux dominent nettement en position d'introduction (71 % et 69 %), alors que le maintien repose principalement sur des anaphores zéro (56 % et 54 %), les pronoms et les SN occupant des positions secondaires. Aucune différence statistiquement significative n'est observée entre les deux groupes (χ^2, $p = 0,79$), ce qui suggère des stratégies référentielles similaires, indépendamment du profil linguistique des élèves.

4.6. Synthèse des corpus d'écrits scolaires italiens

L'examen des corpus scolaires constitués en Italie met en évidence la diversité des approches méthodologiques et des objectifs de recherche poursuivis dans ce contexte scientifique. Les ressources présentées couvrent des niveaux scolaires variés ainsi que différents types de productions, allant des récits d'élèves du primaire aux écrits universitaires, en passant par des archives historiques de cahiers scolaires. Ces corpus se distinguent également par la diversité des

traitements linguistiques appliqués aux données, incluant l'annotation des erreurs, l'étiquetage morphosyntaxique ou encore l'analyse de phénomènes discursifs.

Le Tableau 5 propose une vue d'ensemble synthétique des corpus d'écrits scolaires italiens que nous avons décrits dans cette section. Il en récapitule les principales caractéristiques méthodologiques, notamment leur taille (nombre de textes et/ou volume en nombre de tokens), les types de textes, les niveaux scolaires, les formats de diffusion des données ainsi que les niveaux d'annotation disponibles. Cette présentation comparative permet de situer chaque ressource selon les niveaux scolaires représentés, la diversité diamésique et des genres textuels contenus et la profondeur d'analyse et d'annotation linguistique proposée pour chaque corpus.

Corpus	Année	Taille	Type de textes	Niveaux scolaires	Format de diffusion	Niveaux d'annotation
***ISACCO* (Brunato & Dell'Orletta, 2015)**	2016	21 338 tokens	Oral et écrit - *Retelling* d'un texte narratif	École primaire de 2ème à 4ème classe (du CE1 au CM2).	-	-
***CItA* (Barbagli et al., 2016)**	2016	369 456 tokens[46]	Écrit - essais variés (narratifs, descriptifs, argumentatifs, réflexifs, explicatifs)	Etude longitudinal sur deux années de collège.	Shared Task Grammatical Error Correction 2013 (XML)	-
***CoDiSSc* (Revelli, 2011)**	2011, 2014, 2024	1000 cahiers d'école, 280 écrits scolaires de types variés[47]	Écrit - cahiers scolaires historiques	Ecole primaire italienne sur une période de 100 ans et plus.	Plateforme de consultation des scans en ligne	Transcription (non diffusée)
***Univers-ITA* (Grandi et al., 2023)**	2023	2 137 textes, 810 715 tokens[48]	Écrit - textes formels sur la didactique à distance	Etudiants universitaires, universités et domaines disciplinaires variés.	Plateforme de consultation et d'exploration du corpus en ligne	Lemmatisation, POS tagging, annotation en dépendance (formalisme ISST-TANL).
(Perugini & Zucchini, 2025)	2025	40 textes, 17 692 tokens	Écrit – productions narratives écologiques	Etudiants de collège (13-14 ans)	Non diffusé	Chaines de référence : fonction discursive, typologie et informativité du maillon.

Tableau 5. Tableau récapitulatif des corpus d'écrits scolaires italiens analysés dans ce chapitre.

[46] Ce chiffre est tiré de (Barbagli et al., 2016, p. 89). Dans Barbagli et al. (2015, p. 32) la taille du corpus varie légèrement : « Il corpus, composto da 1.352 testi (366.335 tokens) (…) ».

[47] Ce chiffre est reporté sur la page de caractérisation du corpus, disponible au lien suivant : https://codissc.it/consistenza-archivio. Consulté le 18/08/2025.

[48] Ce chiffre est disponible dans les informations sur le corpus au lien suivant : https://corpora.ficlit.unibo.it/CUSP/crystal/index.html#dashboard?corpname=UniverS-Ita. Consulté le 29/09/2025.

Malgré cette richesse, l'analyse comparative montre que ces ressources présentent certaines limites au regard des objectifs de notre recherche. En particulier, peu de corpus permettent d'observer le développement des compétences d'écriture de manière longitudinale à l'école primaire, et les annotations portant sur la continuité référentielle restent encore rares ou pas diffusées. Les travaux exploratoires récents consacrés à l'annotation des chaines de référence confirment toutefois l'intérêt de cette perspective pour analyser la structuration des textes narratifs produits par les élèves. Ces recherches ouvrent ainsi des pistes méthodologiques et analytiques pertinentes pour l'étude de la gestion de la référence dans les productions écrites d'apprenants.

Au-delà de leur complémentarité, aucune de ces ressources n'a été recueillie de manière longitudinale et à l'école primaire en Italie. En particulier, si le corpus *ISACCO* apporte des données utiles pour l'étude du développement de la narration à l'école primaire, il reste de taille modeste et ne permet pas une observation diachronique complète sur plusieurs années, étant limité à deux niveaux scolaires ayant été recueillis au même temps. L'étude de Zucchini et Perugini adopte une annotation de la continuité référentielle proche de la nôtre, mais appliquée à des niveaux scolaires plus avancés, ce qui en fait un complément pertinent pour appréhender le développement de la continuité référentielle en italien au-delà de l'école primaire. Leur travail repose toutefois sur un nombre restreint de textes et sur une annotation essentiellement manuelle, y compris pour certains traits qui, comme nous le montrerons, pourraient être au moins en partie automatisés, tels que la détermination du type grammatical de la mention (cf. Chapitre 7, §5 pour l'explication de notre solution à cette problématique). Une harmonisation avec les protocoles et guides d'annotation déjà établis favoriserait ainsi la comparabilité des annotations et l'intégration de leurs données dans un ensemble plus large de corpus similaires.

Par contraste, le corpus que nous développons constitue, à notre connaissance, le premier en italien qui combine une collecte longitudinale sur plusieurs niveaux scolaires élémentaires à une partie annotée en continuité référentielle, ce qui ouvre de nouvelles perspectives pour l'étude de la cohérence et de la cohésion textuelles en acquisition de l'écrit.

5. Synthèse de chapitre : corpus scolaires en France et en Italie

Ce chapitre avait pour objectif de situer notre recherche dans le paysage des corpus scolaires existants en France et en Italie, en mettant en évidence les principales ressources disponibles pour l'étude des productions écrites d'élèves. L'analyse des différents corpus présentés montre que les corpus scolaires constituent aujourd'hui des ressources essentielles pour documenter les pratiques d'écriture en contexte d'apprentissage et pour observer le développement des compétences langagières au fil de la scolarité.

Dans le contexte français, plusieurs projets ont contribué à structurer ce domaine de recherche. Les premiers corpus d'écrits scolaires ont posé les bases de la numérisation et de l'exploitation linguistique de ces données, ouvrant la voie à des projets plus récents et de plus grande ampleur. Le projet E-Calm constitue à cet égard une étape importante, en fédérant plusieurs ressources

couvrant différents niveaux scolaires et différents types de productions. Parmi ces corpus, *RésolCo* occupe une place particulière dans la mesure où il propose une annotation systématique de la continuité référentielle dans des textes d'élèves, ce qui en fait une ressource de référence pour l'étude de la gestion des référents dans les écrits scolaires.

L'examen des corpus constitués en Italie met en évidence une situation différente. Si plusieurs ressources existent pour l'étude des productions d'apprenants, celles-ci se distinguent par une grande diversité d'objectifs et de méthodologies : certains corpus visent l'analyse d'erreurs linguistiques, d'autres proposent des ressources diachroniques ou des corpus comparant différentes modalités de production (orale et écrite). Toutefois, malgré cette richesse, peu de ressources permettent d'observer de manière systématique le développement des compétences d'écriture à l'école primaire, et les annotations portant spécifiquement sur les phénomènes de continuité référentielle restent encore limitées.

La comparaison entre les deux contextes nationaux met ainsi en évidence à la fois la complémentarité et les lacunes des ressources disponibles. D'une part, les corpus existants offrent une grande diversité de genres textuels, de niveaux scolaires et de niveaux d'annotation linguistique. D'autre part, ils révèlent certaines limites pour l'étude de la continuité référentielle dans les écrits d'élèves, notamment en ce qui concerne la disponibilité de corpus longitudinaux et la comparabilité des schémas d'annotation. C'est dans ce contexte que s'inscrit le corpus mobilisé dans ce travail. Le corpus *Scolinter* se distingue en effet par sa conception comparative et par son recueil dans plusieurs contextes linguistiques, tout en reposant sur un protocole de collecte et de traitement homogène. En combinant une perspective longitudinale et une annotation dédiée à la continuité référentielle, ce corpus vise à combler certaines des lacunes identifiées dans les ressources existantes et à fournir un matériau empirique adapté à l'étude de la gestion des référents dans les écrits d'élèves.

La présentation détaillée de ce corpus, ainsi que des choix méthodologiques qui ont présidé à sa constitution et à son annotation, fait l'objet du chapitre suivant.

Chapitre 3. La coréférence en TAL et les corpus annotés en coréférence pour le TAL

1. La coréférence en Traitement Automatique des Langues

Les chaines de référence jouent un rôle central dans la construction de la cohérence textuelle. Leur identification constitue par conséquent un enjeu important pour de nombreuses tâches du Traitement Automatique des Langues, où la résolution de la coréférence intervient comme module au sein des chaines de traitement, notamment pour la traduction automatique (Dinarelli et al., 2024; Lopez, 2023), l'extraction d'informations, les systèmes de question-réponse (Jang et al., 2025) ou encore la génération automatique de résumés (Martinelli et al., 2024).

Cette section présente d'abord les principales approches utilisées en TAL pour la résolution de la coréférence, du point de vue théorique et technique. Par la suite nous allons décrire quelques-uns des corpus annotés en coréférence, en précisant leurs objectifs et les notions théoriques sous-jacentes à leur création et annotation. Nous allons détailler leur taille, le schéma d'annotation appliqué et les phénomènes linguistiques pris en compte.

1.1. Définition de la tâche de résolution de la coréférence

La tâche de résolution de la coréférence, telle qu'elle est définie en Traitement Automatique des Langues, consiste traditionnellement à détecter et annoter automatiquement les expressions linguistiques qui désignent un référent spécifique dans l'univers du texte analysé. Ces expressions peuvent être reliées entre elles ou associées au référent qu'elles désignent. Il convient par ailleurs de noter qu'en TAL, la différence entre coréférence, chaine de coréférence et chaine de référence est moins tranchée qu'en linguistique, et ces trois définitions sont utilisées de manière quasiment synonymique : « (…) les spécialistes de TAL assimilent coréférence et chaine de référence en un seul objet d'étude, le plus souvent sous le nom *chaine de coréférences*, où le mot *coréférence* apparait indifféremment au singulier ou au pluriel. » (Landragin, 2021, p. 12).

En TAL, cette tâche a connu des évolutions importantes au cours des dernières années, tant dans les approches théoriques employées pour la conception de modèles de détection que dans la création de ressources annotées en coréférence, nécessaires à l'entrainement et à l'évaluation de ces systèmes. Comme le souligne Stuckard (2016, p. 10) « les progrès considérables réalisés dans le développement des systèmes de résolution des anaphores et des coréférences depuis la moitié des années 1990 n'auraient pas été possibles sans la mise à disposition de ressources appropriées, dont les corpus annotés en relations anaphoriques constituent l'un des éléments clés. »[49] (notre traduction) En effet, même si des approches non-supervisées ont été explorées pour cette tâche, les meilleures performances ont généralement été obtenues par des systèmes d'apprentissage supervisés, qui s'appuient précisément sur les corpus annotés pour l'entrainement et l'évaluation

[49] Dans le texte d'origine, « The huge advances in the development of operational anaphora and coreference resolution systems since the mid of the 1990s would not have been possible without appropriate resources having become available, one key component of which are corpora annotated with anaphoric relations. »

(Ng, 2016, p. 283). La disponibilité de corpus annotés de grande taille apparait ainsi comme un facteur déterminant dans les avancées techniques pour cette tâche.

La résolution de la coréférence a une longue histoire dans le domaine du TAL, comme tâche autonome ou comme partie de chaines de traitement plus larges. La description de la tâche donnée par Karttunen dans « Discourse Referents » (1976) est probablement la première occurrence de formalisation de la nécessité de reconnaissance des entités décrites dans un texte pour en saisir le sens de manière computationnelle :

> « Considérons un dispositif conçu pour lire un texte dans une langue naturelle, l'interpréter et stocker son contenu d'une manière ou d'une autre, par exemple dans le but de pouvoir répondre à des questions à son sujet. Pour accomplir cette tâche, la machine devra au moins remplir les conditions de base suivantes. Elle doit être capable de créer un fichier contenant des enregistrements de tous les individus, c'est-à-dire les événements, les objets, etc. mentionnés dans le texte et, pour chaque individu, d'enregistrer tout ce qui est dit à son sujet. Bien sûr, pour l'instant du moins, un tel interprète de texte ne semble pas être une idée pratique, mais cela ne doit pas nous décourager d'étudier de manière abstraite les capacités que la machine devrait posséder, à condition que notre étude nous fournisse un aperçu du langage naturel en général. » [50] (Karttunen, 1976, p. 2, notre traduction)

Le terme *coréférence* a été popularisé dans le domaine du TAL à partir de la campagne d'évaluation « SemEval » de la Message Understanding Conference-6 (MUC Consortium, 1995b). Cette définition, visant à affranchir la tâche de la simple détection d'anaphoriques, fournit une définition des éléments qu'il faut considérer comme *markables* (c'est-à-dire maillons de la chaine de référence ou mentions). Il faut aussi noter que la définition de la tâche exclut tout type d'anaphore résomptive des expressions à annoter. Les entités nommées sont incluses dans la définition de la tâche.

> « La relation de coréférence sera marquée entre les éléments des catégories suivantes : noms, syntagmes nominaux et pronoms. Les éléments de ces catégories sont marquables. [...] La relation est marquée uniquement entre des paires d'éléments qui sont tous deux *markables*. »[51] (MUC Consortium, 1995a, notre traduction)

[50] Dans le texte d'origine, « Consider a device designed to read a text in some natural language, interpret it, and store the content in some manner, say, for the purpose of being able to answer questions about it. To accomplish this task, the machine will have to fulfill at least the following basic requirement. It has to be able to build a file that consists of records of all individuals, that is, events, objects, etc., mentioned in the text and, for each individual, record whatever is said about it. Of course, for the time being at least, it seems that such a text interpreter is not a practical idea, but this should not discourage us from studying in the abstract what kind of capabilities the machine would have to possess, provided that our study provides us with some insight into natural language in general. »

[51] Dans le texte d'origine, « The coreference relation will be marked between elements of the following categories: nouns, noun phrases, and pronouns. Elements of these categories are markables. [...] The relation is marked only

Depuis l'introduction de ce premier modèle de définition de la tâche, les chercheurs de ce domaine se sont interrogés sur la délimitation du phénomène et sur sa définition, sur le lien entre les expressions et le référent indiqué, ainsi que sur la nature des expressions à considérer comme coréférentielles.

1.2. Modélisation de la tâche

En ce qui concerne la définition de la résolution de coréférence, au moins trois grandes conceptions se distinguent, comme le synthétise Grobol (2020, pp. 13-15) : le modèle issu des campagnes d'évaluation du MUC Consortium, le modèle *pair centric* (ou *mention-pair*) et le modèle *entity centric* (ou entity-mention).

Dans le cadre des campagnes d'évaluation du MUC, la résolution de la coréférence consiste dans le processus de détection de chaque mention m dans un document. Une mention doit appartenir à une catégorie syntaxique déterminée (nom, syntagme nominal, pronom) et doit être coréférente avec au moins une mention m' qu'on trouve précédemment dans le texte.

La définition suivante de la tâche donnée lors du MUC-7 (Hirschman, 1998) reconnait de manière plus explicite le fait que les relations qui incluent des verbes sont « pour le moment » exclues des relations de coréférence et que la fonction primaire de la couche d'annotation en coréférence est d'annoter toutes les mentions d'une entité déterminée :

> « La définition de la tâche a été élaborée afin de saisir un niveau d'information, les expressions coréférentielles, dans le contexte des tâches d'extraction d'informations de la Message Understanding Conference (MUC). La « couche » coréférentielle relie entre elles plusieurs expressions désignant une entité donnée ; pour l'instant, seuls les noms sont reliés, les relations impliquant des verbes étant ignorées. La couche coréférentielle sert à rassembler toutes les mentions d'une entité donnée, y compris celles identifiées dans la tâche « Entité nommée ». Nous pouvons considérer l'annotation de la coréférence comme pouvant potentiellement prendre en charge une sorte de version hyperliée du texte, où les liens relient les mentions d'une entité donnée. »[52] (Hirschman, 1998, Partie 1.1, notre traduction)

La tâche définit ainsi l'objet d'annotation en termes d'*identité d'entité*, critère déjà mobilisé dans l'évaluation des systèmes d'extraction d'information. Dans ce cadre, les relations impliquant des segments discursifs plus larges comme l'anaphore associative ainsi que certaines expressions référentielles de nature verbale, telles que l'anaphore zéro, sont exclues.

---

between pairs of elements both of which are markables. This means that some markables that look anaphoric will not be coded, including pronouns, demonstratives and definite NPs whose antecedent is a clause rather than a markable. »

[52] Dans le texte d'origine, « The task definition has been constructed to capture one level of information, coreferring expressions, in the context of the Message Understanding Conference (MUC) information extraction tasks. The coreference "layer" links together multiple expressions designating a given entity; for now, only nouns are linked -- relations involving verbs are ignored. The coreference layer functions to collect together all mentions of a given entity, including those tagged in the Named Entity task. We can look at coreference annotation as potentially supporting a kind of hyperlinked version of the text, where the links connect the mentions of a given entity.

Cette définition de la tâche de coréférence ne correspond pas tout à fait à la définition de coréférence de la sémantique formelle ou de la linguistique : « Étant donné que l'utilisation du terme « coréférence » en linguistique informatique est désormais courante, nous noterons ici que la tâche de coréférence telle que définie par les directives du MUC n'est pas la même chose que la coréférence au sens de la sémantique formelle ou de la linguistique fonctionnelle. »[53] (Poesio, 2016, p. 39). En linguistique, la notion de coréférence comprend des notions anaphoriques, telles que l'anaphore associative ou encore l'anaphore résomptive que nous avons déjà évoquées ci-dessus, notions que cette définition initiale de résolution de la coréférence tend à exclure. Enfin, depuis la définition donnée lors du MUC, la communauté a fait évoluer les définitions et les ressources : les tâches contemporaines de résolution de la coréférence distinguent généralement deux phases, celle de détection des mentions et celle du regroupement de ces mentions en clusters (ou chaines de coréférence), et des campagnes d'évaluations ultérieures, liées à des jeux de données (*OntoNotes/CoNLL, ARRAU*, etc.) modulent la portée des phénomènes annotés en incluant ou en excluant, selon les objectifs, des cas tels que le bridging, les singletons ou certaines anaphores nominales complexes.

Le modèle *mentions pair* ou *link-centric*, qui est un modèle d'apprentissage supervisé, configure la tâche de résolution de la coréférence comme une tâche de classification binaire, où un classifieur est entraîné à la prise de décision de coréférentialité entre deux mentions à la fois (souvent, des syntagmes nominaux).

Dans le corpus d'entrainement donné en entrée au modèle, chaque paire de syntagmes nominaux est représentée sous forme d'un vecteur contenant des informations telles que la distance entre eux, les traits morphosyntaxiques, des informations lexicales, syntaxiques, sémantiques et une étiquette indiquant la relation (coréférentielle ou pas) entre les deux syntagmes. Ainsi le classifieur binaire peut généraliser son apprentissage à de nouvelles données et décider quelles paires de mentions sont coréférentielles entre elles ou pas. Dans l'étape suivante, ces décisions sont prises en considération pour constituer des chaines de référence par clustérisation (Hoste, 2016). Ce dernier modèle a représenté le modèle *benchmark* tout au long des années 2000, après avoir été popularisé par Soon et al. (2001). Cette approche a été décrite par la suite comme adaptable à différents corpus, dans la mesure où elle repose sur des traits génériques plutôt indépendants du domaine (Hoste, 2016). Cet algorithme était ciblé à la résolution de l'anaphore m_j effectuée de manière telle à trouver l'antécédent m_i le plus probable, selon la formule suivante :

$$\operatorname{argmax}_{m_i} P\left(L \mid m_j, m_i\right)$$

La mise en œuvre du modèle *link-centric* suppose ainsi l'intégration en amont de nombreux modules de pré-traitement linguistique dont la performance est déterminante, notamment pour la segmentation, l'annotation morphosyntaxique et l'identification des groupes nominaux, car

[53] Dans le texte d'origine, « As the C[omputational]L[inguistic] use of the term coreference is here to stay, we will note here that the 'coreference task' as defined by the MUC guidelines is not the same as coreference either in the sense of formal semantics or in the sense of functional linguistics. »

toute erreur ou omission à ce stade du traitement peut affecter directement le résultat de la résolution de coréférence.

Les approches *entity-mention*, en revanche, modélisent directement les liens entre mentions et entités correspondantes, plutôt que les relations entre paires de mentions. Cette orientation se retrouve notamment dans le corpus *OntoNotes* (Hovy et al., 2006) et dans les systèmes déterministes de type *sieve* (ou passoire) décrits par Lee et al. (2013). Dans un modèle *entity-centric*, les mentions qui indiquent la même entité sont modélisées conjointement, ce qui permet au système de résolution d'utiliser l'information contextuelle globale des clusters déjà constitués pour guider les décisions ultérieures de clustérisation et détection des chaines dans le texte.

Lee et al. (2013) décrivent l'un des premiers systèmes largement diffusés et qui a formalisé et popularisé une approche de résolution de la coréférence explicitement *entity-centric*. Leur système se base sur une architecture déterministe et modulaire : une étape de détection de mentions est suivie par un filtrage de toutes expressions non-coréférentielles (par exemple les pronoms impersonnels pléonastiques « *it* is possible that, *it* appears that..) ; les expressions de quantité et les mentions contenant des adjectifs partitifs ou *quantifiers*, etc.) suivies d'un module doté d'une architecture déterministe de type *sieve* (ou passoire) fondée sur une succession de règles à précision décroissante qui détermine le lien entre chaque mention et un référent donné dans le texte.

Dans une taxonomie plus récente, Boiteau et al. (2025) distinguent deux grandes familles de représentation des chaines de référence dans leur travail consacré aux types de modélisation de la coréférence : une représentation relationnelle et une représentation ensembliste. Ces deux approches sont illustrées dans la Figure 10, issue de leur article.

La *modélisation relationnelle* est celle qui représente le mieux la notion de chaine de référence comme décrite par Schnedecker (2021) : lorsqu'un maillon *m* réfère à une entité *e* et un deuxième maillon *m'* fait référence à la même entité *e*, il y a relation de coréférence entre les deux maillons. Autrement dit, ici la chaine de référence est construite à partir de relations explicites entre mentions. Selon Boiteau et al. (2025), cette formalisation est représentée par les annotations des corpus *OntoNotes* (Hovy *et al.*, 2006), *ARRAU* (Poesio et al., 2013), *CO2* (Schang et al., 2011) et *ANCOR* (Muzerelle et al., 2013)[54].

La *modélisation ensembliste* consiste, quant à elle, à annoter « uniquement le lien entre le maillon et son référent » (Boiteau et al., 2025, p. 438). La chaine est conçue comme un ensemble de mentions rattachées à une même entité référentielle. L'annotation faisant partie de notre travail de thèse est identifiée comme exemple de modélisation ensembliste, tout comme celle effectuée sur le corpus *DEMOCRAT* (Landragin, 2020b).

Cependant, dans le texte de l'article, parmi ces schémas (cf. Figure 10), seuls certains sont explicitement liés à des corpus ou des schémas d'annotation existants. Le schéma 3 est présenté comme représentatif du modèle d'annotation du corpus *ARRAU* (Poesio et al., 2013). Le schéma 6

[54] Ces corpus seront décrits plus en détail dans la section 3 de ce chapitre, dédiée aux corpus annotés en coréférence en France.

correspond à l'annotation du corpus *DEMOCRAT* (Landragin, 2020b), tandis que le schéma 7 est censé représenter notre propre approche d'annotation, mise en œuvre dans l'annotation du corpus *Scolinter* (LIDILEM, 2025; Ponton et al., 2021), et rapprochée par les auteurs des propositions de Zhou et Choi (2018) notamment en ce qui concerne l'annotation des mentions de type groupe.

Dans cette perspective, les approches illustrées par les schémas 6 et 7, correspondant respectivement aux corpus *DEMOCRAT* et *Scolinter*, peuvent être qualifiées *d'extensionnelles*, au sens où elles modélisent les chaines de référence comme des ensembles de mentions partageant une même référence, sans imposer de structure relationnelle interne explicite entre les maillons. Enfin, le schéma 8 présente une proposition théorique originale formulée par Boiteau et al. (2025). Cette proposition constitue le point de départ d'un nouveau programme d'annotation de la coréférence, visant à contourner certaines difficultés récurrentes liées à l'annotation des maillons de type groupe. Ces difficultés ont été identifiées par les auteurs à partir de l'analyse critique de plusieurs guides d'annotation existants, dont le nôtre[55].

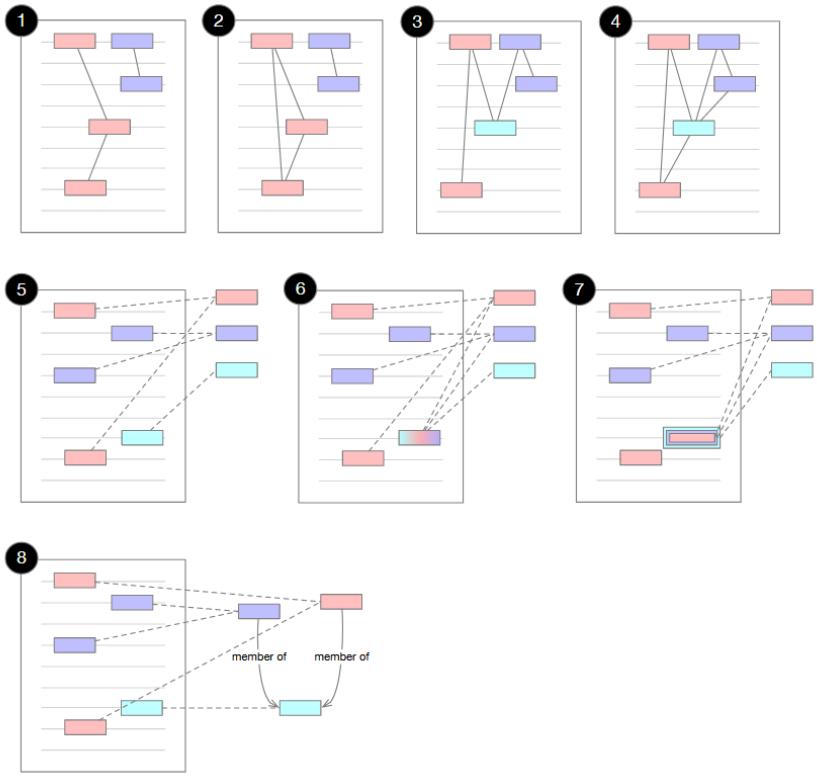


Figure 10. Exemples de modélisations relationnelles (1 à 4) et ensemblistes (5 à 7). Figure tirée de Boiteau et al., 2025, p. 438.

Dans leur publication, Boiteau et al. (2025) mentionnent explicitement le guide d'annotation de notre corpus (Barletta, 2025). Ils soulignent certaines limites de notre annotation, notamment pour les maillons de type *groupe*, identifiés par l'ajout de plusieurs étiquettes correspondant aux entités impliquées, ce qui entraînerait selon eux une perte d'information structurelle. Nous

[55] Les auteurs de l'article avaient accès à notre guide d'annotation, que nous leur avions transmis sur demande avant la publication du document.

considérons au contraire que cet encodage compositionnel permet précisément de conserver la structure groupale, puisque la co-présence des étiquettes rend cette information reconstructible, même si elle a jusqu'à présent été peu exploitée dans nos analyses.

Après avoir présenté certains modèles de résolution de la coréférence, il convient désormais d'examiner les corpus annotés disponibles. Des états de l'art assez complets étant déjà disponibles, à l'international comme en France (entre autres, Grobol (2020b); Lopez (2023); Sukthanker et al. (2020)), nous allons limiter notre analyse à deux des ressources les plus connues, avant d'aborder de manière plus approfondie les ressources françaises existantes.

2. Corpus annotés en coréférence

Les corpus annotés en coréférence constituent une ressource centrale tant pour le développement de modèles d'apprentissage supervisé que pour l'étude empirique des mécanismes référentiels à l'œuvre dans le discours. Au-delà de leur rôle instrumental pour l'entraînement et l'évaluation des systèmes automatiques, ces corpus incarnent des choix théoriques explicites quant à la définition de la référence, à la granularité des annotations et aux phénomènes linguistiques jugés pertinents.

Parmi les corpus les plus importants en littérature, *OntoNotes* et *ARRAU* occupent une place particulière. Bien qu'ils visent tous les deux l'annotation de la référence, ils se distinguent nettement par leur conception de la coréférence et par les phénomènes qu'ils prennent en charge. Cette opposition en fait des points de comparaison privilégiés pour analyser les implications méthodologiques des schémas d'annotation en coréférence.

2.1. OntoNotes

Le corpus *OntoNotes 5.0* (Hovy et al., 2006), aussi connu comme *CoNLL-2021*, avec ses plus de 2.9 millions de tokens sur trois langues différentes (anglais, arabe et chinois), représente l'un des plus grand corpus multilingues annotés pour des tâches sémantiques, incluant la coréférence, la désambiguïsation lexicale, les rôles sémantiques et la syntaxe. Il contient des genres textuels divers tels que la presse écrite, des programmes radio, des transcriptions de conversations téléphoniques, des données tirées d'internet, ainsi que le Nouveau et l'Ancien Testament. L'annotation de la coréférence dans *OntoNotes* repose sur une conception pragmatique et référentielle : elle regroupe sous une même chaine tous les éléments linguistiques qui renvoient à la même entité dans le monde, indépendamment de leur catégorie grammaticale (nom, pronom, etc.). *OntoNotes* se distingue notamment par l'exclusion des coréférences de type cataphorique et des chaines d'événements, qui sont traitées comme des phénomènes distincts. Les singletons sont également exclus de l'annotation.

L'importance d'*OntoNotes* dans la recherche en TAL tient à son rôle de corpus de référence pour l'évaluation des systèmes de résolution de coréférence, notamment dans le cadre du *CoNLL Shared Task* (2011–2012). Son ampleur et sa standardisation en font un support fondamental pour l'entraînement de modèles neuronaux contemporains.

2.2. ARRAU

Le corpus *ARRAU* (*Articulated Reference and Anaphora Annotated Corpus* (Poesio et al., 2013, 2018)) est constitué par plusieurs sous-corpus contenant des textes de types variés : le *RST Discourse Treebank* constitué d'articles du *Wall Street Journal*, le corpus *GNOME*, qui rassemble des textes techniques, médicaux et d'histoire de l'art, *TRAINS*, composé de dialogues orientés tâche, ainsi que *PEAR*, un corpus de récits oraux transcrits. Cette diversité générique et modale vise à évaluer la robustesse des modèles de résolution de la coréférence sur des types de discours contrastés.

Bien qu'il soit plus petit que le corpus *OntoNotes*, le corpus *ARRAU* compte plus de 350 000 tokens (Poesio & Artstein, 2008). Il propose une annotation fine des expressions référentielles, incluant les mentions nominales et pronominales, les cas cataphoriques, les références plurielles ou distributives, ainsi que des phénomènes traditionnellement moins pris en compte dans les corpus antérieurs, tels que le *bridging*, les *singletons* ou certaines anaphores abstraites. Cette extension du schéma d'annotation permet de dépasser une conception strictement centrée sur l'identité d'entité pour intégrer des formes de référence plus discursives, tout en conservant une modélisation principalement fondée sur les relations entre mentions.

3. Corpus annotés en coréférence en France

La linguistique de corpus française s'intéresse depuis des décennies à l'étude et la description des phénomènes de construction de la textualité comme l'anaphore et les chaines de référence (Charolles, 1988, p. 198 ; Chastain, 1975 ; Corblin, 1985, 1995). Déjà dans ces premiers travaux, l'autoanalyse traditionnelle de la compétence interne du linguiste était parfois épaulée par la présence de données empiriques. C'est le cas dans une publication de Charolles (1988a), qui analyse un petit corpus d'écrits d'élèves pour étudier la gestion des risques de confusion entre personnages dans une tâche d'écriture narrative.

Au-delà de ces premières études, effectuées sur des corpus de taille restreinte et dont les données n'étaient pas diffusées, aucun autre exemple de corpus annoté en coréférence n'existait en France avant les premiers travaux, sortis entre 2000 et 2013, qui sont représentés par le corpus *ARCADE* de Tutin et al. (2000) et le corpus *DéDé* de Gardent et al. (2005). Par la suite nous allons décrire plus en détails les corpus annotés en coréférence de grande taille les plus importants en France : *Annodis*, déjà cité lors de la description du corpus *RésolCo*, le corpus *ANCOR*, et le corpus *DEMOCRAT*.

Les projets *Annodis* (Péry-Woodley et al., 2011), *ANCOR* (Garcia-Debanc et al., 2019) pour l'oral spontané, et *DEMOCRAT* (Landragin, 2016a) pour la langue écrite représentent les trois corpus majeurs qui encodent l'anaphore et/ou la coréférence à différents niveaux. Nous allons par la suite présenter les objectifs et la composition de ces corpus, ainsi que les méthodes d'annotations adoptées et les phénomènes qu'ils encodent. Le cas du corpus *RésolCo* (Garcia-Debanc et al., 2019, 2021) ayant déjà été présenté dans le Chapitre 2 (cf. §3.1), ici nous rappellerons seulement que ce dernier représente le seul corpus français d'écrits scolaires de taille moyenne annoté en continuité

référentielle, ce qui le rend similaire en objectifs et en méthodologie au corpus de notre propre travail de recherche.

3.1. Corpus annotés avant 2013 : ARCADE et DéDé

Bien que des corpus annotés en coréférence existent déjà en français, leurs caractéristiques ne les rendent pas globalement représentatifs de la coréférence pour deux raisons principales : soit les annotations codent seulement certains types d'anaphore, comme dans le cas du corpus *ARCADE* (Tutin et al., 2000), soit parce qu'ils représentent seulement un certain type de mention, comme dans le cas du corpus *DéDé* (Gardent et al., 2005).

Le corpus *ARCADE*, développé par l'équipe CRISTAL (Tutin, 2002; Tutin et al., 2000), constitue l'un des premiers corpus français de grande taille annoté en relations anaphoriques, avec environ un million de tokens. Il représente une étape importante dans la mise à disposition de ressources outillées pour l'étude de l'anaphore en français, bien qu'il ne s'agisse pas à proprement parler d'un corpus annoté en chaines complètes de coréférence. Le schéma d'annotation se concentre en effet principalement sur les expressions anaphoriques appartenant à des classes morphosyntaxiques fermées, en plus d'identifier leurs antécédents. Sont notamment annotés les pronoms personnels troisième personne (à l'exception des formes réfléchies), les déterminants possessifs, les pronoms démonstratifs (hors formes neutres telles que *ce, ça, cela, ceci*), les indéfinis et numéraux (comme *un ensemble*, *la plupart*), l'expression multimots « le + faire », certains adverbes anaphoriques (*dedans, dessus*), les ellipses nominales ainsi que des expressions dites *pointers* (*ce dernier, le premier*). Ces unités sont reliées à leur antécédent et les annotations précisent également la nature discursive ou sémantique de la relation (Tutin et al., 2000). En revanche, les syntagmes nominaux lexicaux ordinaires, notamment les descriptions définies, ne font pas l'objet d'une annotation systématique (Gardent et al., 2005; Tutin et al., 2000).

Par conséquent, si *ARCADE* constitue l'un des premiers grands corpus annotés pour l'étude de l'anaphore en français, il ne permet pas la reconstruction exhaustive de chaines de coréférence telles qu'elles sont définies dans les travaux plus récents.

Le corpus *DéDé* (*Definites Descriptions*), de son côté, cible l'annotation des descriptions définies pour en rendre possible l'apprentissage automatique (Gardent et al., 2005). Le but de ce corpus était d'avoir à disposition un corpus annoté pour l'entrainement et l'évaluation d'outils de résolution de l'anaphore, en particulier « en vue de l'interprétation automatique d'expressions définies, c'est-à-dire les expressions de la forme *le/la/les N* » (Gardent et al., 2005, p. 3). Il est constitué d'une sous-partie du corpus *PAROLE*, composée d'articles du journal *Le Monde* datant de septembre 1987, et il comprend 48 360 tokens annotés au niveau morphosyntaxique. 4910 descriptions au total sont annotées dans ce corpus. L'annotation réalisée prend en compte quatre catégories différentes d'expressions à annoter : les descriptions autonomes ; les descriptions coréférentielles ; les descriptions contextuelles et les descriptions non référentielles. Sa méthodologie d'annotation prévoyait l'utilisation d'outils de prétraitement de l'annotation (annotation morphosyntaxique) pour faciliter cette tâche, réalisée par des linguistes expérimentés. Le schéma d'annotation utilisé était affiné par itérations après une première

annotation (Gardent et al., 2005), pratique commune en annotation de corpus, qui prévoit des aller-retour entre le processus d'annotation et la stabilisation du guide d'annotation prévu pour cette tâche.

Néanmoins, malgré leurs limites, ces projets ont contribué à ouvrir la voie aux réflexions sur l'annotation de ces phénomènes sur des corpus de grande taille, ciblés dans certains cas à l'entrainement de modèles *machine learning*.

3.2. Annodis

Le corpus *Annodis* (Péry-Woodley et al., 2011), qui précède le corpus *DEMOCRAT* en élaboration et publication, tout en ayant une taille comparable à ce dernier, ne représente pas un corpus annoté en coréférence car il encode des chaines topicales, c'est-à-dire des segments caractérisés par l'apport d'informations au sujet d'un seul et même référent (Federzoni et al., 2020). Il reste cependant un exemple de corpus de grande taille, qui a inspiré les méthodes d'annotation ensuite utilisées pour annoter le corpus *RésolCo* (Garcia-Debanc *et al.*, 2019, 2021, cf. Chapitre 2, § 3.1). Il contient 156 textes ou extraits de textes pour un total d'environ 687 000 tokens. Ce corpus est composé de textes variés selon plusieurs critères, comme « genre, longueur et le type d'organisation discursive » (Péry-Woodley et al., 2011, p. 72). Il contient cinq sous-corpus : *NEWS*, composé de brèves issues du journal *L'Est Républicain* ; *WIK1* et *WIK2* qui contient des articles Wikipedias ; *LING* contenant 25 articles scientifiques en linguistique et *GEOP* articles ou rapports de géopolitique. Ce corpus « est le résultat de deux types d'annotations manuelles » (Péry-Woodley et al., 2011, p. 72), : une annotation qui part des unités élémentaires du discours pour reconstruire les relations rhétoriques entre unités du texte, et une deuxième annotation « sélective de structure multi-échelle », dont les structures énumératives et les chaines topicales. Dans ce contexte, les chaines topicales sont définies comme « des segments de texte regroupant des phrases reliées par un référent commun. » (Péry-Woodley et al., 2011, p. 74).

3.3. ANCOR

Le corpus *ANCOR* représente « le premier corpus d'oral spontané d'envergure annoté en coréférence » et en anaphore pour la langue française et « distribué librement » (Muzerelle et al., 2013). Il est composé de plusieurs corpus de parole spontanée transcrite. Son objectif était de répondre au manque d'un corpus « suffisant pour apprendre un système de résolution de la coréférence efficace » (Muzerelle et al., 2013), dans un moment où d'autres langues majoritaires étaient déjà dotés de tels corpus.

Le corpus *ANCOR* a fait l'objet d'une annotation déportée (*stand-off annotation*), réalisée à l'aide du logiciel *Glozz* (Widlöcher & Mathet, 2012). Cela permet de dissocier les annotations du texte source et de superposer plusieurs couches d'annotation indépendantes. L'annotation a été réalisée en double par des annotateurs experts, puis validée par supervision, afin de garantir la cohérence des décisions des annotateurs (Muzerelle et al., 2013, p. 556). Une annotation relative à la transcription de l'oral est encodée dans la structure XML du corpus (identification des locuteur, segmentation des tours de parole) (Antoine et al., 2013). Les mentions annotées sont les groupes

nominaux, les pronoms et les groupes prépositionnels, où la préposition n'est pas incluse dans l'annotation mais signalée à l'aide d'un attribut. Le pronom *ça* et ses dérivés ont été exclus de l'annotation, mais les formes explétives du pronom il (par ex., « *Il* pleut ») ont été en revanche annotées, car il est important pour les auteurs « de repérer ces usages non référentiels qui peuvent tromper les systèmes de résolution » (Muzerelle et al., 2013, p. 560).

Les annotations de ces expressions incluent plusieurs couches, notamment des traits morphosyntaxiques (genre, nombre et catégorie grammaticale des éléments, encodés selon les étiquettes pronom, nom ou autre). Elles intègrent également des informations telles que le statut de définitude (défini, indéfini, démonstratif ou explétif), l'appartenance éventuelle à une entité nommée et son type le cas échéant, ainsi que des indicateurs comme l'inclusion dans un syntagme prépositionnel ou des traits discursifs, par exemple « nouvelle entité du discours » pour les mentions. Les relations entre mentions sont elle aussi typées : le corpus distingue non seulement les liens de coréférence stricte, mais aussi des relations anaphoriques plus larges, telles que les anaphores associatives (comme *le village… son clocher*) (Muzerelle et al., 2013).

Le processus d'annotation a suivi une démarche séquentielle, composée de quatre phases distinctes : 1) Le repérage des entités nommées ; 2) La révision croisée des éléments repérés entre annotateurs et la recherche de consensus ; 3) Le repérage et la caractérisation des relations anaphoriques ; 4) La révision finale des relations par un superviseur.

Toutes les mentions successives sont reliées à la première mention de l'entité référentielle présente dans le texte, plutôt qu'utiliser une approche d'annotation en chaine. Un projet mené successivement a enrichi ce corpus avec une couche d'annotation syntaxique automatique (Grobol et al., 2018). Cependant, ce corpus était représentatif de la langue parlée, un corpus annoté représentatif du français langue écrite était encore absent du panorama.

3.4. DEMOCRAT

Le corpus *DEMOCRAT* répond à ce manque. Il constitue « le premier corpus de grande taille librement disponible pour le français écrit » (Landragin, 2021, p. 12). Il contient des textes écrits, appartenant à plusieurs genres, du roman aux articles journalistiques. Il comprend aussi une partie de textes littéraires historiques. L'objectif de ce projet était de constituer un corpus diachronique de textes, écrits entre le 12e et le 21e siècle, relevant de genres textuels variés, « et d'en autoriser des exploitations par des outils de traitement automatique des langues (TAL), plus précisément par des outils faisant appel à de l'apprentissage profond » (Landragin, 2022). Le choix méthodologique fondamental du projet *DEMOCRAT*, qui le démarque de projets existants similaires, réside aussi dans le fait d'avoir annoté pour la première fois en France les chaines de référence dans leur entièreté dans les textes analysés (Landragin, 2021).

Dans le cadre du projet *DEMOCRAT*, une annotation fine des phénomènes de coréférence a été réalisée directement à partir des textes bruts, sans recours à une pré-annotation automatique des mentions. Des outils de traitement automatique des langues ont toutefois été mobilisés en amont afin de faciliter le travail des annotateurs. Le corpus a ainsi été enrichi par une lemmatisation et

un étiquetage morphosyntaxique effectués automatiquement à l'aide de *TreeTagger* (Schmid, 1999) en utilisant un modèle du français écrit contemporain, sans validation manuelle ultérieure. La tokenisation a, quant à elle, été réalisée avec TXM (Heiden, 2010). Pour l'étape d'annotation, les annotateurs pouvaient choisir le logiciel d'annotation parmi *ANALEC* (Landragin et al., 2012), *SACR* (Oberlé, 2018) et *TXM* (Heiden, 2010) (Quignard et al., 2021).

Ce projet a impliqué plusieurs dizaines d'annotateurs. Pour garantir la bonne réussite du projet, un manuel d'annotation a été élaboré, qui propose des exemples concrets et qui se configure comme « un support de description des annotations particulièrement étayé » (Quignard et al., 2021, p. 27), du point de vue technique comme linguistique. Une partie du manuel d'annotation est dédiée aux questions de coréférence (anaphore associative, liens entre référents génériques et particuliers, discours direct, référence floue ou évolutive).

Le processus d'annotation a suivi des étapes déterminées : 1) Le choix des extraits à annoter, par taille homogène, fixée à 10 000 mots ; 2) Des expérimentations d'annotations chronométrées, pour calculer les efforts requis en termes humains, et pouvoir ainsi choisir la meilleure procédure d'annotation en termes de rapport entre qualité d'annotation, accord inter-annotateur et temps d'annotation humaine ; 3) L'application de la procédure choisie à l'ensemble des textes du corpus ; 4) L'utilisation d'un script informatique pour obtenir des annotations complétées automatiquement à partir des annotations manuelles. Plus précisément, les annotateurs humains se sont occupés de la désignation des mentions contenues dans les textes, et de l'association à chaque mention d'un identifiant qui décrit son référent. Le rôle du script était donc de reconstruire les chaines de référence et de coder les différents types d'annotations nécessaires pour des exploitations TAL (Landragin, 2021).

Les annotations du corpus *DEMOCRAT* reposent sur un schéma strict, conçu de manière à limiter autant que possible l'ambiguïté interprétative et le flou définitionnel. Sans entrer dans le détail exhaustif des choix méthodologiques adoptés par le projet, nous présentons ici les principes les plus saillants tels qu'ils sont décrits dans le manuel d'annotation, et qui ont nourri nos propres réflexions sur l'annotation des textes de notre corpus.

L'annotation vise l'exhaustivité des expressions référentielles : ainsi, toutes les mentions sont prises en compte, y compris les singletons. Les propositions relatives ne sont pas considérées comme des mentions à part entière, mais les pronoms relatifs qu'elles contiennent sont annotés. Les prépositions ne sont généralement pas incluses dans les limites des expressions référentielles, sauf dans le cas d'amalgames morphologiques (par exemple « les moustaches [***du*** chat] »). De même, les pronoms impersonnels ainsi que les clitiques réfléchis non référentiels sont exclus de l'annotation. Les verbes ne sont pas traités comme des mentions, sauf lorsque le sujet n'est pas exprimé : dans ce cas, une indication du sujet zéro est ajoutée afin de préserver la cohérence référentielle. Les groupes nominaux comportant un déterminant possessif sont annotés de façon à expliciter la relation de possession, et les référents évolutifs sont maintenus au sein d'une même chaine lorsque la continuité référentielle est jugée pertinente.

Les travaux effectués sur ce corpus ont également contribué à systématiser la définition des indices de description quantitative des chaines de référence. Ce corpus a permis ainsi de mobiliser des opérationnalisations de ces indicateurs de caractérisation des chaines, tels que leur densité ou leur stabilité référentielle (Obry et al., 2017), et offre un cadre d'analyse cohérent pour comparer ces caractéristiques entre corpus différents. Les annotations réalisées pour Democrat ont été incluses dans la partie « Coref » de *Universal Dependencies* et servent aux campagnes d'évaluation internationales CRAC (*Computational Models of Reference, Anaphora and Coreference*).

3.5. Synthèse des corpus français annotés en coréférence

Corpus	Année	Taille	Genres	Mentions	Format de diffusion	Phénomènes annotés
ARCADE (Tutin et al., 2001)	2001	1 000 000 mots	Textes littéraires, journalistiques, narratifs, articles scientifiques en SHS	Pas spécifié – dans une sous-partie du corpus de 74 969 tokens, 2 121 expressions[56]	XML-TEI	Expressions anaphoriques*
DéDé (Gardent et Manuélian, 2005)	2005	48 360 mots	Articles journalistiques	4 910 descriptions définies, dont 567 coréférentielles	XML pour MMAX	Descriptions définies
Annodis	2011	687 000 mots	Brèves journalistiques, extraits et articles Wikipedia, articles de recherche SHS, rapports et articles en géopolitique	572 chaines topicales	XML pour GLOZZ	Structures énumératives, chaines topicales
ANCOR	2013	487 000 mots	Parole spontanée transcrite	115 000 mentions environ	Transcription, XML, format GLOZZ	Chaines de coréférence ou anaphore
DEMOCRAT	2019	560 000 mots, 58 textes	Écrits narratifs et autre genres variés	198 000 mentions, 20 000 chaines	XML TEI, XML-TEI-URS, format TXM	Chaines de référence
RésolCo	2019	97 798 mots, 407 textes	Écrits scolaires	-	Transcription, XML TEI-P5, format GLOZZ	Continuité référentielle sur les entités provoquées par la consigne

[56] Les données sont tirés de Tutin, (2002).

Tableau 7. Tableau de synthèse des caractéristiques des corpus français annotés en coréférence analysés.

Les différents corpus présentés, tout en ayant en commun un intérêt pour les phénomènes de cohérence et de cohésion textuelles, présentent des spécificités, à la fois liées aux genres textuels annotés, aux schémas d'annotation adoptés ainsi qu'aux objectifs que les annotations contribuent à réaliser.

Le Tableau 7 propose une vue d'ensemble des principaux corpus français annotés pour les phénomènes de coréférence ou, plus largement, pour différentes formes de cohésion référentielle que nous avons décrit dans ce chapitre. Il met en évidence à la fois la diversité des objectifs scientifiques poursuivis et l'hétérogénéité des choix méthodologiques adoptés en matière de taille, de genres textuels, d'unités annotées et de formats de diffusion.

Les premières ressources, comme *ARCADE*, s'inscrivent dans une perspective pionnière d'annotation des expressions anaphoriques à grande échelle, avec un volume important de données mais un schéma d'annotation centré essentiellement sur les phénomènes anaphoriques. Le corpus *DéDé* adopte une approche plus ciblée, en se concentrant exclusivement sur les descriptions définies dans des textes journalistiques. *Annodis*, quant à lui, se distingue par une annotation discursive plus large, orientée vers l'identification de chaines thématiques et de structures textuelles globales plutôt que vers la seule coréférence nominale.

Les ressources plus récentes témoignent d'un élargissement progressif des phénomènes pris en compte. Le corpus *ANCOR*, premier grand corpus annoté en coréférence en France, introduit une annotation détaillée des chaines sur des données orales transcrites. *DEMOCRAT* marque une étape supplémentaire en proposant le premier corpus de grande taille représentatif du français écrit, couvrant une large amplitude diachronique et visant une annotation exhaustive des chaines de référence, pensée dès l'origine pour des exploitations en traitement automatique des langues. Enfin, *RésolCo* applique ces principes au domaine des écrits scolaires, en ciblant spécifiquement la continuité référentielle des entités induites par les consignes d'écriture.

Pris ensemble, ces corpus illustrent ainsi une évolution progressive des pratiques : d'annotations partielles et centrées sur certains marqueurs anaphoriques vers des schémas plus complets, intégrant des chaines de référence entières et des formats normalisés favorisant l'interopérabilité et la réutilisation des données. Cette diversité de ressources constitue un panorama complémentaire, au sein duquel s'inscrit le présent travail.

4. Synthèse de chapitre : coréférence en TAL et corpus annotés

Ce chapitre avait pour objectif de présenter les principaux corpus annotés pour les phénomènes de coréférence et de cohésion référentielle utilisés dans le domaine du traitement automatique des langues. À travers l'examen de plusieurs ressources représentatives, nous avons mis en évidence la diversité des approches adoptées pour modéliser et annoter la référence dans les textes.

Les corpus présentés se distinguent d'abord par la variété des genres textuels et des types de données linguistiques qu'ils couvrent. Certaines ressources, comme *ARCADE* ou *DéDé*, se concentrent sur des phénomènes spécifiques tels que les expressions anaphoriques ou les descriptions définies, dans des genres textuels relativement homogènes. D'autres corpus, tels qu'*Annodis* ou *ANCOR*, adoptent une perspective plus large sur l'organisation discursive et la cohésion textuelle, en intégrant l'annotation de structures discursives ou de chaines de coréférence dans des ensembles de données plus diversifiés.

Les ressources plus récentes témoignent d'une évolution vers des schémas d'annotation plus complets et vers une plus grande standardisation des formats de diffusion. *DEMOCRAT* constitue à cet égard un corpus important, en proposant une annotation exhaustive des chaines de référence dans un corpus de français écrit couvrant plusieurs périodes historiques et différents genres textuels. Cette ressource a également contribué à stabiliser certains choix méthodologiques en matière de représentation des chaines et de formats d'annotation, favorisant ainsi la comparabilité des données et leur exploitation dans des applications de traitement automatique des langues.

L'examen de ces corpus met également en évidence certaines limites. D'une part, les choix d'annotation varient sensiblement selon les objectifs des projets, ce qui rend parfois difficile la comparaison directe entre les ressources. D'autre part, les corpus annotés pour les phénomènes de coréférence concernent majoritairement des textes produits par des locuteurs experts et couvrent encore peu les productions d'apprenants. Dans ce contexte, les ressources annotées portant spécifiquement sur les écrits scolaires demeurent relativement rares.

Parmi celles-ci, le corpus *RésolCo* constitue une référence importante, dans la mesure où il applique les principes d'annotation issus de la linguistique de corpus et du TAL à des productions d'élèves. Toutefois, son schéma d'annotation reste limité aux référents introduits par la consigne d'écriture. Le travail présenté dans cette thèse s'inscrit dans la continuité de ces approches, tout en proposant une annotation de la continuité référentielle adaptée à un corpus d'écrits scolaires recueilli dans deux langues.

La section suivante présentera plus précisément les principes méthodologiques adoptés pour l'annotation des chaines de continuité référentielle dans le corpus étudié.

Conclusion

Cette première partie avait pour objectif d'établir le cadre théorique dans lequel s'inscrit notre travail, consacré à l'annotation des chaines de continuité référentielle dans un corpus d'écrits scolaires.

Le premier chapitre a permis de préciser les notions linguistiques nécessaires à l'analyse de la représentation de la référence dans les textes. Nous y avons distingué les phénomènes d'anaphore et de coréférence. Après avoir présenté les principaux types d'anaphore décrits dans la littérature, nous avons examiné les relations qu'entretient cette notion avec celle de coréférence, avant de proposer une définition opératoire de cette dernière.

L'anaphore et la coréférence entretiennent une relation étroite dans la mesure où ces deux notions renvoient à des expressions qui pointent vers un même référent. Elles se distinguent toutefois par la nature de la relation qu'elles établissent entre les expressions concernées. La relation de coréférence se configure comme une relation symétrique : les deux expressions renvoient à la même entité et leur interprétation référentielle peut être considérée comme indépendante. Dans le cas de l'anaphore, en revanche, l'interprétation de l'expression anaphorique dépend de la présence d'un antécédent dans le contexte discursif. L'antécédent est donc nécessaire pour interpréter le sens de l'expression anaphorique, ce qui confère à la relation anaphorique un caractère asymétrique.

Nous avons également approfondi deux notions qui sont étroitement liées à la coréférence et qui s'avèrent particulièrement pertinentes pour l'annotation et l'analyse des textes d'élèves : le statut évolutif des référents et l'ambiguïté référentielle. Ces deux notions se révèlent en effet centrales dans le contexte de notre travail : les récits d'élèves mettent fréquemment en scène des personnages soumis à des transformations, à des changements d'identité ou à des modifications de statut au fil de l'histoire. Ces évolutions, qu'il s'agisse de métamorphoses, de déguisements ou de révélations d'identité, rendent nécessaire une prise en compte explicite du statut évolutif des entités. De même, l'ambiguïté référentielle constitue un phénomène particulièrement saillant dans les écrits scolaires. Elle se manifeste par l'incertitude quant au rattachement d'une mention à une chaine de référence donnée, notamment dans des textes présentant des écarts par rapport aux normes attendues. L'intégration de cette notion dans notre cadre théorique répond ainsi à une exigence méthodologique : rendre compte des « zones de doute » inhérentes aux productions en cours d'acquisition.

Nous avons ensuite rappelé la définition que nous adoptons de la chaine de référence, basée sur celle de Schnedecker. Cette notion est définie dans un premier temps selon un critère numérique : à partir de trois mentions renvoyant au même référent, celles-ci peuvent être considérées comme constituant une chaine de référence. Nous avons également présenté un récapitulatif des indices de description des chaines repérés dans la littérature, que nous réutiliserons en partie dans notre travail de description des chaines au Chapitre 7. Parmi ces indices, nous avons notamment

approfondi les notions de densité et de stabilité référentielle, afin de préciser l'opérationnalisation adoptée dans la suite de ce travail.

Nous avons enfin présenté de manière synthétique deux cadres théoriques qui constituent des points d'appui essentiels pour notre analyse : la théorie de l'accessibilité d'Ariel et la notion de continuité référentielle développée par Givón, qui orientent notre approche de l'analyse qualitative des chaines. La notion de continuité référentielle occupe à cet égard une place centrale dans notre démarche. Elle offre en effet une conception plus souple que celle de la coréférence stricte, ce qui permet de prendre en compte, dans l'annotation, des phénomènes autrement difficiles à représenter, comme par exemple les cas de référents collectifs ou les configurations impliquant une multiplication des entités dans le texte.

Enfin, nous avons présenté une typologie possible des expressions référentielles permettant de distinguer différents types de mentions selon leur contenu en termes grammaticaux. Cette distinction est importante dans la mesure où chaque type de mention joue un rôle spécifique dans la gestion référentielle d'un texte, selon le statut du référent et les attentes associées au genre discursif considéré. Comme nous le verrons par la suite, cette taxonomie rencontre toutefois certaines limites lorsqu'elle est confrontée à l'analyse effective des annotations.

Le deuxième chapitre a ensuite proposé un état des lieux des corpus d'écrits scolaires existants, en particulier dans les contextes français et italien. Après avoir retracé brièvement l'histoire de la constitution des corpus scolaires en France, nous avons complété la synthèse proposée par Wolfarth en décrivant les projets développés après 2019, notamment le projet *E-Calm*. Ce projet rassemble plusieurs ressources importantes, dont *Scoledit*, qui constitue la partie française du corpus *Scolinter*, ainsi que le corpus *RésolCo*, qui a constitué une source d'inspiration importante pour notre travail.

Le troisième chapitre est enfin consacré à la présentation des corpus annotés en coréférence dans le domaine du traitement automatique des langues. Afin de situer notre travail dans cet état de l'art, nous avons principalement examiné des corpus annotés pour le français, en complément de quelques ressources internationales fondatrices. L'absence de corpus natifs de grande taille annotés en coréférence pour l'italien explique ce choix : en dehors de quelques travaux plus anciens et d'initiatives récentes comme la traduction automatique du corpus *OntoNotes* , les ressources disponibles restent encore limitées pour cette langue.

Parmi les corpus examinés, le corpus *DEMOCRAT* occupe une place particulièrement importante pour notre travail. Les recherches menées dans le cadre de ce projet ont en effet proposé un ensemble d'indicateurs de description des chaines de référence qui constituent un point d'appui méthodologique essentiel pour la caractérisation de notre propre corpus. L'adoption de ces indicateurs permet en outre d'inscrire notre travail dans un cadre comparatif avec les analyses déjà présentes dans la littérature.

Dans une perspective de reproductibilité et d'unification des pratiques d'annotation, il apparait en effet souhaitable que les différents projets adoptent, dans la mesure du possible, des schémas d'annotation comparables. L'analyse des ressources existantes nous a ainsi permis d'élaborer un

dispositif d'annotation qui reste compatible avec les pratiques établies, à la fois du point de vue didactique et du point de vue du traitement automatique des langues. Une telle convergence méthodologique pourrait à terme favoriser la constitution de ressources interopérables et l'intégration de différents corpus annotés au sein d'ensembles plus larges, couvrant une diversité de genres textuels et de contextes de production.

Partie 2 – Description du corpus et méthodologie d'annotation

Introduction *109*

Chapitre 4. *Recueil et constitution d'un corpus longitudinal trilingue : contexte, difficultés et enjeux* *111*

1. Lire et écrire au CP 111
2. Corpus Scoledit 115
3. Corpus Scolinter 116
4. Protocole de recueil 117
5. Recueil en Italie et Espagne 118
6. Métadonnées associées aux productions 120
7. Traitement des données : du manuscrit à sa version numérique 121
8. Description du corpus recueilli 139
9. Corpus longitudinaux français et italien : caractérisation 142

Chapitre 5. *Annotation de corpus* *145*

1. Objectifs et description de la chaine de prétraitement 146
2. Méthodologie d'annotation et de sélection du corpus à annoter 155
3. Guide d'annotation 159
4. Le rôle de l'adjudication dans la création d'un corpus *gold* 163

Chapitre 6. *Campagnes d'annotation* *165*

1. Pré-campagnes : plusieurs allers-retours vers le guide et échanges entre enseignantes-chercheuses 167
2. Rodages : campagnes d'annotation en TP, ou le guide à l'épreuve des groupes d'annotateurs de Master 169
3. Première campagne d'annotation : validation méthodologique de la constitution d'un corpus *gold* 171
4. Deuxième campagne d'annotation : constitution d'un corpus *gold* par adjudication 177

Conclusion *183*

Introduction

Cette deuxième partie présente les étapes de constitution et d'annotation qui ont conduit à l'élaboration de la version annotée en continuité référentielle d'une partie du corpus *Scolinter*, lequel constitue le corpus de référence de ce travail de thèse. Elle vise à expliciter le cadre méthodologique général de la constitution de ce corpus, du recueil aux traitements des données, à son annotation. La constitution d'un tel corpus implique généralement plusieurs phases successives : la définition des objectifs scientifiques, la planification et l'organisation du recueil, la collecte des données, leur transformation en formats exploitables, leur enrichissement par des métadonnées, puis la conception et la mise en œuvre du schéma d'annotation. Ces différentes étapes conditionnent directement la qualité, la fiabilité et la réutilisabilité des ressources produites.

Dans le cas de *Scolinter*, certaines de ces phases, en particulier la planification, le recueil des textes du corpus et le traitement d'une grande partie des textes, ont été réalisées (en amont) par les équipes du LIDILEM de l'Université Grenoble Alpes, de l'Université Milano-Bicocca en Italie et de l'Université d'Almería en Espagne, et sont antérieures à notre travail de thèse. Notre travail s'inscrit ainsi dans la continuité de ces efforts collectifs. Il demeure toutefois nécessaire de décrire ces étapes afin de bien comprendre les caractéristiques du corpus.

Un bref historique de sa constitution, la description de ses caractéristiques et des objectifs, ainsi que la démarche de collecte et le processus de numérisation de ces écrits sont illustrés au Chapitre 4. Il faut noter ici que, dans le cas du corpus italien, nous avons pu contribuer aux étapes de numérisation, transcription et normalisation que nous allons décrire.

En ce qui concerne la réalisation des annotations en continuité référentielle, cela représente le cœur de notre travail de thèse : dans le Chapitre 5, nous décrivons la méthode d'organisation adoptée, fondée sur une démarche itérative inspirée des méthodes agiles, qui a permis d'ajuster progressivement le schéma et le guide d'annotation. Cette méthode nous a ainsi permis de revenir au fil des campagnes d'annotations sur le guide fourni aux annotateurs. Le Chapitre 6 décrit ensuite les différentes campagnes d'annotation que nous avons menées, ainsi que les étapes de pré-traitement des textes, et les procédures de validation des annotations mises en place. Dans le Chapitre 6 nous détaillerons davantage la deuxième campagne d'annotation que nous avons réalisée en parallèle sur 148 textes du corpus français et 150 du corpus italien, également distribués entre les niveaux scolaires CE1 et CE2, dont les résultats seront décrits dans la partie 3 de ce travail.

Chapitre 4. Recueil et constitution d'un corpus longitudinal trilingue : contexte, difficultés et enjeux

Le corpus objet de ce travail de thèse, *Scolinter*, est issu de trois projets différents qui se sont déroulés entre 2011 et 2023 : « Lire écrire au CP » (Goigoux, 2016), puis le corpus *Scoledit* (Ponton et al., 2019; Wolfarth, 2019), inclus dans les ressources du projet *E-Calm* (Doquet et al., 2023; Ponton, Doquet, et al., 2022), et finalement le corpus *Scolinter* (Ponton et al., 2021), qui représentent trois étapes d'évolution différentes d'un corpus constitué à partir des mêmes consignes et d'un noyau de protocole de recueil partagé. Le corpus s'est enrichi au fil des années tout en respectant les objectifs initialement déterminés et le protocole de recueil nécessaire à les réaliser.

Le projet « Lire et écrire au CP », initié en 2011 par Roland Goigoux et une nombreuse équipe d'enseignants-chercheurs, marque le début du recueil de données variées sur les pratiques de lecture et écriture au CP. Le recueil de données s'est déroulé entre 2013 et 2015. Le projet de recueil du corpus *Scoledit* reprend la méthodologie du projet et poursuit entre 2016 et 2018 le recueil sur une partie des classes concernées par le recueil de *Lire et écrire*. Le corpus *Scolinter* comprend la partie longitudinale de *Scoledit* et initie un nouveau recueil de données en Italie et en Espagne entre 2018 et 2022, en suivant le même protocole de recueil appliqué en France.

L'objectif initial était de répondre au manque de données de production écrite longitudinales et comparables dans les trois pays, utiles pour quantifier l'évolution de l'écriture à l'école primaire à travers la comparaison diachronique. Le corpus s'est ensuite enrichi de ses contreparties comparables recueillies en Italie et Espagne, ce qui permet d'analyser de manière contrastive ainsi que diachronique les productions écrites et leur variation entre pays différents.

1. Lire et écrire au CP

Le projet « Lire et écrire au CP », coordonné par Roland Goigoux, marque le début du recueil du corpus présenté dans ce travail de recherche. Ce projet a été initié en septembre 2011 et le recueil de données s'est déroulé entre septembre 2013 et juin 2015. Cette recherche, cofinancée par le ministère de l'Éducation nationale (DGESCO) et l'Institut français de l'Éducation (IFÉ–ENS Lyon) et, pour une moindre part, par le laboratoire ACTÉ (université Blaise Pascal, Clermont-Auvergne) (Goigoux, 2016, p. 22) a également reçu le soutien des universités d'appartenance des nombreux enseignants-chercheurs et doctorants qui ont participé à la recherche, dont un groupe d'enseignants-chercheurs du laboratoire LIDILEM de l'Université Grenoble Alpes.

Son objectif initial était d'observer les pratiques d'enseignement de la langue française au CP, et d'évaluer l'effet des pratiques d'enseignement de la lecture et de l'écriture sur les apprentissages des élèves de cours préparatoire (Goigoux, 2016, p. 21) dans 131 classes de CP de 14 académies françaises. Les élèves ont été évalués à travers 16 épreuves qui comportaient « la maîtrise du code grapho-phonétique (connaissance des lettres, conscience phonologique, lecture de mots et de pseudo-mots, fluence) et la compréhension (vocabulaire, compréhension de phrases et de textes

entendus, compréhension de phrases en lecture autonome) » (Brissaud et al., 2016, p. 91). La Figure 11 détaille les différentes épreuves soumises aux élèves.

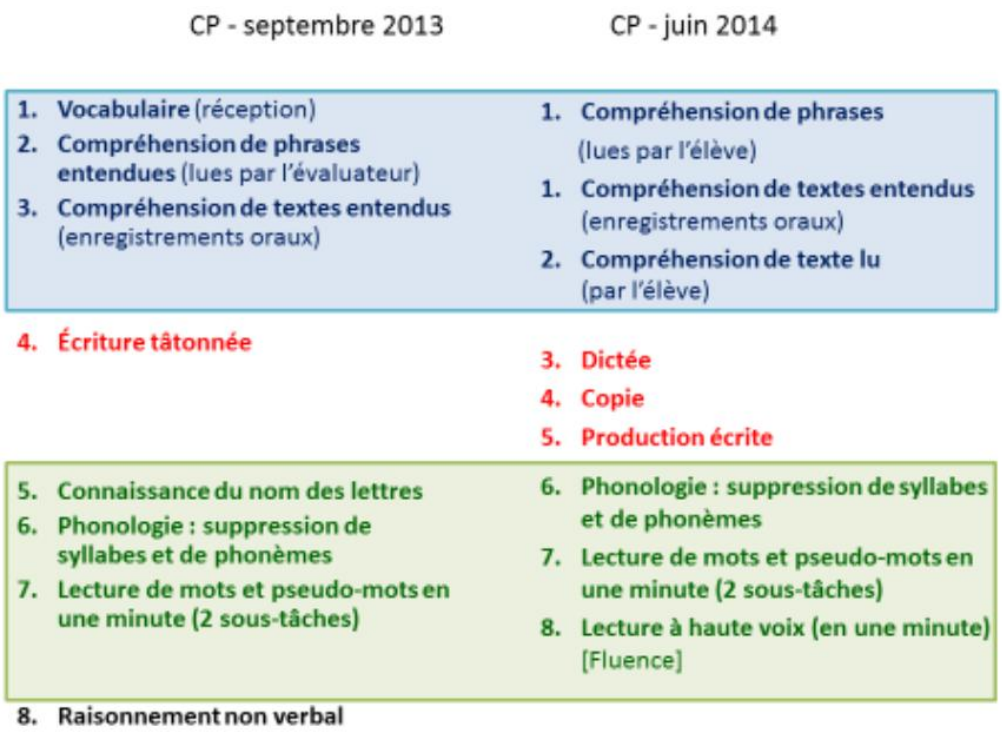


Figure 11. Tableau de synthèse des tâches de lecture-écriture tiré de Goigoux, 2016, p. 71

Le recueil de données couplé à l'observation directe effectuée par les chercheurs dans les classes avait comme objectif d'éviter les biais dérivés de la prise en compte des pratiques déclarées par les enseignants, constatés dans des recherches précédentes (Goigoux, 2016). Pour cette raison, toutes les séances d'enseignement de la lecture et de l'écriture de trois semaines réparties sur l'année (en novembre, mars et mai) ont été enregistrées en vidéo.

Plusieurs questions se posaient sur l'apprentissage de la lecture et de l'écriture au CP, dont des problématiques liées à l'acquisition du code alphabétique et des procédures d'identification des mots, au développement de la compréhension de l'écrit, à l'écriture, l'étude de la langue et l'acculturation à l'écrit. Ces cinq sous-ensembles de questions correspondaient aux domaines de compétences déjà identifiés par les programmes scolaires de la période et répondaient aux préoccupations des acteurs impliqués dans la recherche, enseignants-chercheurs et enseignants (Goigoux, 2016). Dans cette recherche, un paradigme écologique d'observation et recueil de données avait été adopté. L'approche était celle de moindre incursion du monde de la recherche dans l'environnement de classe, pour « étudier des classes ordinaires telles qu'elles fonctionnent quotidiennement, sans manipulation du contexte par le chercheur, c'est-à-dire sans modification de l'action habituelle des enseignants » (Goigoux, 2016, p. 49), d'où le choix d'observer les activités d'enseignement à travers l'observation directe dans les classes.

Le groupe de recherche dédié à l'observation des pratiques d'écriture avait comme objectif de « préciser dans quelle mesure et de quelle manière les pratiques d'écriture influencent les apprentissages de la lecture (code et compréhension) et ceux de l'écriture elle-même » (Goigoux, 2016, p. 36), c'est-à-dire recenser les tâches d'écriture réalisées à l'école et déterminer lesquelles influencent plus fortement l'apprentissage de l'écriture.

L'évaluation de l'écriture était réalisée en début et en fin de CP. Au début du CP, elle reposait sur trois tâches : l'écriture du prénom, une dictée de mots et d'une courte phrase ; en fin de CP, les épreuves reprenaient à l'identique celle réalisée en début du CP augmentée d'une phrase à la dictée de phrases, une épreuve de copie et une production de texte narratif. Au début du CP, la passation des épreuves était individuelle ; en fin de CP, la dictée et la production écrite narrative ont été menées en passations de petits groupes de 6-12 élèves, en dehors de la classe, dans des locaux sans textes écrits affichés pour ne pas influencer les enfants. L'épreuve de copie était quant à elle menée en passation individuelle. En fin de CE1, les mêmes groupes d'élèves ont participé à des tâches ultérieures, dont la production d'un texte narratif dans le respect d'une consigne donnée.

De cette manière, un texte issu de la tâche de production écrite a été recueilli pour chaque élève en fin de CP en juin 2014 puis en fin de CE1 en juin 2015. Ces recueils constituent les parties du corpus de CP et CE1 successivement incluses dans le corpus *Scoledit* puis *Scolinter*.

1.1. Epreuve de fin de CP – juin 2014

Cette épreuve de production écrite d'une courte histoire narrative présente une consigne visuelle, composée de quatre images, visible dans la Figure 12. Ces images représentent le déroulement dans le temps d'une courte histoire avec des chats comme protagonistes. Dans la première vignette nous voyons un chat adulte qui dors avec des petits chats blottis contre lui, pendant qu'un petit chaton s'éloigne du groupe. Dans la deuxième vignette, le petit chat tombe d'une marche ou d'un rebord, puis dans la troisième vignette il miaule, ce qui réveille le chat adulte et attire son attention. Dans la dernière vignette, le chat adulte récupère le petit chaton et le ramène vers le groupe de petits chats. Ces vignettes sont donc censées représenter une situation initiale, son évolution et sa résolution finale.

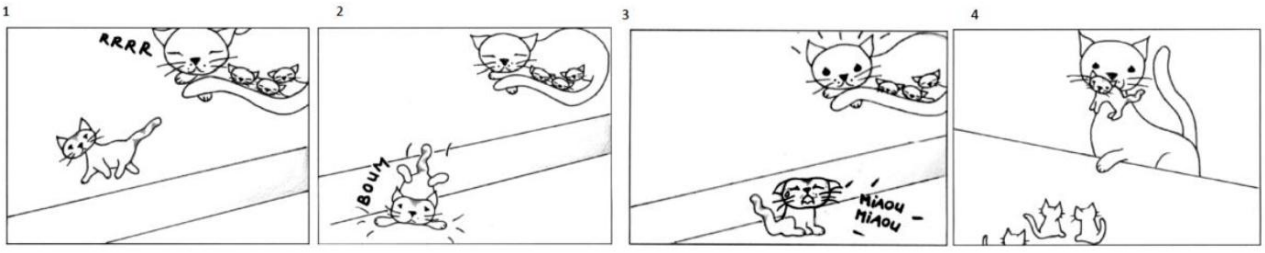


Figure 12. Consigne visuelle utilisée au CP pour le recueil des données du le projet « Lire et écrire au CP ».

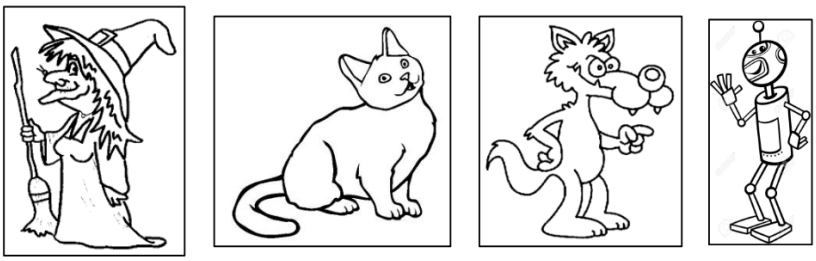
Le rapport de Goigoux (2016) décrit de manière détaillée le protocole de passation à mettre en place lors de cette épreuve de production écrite : « Cette épreuve consiste en la production d'une histoire sur la base de quatre images montrées par l'évaluateur. L'épreuve dure 15 minutes (une

fois les consignes données). Les affichages muraux susceptibles d'aider les élèves sont retirés (le mot *chat*, des listes de verbes, etc.). » Quand les élèves sont tous attentifs, l'évaluateur annonce aux élèves qu'ils vont écrire chacun l'histoire d'un petit chat. Chaque vignette, en grand format, est affichée successivement au tableau. « Puis l'évaluateur laisse aux élèves le temps de « mettre l'histoire dans leur tête ». Quand ils sont prêts, il leur distribue une feuille avec lignes Seyes et leur précise que s'ils ont oublié l'histoire, ils peuvent retourner la feuille pour retrouver les dessins. » (Goigoux, 2016, p. 67-68)

Dans la recherche « Lire et écrire au CP » ces textes étaient recueillis puis évalués selon six critères d'évaluation ou analyseurs : longueur du texte en termes de « lettres » ou plutôt de caractères, segmentation des mots, lisibilité (si au moins une fraction du texte est lisible ou déchiffrable à haute voix sans effort), présence de séparateurs d'idées (majuscules, ponctuation forte, retours à la ligne volontaires, connecteurs), quantité d'informations (au moins une information par image dans le cas de la consigne du CP), traces de narration comme la présence d'un titre, d'une formule d'ouverture comme « il était une fois », présence de verbes au passé, d'une formule de fermeture et d'au moins une reprise pronominale du personnage principal protagoniste de la consigne. Pour cette épreuve, les chercheurs ont recueilli 2507 textes au total, un pour chaque élève autorisé par ses parents et présent au début et à la fin de l'année des 131 classes participant au recueil.

1.2. Épreuve de fin de CE1 – juin 2015

Comme mentionné auparavant, des épreuves en guise de post-tests étaient proposées en fin de CE1, « pour savoir si les progrès sont variables, amples et durables » (Goigoux, 2016, p. 49). L'épreuve de production écrite libre proposée en CE1 consiste en l'écriture d'une histoire sur la base de quatre images représentant chacune un personnage : une sorcière, un chat, un robot et un loup. Les images sont accompagnées d'une consigne écrite, visible dans la Figure 13, qui spécifie de choisir un ou deux des personnages figurant dans les images pour raconter une histoire les concernant. La durée de l'épreuve est de vingt minutes une fois les consignes données à voix haute par l'enseignant à toute la classe. Comme dans le cas de l'épreuve de CP, tout affichage mural susceptible d'aider les élèves a été retiré et les enseignants ne peuvent apporter aucune aide aux élèves lors du déroulement de l'épreuve. Pour cette épreuve, les chercheurs ont recueilli 2049 textes.



1 2 3 4

Voici 4 personnages. Choisis un ou deux personnages et raconte une histoire.
Entoure le ou les personnages que tu as choisis.

Figure 13. Consigne visuelle utilisée du CE1 au CM2.

2. Corpus Scoledit

Le corpus *Scoledit* inclut les textes recueillis en CP et CE1 au sein du projet « Lire écrire au CP » et en reprend la même méthodologie (et la même consigne composée de quatre personnages) pour poursuivre le recueil de textes de manière longitudinale en CE2, CM1 et CM2 auprès d'une partie des élèves. Si le projet précédent avait rassemblé des chercheurs d'un grand nombre d'universités réparties dans quatre académies, dans le cas du corpus *Scoledit*, le recueil a été réalisé principalement par des membres du laboratoire LIDILEM, aidés par deux collègues de l'université de Toulouse. Parmi les écoles ayant participé au projet *Lire écrire au CP*, 42 classes issues de 4 académies (Grenoble, Clermont-Ferrand, Lyon et Toulouse) ont permis de poursuivre le recueil, auxquelles 3 nouvelles écoles ont été ensuite rajoutées pour un total de 676 élèves participants, dont 53,5% de genre féminin et 46,5% de genre masculin. En plus des textes de production écrite libre dont nous avons déjà détaillé les consignes (cf. §1.1, §1.2), les chercheurs ont aussi collecté des dictées produites par les élèves, qui n'ont été que partiellement traité jusqu'à présent.

Académie	***Classes***	***Elèves***
Clermont-Ferrand	16	282
Grenoble	10	167
Lyon	11	174
Toulouse	5	53
Tot.	*42*	*676*

Tableau 8. Distribution des élèves participants au recueil en France.

Le corpus *Scoledit* représente actuellement le seul corpus longitudinal d'écrits scolaires recueillis du CP au CM2 entièrement transcrit et librement accessible en ligne (Ponton et al., 2021). Il est actuellement disponible sur Ortolang, de même que les différents corpus du projet E-CALM[57], ainsi que sur une plateforme dédiée à son exploration, conçue par Claude Ponton[58].

Il a été conçu par les chercheurs du laboratoire LIDILEM de l'Université Grenoble Alpes. Grâce aux outils d'exploitation de ce corpus développés par cette équipe, et en particulier par Claire Wolfarth dans le cadre de son doctorat (2019), il a été possible de mener des analyses sur les niveaux orthographique, lexical et morphosyntaxique et dans une perspective longitudinale (Ponton et al., 2019; Wolfarth, Brissaud, et al., 2018b; Wolfarth et al., 2017; Wolfarth, Ponton, et al., 2018). Le produit de ces deux recueils de données est donc un grand corpus numérique longitudinal de textes narratifs scolaires produits par les mêmes élèves entre 6 et 11 ans. L'objectif initial était celui de « réaliser une description linguistique des structures utilisées par les élèves au cours de la construction de leurs apprentissages de l'écrit (morphographie, syntaxe, lexique, orthographe, acquisition des conjugaisons, structuration du discours), ainsi que de l'évolution des procédés d'écriture à différents moments de la scolarisation à l'école primaire. » (Wolfarth, 2019, p. 107), à travers une ressource standardisée suffisamment large pour servir à réaliser des études

[57] La ressource E-CALM est disponible au lien suivant : https://www.ortolang.fr/market/corpora/e-calm

[58] La plateforme Scoledition permet de consulter les textes du corpus et de comparer leur scan avec la transcription de la production écrite. Elle est disponible au lien suivant : https://scoledit.org/scoledition/

statistiquement significatives. Notre corpus *Scolinter* est une extension à l'Italie et à l'Espagne des méthodes de recueil et de traitement déjà utilisées pour la constitution du corpus *Scoledit*. Nous allons introduire maintenant plus en détail quels sont les enjeux et les étapes nécessaire pour mettre en place le recueil d'un tel corpus.

Le recueil d'un corpus d'écrits scolaires demeure une opération complexe et de longue haleine. D'un côté, sur le plan logistique, un tel recueil nécessite la création d'un véritable réseau de contacts avec les différentes institutions, écoles et enseignants ; de l'autre côté, d'autres problématiques se présentent dans la gestion des données recueillies, que ce soit la numérisation des productions écrites ou le stockage des données personnelles des élèves, leur anonymisation, etc. En effet, la disponibilité des directeurs d'école, des enseignants, des parents d'élèves, ainsi que la possibilité pour les chercheurs de s'engager dans ces activités de recueil de longue durée représentent seulement une partie des différents facteurs qui influencent le long processus de récolte des données longitudinales dans les écoles. Ce travail crucial est cependant peu abordé à cause des différentes difficultés que cela comporte et du temps qu'il nécessite. Ces difficultés ont déjà été approfondies dans le travail de Wolfarth (2019), qui a aussi décrit dans le détail le processus de recueil du corpus *Scoledit*, partie française du corpus *Scolinter*. Dans ce chapitre nous décrivons le protocole de recueil adopté pour la constitution de ce corpus comparable en français, italien et espagnol. L'autrice de ce manuscrit n'a pas directement participé à la phase de récolte des données des trois parties du corpus. La description ici fournie est le résultat d'une enquête menée en 2024 auprès des équipes qui ont effectué le recueil en Italie et en Espagne. Dans le cas du corpus français, le protocole de recueil a déjà été décrit plus en détail par Wolfarth (2019), raison pour laquelle nous allons davantage parler du travail effectué en Italie et en Espagne, car encore peu décrit.

3. Corpus Scolinter

L'objet du présent travail de thèse peut se résumer à l'étude de la continuité référentielle et de ses caractéristiques dans des textes de scripteurs en phase d'apprentissage du français langue première, dont l'analyse s'appuie sur l'annotation d'un corpus d'écrits scolaires. Cependant, ce travail n'aurait pas pu être mené sans l'effort collectif des chercheurs, doctorants et étudiants de l'Université Grenoble Alpes, de l'Université Milano-Bicocca en Italie et de l'Université d'Almería en Espagne, qui ont participé à l'organisation, au recueil, à la numérisation et à l'annotation des textes constituant le corpus *Scolinter* (LIDILEM, 2025).

Le projet *Scolinter* regroupe l'équipe de chercheurs du laboratoire *LIDILEM* de l'Université Grenoble-Alpes déjà impliquée dans la recherche *Scoledit* (Wolfarth, Brissaud, et al., 2018a) ainsi qu'une équipe italienne du département de Sciences Humaines pour l'Education « Riccardo Massa » de l'Université de Milano-Bicocca, coordonnée par Lilia Teruggi et Elisa Farina, et une

équipe espagnole du département d'éducation de l'Université de Almería dirigée par Rafaela Gutierrez Cacéres[59].

Tout comme en France avant le projet E-Calm, en Italie et Espagne on pouvait constater le manque de corpus scolaires longitudinaux, numérisés et librement accessibles. Si des projets ont tenté de combler ce manque en Italie (Brunato & Dell'Orletta, 2015; Revelli, 2011), aucun de ces corpus n'a été constitué comme véritablement longitudinal. De ce constat, des recueils ont été initiés en Italie et en Espagne pour constituer un corpus de grande taille comparable aux données déjà disponibles en français. Les textes sont produits par les élèves à partir de la même consigne dans les trois langues (français, dont les textes avaient déjà été recueillis dans le corpus *Scoledit*, italien et espagnol), et les protocoles de recueil et de traitement des textes demeurent uniformes, ce qui rend les données comparables. Des analyses contrastives ont pu être ainsi menées sur les deux premières années du primaire autour de la segmentation des mots et du développement de la morphologie verbale (Farina & Teruggi, 2021; Ponton et al., 2021). En outre, des études préliminaires ont porté sur la continuité référentielle et son annotation (Barletta, 2024; Barletta & Ponton, 2025a), objet de ce travail de thèse. Vu que le recueil du corpus français et les difficultés rencontrées ont été détaillées dans le travail de thèse de Wolfarth, ici nous allons nous limiter à décrire les expériences des équipes italienne et espagnole, qui demeurent cependant très similaires à celles conduites en France.

4. Protocole de recueil

La méthodologie de recueil adoptée était la même pour toutes les classes à partir du CE1 jusqu'au CM2 dans les trois pays. Le protocole adopté en CP étant déjà abordé de manière approfondie dans la section concernant le projet *Lire écrire au CP* (cf. ce chapitre, §1), nous décrivons ici davantage le protocole adopté par l'équipe du laboratoire LIDILEM pour le recueil des textes du CE1 au CM2, produits à partir de la même consigne visuelle utilisée pour le recueil de textes narratifs de CE1 du projet *Lire écrire au CP* (visible dans la Figure 13). Après avoir distribué aux élèves des feuilles contenant la consigne visuelle, et après avoir obtenu l'attention des élèves, ceux-ci entendaient la consigne suivante : « Aujourd'hui, vous allez écrire chacun une histoire avec un ou deux personnages. » Après avoir demandé à la classe de nommer chaque personnage dans la consigne visuelle, le chercheur ou la chercheuse disait : « Je vais vous lire la consigne qui est écrite sous les images, écoutez bien la consigne. Voici quatre personnages. Choisis un ou deux personnages et raconte une histoire. Entoure le ou les personnages que tu as choisis. ». Le chercheur ou la chercheuse précisait enfin : « Avant de commencer à écrire, mettez bien dans votre tête l'histoire de ce personnage ou de ces deux personnages. Vous avez 20 minutes pour écrire cette histoire. Vous allez travailler seul ; personne ne vous aidera, par exemple à écrire un mot. ». (Consigne tirée de Wolfarth, 2019, p. 114)

[59] Dans ce travail de thèse, nous utilisons le féminin pour décrire les équipes de recherche responsables du recueil en Italie et Espagne car exclusivement composés d'enseignantes-chercheuses. Ces deux équipes n'ont pas reçu de financements dédiés à la collecte ou au travail de traitement du corpus.

Même si le protocole a été globalement appliqué de manière uniforme dans les trois pays, on peut retrouver quelques différences notamment sur les caractéristiques du support fourni pour l'écriture: les lignes Seyes utilisées en France (Figure 14) ont été remplacées en Italie par un support à ligne plus larges (Figure 15). Dans le cas des classes du CE1 au CM2 en France, le support préconisé était plus libre (grands carreaux) mais cela pouvait varier dans le corpus au cas par cas. En Espagne, de la même manière, le support utilisé n'est pas uniforme, il peut varier entre l'utilisation des lignes comme celles du support italien et celles représentées dans la Figure 16.

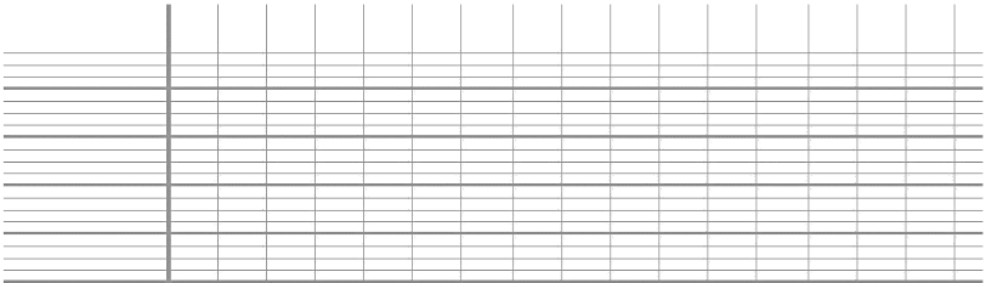


Figure 14. Exemple de support d'écriture utilisé en France au CP - lignes Seyes.

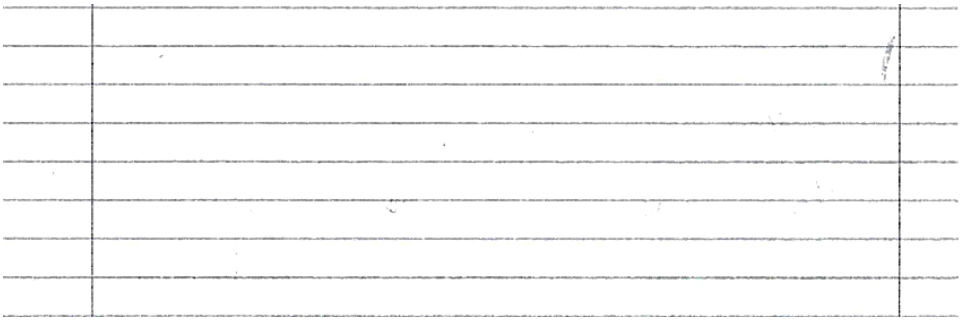


Figure 15. Exemple de support d'écriture utilisé en Italie - lignes larges.

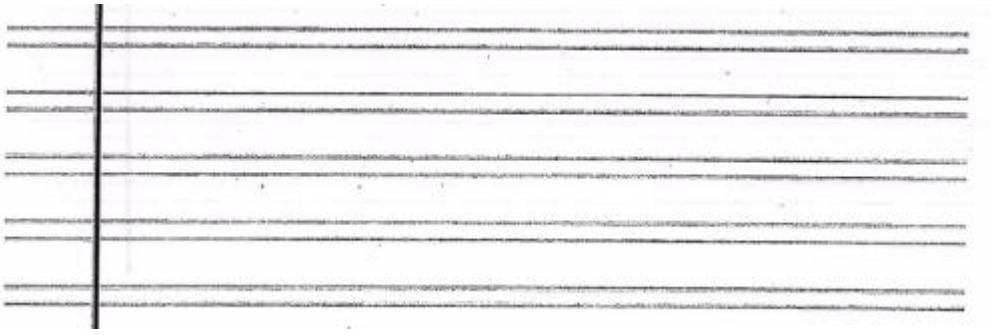


Figure 16. Exemple de support d'écriture utilisé en Espagne.

5. Recueil en Italie et Espagne

Les recueils en Italie et en Espagne ont débuté en 2018, soit au moment où s'achevait le recueil mené en France (réalisé entre 2016 et 2018). Le fait que le recueil de ces deux corpus ait commencé alors que l'expérience française touchait à sa fin, avec des protocoles de recueil et de traitement des textes déjà bien définis, a grandement facilité la constitution d'un corpus comparable pour les trois langues. Toutefois, la période de recueil a malheureusement été perturbée par des

circonstances exceptionnelles. En effet, les années scolaires 2019-2020 puis 2020-2021 dans la quasi-totalité du monde, ont été fortement impactées par la pandémie mondiale de Covid-19. Cette période a été qualifiée de crise scolaire, en plus d'être une crise sanitaire globale, car cette dernière a provoqué la fermeture totale des institutions d'enseignement de tous niveaux dans plus de 200 pays et territoires entre mars 2020 et février 2021. Ces fermetures ont impacté environ 168 millions d'enfants dans le monde pendant une année scolaire entière (UNICEF, 2021). Dans la plupart des cas, l'arrêt des activités d'enseignement en présentiel a entrainé l'adoption de pratiques d'enseignement en distanciel, chose tout à fait nouvelle pour certains niveaux d'enseignement comme dans le cas de l'école primaire.

Si les effets de cet événement sur l'éducation font encore l'objet d'analyses (Betthäuser et al., 2023), l'une des conséquences les plus immédiates pour notre recherche a été l'impossibilité de collecter des données pendant cette période, y compris après la réouverture partielle des établissements à la fin des années 2020 et 2021. La collecte était déjà achevée en France depuis 2018, mais elle était encore en cours en Italie et en Espagne. Dans ces deux pays, les fermetures scolaires ont fortement perturbé le recueil : en Italie, aucune donnée n'a pu être collectée durant ces années scolaires, ce qui a entrainé la perte des données longitudinales pour les classes 3e et 4e (CE2 et CM1) ; en Espagne, bien que certaines écoles primaires aient ouvert à nouveau en septembre 2020, plusieurs établissements initialement engagés dans le projet n'ont pas renouvelé leur participation en 2020-2021, entraînant une rupture partielle du suivi longitudinal. Ces manques ont été compensés par l'intégration de nouvelles écoles volontaires.

Le recueil de la partie italienne du corpus s'est déroulé ainsi entre 2018 et 2022. Pendant ces années, l'équipe a pu collecter des productions écrites d'élèves suivis du CP, au CE1 puis en CM2 ainsi que d'un autre groupe comprenant les mêmes élèves en CE1 et CE2. Pour cette raison, le corpus se divise en deux parties chacune longitudinale : une partie contenant les textes des mêmes élèves au CP, CE1 et CM2, recueillis en 2018, 2019 et 2022, la partie que l'on caractérise de « corpus longitudinal » ; le « mini-corpus », recueilli dans les mêmes classes en CE1 et CE2 en 2018 puis en 2019.

L'organisation du recueil italien a pu s'appuyer sur les réseaux de connaissances des membres du groupe de recherche, dirigé par Lilia Teruggi et Elisa Farina, et composé au total de huit enseignants intervenant pour la plupart dans la formation des enseignants à l'Université Milano-Bicocca. Le groupe a ainsi contacté de manière directe des enseignants d'école déjà connus et parfois aussi formés à l'université par des membres du groupe de recherche, enseignantes-chercheuses du Département de Sciences Humaines pour la Formation de l'Université Milan-Bicocca. Ces enseignants ont permis d'avoir un point d'entrée dans les établissements, avant de procéder à une prise de contact avec les directeurs des écoles (*dirigenti scolastici*). Ces derniers ont été contactés par la responsable du projet côté italien à travers une lettre modèle (disponible en Annexe B). Le directeur se chargeait ensuite de communiquer avec les enseignants ainsi qu'avec les parents des élèves par rapport à l'autorisation des élèves à participer à la recherche. Du côté italien, c'était la direction de l'école qui s'occupait de fournir aux parents d'élèves les autorisations à signer sous forme de circulaire interne à l'école. Les autorisations signées ont été recueillies par les enseignants

et stockées par la direction de chaque école, qui se chargeait ensuite de transmettre aux chercheuses la liste des élèves bénéficiant d'une autorisation signée par un ou deux parents (le nombre de signatures demandées variant selon l'école et le directeur). Une autorisation a été signée par les tuteurs légaux de l'élève chaque année et donc pour chaque texte collecté. Le recueil des données s'est déroulé en totalité en Lombardie, sur trois provinces différentes : Milan, Bergame et Sondrio. Au total, le recueil a été réalisé dans 9 établissements différents, répartis entre ces provinces.

Les chercheuses du groupe se sont partagé le travail de recueil à hauteur d'environ deux écoles chacune, et elles ont cherché à distribuer les écoles selon le critère de proximité géographique entre l'école et le domicile de la chercheuse concernée afin de réduire au maximum le temps nécessaire pour rejoindre chaque école. Au total, huit des enseignantes-chercheuses rattachées au département de Sciences de la formation de l'Université Milan-Bicocca ont participé au processus de recueil des données.

En comptant tous les déplacements effectués par les chercheuses, y compris la nécessité de revenir dans la même école à plusieurs reprises pour effectuer le recueil dans des classes différentes, nous considérons que le temps de travail nécessaire était d'environ trois heures par classe dont la demi-heure de passation. Ainsi nous arrivons à un total d'environ 230 heures de travail dédiées au recueil. Nous n'avons pas inclus ici le temps passé pour la prise de contact entre le groupe de recherche et les écoles, ainsi que les temps effectifs de déplacement nécessaire pour rejoindre certaines écoles.

Le groupe responsable du recueil s'est réuni à plusieurs reprises y compris avec des chercheurs du laboratoire grenoblois LIDILEM ainsi qu'avec la chercheuse de l'équipe espagnole pour s'entrainer au protocole de recueil, pour que les conditions soient les plus similaires possibles dans toutes les écoles des trois pays. Deux protocoles de recueil similaires ont été établis, un protocole à suivre dans les classes de première année (équivalentes au CP) et un protocole à suivre dans les autres classes (de la 2e à la 5e année – du CE1 au CM2), la principale différence portant surtout sur le temps de passation prévu pour l'épreuve.

Le recueil en Espagne a été réalisé par une seule enseignante-chercheuse, Rafaela Gutierrez Cacéres, qui s'est occupée des étapes de recueil, numérisation et de constitution du corpus. Un membre de l'équipe italienne, également hispanophone, Lilia Teruggi, a pu accompagner la chercheuse pour l'étape de normalisation d'une partie des textes. Le recueil de la partie espagnole du corpus a été réalisé entre mai-juin 2018 et mai-juin 2023, et a eu lieu en Andalousie, en particulier dans la ville d'Almería ainsi que dans les provinces d'Almería et de Granada.

6. Métadonnées associées aux productions

Les métadonnées sont les informations qui accompagnent les données linguistiques qui forment un corpus écrit. Leur rôle est, entre autres, de documenter la réalisation du corpus, permettre de sélectionner des sous-parties du corpus selon des variables définies, et de faciliter la réutilisation du corpus par des équipes de recherche et chercheurs autres que leurs auteurs. Le jeu de métadonnées est habituellement commun à l'entièreté du corpus : cela permet de sélectionner de

manière cohérente les données lors des analyses. Cependant, entre les trois langues du corpus, des différences existent dans les métadonnées recueillies.

Les métadonnées que nous avons à disposition pour le corpus entier sont :

- l'identifiant de l'élève, qui permet de suivre le développement de l'écriture des élèves présents tout au long ou dans une partie du corpus ;
- le niveau scolaire de production du texte ;
- l'année de recueil ;
- les personnes responsables du recueil des textes ;
- la consigne proposée aux élèves pour la production écrite.

Dans la partie française et espagnole du corpus, nous avons en plus le genre de l'élève, donnée qui est absente pour l'italien. Si dans le projet français nous avons des informations par rapport au milieu socio-culturel des écoles du recueil et des parents d'élèves, données héritées de la recherche *Lire Ecrire au CP*, ces informations sont absentes pour l'italien et indiquées seulement dans des cas de milieu social très défavorisé de l'enfant pour l'espagnol.

Dans le corpus français, nous disposons de l'indication du nombre de langues parlées dans le contexte familial pour tous les élèves. Nous avons aussi des informations sur les langues autres que la langue du pays d'éducation, parlées dans la famille, dans les parties italienne et espagnole du corpus, ainsi que des informations sur d'autres éléments susceptibles d'influencer le parcours d'apprentissage, tels que la présence de troubles de l'apprentissage ou d'autres conditions particulières. Ces données, fragmentaires et à normaliser, ont été fournies aux chercheuses pour la plupart par les enseignants des classes qui ont participé au recueil. À cause de leur nature fragmentaire et du fait que ces données n'ont pas été normalisées de manière constante, nous n'allons pas les exploiter dans cette recherche. À terme, nous prévoyons d'effectuer une uniformisation et normalisation de ces métadonnées qui puisse nous permettre de les exploiter pour obtenir une caractérisation plus approfondie de la population qui a produit notre corpus.

7. Traitement des données : du manuscrit à sa version numérique

La constitution d'un corpus d'écrits scolaires représente une tâche complexe à plusieurs égards : d'une part, le recueil nécessite souvent l'implication de plusieurs chercheurs et enseignants, en plus des élèves et des enseignants ; d'autre part, une fois le recueil effectué, il est souvent nécessaire de numériser ces écrits pour les rendre exploitables à l'aide d'outils de Traitement Automatique des Langues, notamment lorsque les textes sont manuscrits. Cette étape de numérisation peut consister en la transcription et ultérieurement en la normalisation de ces textes.

La transcription d'un texte manuscrit représente une étape complexe, et la transcription enrichie « constitue un accompagnement nécessaire au travail du chercheur » (Lebarbé, 2009, p. 222) : du choix des éléments à représenter dans la version « finale » du texte, aux conventions choisies pour codifier ces éléments, le travail de transcription présente plusieurs aspects complexes à aborder par les chercheurs, cela dans le but d'obtenir une version du texte qui soit le plus possible représentative du texte de départ et de ses caractéristiques génétiques ou graphiques, et en même

temps lisible et exploitable à des fins de recherche. Si les outils de l'OCR (*Optical Character Recognition*) pourraient être utilisés pour au moins faciliter cette étape, leur application aux textes non normés présente encore des limites importantes, comme nous allons le montrer par la suite. La normalisation, quant à elle, est entendue dans le contexte de notre travail comme une réécriture proche de la norme de la langue représentée dans les textes, selon la définition de Wolfarth (2019)[60]. Ces traitements, bien qu'essentiels, demandaient un investissement conséquent en termes de temps, de ressources humaines et de moyens techniques, ce qui constitue un frein majeur à leur mise en œuvre systématique et sur de grandes quantités de données. Si au début du projet la normalisation des textes était effectuée manuellement par des experts, les technologies numériques (et notamment les grands modèles de langage) actuellement disponibles pourraient être exploités à terme pour réaliser cette dernière étape et pallier la limite de ressources disponibles pour normaliser des grandes quantités de textes.

Nous pouvons donc identifier trois étapes de traitement différentes nécessaires à la numérisation et mise à disposition du corpus *Scolinter* : 1) la numérisation des textes écrits sous forme d'image à travers le scan du texte ; 2) la transcription de ces textes numérisés et 3) leur normalisation. Si l'étape de numérisation des textes a été principalement réalisée par les chercheurs responsables du projet, la transcription et la normalisation ont été effectuées par des chercheurs et des étudiants, souvent rémunérées pour effectuer ces tâches. Les personnes chargées des étapes de transcription et normalisation travaillaient par équipe de deux, « afin de faciliter les échanges pour une meilleure qualité des données » (Wolfarth, 2019). Les textes ainsi traités ont été ensuite vérifiés et validés, après modifications si nécessaires, par les enseignants-chercheurs responsables de projet. Dans notre cas, nous avons pu contribuer en ce sens avec le scan, la transcription et la normalisation de 100 textes de niveau CE2 du corpus italien, qui n'étaient autrement pas disponibles.

Dans le cas de notre corpus, un processus d'anonymisation a été appliqué à toutes les copies, selon règlement RGPD et selon les autorisations de diffusion de données signées par les tuteurs légaux des élèves, qui précisaient que les textes allaient être anonymisés. Ce processus est détaillé dans Wolfarth (2019) et il a été ensuite aussi appliqué tel qu'il est décrit pour l'anonymisation des parties italienne et espagnole du corpus.

7.1. Les scans : l'image du texte

La numérisation sous forme d'image représente la première étape nécessaire à la pérennisation des écrits du corpus. Déjà à cette étape, il est nécessaire d'effectuer une première anonymisation des textes. Cette opération est importante, « de sorte que l'élève qui a produit le texte diffusé ne

[60] La normalisation, dont les conventions utilisées sont détaillées dans §7.3, représente une deuxième version du texte, conçue comme un rapprochement du texte d'origine à la norme de la langue cible, tout en respectant au mieux les choix effectués dans l'écrit d'origine. Comme nous l'expliquerons plus en détail par la suite, cette version permet d'appliquer des techniques d'exploration de la linguistique outillée aux textes ainsi qu'effectuer des analyses à travers des outils TAL (tokenisation, parsing morphosyntaxique etc.).

puisse pas être identifié » (Wolfarth, 2019, p. 104). Pour ce faire, chaque élève est identifié à travers un code, souvent alphanumérique, qui représente alors le seul repère utilisé pour manipuler les productions. Par exemple, dans le cas du corpus italien, chaque texte comportait un en-tête demandant à l'élève d'écrire son nom, prénom et sa classe d'appartenance. Cette bande supérieure permet d'associer le bon identifiant avec la copie et il a été éliminé des scans avant de passer à l'étape de transcription du texte. De même, les textes composés de plusieurs images scannées nécessitent une étape de recomposition des différentes images dans une seule représentation du texte qui sera ensuite utilisée pour les traitements successifs et diffusée.

Cette « édition fac-similé » (Testenoire, 2017), obtenue à l'aide de procédés photographiques et/ou numériques, a comme objectif la reproduction du texte source de manière la plus fidèle possible. Cette version rend possible la consultation et la copie du contenu du texte à l'aide de techniques variées (OCR ou transcription manuelle) ; elle reste consultable mais n'est pas directement exploitable à l'aide d'outils numériques de traitement de corpus. Elle facilite la transcription des copies, car elle est directement accessible sur écran et permet « un retour à la copie d'origine et offre ainsi une iconicité maximale » (Boré & Elalouf, 2017, paragr. 14): par exemple, sa présence permet à tout moment de la chaine de traitement des textes de revenir à l'original et de vérifier si des erreurs se sont produites lors des étapes de transcription ou lors des étapes successives.

7.2. De « l'image du texte » à sa version numérique : transcription des textes

C'est seulement à travers l'étape de transcription des textes, qui nous en fournit une représentation dactylographique, que ceux-ci deviennent manipulables à l'aide d'outils informatiques (Wolfarth, 2019). Dans cette étape de représentation, néanmoins, des choix doivent être effectués. Tout d'abord, il faut choisir le modèle de transcription à adopter. Une fois le modèle le plus adapté choisi, il faut sélectionner les caractéristiques des textes que l'on souhaite représenter pour que la transcription soit efficace, lisible et exploitable. Enfin, il faut choisir un outil de transcription qui, idéalement, permet de consulter le scan tout au long du processus et qui facilite l'insertion des éléments iconographiques retenus.

Dans le cas de la constitution d'un corpus scolaire, le but de la transcription réside dans la représentation de la production écrite la plus possible fidèle au texte d'origine dans ses caractéristiques considérées comme saillantes. Selon les projets de recherche et le corpus à développer, cela peut comprendre les segmentations erronées des mots et des phrases, les déviations de la norme orthographique, certaines caractéristiques génétiques et visuelles du texte : « les éléments de mise en forme, les marques de paragraphe, les traces de révisions, les interventions de l'enseignant, parfois même certaines informations orthographiques et certaines autres informations iconiques comme les changements de stylo ou de tracé.» (Wolfarth, 2019, p. 70). Les transcriptions doivent rendre ces phénomènes sélectionnés visibles selon les choix de représentation effectués tout en autorisant l'exploitation informatique des textes, en choisissant les phénomènes à observer selon « La nécessité de catégoriser les observables, condition de leur exploitation » qui « est un des rôles de la transcription, qui s'accompagne toujours d'une perte » (Doquet et al., 2021, p. 15). Au moins dans le cas de corpus scolaires, habituellement les

transcriptions doivent surtout reproduire de manière fidèle les choix de l'apprenant en termes de ponctuation, d'orthographe, syntaxe et lexique.

Les choix de transcription sont fortement influencés par les objectifs de recherche auxquels un corpus se rapporte, ou par ceux qui ont été définis lors de sa constitution. Un corpus peut également être utilisé pour d'autres finalités que celles initialement prévues, que ce soit en linguistique ou en informatique. Ces choix de transcription peuvent évoluer tout au long du processus de numérisation d'un corpus donné, comme c'était le cas lors de la transcription des textes du corpus *Scoledit* (Wolfarth, 2019, p. 112). Cette étape doit également permettre de sélectionner de manière raisonnable les éléments à reproduire, car une surcharge d'éléments iconiques risque de rendre le texte peu lisible ou difficile à exploiter automatiquement. Par ailleurs, plus le niveau de finesse de la transcription est élevé, plus le temps nécessaire à sa réalisation augmente ; parallèlement, l'augmentation du niveau de détail à prendre en compte peut également entraîner un risque d'erreurs majeur. Wolfarth (2019) prend l'exemple des changements de couleurs d'encre entre texte initial et révisions, les incertitudes entre deux lettres, les variations de taille de caractères etc. : bien que ces éléments puissent être intéressants, leur sur-représentation pourrait rendre les textes peu lisibles ou difficilement exploitables.

L'étape de transcription prévoit aussi d'une part de limiter le plus possible son interprétation du contenu textuel et d'autre part de le reproduire le plus possible à l'identique. Dans le cas d'écritures non normées, ce critère de limitation de l'interprétation humaine est particulièrement difficile à équilibrer : certaines productions nécessitent un effort parfois créatif pour saisir la signification d'origine ou même la graphie finale d'un mot donné.

Bien que la transcription soit souvent considérée comme une représentation du texte, alors que l'annotation apporte habituellement des informations métalinguistiques, certaines opérations de transcription ressemblent fortement à l'annotation, car des informations métalinguistiques sont rajoutées au contenu linéaire du texte lui-même, selon un symbolisme défini au préalable et selon des objectifs définis de recherche et d'exploitation du corpus. « Toute transcription est donc un processus au cours duquel s'effectuent à la fois une perte d'information, notamment une perte des informations visuelles et manuscrites, et un ajout d'information, puisqu'elle introduit une certaine part d'interprétation et d'analyse » (Testenoire, 2017, cité par Wolfarth, 2019, p. 70).

La transcription peut être réalisée selon trois modèles possibles, décrits par Testenoire (2017) : le modèle de transcription diplomatique, le modèle linéarisé, ou le modèle chronologique. Si la transcription diplomatique vise la reproduction exacte d'un manuscrit donné, y compris dans la copie de la position des différents éléments graphiques sur une page donnée, la transcription linéarisée ramène tous ces « signifiants graphiques » sur une même ligne continue de signes, et encode de manière typographique tous les éléments qui sont hors ligne, grâce à des conventions déterminées au préalable (par exemple, le retour à la ligne est encodé par un anti-slash). La transcription chronologique ou génétique, quant à elle, codifie les quatre opérations fondamentales d'écriture (ajouts, suppressions, remplacements et déplacements) et décrit un historique de la création d'un manuscrit donné à travers leur représentation. Dans le cas du projet *Écriscol* (Doquet, 2020), Testenoire avait défini la typologie de transcription adoptée comme

« linéarisée semi-diplomatique » : elle emprunte des éléments des deux types de transcription, car elle « reproduit l'interruption des lignes liées au support et cherche à transcrire tous les signifiants graphiques du manuscrit ; elle est linéarisée car elle insère chaque évènement d'écriture dans une chaine de caractères » (Testenoire, 2017, p. 8).

Dans le cas du projet *Scoledit*, une transcription dite « linéaire semi-diplomatique » a été retenue : elle est principalement linéaire car tous les éléments sont représentés de manière linéaire, à travers un ensemble de symboles et de balises XML, semi-diplomatique car certains éléments génétiques et visuels ont été retenus, dans un souci de représentation fidèle du texte final. Dans ce cas, l'objectif principal de constitution du corpus étant une description linguistique des productions finales, les choix de transcription effectués étaient alignés avec les objectifs poursuivis.

Le travail de transcription, détaillé dans la thèse de Wolfarth (2019, p. 104-117), a été effectué sur un outil de transcription conçu pour la partie française du corpus et par la suite adapté à la transcription de l'italien et de l'espagnol, sur le site internet http://scoledit.org/scoledition/ puis https://scoledit.org/scolinter/[61]. Ce site a comme but l'amélioration des étapes de transcription et de normalisation. Il permet aussi de faciliter la sauvegarde de ces versions des textes dans une base de données centralisée, ainsi que de visualiser les différentes versions des productions écrites.

Comme mentionné auparavant, la transcription des textes elle-même peut être déjà considérée comme une forme d'annotation : lors de cette étape, les transcripteurs doivent effectuer des choix par rapport aux phénomènes à représenter dans le texte transcrit, selon des formalismes établis au préalable pour assurer la cohérence interne du corpus et des phénomènes représentés. Ces choix ont été formalisés dans un protocole de transcription qui est détaillé dans la thèse de Wolfarth (2019) et qui a été décrit dans un guide de transcription. Ce document (disponible en Annexe D) était accessible aux transcripteurs pour référence pendant le processus de transcription.

[61] Les sites web du projet ont été réalisés par Claude Ponton. La partie « édition » utilisée pour élaborer le corpus est accessible uniquement aux utilisateurs enregistrés sur la plateforme.

Les éléments à renseigner dans la transcription que nous détaillons par la suite, ont été codés à travers des balises XML ou des éléments de ponctuation spécifiques. Les retours à la ligne ont été signalés à l'aide de ponctuation, alors que d'autres éléments sont, quant à eux, représentés à travers des balises XML spécifiques. C'est le cas des éléments génétiques comme les révisions et les ajouts, des éléments visuels tels que les dessins ou des choix spécifiques en termes d'indication des guillemets pour l'indication des répliques de dialogues, ou encore des éléments métatextuels, par exemple les titres et la présence de paragraphes dans la production écrite. Sur la plateforme en ligne, ces éléments sont corrélés à des modes d'affichage qui permettent une lecture plus accessible du texte, comme montré dans la Figure 17. Ici, le symbole **[x]** indique un phénomène de révision du texte par l'élève, comme une rature.



Figure 17. Texte visualisé sur la plateforme Scoledit *(corpus français, CE1, 51)*.

1) Eléments textuels

Dans la transcription, le texte est restitué dans sa linéarité, « en accord avec les objectifs de description linguistique du projet » (Wolfarth, 2019, p. 105). Comme déjà décrit par Wolfarth (2019), les choix textuels effectués par l'apprenant sont repris à l'identique, en termes de segmentation des mots et des phrase, d'orthographe, syntaxe et lexique. Les choix erronés des accents (qui constituent une déviation de la norme orthographique en français et en espagnol) sont également reproduits (pour un exemple de transcription visualisé sur la plateforme dédiée, voir Figure 18). Certains choix graphiques sont également repris, comme le choix des majuscules accentuées ou non, ainsi que l'usage de guillemets spécifiques ou d'autres indications de dialogue comme les tirets.

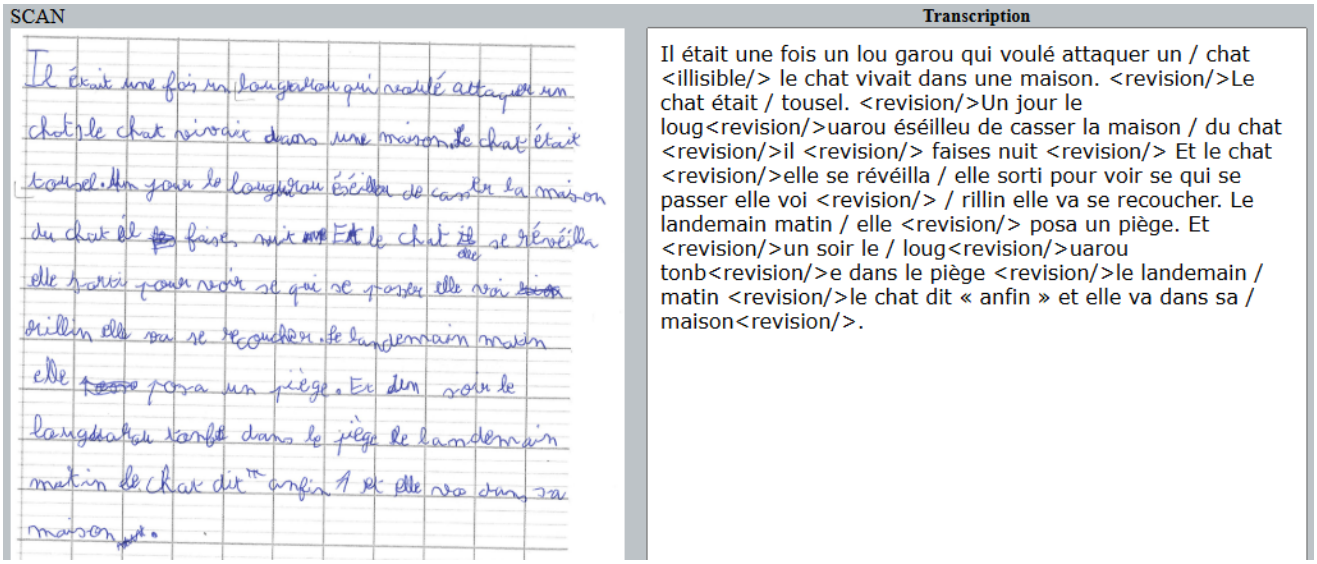
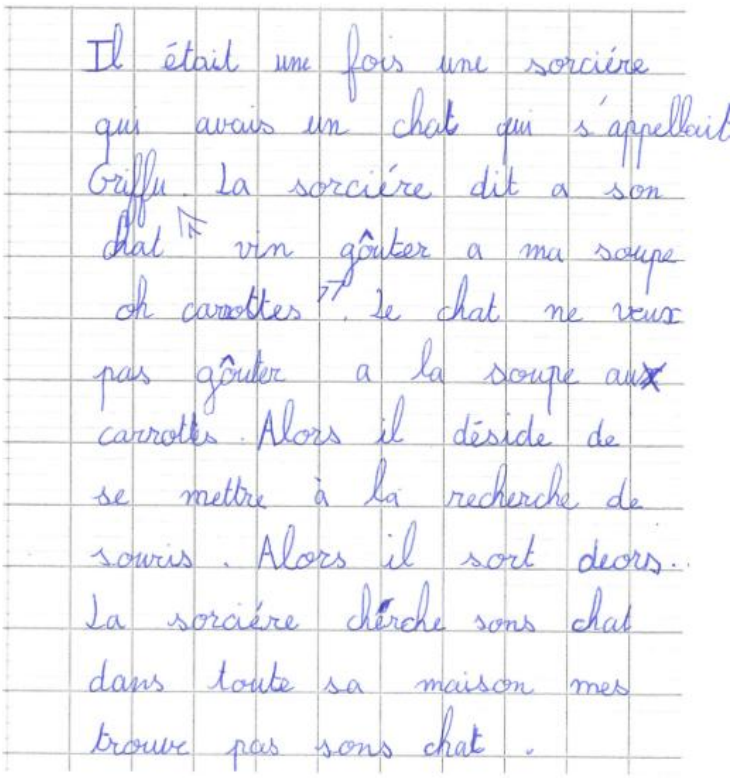


Figure 18. Visualisation côté transcripteur du scan et de la transcription du texte sur la plateforme Scolinter *(corpus français, CE2, 2983).*



Transcription :
Il était une fois une sorciére /
qui avais un chat qui s'appellait /
Griffu. La sorciére dit a son /
chat « vin gôuter a ma soupe /
oh carrottes ». Le chat ne veux /
pas gôuter a la soupe au<revision/> /
carrottes. Alors il déside de /
se mettre à la recherche de /
souris. Alors il sort deors . /
La sorciére cherche sons chat /
dans toute sa maison mes /
trouve pas sons chat. //#

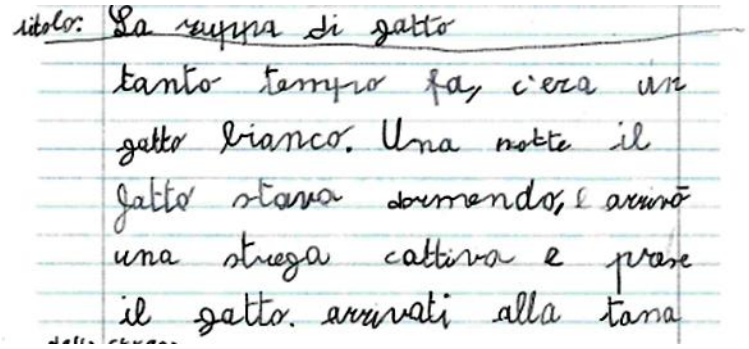
Figure 19. Extrait du scan et de la transcription d'un texte du corpus Scolinter *(corpus français, CE1, 52).*

Des éléments de la mise en page sont également retenus, comme les retours à la ligne. Les retours à la ligne peuvent être utilisés pour définir une nouvelle phrase ou encore une nouvelle unité syntaxique ou sémantique dans des phases d'apprentissage de l'écriture, c'est pour cette raison que les retours à la ligne en fin de support d'écriture ont été distingués des retours à la ligne « intentionnels », c'est-à-dire opérés avant la fin du support d'écriture, en utilisant deux symboles différents, **/** un slash dans le premier cas, **//** double slash dans ce deuxième cas. Le changement de page a été également indiqué par le symbole dièse **#**.

L'adoption de ces règles de transcription permet aussi l'affichage d'une transcription semi-diplomatique, « alignée avec le scan afin d'en faciliter la lecture » (Wolfarth, 2019, p. 110), comme montré dans la Figure 19.

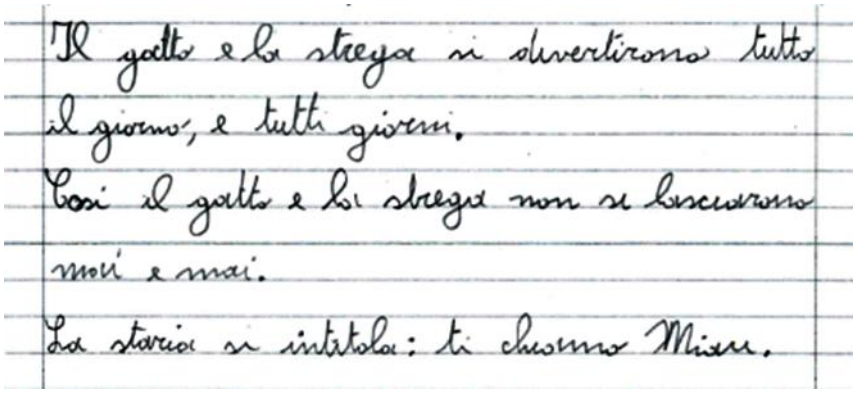
2) Eléments métatextuels

Des éléments métatextuels ont été également retenus dans la transcription, par exemple les titres. Ceux-ci sont indiqués à travers des balises XML dédiées, **<titre> </titre>**. Même si cet élément est habituellement situé en début de texte, dans le cas des textes en italien, ce n'est pas rare de retrouver des titres précédés d'une indication métatextuelle, comme « Titolo » (Figure 20) ou des expressions contenant le verbe « intitolarsi », comme dans le cas de « la storia si intitola », y compris en fin de texte, comme dans le cas du texte dans la Figure 21.



Transcription :
<meta> titolo: </meta> **<titre>** La zuppa di gatto **</titre>**
tanto tempo fa, c'era un /
gatto bianco. Una notte il /
gatto stava dormendo, arrivò /
una strega cattiva e prese /
il gatto. arrivati alla tana /
[…]

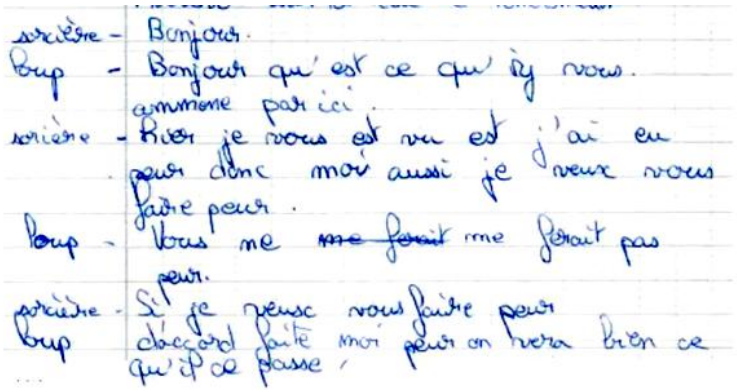
Figure 20. Extrait de scan et transcription du même texte contenant un titre et son indication métatextuelle (*corpus italien, CE1, 735).*



Transcription :
Il gatto e la strega si divertirono tutto /
il giorno, e tutti giorni. //
Cosi il gatto e la strega non si lasciarono /
mai e mai.
<meta> La storia si intitola: </meta>
<**titre**> ti chiamo Miau </**titre**>

Figure 21. Scan et transcription du même texte *(corpus italien, CE2, 826)*. Le titre se trouve en fin de texte, précédé d'une indication métatextuelle.

D'autres éléments métatextuels sont également indiqués, comme la présence d'éléments visibles dans le scan mais à ne pas prendre en compte dans l'analyse du texte :par exemple, un paragraphe rédigé par l'enseignant ou les indications du locuteur en début de ligne en correspondance des tours de parole du ou des personnages. Ces éléments ont été indiqués par la balise **<meta> (**Figure 22**)**. Cette balise permet d'ailleurs de ne pas prendre en compte l'élément lors d'une analyse automatique du texte.



Transcription :

[...] **<meta>**sorcière**</meta>** - Bonjour. / **<meta>**loup **</meta>** - Bonjour qu'est ce qu'<unsure>iy</unsure> vous / ammene par ici. / **<meta>**sorcière**</meta>** - hier je vous est vu est j'ai eu / peur donc moi aussi je veux vous / faire peur. / <meta>loup</meta> - Vous ne <revision/> me ferait pas / peur / **<meta>**sorcière**</meta>** - Si je veux vous faire peur / **<meta>**loup**</meta>** d<ajout>'</ajout>accord faite moi peur on vera bien ce / qu'il ce passe. /# [...]

Figure 22. Extrait d'un scan et de la transcription du même texte (*corpus Scoledit, CM2, 103).* Exemple tiré de Wolfarth, 2019.

3) Eléments non textuels

Une balise **<dessin/>** peut indiquer la présence de dessins dans le texte, y compris dans le cas où le texte a été embelli par un petit dessin, comme dans le cas de la clé à molette, lié au thème de l'histoire qui en accompagne le titre « Le chat constructeur » dans la Figure 23 ; cela peut aussi concerner des lettres ou chiffres dessinés, comme dans la Figure 24, où le nom d'un des personnages a été écrit en grandes lettres dessinées.



Transcription:

<titre> Il gatto costruttore </titre> **<dessin/>**//

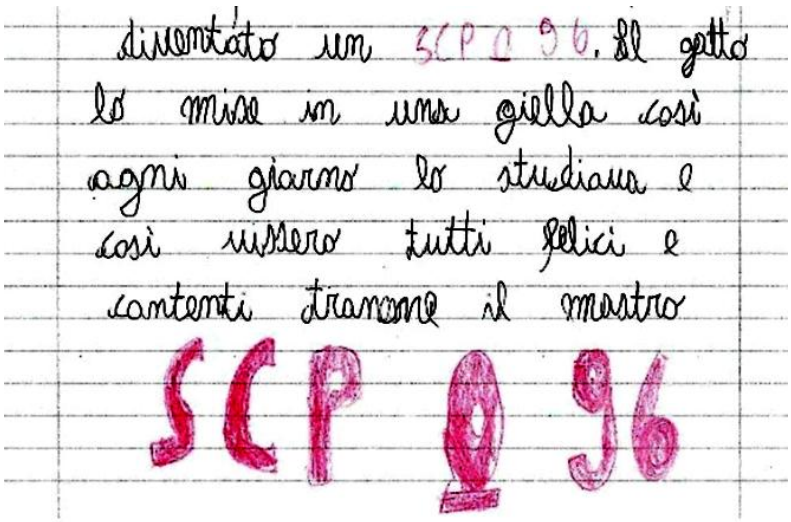
C'era una volta un gatto /

costruttore, quel gatto era così /

bravo a costruire che un /

giorno ha costruto un robot. / […]

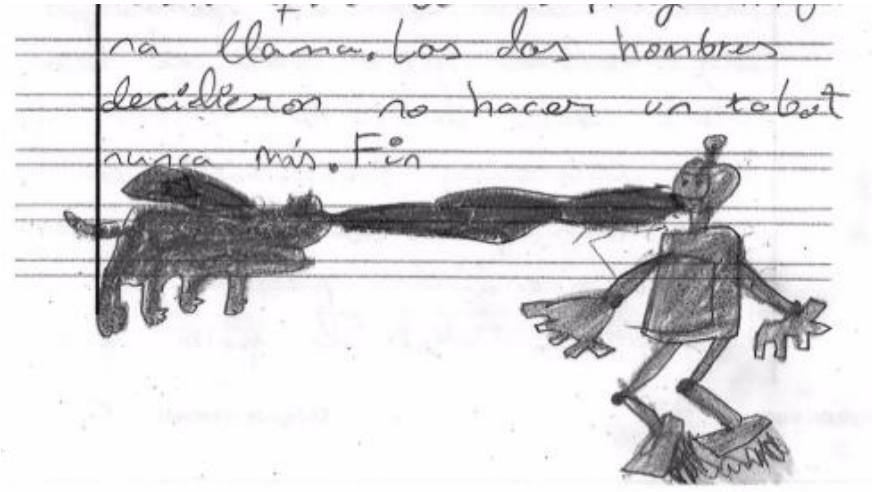
Figure 23. Extrait d'un scan et de la transcription du même texte *(corpus italien, CE2, 769).*



Transcription :
[…] / diventato un SCP O 96. Il gatto / l
o mise in una <unsure>giella </unsure>così lo studiava e /
cosi vissero tutti felici e contenti tranne il mostro
<dessin/>SCP O 96**<dessin/>**

Figure 24. Extrait du scan et de la transcription du même texte *(corpus italien, CE2, 769).*

Bien que la présence de dessins ait été découragée afin de laisser plus de place à la tâche d'écriture, certains textes contiennent de véritables illustrations, en lien avec les personnages et les cadres évoqués. Dans la Figure 25, le robot dessiné par l'auteur espagnol du texte représente probablement le robot protagoniste de l'histoire racontée (où un robot détruit le laboratoire de ses créateurs).



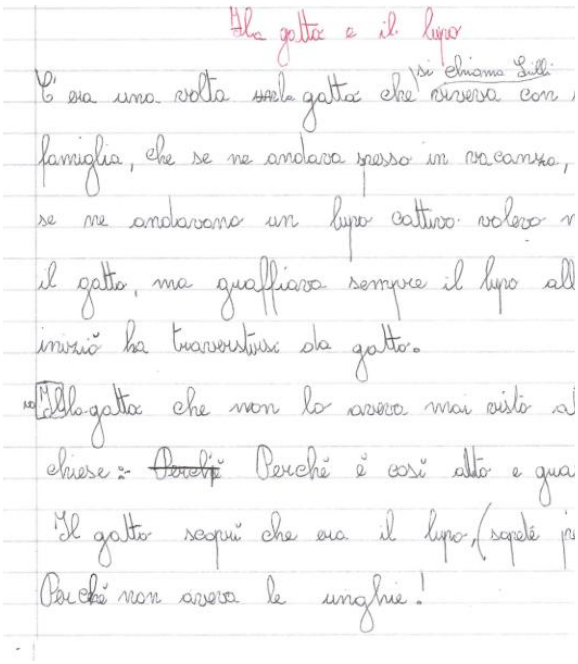
Transcription :
[…] Los dos hombres /
decidieron no hacer un robot /
nunca más. Fin **<dessin/>**

Figure 25. Extrait du scan et de la transcription du texte du corpus espagnol *(CE1, 988).*

4) Eléments génétiques

Etant donné les objectifs du projet, principalement dédié à la description linguistique, les aspects génétiques sont donc peu détaillés dans les transcriptions des textes du corpus. La génétique textuelle (comme déjà mentionné dans le Chapitre 2, §3) codifie habituellement les opérations d'écriture de base d'ajout et de suppression, dont la combinaison permet de coder les remplacements et les déplacements (Doquet et al., 2017). Dans le projet *Scoledit* puis *Scolinter*, seules les opérations d'ajout ont été retenues, et donc les révisions ou ajouts de mots dans un texte sont codifiés. La position des révisions dans le texte est maintenue, codifiée à travers les balises

<revision/> (et non pas le contenu supprimé), alors que les balises **<ajout> </ajout>** ont été utilisées pour entourer le mot rajouté dans le texte par l'élève.

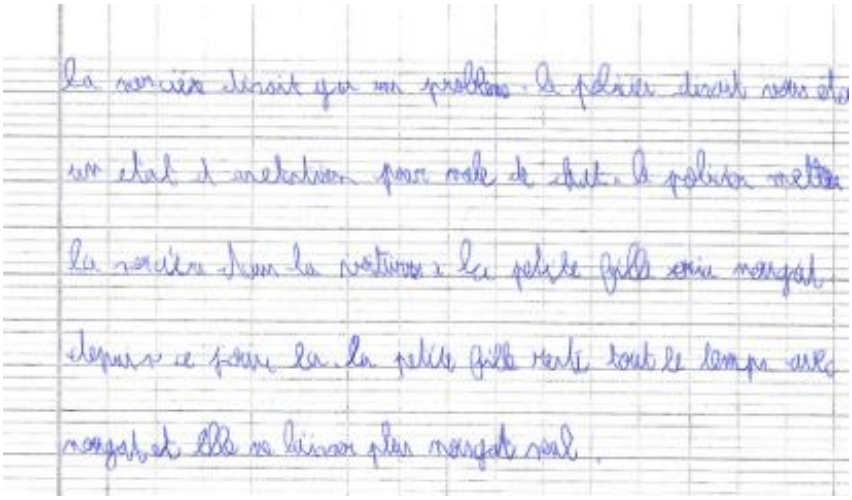


Transcription :
<titre>La gatta e il lupo</titre>
C'era una volta **<revision/>** la gatta che **<ajout>**si chiamava Lilli**</ajout>** viveva con una /
famiglia, che se ne andava spesso in vacanza, quando /
se ne andavano un lupo cattivo voleva mangiarsi /
il gatto, ma graffiava sempre il lupo allora /
iniziò ha traverstirsi da gatto. //
<revision/> la gatta che non lo aveva mai visto allora si /
chiese:- **<revision/>** Perché è così alto e grasso? - /
Il gatto scoprì che era il lupo, (sapete perché) /
Perché non aveva le unghie! // /# […]

Figure 26. Extrait du texte du corpus italien contenant des révisions et un ajout (*corpus italien, CE2, 826).*

5) Indications d'incertitude dans la transcription

Deux balises sont utilisées dans le cas d'incertitude dans la transcription, **<unsure>** et **<illisible/>** pour indiquer des cas où le transcripteur est en doute dans la transcription à effectuer ou dans les cas où l'identification des éléments linguistiques dans le texte est difficilement possible. Les productions non alphabétiques ou vides ont été également indiquées à travers une balise dédiée, **<empty/>**.



Transcription : […]
la sorciére disait ya un probleme. le policier disait vous était /
un etat d'aretation pour vole. le policier mett<revision/>e /
la sorcière dans la voiturr<revision/>
<illisible/>. la petite fille <revision/> cria nougat /
depuis ce joure la. la petite fille rente toute le temps avec /
nougat et elle ne laisser plus nougat seul.

Figure 27. Extrait du scan et de la transcription contenant un fragment illisible (*corpus français, CE1, 1926).*

Le même protocole de transcription a été adopté pour les corpus français, italien et espagnol. Le corpus français a été transcrit par les enseignants-chercheurs du projet ainsi que par des étudiants, rémunérés grâce à des vacations recherche (Wolfarth, 2019). Les textes italiens ont principalement été transcrits par des enseignantes-chercheuses responsables du recueil et des étudiants, de Licence ou de Master, bénévoles. Les textes du corpus espagnol ont été transcrits uniquement par les chercheuses du projet[62]. La vérification est effectuée par une personne autre que celle qui a transcrit les mêmes textes ou par la même personne après un certain temps pour réduire les erreurs. Cette vérification permet de réduire le taux d'erreurs possibles dans des transcriptions difficiles comme celles des textes déviant de la norme et qui présentent des graphies parfois difficiles à décoder.

Cependant, comme remarqué par Wolfarth, « malgré ce processus de transcription et de vérification, il est nécessaire de garder à l'esprit que les erreurs subsistent et que le corpus est amené à être révisé sans cesse » (Wolfarth, 2019, p. 135). Même après plusieurs années et le travail de nombreux chercheurs et vacataires sur le corpus, il subsiste parfois des erreurs résiduelles nécessitant des corrections. Ce phénomène s'est d'ailleurs présenté de façon occasionnelle au cours de notre travail de thèse. Ainsi, au cours du processus d'annotation des textes, il a parfois été nécessaire d'apporter des corrections aux transcriptions existantes, qui ont ensuite été intégrées dans le corpus.

7.3. Conventions de normalisation

La normalisation aboutit à une deuxième version du texte, conçue comme un rapprochement du texte d'origine la norme de la langue cible, tout en respectant au mieux les choix expressifs

[62] En fait, dans les équipes française et italienne, nous pouvons retrouver des chercheuses hispanophones natives ou à un niveau avancé de connaissance de la langue, qui ont pu donc épauler l'équipe espagnole dans les opérations de traitement du corpus.

effectués dans l'écrit original. Selon les projets, cette normalisation peut être embarquée, donc intégrée dans le texte lui-même, par exemple à travers des balises qui indiquent les formes correctes vis-à-vis des formes fortement déviantes, ou déportée, c'est-à-dire constituer une donnée à part entière par rapport aux transcriptions. Par exemple, dans le cadre du projet *Écriscol* (Doquet, 2020) (cf. Chapitre 2, §3) c'est cette dernière méthodologie qui avait été choisie.

La normalisation du texte peut donc être considérée comme une forme d'annotation déportée du texte d'origine. Il s'agit de reconstruire une version du texte qui respecte autant que possible l'original tout en étant conforme aux normes de la langue cible. Cependant, cette opération implique souvent une certaine interprétation du sens du texte, surtout dans le cas où l'écriture de l'élève ne peut pas être immédiatement décodée. Dans le cas de notre corpus, la normalisation porte essentiellement sur les niveaux textuel, orthographique et morphologique des textes. Ces normalisations ont été réalisées avec le double objectif de permettre l'exploitation des textes à l'aide du TAL et de reproduire sous cette forme les choix faits par les élèves en termes lexical et stylistique. En outre, il est toujours possible d'établir un lien entre transcription et normalisation, ce qui fait de la normalisation « l'interface entre la transcription de la production et les outils d'exploitation » (Wolfarth, 2019, p. 122)

Lors de la conception du corpus *Scoledit*, le choix avait été fait de « produire une normalisation déportée, indépendante de la transcription, à l'instar de celle pratiquée dans les projets de Fairon et al. (2006) » (Wolfarth, 2019, p. 140). Cela à cause du grand nombre des variations à la norme rencontré dans ce type d'écrits, qui aurait rendu une normalisation embarquée très peu lisible. De plus, l'objectif était de pouvoir traiter de manière comparable et à travers les mêmes outils les productions écrites narratives et le corpus des dictées, qui comprenait la version transcrite ainsi que la normalisation de ces textes. Ces normalisations, étant partie intégrante de l'analyse dans le projet *Scoledit* puis dans le corpus *Scolinter*, il est important de décrire leurs caractéristiques et les choix effectués lors de cette étape, choix également codifiés dans un protocole de normalisation établi et appliqué aux trois langues. Dans cette version, certains éléments sont reconstitués dans le texte lui-même, alors que l'insertion d'éléments absents est faite à travers l'utilisation de balises au format XML. Certains éléments déjà présents dans la transcription sont également reconduits dans la normalisation (par exemple, l'indication du titre et les retours à la ligne intentionnels) alors que d'autres éléments ne sont pas maintenus car spécifiques à la transcription (la présence de dessins, révisions, ajouts, retours à la ligne dictés par la fin du support physique et changements de page). Le processus de normalisation a été codifié dans un guide qui spécifie les choix de normalisation à effectuer, et qui était consultable par les responsables des normalisations tout au long de ce processus. La version française et italienne du guide est disponible en Annexe E.

Les éléments suivants ont été reconstitués dans les normalisations :

1) La segmentation des mots attendue et leur orthographe standard sont normalisées – reconstituer ces éléments permet l'analyse automatique à l'aide d'outils TAL des mots du texte. Cela permet aussi d'étudier leurs variations orthographiques (Wolfarth, 2019, p. 122). Les majuscules sont reconstituées après les points et en début de noms propres.

	Normalisation[64] *:*
Ses lisiar dune sorstir cui /	C'est l'histoire d'une sorcière qui
sapelle **morgule** cui avér un chat. /	s'appelle **Morgule** qui avait un chat
Cui sapelle minou. (…)	qui s'appelle **Minou**. (…)
(corpus français , CE1, S205)[63]	

2) Le choix du tiroir verbal effectuée par l'élève est maintenu mais le verbe est corrigé, comme dans le cas de l'exemple suivant (CP, 6), tiré du corpus italien. Les verbes *cadde*, *si fece (male)*, *miagolò*, *si preoccupò* et *si rimpicciolì* ont été reconduits à leurs formes attendues ;

		Normalisation :
CIERA·UNA·VOLTA·UN	/	C'era una volta un
GATINO·CHE·MENTRE·/		gattino che mentre
LA·MAMMA·DEL·GATINO·	/	la mamma del gattino
DORMIVA·ILGATINO·**CADE**·	/	dormiva il gattino **cadde**
E·**SI·FECIE**·MOLTO·MALE	/	e **si fece** molto male
POI·IL·GATINO·**MIAGOLO**·/		poi il gattino **miagolò**
E·TUTTI·SI·SVEGLIARONO	/	e tutti si svegliarono
E·LA·MAMMA·**SI·PRUCOPO**·	/	e la mamma **si preoccupò**
PERCHE·IL·GATINO·**SI**·/		perché il gattino **si**
RINPICIOLO		**rimpicciolì**.
(corpus italien, CP, 6)		

Les éléments suivants, quant à eux, ont été indiqués à travers l'ajout de balises XML dans la normalisation :

1) l'indication des titres, déjà présente en transcription, est maintenue à travers les balises **<titre> </titre>** ;
2) les retours à la ligne intentionnels ont été marqués en tant que paragraphes à travers la balise **<p/>** ;

	Normalisation :
<titre>LA STREGA FELICE**</titre>** //	**<titre>**La strega felice**</titre>**
Un giorno che non era ancora giorno una /	<p/> Un giorno che non era ancora giorno una
vecchia strega di nome Riunione si sveglio /	vecchia strega di nome Riunione si svegliò
alle 7:00 per vestirsi fare colazione e /	alle 7:00 per vestirsi, fare colazione e
lavare i denti. //	lavare i denti. <p/>
Ma voleva una cosa. //	Ma voleva una cosa.<p/>
Cosa? //	Cosa? <p/>
Un amicetto per pasare il tempo! // (…)	Un amichetto per passare il tempo! <p/> (…)
(CE1, S11, corpus italien)	

[63]Pour améliorer la lisibilité de la transcription, les nombreuses marques de révision indiquées dans ce texte ont été éliminées de l'exemple.

[64] Les normalisations ne contiennent pas des marques de retour à la ligne déterminés par la fin du support d'écriture. Cependant, nous avons fait le choix de reproduire l'alignement graphique de transcription et normalisation dans nos exemples pour en améliorer la lisibilité

3) l'ajout d'éléments marquant la séparation entre fragments textuels quand absents dans les textes, tels que les marques de ponctuations fortes ou faibles ; ceux-ci sont indiqués à l'aide de balises **<segmentation/>** dans le cas de signes de ponctuation indiquant un changement thématique / d'épisode / de situation, et de la balise **<s/>** dans le cas où il semble nécessaire de rajouter un signe de ponctuation tels que la virgule, les deux points ou encore la conjonction « et ». Les points de fin de texte ont été toujours directement reconstitués, ainsi que les majuscules de début de production ;
4) les omissions de mots présentes dans les textes originaux : dans le cas où des mots susceptibles de contribuer de manière significative à la structure syntaxique de la phrase auront été oubliés par l'élève, une balise **<omission/>** a été insérée dans la normalisation. Cette balise contient un attribut '**type'** indiquant la catégorie morphosyntaxique possible du mot omis, reconstruite à partir du contexte par le chercheur ou la chercheuse responsable de la normalisation ;
5) l'indication des segments de discours direct, indiqués à travers les balises **<dialogue> </dialogue>**, y compris dans les cas où ces segments sont indiqués à travers des signes de ponctuation appropriés

	Normalisation :
(…)	(…)
Il gatto in vecie poteve /	Il gatto invece poteva
traspormarsi in quallunque cosa /	trasformarsi in qualunque cosa **<segmentation/>**
il gatto disse alla stega io /	il gatto disse alla strega **<dialogue>**io
voglio essere ilbero la stega /	voglio essere libero**</dialogue>** la strega
disse basta melo chiedi sempre /#	disse **<dialogue>**basta me lo chiedi sempre
e il gatto scapo.//	**</dialogue>** e il gatto scappò. <p/>
La stega arabiata trovo il gatto /	La strega arrabbiata trovò il gatto
e lo mette nella gabia e /	e lo mette nella gabbia e
la stega dispiagiuta libero la /	la strega dispiaciuta **<omission type="pronom"/>** liberò **<segmentation/>** la
stega disse scusa se sono /	strega disse **<dialogue>**scusa se sono
stata cattiva e il gatto e la /	stata cattiva**</dialogue>** e il gatto e la
stega fanno la pacie e la /	strega fanno la pace e la
stega era più buona (…)	strega era più buona
(CE1, S14, corpus italien)	(…)

6) les balises **<unsure> </unsure>** indiquent les cas où il y avait un doute dans la transcription ou un doute est survenu dans la phase de normalisation.
7) La balise **<incomprehensible/>** qui indique les cas de parties où le texte est difficilement interprétable de la part du transcripteur/normaliseur.
8) La balise **<nonfini/>** dans le cas de texte incomplet ou non achevé.
9) La balise **<empty/>** , reprise depuis les transcriptions, qui indique les textes vides mais qui ont été retenus dans le corpus.

7.4. L'importance des traitements manuels dans la constitution d'un corpus scolaire

Les étapes décrites ci-dessus sont primordiales pour la constitution et l'exploitation du corpus, selon les objectifs de recherche du projet, c'est-à-dire étudier et décrire les spécificités des écrits scolaires à l'école primaire. De plus, il est important que les résultats de ces étapes du travail garantissent la meilleure qualité possible des textes obtenus, pour assurer une réelle représentativité des phénomènes observés.

Déjà lors des précédentes étapes de constitution du corpus, des expérimentations ont exploré les possibilités de transcription automatique des textes à partir des scans, en utilisant les techniques de reconnaissance optique de caractères, ou OCR. Les outils disponibles en utilisation libre qui exploitent ces techniques, que nous avons testés dans des temps plus récents, restent encore très fautifs sur des écritures manuscrites peu normées et très variées, comme c'est le cas dans un échantillon composé de textes de nombreux élèves. Pour cela, toutes les transcriptions ont été effectuées manuellement par plusieurs enseignants-chercheurs, étudiants et vacataires.

La centralisation de leur travail a été possible grâce à la plateforme *Scolinter*[65]. Ce site web permet de transcrire les textes en gardant le scan à côté de la fenêtre de transcription et de la fenêtre de normalisation, de telle manière que les trois versions du texte soient toujours sous l'œil de la personne chargée du traitement du texte. La visualisation du scan est dotée d'une loupe virtuelle qui permet de cibler des détails peu lisibles de l'image du texte. Des boutons placés à côté des fenêtres permettent l'insertion des balises spécifiées par le guide pour chacune des deux versions tapuscrites du texte. Le site web intègre également un outil de vérification de la bonne formation des balises avant de valider l'insertion d'un nouveau texte dans la base de données du corpus. Ce site web représente un des outillages possibles pour effectuer ce travail : il permet de mieux visualiser les données pour les chercheurs qui souhaitent explorer le corpus, et permet aussi de vérifier la qualité des textes du côté des chercheurs responsables du projet (car le site permet de mieux voir les textes, avec leur original en face, permet aussi de vérifier les balises mal formées ou manquantes etc.).

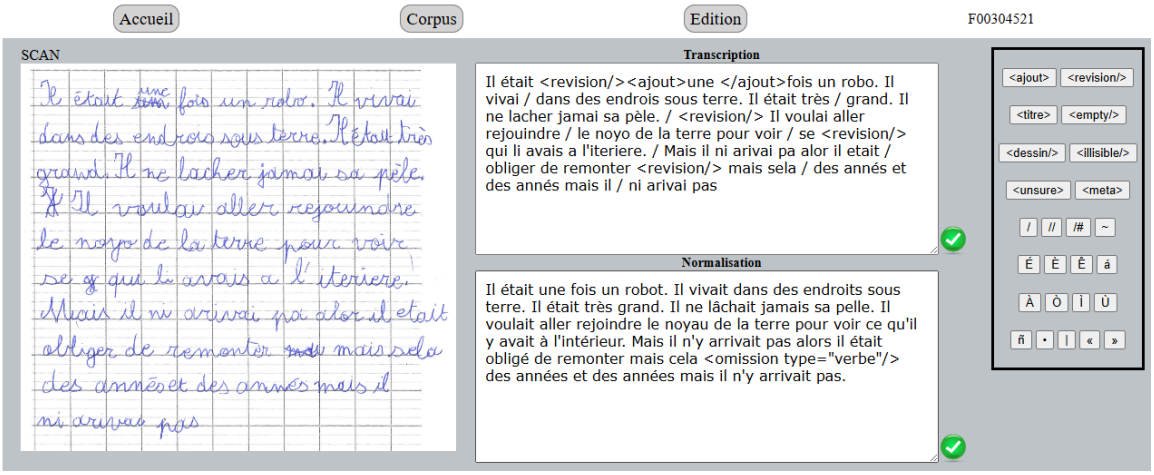


Figure 28. Page d'édition d'un texte du corpus sur la plateforme Scolinter.

[65] La plateforme est disponible au lien suivant : https://scoledit.org/scolinter/corpus.php

7.5. Traitements automatiques et traitements manuels

La transcription

Les traitements qu'on vient de décrire restent couteux en termes de temps et de ressources humaines et sont parfois assez rébarbatifs à opérer, entrainant notamment une fatigue de l'annotateur et donc des risques d'erreur accrus. Néanmoins, ils ont été réalisés manuellement tout au long du projet pour plusieurs raisons. Lors du début du projet (en 2016 pour le corpus *Scoledit* et en 2018 pour le corpus *Scolinter*), les outils de reconnaissance automatique des caractères (OCR, *Optical Characters Recognition*) disponibles et prêts à l'emploi ne pouvaient pas traiter efficacement l'écriture manuscrite des enfants, particulièrement avec un corpus comportant de nombreux scripteurs distincts (Wolfarth, 2019). La reconnaissance automatique des caractères à partir d'images des textes, ou reconnaissance optique des caractères, reste une des tâches clés dans les problèmes du traitement automatique de la langue (Alhamad et al., 2024). Même si des avancées importantes ont été réalisées dans ce domaine et permettent aujourd'hui d'obtenir des résultats satisfaisants dans certains contextes, notre corpus demeure particulièrement difficile à traiter par ces outils.

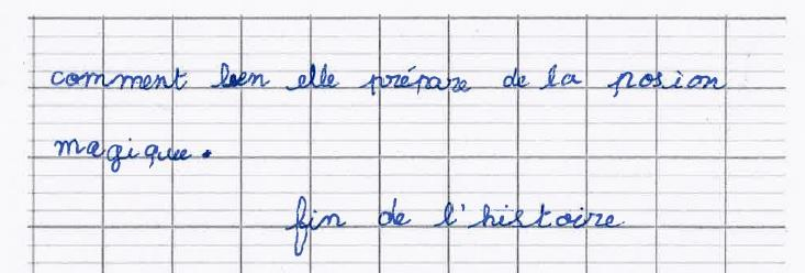
En effet, les scans de notre corpus présentent des caractéristiques très variables, en termes de support matériel, d'outil graphique (le crayon, parfois employé, produit un tracé moins net que le stylo, dont l'usage était pourtant préconisé lors du recueil), ainsi qu'en termes de résolution de l'image et de graphie de l'élève. L'ensemble de ces variations introduit du bruit dans le traitement automatique des textes. À cela s'ajoute une difficulté spécifique liée à la nature du corpus : les textes ont été produits par un grand nombre d'élèves, ce qui implique une forte variabilité. Même dans le cas de données longitudinales, où un même élève est suivi dans le temps, les productions sont espacées d'environ une année, ce qui peut entraîner des variations importantes de la graphie, en particulier dans un contexte d'apprentissage de l'écriture.

Enfin, ces outils doivent faire face à une complexité majeure dans le traitement de textes en écriture cursives, qui constitue le cas majoritaire dans notre corpus (notamment pour les textes en français). La continuité du tracé, le fait que les lettres soient reliées entre elles (Alhamad et al., 2024), ainsi que la variabilité des formes graphiques, rendent leur reconnaissance plus difficile que dans le cas de l'écriture en script. Tous ces facteurs, combiné à la variabilité inter-scripteurs, rendent l'utilisation des outils d'OCR peu efficiente pour nos objectifs de traitement automatique.

Afin de vérifier empiriquement ces limites, nous avons testé l'application de ces outils sur un échantillon de cinq textes issus du corpus français et cinq textes issus du corpus italien. Pour ce faire, nous nous sommes servi de la version du logiciel Tesseract (Smith, 2007) disponible grâce au réseau HumaNum[66]. Les tests ont été réalisés une première fois au début de l'année 2023, puis à nouveau en avril 2025. Nous présentons ici un seul exemple de résultat par langue, obtenus en 2025, car l'ensemble des transcriptions produites est fortement affecté par le bruit, ce qui les rend difficilement exploitables. Néanmoins, les résultats obtenus en italien apparaissent légèrement meilleurs que ceux observés pour les textes en français, pour lesquels certains essais d'OCR

[66] https://www.huma-num.fr/

aboutissent à des transcriptions vides. Ces résultats confirment les difficultés évoquées précédemment, liées à la variabilité graphique et aux caractéristiques des écrits d'élèves.



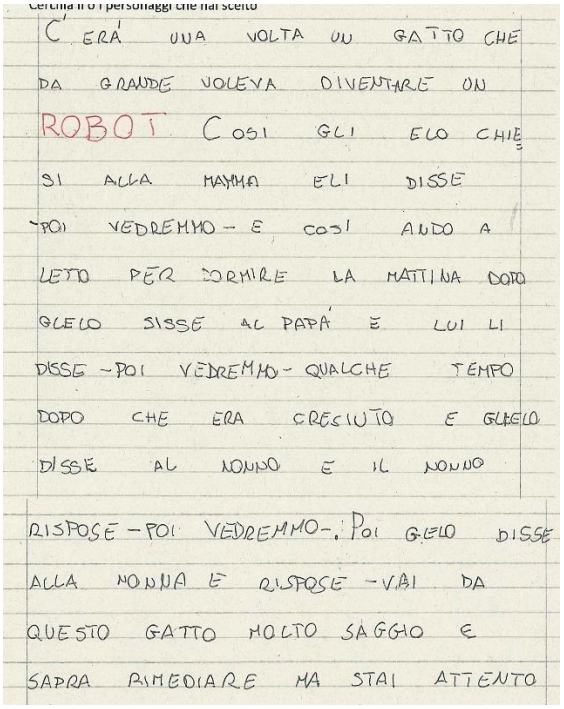
db fr L fa| noi in)

dr ble. hits

Transcription attendue :

(…) Comment b[x]en elle prépare de la posion /magique. // fin de l'histoire

Figure 29. Test d'OCR effectué sur un extrait du texte du corpus français (CE1, 828), image et transcription attendue.



LUI UNITO. UT PETSUTlabpl LUTTO TIdl SUCITO

Coi oi
DALL EPM VOR DIVEVTRE VU

ROBON Cosi GELO SAIL

Si ALLA: pun ELI pisshi t1
SO) VEDREMMO — E così ALDI A

Lema Pao Sole LA HAT So

Serio ASI Lol
DISSE —Pol VERREMM- QUALCHE TEMPO |

DOPo — CHE ERA CRECIUIO E GUEIO

Disk Rui Au paoli.

IMISPOGE — Tor vengeMMO-. Pai CA iii

ALLA MObUA ose Ve Da
QUESTO GATTO MOLTO SAGGIO €

SAPLA BIMEDIALE MASIAI ATTENTO. (…)

Transcription attendue :

C'era una volta un gatto che / da grande voleva diventare un / robot Cosi gli elo chie~/si alla mamma eli disse / – poi vedremmo – e cosi ando a / letto per dormire la mattina dopo / glielo disse al papà e lui li / disse – poi vedremmo – qualche tempo / dop che era cresiuto e glielo / disse al nonno e il nonno / rispose – poi vedremmo –. Poi glielo disse / alla nonna e rispose – vai da / questo gatto molto saggio e /sapra rimediare ma stai attento (…)

Figure 30. Extrait du test d'OCR sur le texte du corpus italien (CE2, 747), image et transcription attendue.

La normalisation

Dans le cas de la normalisation, le guide (cf. ce chapitre, §7.3, disponible en Annexe E) précise les éléments à garder et ceux à normaliser. La nécessité de rester toujours au plus près de la production de l'élève tout en corrigeant les maladresses orthographiques ainsi que la segmentation en mots, la volonté de garder les caractéristiques lexicales et les choix de l'élève en termes de

morphologie verbale bien que maladroites, etc. sont des choix effectués par les chercheurs du projet pour répondre à des objectifs de recherche spécifiques (cf. ce chapitre, §7.3).

Lors du début du projet en 2015, les LLM génératives n'étaient pas diffusées ni prêtes à l'emploi comme c'est le cas aujourd'hui. Celles existantes ne semblaient pas en mesure de normaliser de manière efficace des textes éloignés de la norme, alors que les avancées récentes de ces technologies rendent potentiellement possible la réalisation de normalisations automatiques.

Des tâches très similaires sont déjà abordées dans des défis, comme dans le cas du défi *Multilingual Grammatical Error Detection*[67]. Même si la description de cette tâche se rapproche de celle du processus de normalisation, les objectifs poursuivis sont différents : bien que la tâche de correction minimale préconisée pour cette campagne puisse se rapprocher des besoins de notre projet et permettre d'identifier, dans les systèmes proposés, des pistes d'automatisation de ce processus, elle reste encore éloignée de nos objectifs de recherche actuels.

En effet, la résolution de cette tâche préconise la correction grammaticale et syntaxique des écrits, y compris la reformulation des formes déviantes de la norme, ce qui donne lieu à des résultats « sur-corrigés ». Pour cette raison, certains choix de normalisation effectués pour nos objectifs de recherche nous ne semblent pas encore totalement intégrables dans un processus de normalisation automatisé. Cependant, les modèles de type LLM pourraient constituer un support pertinent pour assister le processus de normalisation, à condition d'être intégrés dans une démarche semi-automatique incluant une validation et un contrôle humains.

8. Description du corpus recueilli

Comme déjà mentionné, un recueil longitudinal est un processus complexe, qui dépend aussi des écoles impliquées. Dans le cas des différents corpus, une partie des écoles a accepté de participer tout au long du recueil, ce qui a permis d'obtenir des données longitudinales sur 5 années en France et des données semi-longitudinales en Espagne et Italie, où l'année de crise sanitaire a interrompu le processus de recueil. Le recueil a eu lieu en France dans 38 écoles, dans 9 établissements en Italie et dans 13 en Espagne. Si en France le recueil était étendu à différentes académies (cf. Chapitre 4, §2), en Italie et Espagne le recueil était limité à la seule région d'appartenance des équipes de chercheuses participant au projet : le manque d'un financement dédié à la collecte du corpus et le peu de moyen humains disponibles pour la réalisation du recueil ont grandement limité les possibilités d'étendre le recueil à d'autres parties du pays.

8.1. France

Sur les 38 écoles qui ont participé au recueil en France, 373 élèves étaient présents du CP au CE1 (Wolfarth, 2019) avec une baisse à 369 en CE2 et CM1 puis 337 en CM2. La période de recueil, de 2014 à 2018, correspond pour les deux premières années au recueil effectué dans le cadre du

[67] Plus d'informations sur le défi MultiGEC sont disponibles au lien suivant https://spraakbanken.github.io/multigec-2025/shared_task.html

projet *Lire écrire au CP* entre 2014 et 2015, puis continué par les chercheurs du LIDILEM entre 2016 et 2018.

Niveaux scolaires	Textes recueillis	Textes longitudinaux	Année scolaire
CP – 1	2 507	373	Juin 2014
CE1 – 2	2 049	373	Juin 2015
CE2 – 3	1 135	369	Mai-juin 2016
CM1 – 4	1 132	369	Mai-juin 2017
CM2 – 5	1 030	337	Mai-juin 2018
CP-CM2 – 1-5	**7 853**	**1 820**	**2014-2018**

Tableau 9. Textes recueillis pour la partie française du corpus - corpus Scoledit.

8.2. Italie

En Italie, le recueil a été interrompu par la crise sanitaire. Cependant, les chercheuses du projet ont pu recueillir des données semi-longitudinales grâce à l'organisation de deux campagnes de recueil parallèles : une première réalisée en CP (1e année) et en CE1 (2e année) en 2018 en fin d'année scolaire (juin), et une seconde organisée dans les mêmes classes l'année suivante, en 2019, à la même période. Cela nous a permis d'obtenir des données véritablement longitudinales sur des classes de niveaux 2 et 3, qui auraient autrement été perdues en raison des deux années de fermeture des écoles dans la région concernée par le recueil.

Ainsi nous avons à disposition les textes de 466 élèves avec des productions pour le CP, CE1 et CM2 ainsi que les textes de 691 élèves dont on a pu récolter des textes en CE1 et CE2.

Niveaux scolaires	Textes recueillis	Textes longitudinaux	Année scolaire	Textes recueillis	Textes longitudinaux	Année scolaire
CP – 1	715	447	2018	-	-	-
CE1 – 2	742	447	2019	726	691	2018
CE2 – 3	-	-	-	698	691	2019
CM1 – 4	-	-	-	-	-	-
CM2 – 5	740	447	2022	-	-	-
CP-CM2 – 1-5	**2 197**	**1341**	**2018-2022**	**1 433**	**1382**	**2018-2019**

Tableau 10. Textes recueillis pour la partie italienne du corpus. Les deux recueils de 2018 et de 2019 ont été réalisés en parallèle.

8.3. Espagne

En Espagne le contexte sanitaire a aussi interrompu la collecte en 2020. Néanmoins, une collecte parallèle de CE1 et CE2 a été réalisée en 2020, ce qui nous permet d'avoir des données de 2ème niveau même si elles ne sont pas intégrées dans un suivi longitudinal. Le recueil a concerné 1122 élèves et a permis de collecter 2203 textes.

Niveaux scolaires	Textes recueillis	Textes longitudinaux	Année scolaire	Textes recueillis	Année scolaire
CP – 1	382	131	2018	-	-
CE1 – 2	-	-	-	385	2021
CE2 – 3	394	131	2021	-	-
CM1 – 4	457	131	2022	-	-
CM2 – 5	585	131	2023	-	-
CP-CM2 – 1-5	**2 203**	**524**	**-**	**385**	**-**

Tableau 11. Distribution des textes recueillis pour le corpus espagnol.

Niveaux scolaires	Écoles	Classes	Textes	Année scolaire
CP – 1	8	18	382	2018-2019
CE1 – 2	10	18	385	2020-2021
CE2 – 3	10	20	394	2020-2021
CM1 – 4	11	22	457	2021-2022
CM2 – 5	13	28	585	2022-2023
CP-CM2 – 1-5	**-**	**-**	**2 203**	**2018-2023**

Tableau 12. Distribution des écoles et des années scolaires - corpus espagnol.

8.4. Textes du corpus à disposition

Le traitement des textes (transcription et normalisation, cf. §7.2 et §7.3) des parties italienne et espagnole du corpus est actuellement encore en cours. Les chiffres indiqués ici sont ceux observés en avril 2025. À terme, les textes de CM2 italiens ainsi que ceux de CM1 et CM2 espagnols seront aussi disponibles sur la plateforme *Scolinter*.

Niveaux scolaires	Transcription			Normalisation		
	français	**italien**	**espagnol**	**français**	**italien**	**espagnol**
CP – 1	373	702	382	373	702	382
CE1 – 2	373	672	384	373	672	384
CE2 – 3	369	96	394	369	96	394
CM1 – 4	369	-	457	369	-	En cours
CM2 – 5	337	En cours	585	337	En cours	En cours
CP-CM2 – 1-5	**1 820**	**1 487**	**2 203**	**1 820**	**1 487**	**1 160**

Tableau 13. Textes disponibles sur la plateforme Scolinter.

Bien que le corpus espagnol fasse partie des données collectées dans le cadre du projet, son annotation n'a pas été intégrée à la présente étude. Ce choix s'explique principalement par des contraintes organisationnelles et méthodologiques. D'une part, le temps de cette recherche n'a pas permis de mettre en place une campagne d'annotation dédiée dans des conditions comparables à celles des corpus français et italien. D'autre part, l'annotation de chaines de référence requiert des annotateurs disposant à la fois d'une compétence linguistique avancée dans la langue concernée et d'une formation spécifique aux principes de l'annotation en linguistique de

corpus. Or, de tels profils n'étaient pas disponibles pour l'espagnol au moment de la réalisation de l'étude.

9. Corpus longitudinaux français et italien : caractérisation

9.1. Caractérisation du corpus longitudinal français

Le corpus *Scolinter* comprend ainsi la partie longitudinale du corpus *Scoledit* (cf. ce chapitre, §2). L'échantillon est composé de 337 élèves, dont 57,27% sont des filles et 42,73% des garçons ; ces élèves proviennent majoritairement de la région Auvergne Rhône Alpes (98,22%), avec une très faible représentation de la région Occitanie (1.78%). Le Tableau 14 précise la distribution des élèves selon le type d'établissement scolaire fréquenté. La majorité des participants provient d'écoles publiques dites « mixtes » (63,5 %), c'est-à-dire d'établissements ne relevant pas d'un dispositif spécifique d'éducation et accueillant un public socialement varié. Une part importante des élèves est également issue d'écoles situées en contexte défavorisé (17,21 %), tandis que les établissements favorisés représentent seulement 8,31 % de l'échantillon.

Les autres élèves sont scolarisés dans des écoles relevant de dispositifs d'éducation prioritaire (EP), qui regroupent des établissements accueillant des populations socialement plus vulnérables et bénéficiant de moyens pédagogiques supplémentaires. Parmi ces dispositifs, 7,72 % des élèves proviennent d'écoles classées REP (réseaux d'éducation prioritaire), 2,08 % d'écoles REP+ (écoles dont les élèves se retrouvent dans une situation davantage défavorable par rapport aux élèves en REP), et 1,19 % d'écoles identifiées plus largement comme relevant de l'éducation prioritaire.

Type d'école	% élèves
école publique mixte	63,50
école publique défavorisée	17,21
école publique favorisée	8,31
école publique EP – REP	7,72
école publique EP – REP+	2,08
école publique EP	1,19

Tableau 14. Distribution des élèves par école d'appartenance dans le corpus longitudinal.

Le Tableau 15 précise la distribution des textes par niveau et par longueur en termes de tokens dans chaque niveau. Il est composé de 1685 textes, 337 par niveau scolaire du CP au CM2. Au total, ces 1 685 textes comportent 178 183 tokens, avec une longueur moyenne de 105,75 tokens par texte. La longueur moyenne des textes semble augmenter de manière progressive avec le niveau scolaire, bien qu'un écart plus marqué soit observé entre le CE1 et le CE2, deux niveaux intermédiaires retenus pour nos annotations.

Niveaux scolaires	Année scolaire	Nombre de textes	Nombre total de tokens	Nombre moyen de tokens	Étendue
CP – 1	2014	337	8 931	26,53	3-65
CE1 – 2	2015	337	22 241	66	6-212
CE2 – 3	2016	337	39 582	117,45	6-349
CM1 – 4	2017	337	48 357	143,49	25-398
CM2 – 5	2018	337	59 072	175,28	19-591
CP-CM2 – 1-5	**2014-2018**	**1 685**	**178 313**	**105,75**	**3-591**

Tableau 15. Caractérisation des textes du corpus français longitudinal.

9.2. Caractérisation du corpus longitudinal italien

Comme décrit dans le chapitre 2, le recueil du corpus italien s'est achevé en 2022, avec un recueil parallèle des niveaux scolaires 1e et 2e (CE1 et CE2) puis 2e et 3e (CE1 et CE2) réalisé respectivement en 2018 et 2019 (voir ce chapitre, §8.2, pour le détail du recueil de données). La phase de traitement des textes de niveau 5 (CM2) est en cours, et la période de fermeture des écoles due à la crise sanitaire de Covid-19 n'a pas permis de collecter des données longitudinales de 3e et 4e (CE2 et CM1) lors des années scolaires 2019-2020 et 2020-2021. Ici nous décrivons les données relatives aux textes dont la phase de traitement a déjà été réalisée (scan, transcription et normalisation des textes, phases décrites en ce chapitre, §7).

Niveaux scolaires	Année scolaire	Nombre de textes	Nombre total de tokens	Nombre moyen de tokens	Étendue
CP – 1	2018	702	27 505	39,18	2-165
CE1 – 2	2018-2019	672	72 182	107,41	12-552
CE2 – 3	2019	96	15 812	164,71	42-451
CM2 – 5	*2022*	*19*	*4 374*	*257,29*	*134-394*
CP-CM2 – 1-5	**-**	**1 487**	**119 873**	**80,61**	**2-552**

Tableau 16. Caractérisation des textes du corpus italien longitudinal.

Le Tableau 16 décrit la distribution des textes du corpus italien longitudinal selon le niveau scolaire, l'année de recueil et détaille les longueurs en termes de tokens.

Étant donné que seulement les textes de 1re, 2e (CP et CE1) et un petit échantillon des textes de 5e (CM2) étaient déjà transcrits et normalisés par l'équipe des chercheuses italienne lors du début de notre travail de recherche, nous avons réalisé les étapes de traitement du corpus sur la totalité des 96 textes de niveau 3 (CE2)[68] pour pouvoir ensuite annoter un échantillon similaire à celui choisi pour le français (sur les niveaux CE1 et CE2). À ce jour, 1 487 textes ont ainsi été traités, dont 702 de 1re (CP), 672 de 2e (CE1), 96 de 3e (CE2) et 17 de 5e (CM2), les niveaux CE2 et CM1 étant encore en phase de traitement (les statistiques ont été incluses à titre informatif car pas encore représentatives de la totalité de ces deux niveaux). Sur un total de 1 487 textes disponibles pour 119 873 tokens, la longueur moyenne des textes en italien est alors de 80,61 tokens. Il faut noter

[68] L'étape de transcription de la moitié de ces textes a été réalisée par une étudiante en M1 Sciences du Langage, supervisée par la doctorante autrice de ce manuscrit. Les deux sont de langue maternelle italienne et formées à l'enseignement de l'italien. Les étapes de scan et normalisation ont été réalisées par la doctorante.

ici que, comme dans le cas du corpus français, nous pouvons observer une augmentation sensible de la longueur des textes en termes de tokens au fur et à mesure des années. Cela est visible dans la Figure 31. La Figure 32 compare les longueurs moyennes des textes dans ces deux corpus longitudinaux, français et italien, aux niveaux 2e et 3e (CE1 et CE2). Ici nous pouvons remarquer la différence entre les deux niveaux scolaires ainsi que pour chaque langue, les textes italiens, montrant une longueur moyenne plus importante que les textes français.[69]

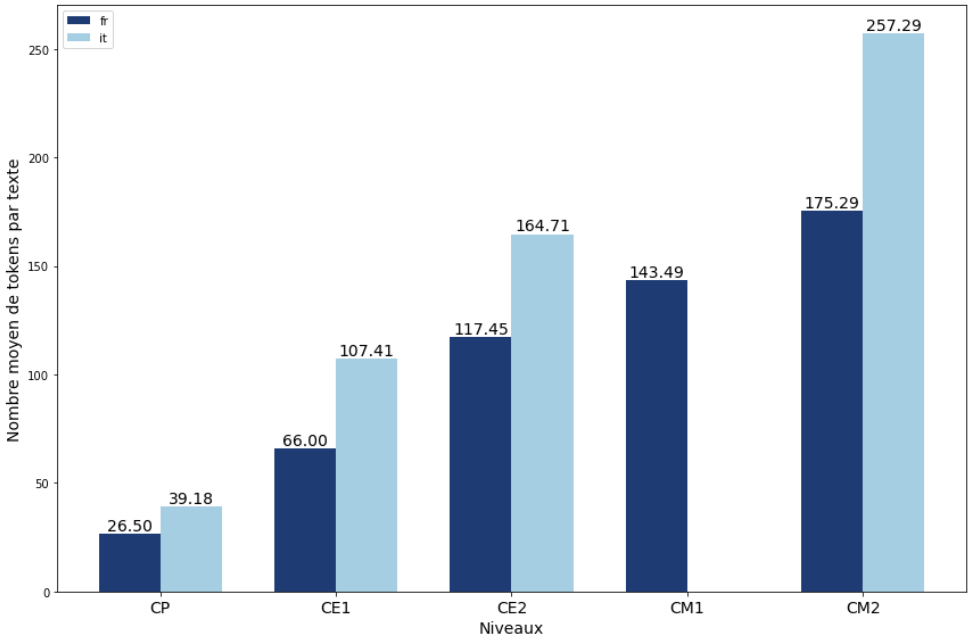


Figure 31. Nombre moyen de tokens par texte – comparaison entre corpus longitudinaux français et italien.

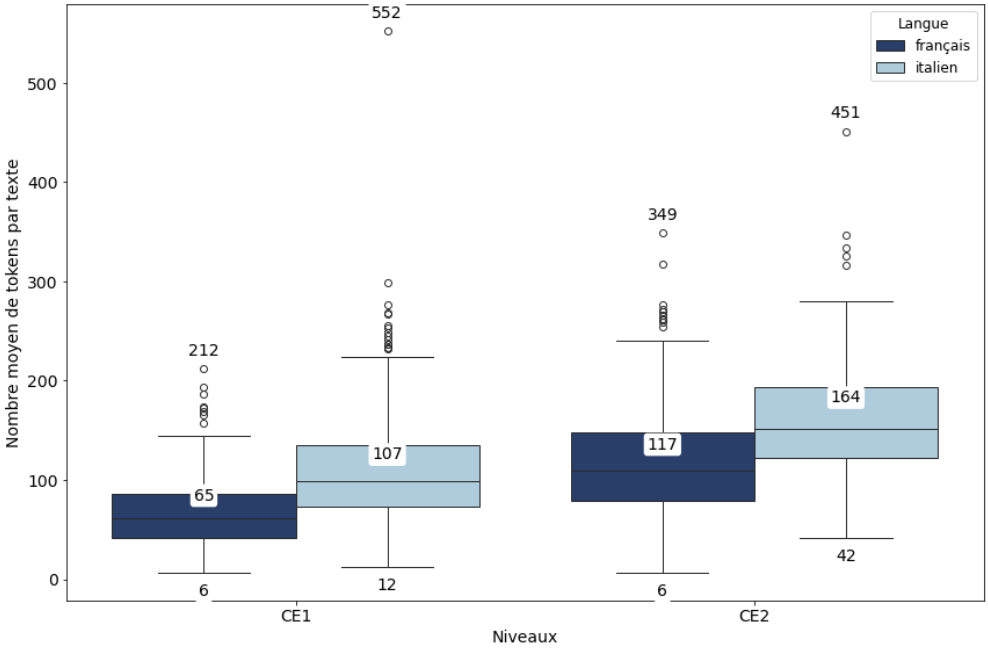


Figure 32. Longueur des textes dans le corpus longitudinal en français et en italien. Comparaison effectuée entre CE1 et CE2.

[69] Nous avançons des hypothèses empiriques qui puissent expliquer cette variation de longueur entre langues dans le Chapitre 8, §4.

Chapitre 5. Annotation de corpus

Dès le début de ce projet de thèse, nous avons été confrontée à des contraintes de temps et de ressources humaines disponibles, inhérentes à tout projet d'annotation de corpus. Ce sont ces contraintes qui ont orienté certains choix méthodologiques, notamment dans la manière de conduire le processus d'annotation et dans la sélection de la sous-partie du corpus à annoter.

Ce projet étant pour le moment restreint au cadre du travail doctoral ici présenté, la contrainte temporelle est intrinsèquement liée au temps d'une thèse de doctorat et aux objectifs fixés initialement, c'est-à-dire d'achever la plupart du travail d'annotation et d'analyse au cours des trois ans de financement doctoral, ou du moins de constituer une ressource annotée bien que limitée en taille à la fin de cette période. En termes de ressources humaines, nous avons pu profiter des financements dédiés aux vacations de recherche par le laboratoire LIDILEM. Dans le chapitre suivant (cf. Chapitre 6), nous allons décrire plus en détail les actions menées grâce à ce soutien et comment elles ont eu un impact sur le processus d'annotation. Ces ressources nous ont ainsi permis de réaliser les annotations manuelles sur une partie du corpus français et italien, selon la méthodologie que nous allons présenter dans la suite de ce chapitre[70].

Dans le prolongement direct de ces contraintes méthodologiques et organisationnelles, se pose la question de la représentativité du corpus annoté. En effet, la représentativité est une des caractéristiques essentielles à atteindre dans la constitution d'un corpus, à partir de sa conception jusqu'à la codification des phénomènes à représenter. La phase de conception du corpus « comprend la sélection d'un échantillon représentatif de locuteurs et de matériels appropriés pour la récolte de données » [71] (Voormann & Gut, 2008, p. 237). Cette exigence s'applique également à la sélection d'une sous-partie du corpus à annoter.

Or, la tâche d'annotation de corpus étant assez chronophage, la nécessité de limiter la taille des échantillons à annoter manuellement se heurte à la volonté de créer des ressources les plus représentatives possibles du corpus de départ, et dont les annotations soient aussi valides, fiables et cohérentes avec les objectifs de recherche visés. Lorsque la délimitation et l'annotation d'un corpus doivent être réalisées dans un contexte contraint en temps et en ressources, il devient donc nécessaire de s'interroger sur la manière de préserver cette représentativité de l'ensemble du corpus, sachant que le corpus de départ constitue déjà une sélection limitée de la réalité de la diversité linguistique recherchée. Comment peut-on garantir cette représentativité de notre sous-échantillon vis-à-vis du phénomène que l'on souhaite décrire ?

Il n'est pas possible d'apporter une réponse univoque à cette question : les choix à effectuer dans ce cas-là dépendent des objectifs de recherche et des besoins auxquels l'annotation doit répondre. Nous pouvons néanmoins détailler les critères que nous avons adoptés pour répondre à cette

[70] Au moment de l'écriture de cette thèse, des ressources ultérieures ont été dédiées à l'annotation d'une partie du corpus espagnol, qui est en cours.

[71] Dans le texte d'origine, « (…) involves the selection of a representative sample of speakers and appropriate materials for the data collection ».

question dans notre travail, pour garantir une certaine représentativité par rapport à notre objectif de recherche.

Dans ce chapitre, nous présentons tout d'abord, de manière synthétique, les prétraitements appliqués aux textes du corpus afin de préparer l'étape d'annotation. La section 1 est ainsi consacrée à la description de la chaine de traitement mise en place, en particulier aux choix effectués pour pallier l'absence de mots dans certaines phrases du corpus. Nous présentons ensuite dans la section 2 les principes de l'annotation agile, ainsi que la manière dont cette approche, inspirée des méthodes de développement logiciel et largement répandue aujourd'hui, a été mise en œuvre dans notre travail. Enfin, nous abordons dans la section 3 le guide d'annotation élaboré pour nos campagnes : les phénomènes qu'il cible, ses modalités d'utilisation, ses limites, ainsi que le rôle de l'adjudication dans la constitution d'un corpus gold, illustré dans la section 4.

1. Objectifs et description de la chaine de prétraitement

Si la normalisation des textes représente une interface entre la transcription et les traitements automatiques, la présence des éléments introduits dans la normalisation, tels que les balises XML, utilisées pour signaler des phénomènes variés, posait problème lors de l'étape d'annotation et de *parsing* des textes. Deux approches différentes étaient pouvaient être envisagées afin d'obtenir une sortie textuelle plus facilement exploitable pour l'annotation :

- constituer une version parsée des textes dans un format interopérable (CoNLL-U) qui prenait déjà en compte la position des balises présentes dans la normalisation et la reporte sur le ou les tokens correspondants ;
- produire une version des textes épurée de toute balise et au format textuel simple (extension txt) pour ensuite effectuer le *parsing* sur cette version et récupérer dans un second temps les informations issues de la normalisation dans le format d'annotation obtenu.

Dans ce contexte, la deuxième approche a été retenue, afin de disposer rapidement de versions de textes directement exploitables pour l'annotation, tout en conservant la possibilité de réintroduire ultérieurement les phénomènes signalés lors de la normalisation dans version annotée.

La chaine de traitement adoptée prenait ainsi en entrée les textes normalisés au format XML et donnait en sortie un fichier par texte au format txt. À partir des exports au format XML issus de la plateforme *Scolinter*, nous avons extrait la partie de fichier contenant la production écrite, en excluant les métadonnées et les balises XML. Une étape de pré-traitement a ensuite été appliquée afin de traiter ces balises (cf. Chapitre 4, §7.3), dont la présence entrainait des erreurs lors du *parsing* automatique. Au même temps, leur exclusion du *parsing* aurait entrainé du silence vis-à-vis des éléments que ces balises encodaient, notamment la ponctuation des textes qu'elles représentent.

Ce prétraitement consiste en une série d'opérations de remplacement, effectuées sur ces textes selon la table de conversion présentée dans le Tableau 17. Les balises montrées correspondent à celles générées par la plateforme qui héberge les textes lors de l'export au format XML ; pour cette

raison elles diffèrent légèrement de celles décrites dans la section consacrée à la normalisation des textes. Comme on peut l'observer dans le Tableau 17, les balises concernées portent principalement sur la ponctuation des textes, qui a été ainsi reconstituée : les balises de segmentation forte ont été remplacées par des points, celles de segmentation faible par des virgules, et les balises indiquant la présence d'un dialogue par des guillemets français, lorsque ceux-ci n'étaient pas déjà présents dans le texte. D'autres balises, jugées non pertinentes pour l'analyse, ont été supprimées. À terme, ces informations issues de la normalisation ont vocation à être réintégrées dans le format d'annotation final, afin de permettre la reconstruction des éléments introduits de manière « artificielle » dans les productions écrites des élèves.

Balise XML normalisation	**Balise XML export**	**Élément de remplacement**
<s/>	<seg type="segm. faible"/>	,
<segmentation/>	<seg type="segm. forte"/>	.
<dialogue>	<u>	«
</dialogue>	</u>	»
//	</*p>	espace
<unsure> </unsure> ????	<gap/>	éliminée
</unclear>	</unclear>	éliminée
<titre> </titre>	</*head>	éliminée
Omission	<omission type="([a-zA-Z])*"/>	Voir ce chapitre, §1.1, pour le traitement des omissions

Tableau 17. Correspondances entre balises XML présentes dans les normalisations (côté transcripteur), leur équivalent dans le fichier exporté et l'élément utilisé pour remplacer chaque balise dans notre chaine de prétraitement.

En ce qui concerne le prétraitement des balises d'omission, elles ont été traitées automatiquement à l'aide des grands modèles de langage, qui ont remplacé la balise par le mot le plus probable selon son contexte lors d'une expérimentation, que nous décrirons dans la section suivante (cf. 1.1).

1.1. Étude préliminaire. Le traitement des omissions dans le corpus gold

La présente section s'appuie largement sur des analyses que nous avons précédemment publiées dans (Barletta & Ponton, 2025b). Au cours des premières phases du processus d'annotation, la question du traitement des balises d'omission dans notre corpus longitudinal s'est posée. Dans la mesure où l'annotation est effectuée sur les versions normalisées des textes[72] (cf. chapitre précédent, §7.3 pour la définition de la normalisation), ces dernières comportent parfois des éléments nécessitant une prise de décision ultérieure afin d'aboutir à une version totalement normalisée. Comme discuté dans la section dédiée à la normalisation des textes du corpus, ces versions de nos textes proches de la norme comprennent l'insertion d'éléments absents dans le texte d'origine mais qui sont nécessaires pour reconstruire sa signification. L'utilisation des balises XML permet ainsi de signaler des omissions de ponctuation, de marques de dialogue ou encore des mots oubliés par l'élève, comme dans l'exemple suivant (63). Le scan de cette phrase dans le texte d'origine est visible dans la Figure 33.

[72] Pour rappel, la normalisation porte essentiellement sur les niveaux orthographique et morphologique des textes.

(63) Il étais une fois un robos étais seul au monde
(transcription, corpus français, CE1, 1350)
Il était une fois un robot **<omission type='pronom'/>** était seul au monde
(normalisation, corpus français, CE1, 1350)

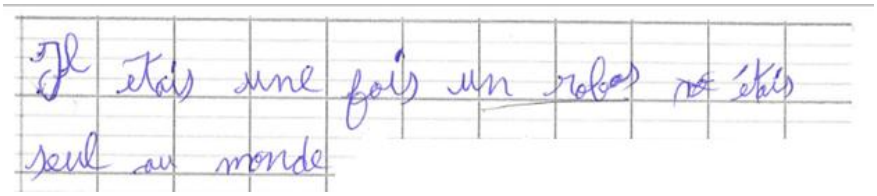


Figure 33. Scan d'un extrait de texte contenant une omission (corpus français, CE1, 1350).

Dans le corpus, les mots omis par les élèves sont signalés au moyen d'une balise *<omission>*, qui indique la catégorie grammaticale la plus probable de l'élément manquant, telle qu'elle peut être inférée à partir du contexte de la phrase ou du texte. Cette catégorisation est réalisée manuellement par les chercheurs-linguistes du projet lors de la constitution du corpus et encodée dans l'attribut de la balise insérée dans le texte.

Ce marquage des omissions implique un choix méthodologique lorsque les textes sont exploités pour des analyses ultérieures : soit ignorer ces éléments omis et effectuer des traitements TAL sur le texte en l'état, soit remplacer la balise par un token probable pour que le traitement n'échoue pas. Ces choix sont déterminants car les erreurs de traitement peuvent ensuite avoir des répercussions sur les traitements de plus haut niveau, comme l'annotation et l'analyse de la continuité référentielle qui sont les sujets de notre travail de recherche (Barletta, 2024; Barletta & Ponton, 2025a). Ces erreurs peuvent être de deux types : des erreurs de bruit (mention détectée qui n'en est pas une) ou des erreurs de silence (mention non détectée). Pour réduire les erreurs provoquées par la présence de ces balises dans notre corpus, nous avons conçu une méthodologie visant à traiter automatiquement ces balises d'omission dans le corpus longitudinal.

Pour mener ces expérimentations, nous avons extrait un sous-corpus de 54 textes à partir des 1 684 textes du corpus longitudinal français (un texte composé d'un seul token ayant été exclu des 1 685 initialement disponibles). Ce sous-corpus, couvrant les niveaux CE1 à CM2, comprend uniquement des textes contenant au moins une balise d'omission. Afin de conserver une répartition comparable à celle du corpus initial en termes de longueur des textes, nous avons partitionné les données en sept tranches et sélectionné un nombre de textes proportionnel à la distribution observée (cf. Tableau 18). Notons que la tranche 0-20 tokens n'a pas été prise en compte car elle propose des textes trop courts et souvent peu intéressants pour notre étude.

Nombre de tokens	Corpus longitudinal		Corpus *gold*	
	Nb de textes	% de textes	Nb de textes	% de textes
0-20	126	7,48 %	0	0
21-50	372	22,09 %	6	11,11 %
51-100	414	24,58 %	17	31,48 %
101-150	343	20,37 %	12	22,22 %
151-200	237	14,07 %	6	11,11 %
201-250	111	6,60 %	5	9,26 %
251-plus	81	4,81 %	8	14,81 %
Tot.	1 684	100 %	54	100 %

Tableau 18. Comparaison entre distribution sur le corpus longitudinal et le sous-corpus *gold* de test sélectionné pour l'expérimentation.

Dans ce sous-corpus de textes normalisés, les balises d'omission ont été remplacées manuellement par le token jugé le plus probable dans le contexte du texte, tout en respectant la catégorie grammaticale proposée lors de l'étape de normalisation et présente en attribut de la balise d'omission. Par exemple, dans l'exemple de la Figure 33, la catégorie proposée était celle de pronom, et nous pouvons faire l'hypothèse que le token omis le plus probable pour reconstruire le sens de la phrase dans laquelle il est inséré soit le pronom relatif « qui », qui a ainsi été retenu comme candidat le plus plausible pour reconstruire la structure de la phrase. L'annotation de ces omissions a été ensuite discutée et validée avec les chercheurs-linguistes du projet. Aucun de ces choix n'a suscité de désaccord mais, dans le cadre d'une expérimentation plus large, il serait certainement nécessaire de procéder à un calcul d'accord inter-annotateur.

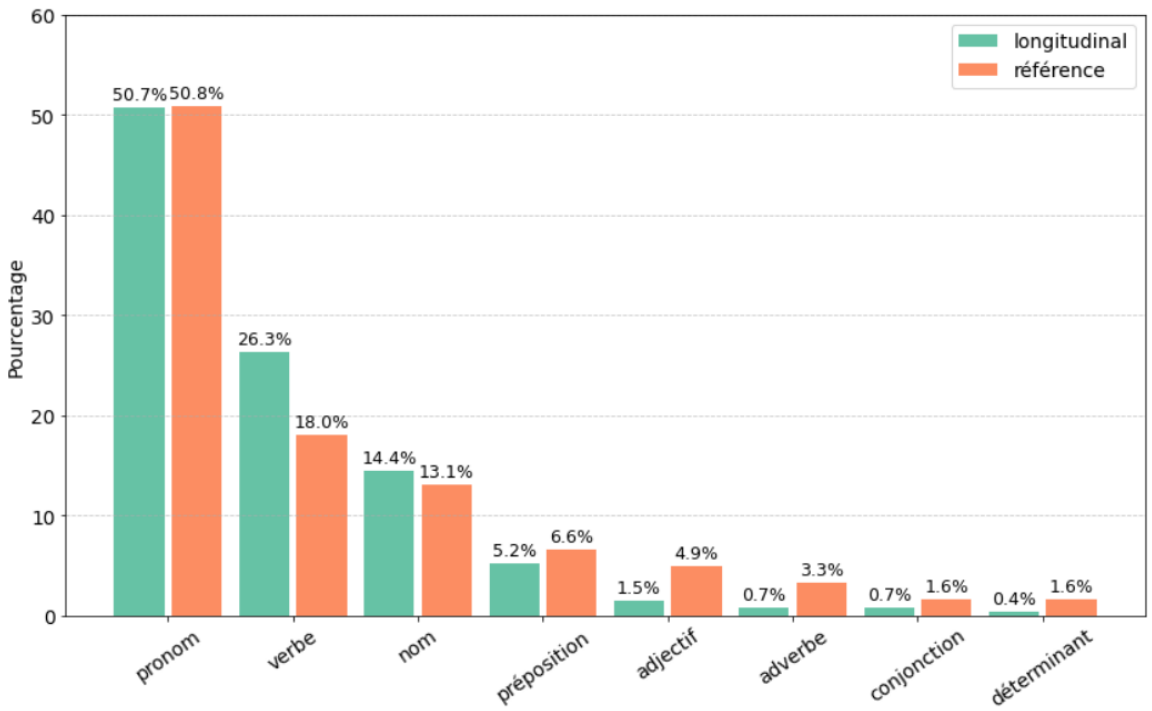


Figure 34. Distribution des omissions par étiquette syntaxique -comparaison entre corpus longitudinal et corpus gold (ou de référence).

Enfin nous avons comparé la distribution des omissions par catégorie syntaxique (POS ou *part of speech*) entre les 274 omissions du corpus longitudinal et les 61 occurrences du corpus de référence. Les catégories utilisées pour ces analyses correspondent à celles indiquées en attribut des balises d'omission. Les résultats de cette comparaison sont présentés dans la Figure 34.

La distribution des omissions par catégorie syntaxique est globalement similaire entre les deux corpus sur les catégories les plus fréquentes, ce qui permet de considérer le sous-corpus de référence comme suffisamment représentatif de notre corpus complet. L'analyse des distributions par POS dans les corpus longitudinal et de référence montre que la catégorie la plus fréquemment omise est celle du pronom. Les pronoms représentent ici plus de 50 % des tokens omis, et selon des observations empiriques, l'essentiel de ces pronoms omis correspond à des pronoms relatifs. Cette catégorie est suivie par les verbes (26,3 % dans le corpus longitudinal, qui descend à 18 % dans celui de référence) et les noms (14,4 % dans le longitudinal et 13,1 % dans le corpus de référence), alors que d'autres catégories comme prépositions, adjectifs, adverbes, conjonctions et déterminants représentent au total 8,5 % des omissions dans le corpus longitudinal et 18 % dans le corpus de référence. À noter que lors de la normalisation, les chercheurs ont décidé de regrouper les deux catégories d'auxiliaire et de verbe sous la même étiquette.

Si des éléments tels que les prépositions peuvent être peu pertinents pour notre recherche, d'autres éléments tels que les pronoms et les noms sont susceptibles d'être annotés comme partie des chaines de continuité référentielle[73], comme dans le cas de l'exemple suivant (cf. Figure 35). Dans cette phrase, l'absence du pronom relatif « qui » impacte la structure de la phrase qui le contient. Les verbes font aussi partie des tokens qui peuvent porter une marque de présence du référent dans le texte et pourraient donc être susceptibles d'être annotés. Par exemple dans le cas de verbes qui se trouvent dans des phrases coordonnées où le sujet n'est pas répété[74]. Ces observations préliminaires confirment la nécessité de traiter au mieux ces tokens absents pour assurer la réussite de notre travail d'analyse du corpus *Scolinter* à travers l'annotation et l'analyse des chaines de continuité référentielle.

[73] En particulier, dans notre travail nous allons annoter les maillons des chaines faisant référence aux personnages animés dans les écrits des élèves, ce qui délimite la tâche d'annotation, mais qui n'exclut pas le fait que ces omissions concernent très probablement des personnages à annoter.

[74] Ou de manière plus fréquente, dans les textes des langues où le sujet explicite peut être omis comme l'italien et l'espagnol, langues présentes dans notre corpus.

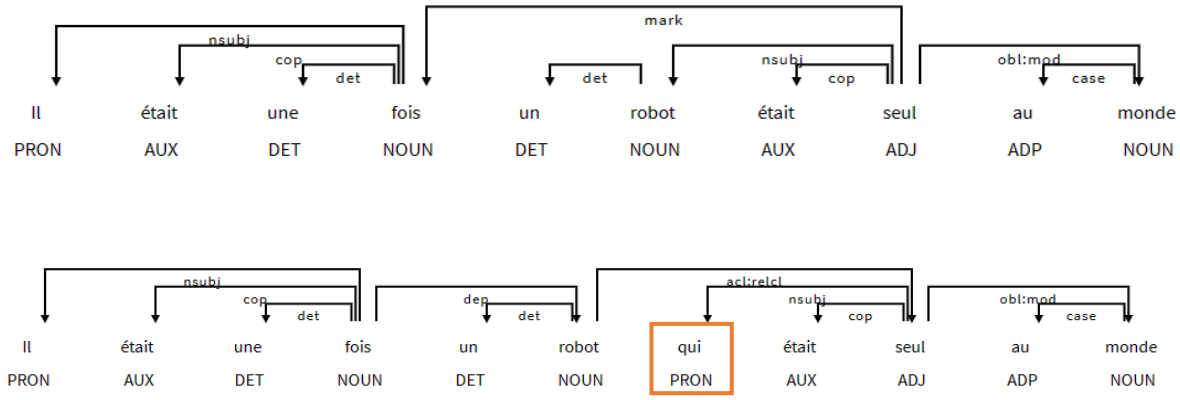


Figure 35. Analyse en dépendance d'une phrase tirée du texte CE1, S1350, générée par Spacy (v.3) en utilisant le modèle fr_core_news_lg. En haut, la phrase sans le mot omis, en bas la phrase avec le token « qui » réintroduit (entouré dans la représentation).

1.2. Méthodologie

D'un point de vue méthodologique, nous comparons trois méthodes de traitement automatique afin d'identifier la plus pertinente en termes de temps et de précision, en nous appuyant sur le corpus de référence élaboré manuellement par trois chercheurs linguistes du projet. Dans ce corpus, les mots omis ont été remplacés par le mot candidat jugé le plus probable en contexte. Ces trois méthodes ont été appliquées au sous-corpus de référence, puis évaluées à l'aide des métriques couramment employées pour l'analyse des performances des parseurs en dépendance. L'impact de chaque méthode est ainsi mesuré à partir des analyses produites par le parseur morphosyntaxique en dépendance SpaCy v.3 (Honnibal et al., 2020).

La première méthode envisagée consiste à remplacer chaque balise *<omission>* par un token « masque » identique, en l'occurrence le token de forme xxx. Ce token est habituellement utilisé en TAL en tant que token « masque » dans d'autres applications[75]. La deuxième méthode repose sur un remplacement de chaque balise *<omission>* par un token choisi *a priori* pour chaque catégorie grammaticale indiquée dans la balise.

La deuxième méthode, dite « prototypique », consiste à remplacer chaque catégorie morphosyntaxique indiquée dans les balises d'omission par un même token, que nous avons qualifié de « prototypique ». Ce token est défini comme un mot représentatif de la catégorie concernée, sélectionné selon plusieurs critères visant à minimiser son impact sur l'analyse morphosyntaxique : une appartenance catégorielle non ambiguë, afin d'éviter toute variation d'étiquetage possible ; une fréquence d'usage relativement élevée, permettant de limiter les effets liés à la rareté lexicale ; une forme morphologique simple et non marquée, afin de ne pas introduire de traits grammaticaux supplémentaires susceptibles de biaiser l'analyse. Certes, tout token est susceptible d'influencer la sémantique du texte ; néanmoins, dans la mesure où notre réflexion porte principalement sur les effets au niveau morphosyntaxique, nous avons volontairement restreint notre analyse à cet aspect. Il s'agit d'une méthode semi-automatisée : à l'issue d'une

[75] Par exemple, dans l'anonymisation des données personnelles dans les corpus (Kodandaram et al., 2021).

phase de concertation entre annotateurs experts, visant à sélectionner des tokens répondant aux critères précédemment définis, nous avons établi une liste de correspondances (entre catégorie grammaticale et token), utilisée ensuite pour le remplacement automatique des balises (cf. Tableau 19). La fréquence de ces tokens dans le corpus était prise en compte pour effectuer ces choix.

Catégorie	Token
conjonction	et
adverbe	vraiment
adjectif	méchant
préposition	avec
nom	maison
verbe	mange
pronom	qui

Tableau 19. Tokens choisis pour remplacer chaque catégorie morphosyntaxique proposée lors de la normalisation.

La troisième méthode exploite un modèle de langage de type Transformeur, FlauBERT (Le et al., 2020), afin de prédire automatiquement le mot le plus probable pour chaque omission (un token par balise). Nous avons utilisé le modèle FlauBERT pré-entrainé (FlaubertLMHeadModel) et son tokeniseur sur les textes du corpus. Chaque balise *<omission>* est remplacée par un mot masque propre au tokeniseur, qui permet de prévoir la meilleure substitution possible, option qui est ensuite intégrée dans le texte à la place de la balise *<omission>*. Du point de vue technique, nos textes ont fait l'objet de plusieurs traitements, dont une transformation du format XML d'origine au format txt, avec les balises d'omission remplacées selon le scénario concerné.

Tous les textes ont été ensuite parsés au format CoNLL-U en utilisant le parseur *Spacy v. 3* (Honnibal et al., 2020) et le modèle pour le français *fr_core_news_lg*.[76] Dans le Tableau 20, nous comparons un extrait issu du corpus de référence avec les résultats obtenus après l'application de chaque méthode.

Gold	Méthode xxx	Méthode prototypique	Méthode FlauBERT
(…) Toma va montrer à l'entraîneur de l'équipe de badminton. Il lui **<omission type="verbe" prop="dit"/>** « tu as 3 min pour me montrer, tu vas jouer contre Tom » (…)	(…) Toma va montrer à l'entraîneur de l'équipe de badminton. Il lui **xxx** « tu as 3 min pour me montrer, tu vas jouer contre Tom » à la fin du match Toma met le dernier point et remporte le match. (…)	(…) Toma va montrer à l'entraîneur de l'équipe de badminton. Il lui **mange** « tu as 3 min pour me montrer, tu vas jouer contre Tom » à la fin du match Toma met le dernier point et remporte le match. (…)	(…) Toma va montrer à l'entraîneur de l'équipe de badminton. Il lui **dit** « tu as 3 min pour me montrer, tu vas jouer contre Tom » à la fin du match Toma met le dernier point et remporte le match. (…)

Tableau 20. Extrait du texte CE1, 2449 dans sa *version* gold et dans les résultats obtenus à travers les trois méthodes testées.

[76] Nous avons choisi d'adopter la librairie Spacy pour différentes raisons. Au début de ce travail de recherche (2023), nous avons comparé les performances de Spacy et de Stanza sur le français et sur l'italien. Selon nos observations empiriques, Spacy obtenait de meilleures performances sur la langue italienne. De plus, Spacy avait le mérite de ne pas séparer les articles contractés en deux tokens (*du = de + le*), contrairement à Stanza, qui tokenisait ainsi les amalgames. Cela rajoutait une complexité supplémentaire lors de l'évaluation manuelle des sorties des parseurs, qui consistait à comparer une tokenisation et une annotation en POS effectuées manuellement avec la sortie du tokeniseur automatique.

Afin d'évaluer la capacité des différentes méthodes à choisir un token afin d'obtenir une analyse en dépendance aussi proche que possible de celle d'une référence qui inclut un token remplaçant l'omission choisie selon son contexte, nous avons opté pour l'utilisation des métriques couramment employées dans la littérature pour évaluer les analyseurs en dépendance (Jurafsky & Martin, 2024, pp. 431-432) à savoir UAS, LAS, LS et la précision sur la partie du discours. Ces métriques ont été ensuite appliquées uniquement aux phrases contenant des tokens omis. Le LAS (*Labeled Attachment Accuracy*) fait référence à l'assignation correcte d'un mot à sa tête ainsi qu'avec la relation de dépendance correcte, alors que le UAS (*Unlabeled Attachment Score*) considère seulement le rattachement à la tête (Jurafsky & Martin, 2024, pp. 431-432). Le LS (*Labeled accuracy Score*) considère le nombre de tokens auxquels a été assignée l'étiquette en dépendance correcte. La précision correspond ici au nombre de tokens omis dont le POS est identique à celui prévu dans l'annotation de référence.

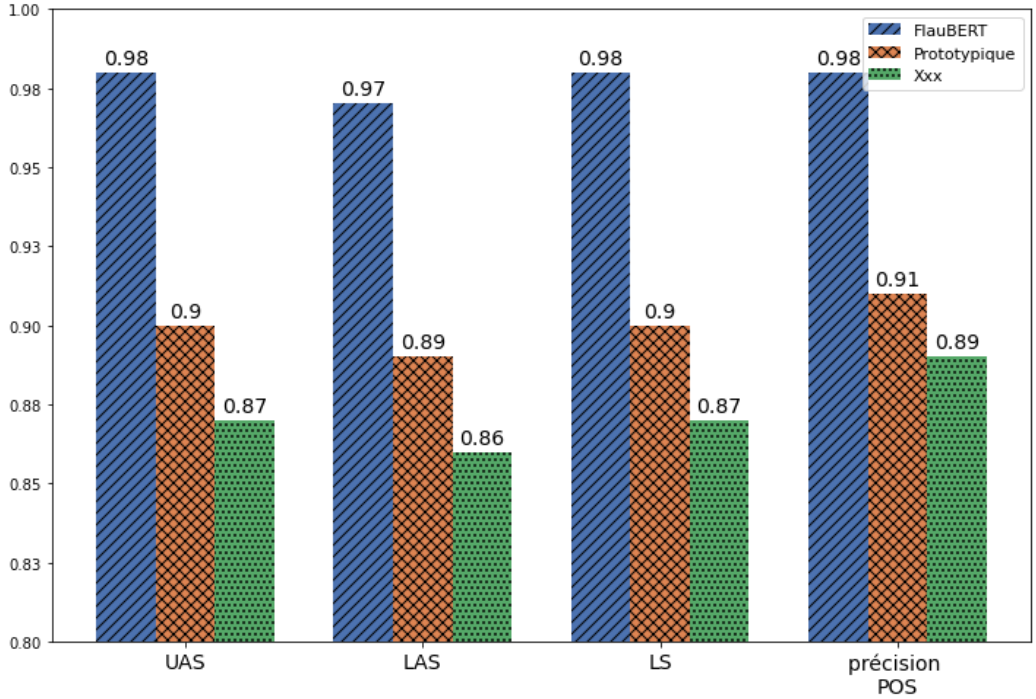


Figure 36. Métriques UAS, LAS, LS et précision sur le POS calculé en comparaison avec les 54 textes du *gold*.

1.3. Résultats obtenus

Il convient de préciser que cette évaluation ne porte que sur les tokens omis, et non sur le parsing de l'intégralité des textes.

Tout en ayant exploré l'application de méthodes semi-automatisées comme celle dite "prototypique" ou totalement automatisées comme la méthode "xxx", celles-là ont obtenu des scores moins satisfaisants par rapport à la méthode qui exploite le Transformeur. Cette méthode a atteint des résultats de 0.98 pour UAS, le LS et la précision sur POS, ainsi que de 0.97 pour le LAS, comme visible dans la Figure 36 ci-dessus.

Notre analyse qualitative se focalisera principalement sur la méthode utilisant FlauBERT, étant donné que les autres alternatives présentent des résultats plus faibles. En analysant plus en détail les données obtenues, nous avons pu observer des résultats très encourageants surtout sur le

remplacement des pronoms. Parmi les 35 textes qui contiennent des balises de pronoms omis, 49 % ont obtenu le score de 1 aux quatre métriques. Cela peut être expliqué par le fait que les pronoms relatifs représentent 51 % des pronoms omis. Ces pronoms relatifs ont été correctement reconstruits par rapport à leur POS et leur rôle dans le texte à 100 % des cas par le modèle FlauBERT[77].

Gold	**FlauBERT**
Il était une fois une sorcière **qui** était chez elle et elle s'est dit qu'elle n'avait rien à faire donc elle est partie dans sa salle magique (…)	Il était une fois une sorcière **qui** était chez elle et elle s'est dit qu'elle n'avait rien à faire donc elle est partie dans sa salle magique (…)

Tableau 21. Versions *gold* et issue de la méthode FlauBERT du même texte (CM1, 3012, corpus français).

Dans certains cas, le modèle n'a pas proposé de tokens cohérents avec le contexte de la phrase analysée, donnant lieu à des phrases « incomplètes » dont la structure reste déviante à la norme. C'est le cas de l'exemple ci-dessous, où à la place du verbe qui avait été proposé, le modèle FlauBERT a proposé le token « tout » en fonction d'adverbe (erronément catégorisé comme adjectif dans l'analyse en dépendance obtenue en sortie).

Gold	**FlauBERT**
Un chat **était** seul mais un jour il rencontra un loup alors il courut puis se reperdit.	Un chat **tout** seul mais un jour il rencontra un loup alors il courut puis se reperdit.

Tableau 22. Versions *gold* et issue de la méthode FlauBERT du même texte (corpus français, CE1, 2984).

1	Un	DET	2	det	1	Un	DET	2	det
2	chat	NOUN	4	nsubj	2	chat	NOUN	9	nsubj
3	était	AUX	4	cop	3	tout	ADJ	4	advmod
4	seul	ADJ	14	advmod	4	seul	ADJ	2	amod
5	mais	CCONJ	9	cc	5	mais	CCONJ	9	cc
6	un	DET	7	det	6	un	DET	7	det
7	jour	NOUN	9	nsubj	7	jour	NOUN	9	nsubj
8	il	PRON	9	nsubj	8	il	PRON	9	nsubj
9	rencontra	VERB	4	conj	9	rencontra	VERB	0	ROOT
10	un	DET	11	det	10	un	DET	11	det
11	loup	NOUN	9	obj	11	loup	NOUN	9	obj
12	alors	ADV	14	advmod	12	alors	ADV	14	advmod
13	il	PRON	14	nsubj	13	il	PRON	14	nsubj
14	courut	VERB	0	ROOT	14	courut	VERB	9	advcl
15	puis	CCONJ	17	cc	15	puis	CCONJ	17	cc
16	se	PRON	17	expl:pass	16	se	PRON	17	expl:pass
17	reperdit	PROPN	14	conj	17	reperdit	PROPN	9	conj
18	.	PUNCT	14	punct	18	.	PUNCT	9	punct

Figure 37. Analyse en dépendance des deux versions effectué automatiquement à l'aide de SpaCy v. 3 et du modèle fr_core_news_lg

Certains cas restent néanmoins difficiles à traiter, de manière manuelle comme à l'aide d'un modèle de langage. Il s'agit surtout des cas où le choix du terme à insérer est sujet à diverses interprétations, à cause de la nature fortement ambiguë du texte traité ou du contexte. Par exemple, dans le texte suivant, la proposition faite par les chercheurs ne correspond pas à celle

[77] Le seul cas de texte qui n'a pas obtenu des métriques à 1 sur la totalité du texte tout en contenant une omission de type pronom relatif est dû à la présence d'une deuxième balise d'omission dans le texte c'est-à-dire d'une deuxième omission présente dans une autre phrase du texte.

faite par le modèle. Le choix du token dans la référence peut être questionnable, notamment en raison de l'incomplétude de la phrase, qui contient seulement un syntagme nominal en contexte gauche. La phrase semble ensuite interrompue par la présence d'une conjonction de coordination, « mais ». La relative pauvreté du contexte ici ne permet donc pas de reconstruire un token omis assez probable, et cela est confirmé dans la proposition de pseudo-token aussi erronée faite par FlauBERT dans ce cas (« dor »).

Gold	**FlauBERT**
C'est l'histoire d'un chat qui se faisait toujours embêter par un loup. Ce méchant loup n'arrêtait pas d'aller chez le chat et de mettre le bazar. Ce pauvre animal **fatiguait** mais un jour Rouqui le chat décida de mettre des pièges de partout.	C'est l'histoire d'un chat qui se faisait toujours embêter par un loup. Ce méchant loup n'arrêtait pas d'aller chez le chat et de mettre le bazar. Ce pauvre animal **dor** mais un jour Rouqui le chat décida de mettre des pièges de partout.

Tableau 23 Versions de référence et méthode FlauBERT du même texte (corpus français, Texte CE2, 2983).

Cette étude nous a permis d'explorer différents scénarios envisageables pour le traitement des tokens omis dans le corpus *Scolinter*. Cela nous a permis de déterminer quelle méthode peut être la plus efficace pour reconstruire les tokens omis afin de minimiser le bruit sur les analyses morphosyntaxiques et en dépendance ultérieures. Ces résultats sont prometteurs, avec des performances atteignant entre 97 % et 98 % selon les différentes métriques utilisées pour évaluer les résultats de l'étude, c'est-à-dire UAS, LAS, LS et précision sur le POS limitées aux phrases contenant des omissions. Néanmoins, cette méthode présente également des limites, notamment lorsque les substitutions suggérées ne semblent pas cohérentes avec le contexte du texte analysé, lorsque l'éventuel token à remplacer n'est pas évident à prévoir même pour les chercheurs humains, ou en présence de « disfluences » qui portent notamment sur le niveau syntaxique dans le texte.

Le scénario qui exploite un modèle de langage de type Transformeur a ainsi été retenu pour la suite de notre travail, et utilisé pour compléter les versions normalisées des textes du corpus *Scolinter*. Cette combinaison entre l'automatisation et le processus de réflexion déjà entrepris par les chercheurs nous semble être la solution la plus efficace à la fois d'un point de vue technique et temporel pour répondre à la problématique du bruit causé par ces omissions de mots dans les textes du corpus.

2. Méthodologie d'annotation et de sélection du corpus à annoter

2.1. Annotation agile

Comme nous avons déjà mentionné, les pré-requis pour la construction d'un corpus qui puisse effectivement aider à décrire le fonctionnement de la langue sont « la représentativité, la taille déterminée et leur fonction en tant que référence du standard »[78] (Voormann & Gut, 2008, p. 237, notre traduction). La richesse mais aussi la précision et la fiabilité des annotations sont cruciales

[78] Dans le texte d'origine, « their representativeness, their finite size and their function as a standard reference ».

pour déterminer la qualité des résultats obtenus suite à l'analyse d'un corpus annoté (Voormann & Gut, 2008).

En ce qui concerne la méthode suivie pour le processus d'annotation, nous avons décidé d'adopter une méthode d'annotation agile (Fort, 2012). Cette approche dérive des processus de développement de logiciel du même nom. En linguistique de corpus et en TAL, ce type d'organisation du travail a été décrit pour la première fois par Voormann & Gut (2008), en contraste avec les méthodes de création de corpus « à cascade ». La création de corpus dite « en cascade », adoptée traditionnellement, se déroulait par étapes successives. Après avoir déterminé les critères de sélection ou de recueil des textes, le protocole d'annotation était décrit, le guide établi et une campagne d'annotation était effectuée sur la partie de corpus choisie ou sur son entièreté. Cette méthode impliquait plusieurs facteurs : la validité absolue du guide d'annotation tel qu'il avait été pensé avant le début de la campagne d'annotation elle-même ; le postulat d'une cohérence intrinsèque du guide avec la tâche d'annotation prévue et avec la perception de la tâche par les annotateurs. Ces facteurs, sans pour autant remettre en cause la possible faillibilité des concepteurs de la campagne, rendent difficile de déterminer dans quelle mesure les incohérences du guide ou de l'annotation se répercutent sur la qualité des données annotées. Or, comme remarqué par Voormann & Gut (2008), un guide d'annotation peut se révéler implicitement incohérent à cause de la variabilité de la perception humaine de certains phénomènes ou à cause d'autres facteurs (les instructions contenues ne permettent pas de généraliser le matériau langagier effectivement présent dans le corpus, les exemples contenus dans le guide ne sont pas suffisants, etc.) ; ou encore, le schéma d'annotation adopté peut se révéler peu consensuel, ce qui va influencer l'accord inter-annotateur de manière négative ; et enfin, les caractéristiques individuelles des annotateurs peuvent elles aussi influencer l'annotation (éducation, motivation, fatigue, etc.). Dans un développement de corpus à cascade, ces facteurs sont peu ou pas contrôlables : une fois la campagne d'annotation démarrée, ni le guide ni le schéma d'annotation ne peuvent être rectifiés en cours de route, sous peine de rendre les annotations incohérentes.

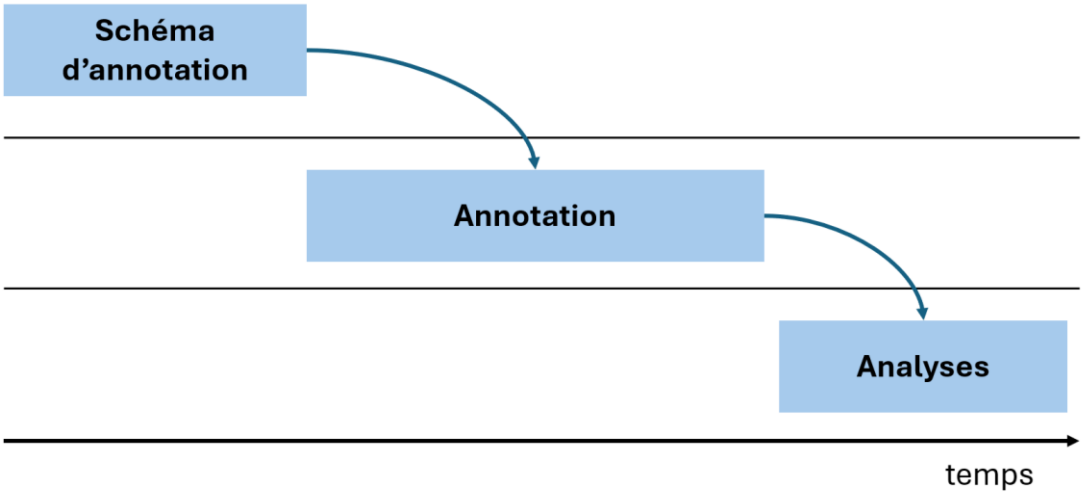


Figure 38. Phases de la création de corpus traditionnelle ou "à cascade". Image inspirée de Voorman et Gut (2008).

L'application d'une méthode de constitution agile permet d'avoir un certain degré de contrôle sur ces différents facteurs. Inspirée de la philosophie de développement logiciel, cette méthode consiste en des phases qui restent les mêmes que pour la constitution de corpus annotés classique mais qui sont plus courtes, et sur lesquelles on revient de manière cyclique et itérative. Cette méthode reste couteuse et difficilement applicable à une campagne de grande envergure (Guillaume & Fort, 2013). Dans le cas de notre travail, où les campagnes d'annotation restent assez limitées en termes de taille du corpus, cette méthode nous a semblé la plus adaptée. En particulier, nous avons adopté la méthode agile décrite, par laquelle la mise au point du guide est réalisée en parallèle à l'annotation des mini-références.

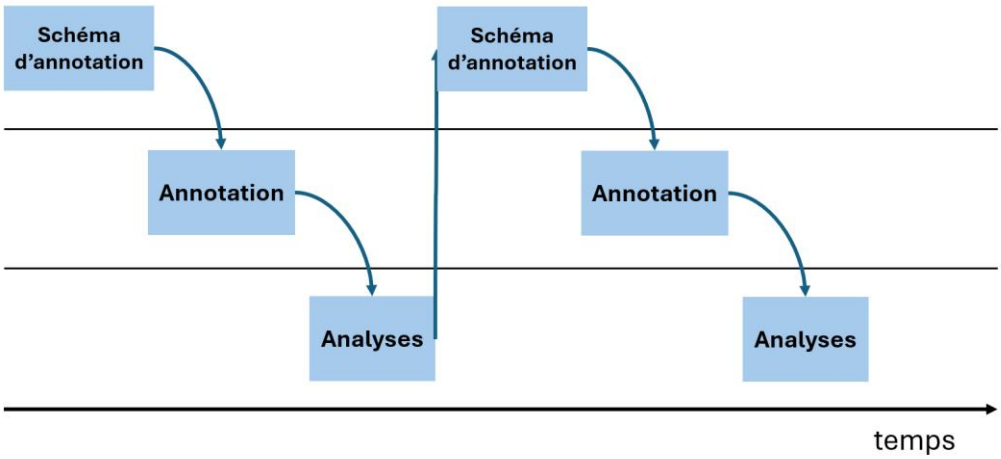


Figure 39. Phases de la création d'un corpus annoté selon la méthode agile. Figure inspirée de Voorman et Gut (2008).

Ce mode de fonctionnement itératif a permis à la fois une certaine qualité des annotations et également d'assurer un plus grand degré de contrôle sur leur cohérence interne. En ce qui concerne la perception de la tâche par les annotateurs, nous avons pu vérifier cela à travers l'accord inter-annotateur mesuré tout au long de la dernière campagne d'annotation. La qualité des annotations a été vérifiée à travers des adjudications réalisées tout au long de la campagne. Plutôt que mettre en place une séquence d'étapes figées, qui ne laisse pas de place à l'erreur, nous avons eu la possibilité de revenir périodiquement sur la partie réduite de textes déjà annotés pour vérifier la fiabilité à travers l'accord inter-annotateur et pour échanger entre annotateurs sur les points de doute rencontrés lors de l'annotation. Cela a été possible grâce aux fonctionnalités de la plateforme d'annotation et grâce à la mise en place d'un document partagé entre les annotateurs et utilisé comme journal de bord.

2.2. Critères de représentativité et choix de sélection du corpus

Une campagne d'annotation doit tenir compte des contraintes de temps et de ressources disponibles. Dans une approche agile, ces limitations imposent de définir des critères de sélection pour constituer des mini-corpus à annoter. L'objectif est de concilier ces contraintes avec la représentativité des données par rapport au corpus global et à la réalité langagière étudiée.

Trois exigences principales guident ainsi la constitution d'une ressource annotée : 1) un équilibre entre taille et représentativité du corpus, 2) la richesse des annotations et 3) une fiabilité suffisante de celles-ci. Dans cette perspective, nous avons retenu deux critères de sélection cohérents avec les objectifs de notre recherche : le niveau scolaire, limité aux niveaux CE1 et CE2 (2e et 3e), et la longueur des textes, mesurée en nombre de tokens. Ces deux niveaux ont été privilégiés à la fois pour des raisons de disponibilité des données et pour leur comparabilité, puisqu'ils partagent une même consigne d'écriture, impliquant la présence de référents communs.

La disponibilité des données variait toutefois selon les contextes. Pour le corpus français, les données avaient déjà été recueillies et prétraitées au début de notre travail de thèse, ce qui a permis une sélection systématique des textes. En revanche, en Italie, le recueil et le prétraitement étaient encore en cours. L'absence de financement dédié et le nombre restreint de chercheurs impliqués ont ralenti ces opérations, limitant temporairement la quantité de textes exploitables. Dans ce cadre, nous avons contribué directement à la numérisation, à la transcription et à la normalisation des textes. Avec l'aide d'une étudiante de master italophone, 25 heures ont permis de transcrire et vérifier 50 textes de CE2, auxquelles se sont ajoutées environ 80 heures pour compléter la transcription, la normalisation et la vérification de 100 textes supplémentaires.

Concernant la taille de l'échantillon, la campagne d'annotation finale a été menée sur 150 textes par langue, répartis équitablement entre CE1 et CE2. Ce volume correspondait à une estimation réaliste au regard du temps d'annotation disponible. Dans le cas du français, deux textes ont été exclus de l'échantillon en phase d'annotation pour des raisons de cohérence, ce qui a réduit l'échantillon français à 148 textes.

Pour le corpus français, la sélection s'est appuyée sur le corpus longitudinal, qui comprend les productions des mêmes 337 élèves du CP au CM2. Afin de préserver la représentativité, nous avons d'abord calculé la distribution du nombre de tokens par texte pour chaque niveau, puis reproduit cette distribution dans le corpus à annoter. Les textes de moins de 20 tokens (21 textes) ont été exclus, car jugés trop courts pour permettre une analyse pertinente de la cohérence textuelle. Les textes restants ont été répartis par tranches de longueur (20–50, 51–100, 101–150 tokens, etc.), puis échantillonnés aléatoirement en respectant les proportions observées dans le corpus global. Dans certains cas, ce critère de représentativité a été complété par des choix guidés par nos objectifs d'annotation. Étant donné que notre étude porte sur des phénomènes discursifs, certains textes particulièrement longs ont été intégrés afin de mieux observer des chaines référentielles étendues, tandis que les textes les plus courts ont été systématiquement exclus.

La méthodologie de sélection adoptée repose ainsi sur deux étapes complémentaires : un prétraitement concernant les textes du corpus longitudinal (décrit précédemment, cf. ce chapitre, §1) et une étape de caractérisation du corpus longitudinal permettant de définir les critères de sélection. Cette approche est restée globalement stable au cours des différentes campagnes d'annotation. Les différences dans les traitements appliqués selon la campagne seront décrites dans le chapitre suivant (cf. Chapitre 6). La première étape de ce processus repose sur le prétraitement des textes du corpus longitudinal, que nous avons détaillé précédemment (cf. §1). Pour rappel, à partir des exports des textes normalisés au format XML, nous avons extrait la partie

contenant les productions, en excluant les métadonnées, et remplacé les balises XML présentes dans les productions par la ponctuation correspondante.

3. Guide d'annotation

Bien que la consigne à l'origine des textes du corpus ne soit pas spécifiquement conçue pour l'étude des niveaux de cohérence et de cohésion textuelles, contrairement à la consigne utilisée pour la collecte du corpus *RésolCo* (cf. Chapitre 2, §3.1), la présence imposée d'au moins un ou deux des quatre personnages de la consigne (mentionnée auparavant, cf. Chapitre 4, §2) rend possible la réalisation d'annotations comparables entre textes de niveaux scolaires différents et entre langues. Cela permet également de réaliser des annotations de la continuité référentielle limitées aux référents présentés dans la consigne, qui peut être élargie aux référents animés présents dans les textes.

Dans le cadre de notre travail doctoral, nous avons choisi de réaliser les annotations sur les versions normalisées des textes, afin de pouvoir, à terme, les aligner avec leurs versions transcrites. Étant donné que nous disposons déjà de ces versions normalisées, conçues spécifiquement pour permettre des traitements automatiques et plus facilement lisibles que des simples transcriptions, il était presque impératif de les utiliser dans le processus d'annotation. Ce choix visait également à rendre le travail d'annotation moins exigeant mentalement et au même temps permettre d'exploiter les traitements automatiques comme le *parsing* morphosyntaxique pour réaliser des descriptions des mentions annotées plus approfondies déjà à partir du texte annoté, comme nous verrons au Chapitre 7.

Tout projet d'annotation devrait être accompagné d'un guide d'annotation. Ce document décrit la nature de la tâche d'annotation à réaliser, ses objectifs, et comme son nom l'indique, guide les annotateurs dans le processus de prise de décision à effectuer tout au long de l'annotation. Ces lignes directrices permettent de formaliser quels sont les éléments de langage objets d'une recherche et à quelles étiquettes ils doivent être associés. Ces documents sont spécialement nécessaires lorsqu'une campagne d'annotation fait intervenir plusieurs annotateurs et annotatrices.

Dans notre projet, nous avons fourni aux annotateurs un guide qui a évolué de manière agile (voir ci-avant cf. §2.1) tout au long des campagnes réalisées. Ce guide était inspiré de ceux existants pour l'annotation de la coréférence et de la continuité référentielle en français des corpus *DEMOCRAT* (Landragin, 2019) et *RésolCo* (Federzoni, Rebeyrolle, et al., 2021), tout en étant adapté à la consigne de production des textes de notre corpus et à nos objectifs de recherche. Le guide ici présenté a subi des changements et des évolutions tout au long du travail d'annotation, grâce aux observations des annotateurs et des enseignants-chercheurs participant au projet. Ce document dans sa dernière version intégrale est disponible en Annexe (cf. Annexe G) et en ligne sur HAL (Barletta, 2025).

Comme déjà décrit par Fort (2012), un guide d'annotation constitue aussi la documentation qui doit accompagner le « produit final », c'est-à-dire le corpus annoté et diffusé. Nous reportons ici

ses recommandations sur l'écriture d'un guide, que nous avons cherché à appliquer lors de la rédaction du nôtre :

- Indiquer ce qu'il faut annoter plutôt que comment annoter ;
- Ne pas exclure a priori ce qui serait douteux ou trop difficile à reproduire par un outil automatique ;
- Donner aux annotateurs une vision claire de l'application visée ;
- Ajouter des définitions précises, justifier les choix méthodologiques effectués et donner la logique de l'annotation visée (ne pas se contenter de donner des exemples).

Ces indications doivent aussi laisser de l'autonomie à l'annotateur dans son processus de prise de décision. Le guide ne doit pas « sur-spécifier », d'un côté pour un souci de lisibilité (un guide ne peut pas faire 50 pages, ça doit être un document facile à appréhender ou dans lequel les annotateurs puissent retrouver facilement et rapidement les informations qui les intéressent), de l'autre pour ne pas empêcher les annotateurs experts d'apporter « une vraie valeur ajoutée au corpus » (Fort, 2012, p. 74).

Le document conçu pour notre projet introduit brièvement le contexte de l'annotation (un corpus d'écrits scolaires longitudinal dont on annote les normalisations), pour ensuite définir la continuité référentielle en tant qu'objectif principal de l'annotation. Le jeu d'étiquettes présent sur la plateforme d'annotation est illustré ainsi que son utilisation. Suit une description des expressions à annoter dans les textes. Cette partie, comme déjà mentionné, reprend les indications des guides existants en les illustrant à travers des exemples d'annotations issus de notre corpus. Il comprend également des indications plus ponctuelles sur l'annotation de certains éléments, issues des observations des annotateurs qui ont participé aux campagnes précédentes. Nous allons par la suite développer certaines des notions encodées dans ce guide et la réflexion méthodologique qui a mené à certains choix d'annotation.

3.1. Les entités à annoter, le jeu d'étiquettes et les référents évolutifs

Pour répondre à nos objectifs de recherche, nous avons initialement décidé de limiter l'annotation aux entités présentes dans la consigne. Ce choix opératoire avait déjà été adopté dans l'annotation du corpus *RésolCo*. Nous avons cependant décidé d'élargir cette annotation aux référents animés présents dans les textes. Cette limitation aux entités animées (à la différence de l'annotation de toutes les entités, comme cela était fait dans le cas du corpus *DEMOCRAT*) sert d'un côté à rendre les annotations de textes différents le plus possible comparables entre elles, de l'autre à simplifier le travail d'annotation et d'analyse ensuite. Nous partons du postulat que tout élève qui a participé au recueil a inclus au moins un des personnages imposés par la consigne, et donc chaque texte contiendra au moins une entité à annoter parmi le chat, la sorcière, le robot ou le loup ; l'élargissement des annotations à des entités autres que celles imposées par la consigne, à la différence de ce qui était fait dans le cas du corpus *RésolCo*, concerne toutes les entités animées (et non pas seulement les entités imposées par la consigne) car celles-ci sont les plus susceptibles de « participer » en tant que personnages actifs à l'intérieur de la narration. Cela nous permet aussi d'une part de ne pas limiter la représentation de la richesse lexicale et la variation présente dans

les chaines et, d'autre part, cette méthodologie d'annotation nous permet d'inclure les référents qui apparaissent dans les mentions plurielles, le chevauchement des chaines, etc.

Le choix de limiter notre annotation aux seuls personnages animés fait écho à d'autres travaux qui ont limité l'annotation en coréférence aux seuls personnages des textes narratifs. Entre autres, nous pouvons citer le projet *Multilingual BookNLP* (Mélanie-Becquet et al., 2024 pour la partie française du projet), dirigé par David Bamman à l'Université de Berkley, dont l'objectif était de permettre l'annotation massive des chaines de références des personnages et d'entités saillantes dans les romans ; le travail déjà cité de Pallotti et al. (2021) et l'étude de Zucchini et Perugini (2025) sur les chaines de référence des personnages dans les écrits narratifs d'élèves que nous avons déjà décrits (cf. Chapitre 2, §4.5).

Le jeu d'étiquettes que nous avons conçu et adopté est simple – chaque référent de la consigne est indiqué à travers son nom en anglais[79], c'est-à-dire *cat, wolf, robot* et *witch*. Les autres personnages animés sont indiqués à travers l'étiquette *extN*, où *N* correspond à l'ordre d'apparition du personnage dans le texte (*ext1* si c'est le premier « personnage externe » puis *ext2* pour le deuxième, etc.) Dans le cas des personnages de la consigne, si par exemple plusieurs chatons sont présents dans un texte, nous annotons comme *cat* le premier chat à paraitre, puis *cat1*, *cat2* etc. pour ceux qui interviennent par la suite dans le texte. Ce choix découle des deux possibilités d'annotations de la coréférence, comme elles ont été détaillées par Landragin (Landragin, 2019) : « soit on saisit, pour chaque expression, un identifiant du référent, soit on regroupe les expressions en chaines (anaphoriques et/ou coréférentielles), ce qui fait l'économie des identifiants des référents mais nécessite de construire des chaines, autrement dit des objets non liés à un et un seul marquable » (Landragin, 2019, p. 5).

C'est à un type de modélisation, dit parfois *ensembliste* (cf. Chapitre 3, §1.2), qu'appartient notre travail, ainsi que celui effectué sur le corpus *DEMOCRAT*[80]*.* Le choix de ce type de modélisation a été fait d'une part pour assurer une certaine comparabilité avec les annotations déjà effectuées sur *DEMOCRAT* et sur le corpus *RésolCo* (qui suivent elles aussi ce modèle d'annotation), et d'autre part pour un souci de simplification de la tâche d'annotation. L'objectif de notre travail étant d'analyser les chaines de continuité référentielle, nous avons réduit le plus possible l'effort de la tâche d'annotation à la sélection et délimitation des expressions indiquant un référent donné puis à l'association de chaque expression identifiée à un identifiant de référent[81].

En ce qui concerne les référents évolutifs (cf. Chapitre 1, §3.1) , nous suivons les indications données pour l'annotation du corpus *DEMOCRAT*. Ainsi, les référents évolutifs sont annotés dès leur première occurrence à travers l'étiquette qui identifie leur première « forme ». Par exemple,

[79] Un jeu d'étiquettes en anglais a été retenu afin d'éviter de privilégier l'une des trois langues du projet et en raison de la concision offerte par cette langue.

[80] Boiteau et al. (2025), que nous avons déjà cité, citent notre travail et *DEMOCRAT* comme exemples d'approches ensemblistes.

[81] Cela diffère de la description donnée par Boiteau et al. (2025), selon lesquels « La particularité de cette approche est qu'elle dispense l'annotateur d'une quelconque mise en relation, et réduit son effort à une tâche de catégorisation. » (p. 439)

dans le texte suivant, la souris se transforme en monstre. Pour indiquer la continuité de référent entre les deux expressions, nous les avons annotées en utilisant la même étiquette. Ce choix a été retenu, en contraste avec une possible approche qui annote chaque étape d'évolution d'un référent donné en tant que référent à soi-même. Cela pour deux raisons : nous avons considéré que la majorité des métamorphoses dans nos textes ce sont de type endomorphiques et radicaux (changement net entre identité de départ et identité en sortie), et en même temps la tendance des auteurs de ces textes est de considérer chaque personnage selon son identité de départ. Si dans notre travail le but est de reconstruire le développement des chaines et donc suivre les actions et les états des personnages tout au long du texte, il nous semble raisonnable d'adopter un tel modèle, nous permettant de reconnecter le personnage avec tous ses états possibles.

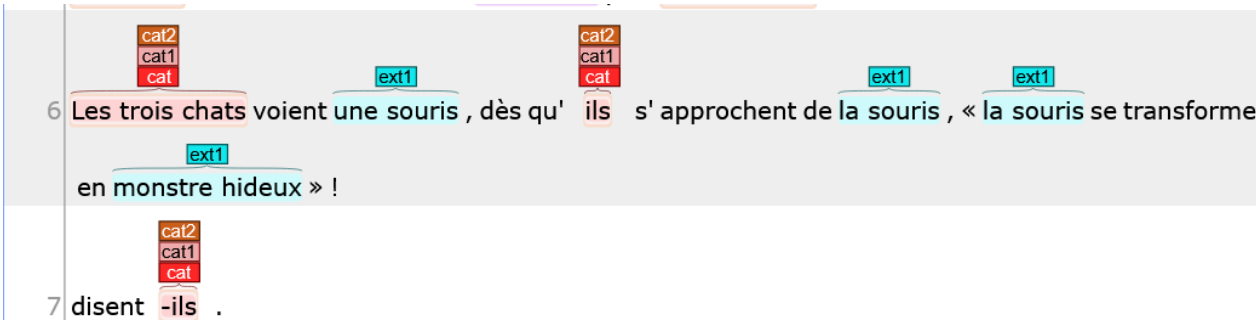


Figure 40. Texte contenant un référent évolutif (corpus français, CE2, 1851).

3.2. Les mentions à annoter

Les mentions annotées correspondent à celles définies dans le guide d'annotation de *RésolCo*, auxquelles s'ajoutent, pour partie, certains choix retenus dans le guide du corpus *DEMOCRAT*. L'annotation porte ainsi sur plusieurs catégories d'expressions référentielles : les syntagmes nominaux, les noms propres, les pronoms personnels (sujets ou objets) ainsi que les pronoms relatifs, les déterminants possessifs et les anaphores zéro.

Les syntagmes nominaux, ou groupes nominaux, regroupent les formes définies, indéfinies et démonstratives (voir le Chapitre 1 pour une description détaillée). Le périmètre du maillon comprend le groupe nominal dans son ensemble, incluant les déterminants éventuels ainsi que les modifieurs, qu'ils soient antéposés ou postposés. Les adjectifs, qu'ils précèdent ou suivent le nom, ainsi que les participes employés comme adjectifs avec leurs compléments, sont donc intégrés au maillon. Les syntagmes prépositionnels internes au groupe (par exemple *une sorcière de cent ans*), les appositions, les noms propres immédiatement associés à un nom commun (par exemple *le chat Victor*), ainsi que les possessifs, sont également inclus. En revanche, les propositions relatives constituent un type de mention distincte et ne sont pas intégrées au syntagme nominal, de même que les prépositions placées en amont du groupe.

Les noms propres sont eux aussi annotés comme des mentions à part entières. Lorsqu'un nom propre est adjacent à un syntagme nominal, les deux éléments ont été annotés dans une seule mention.

Concernant les pronoms, sont pris en compte les pronoms personnels sujets et objets, les pronoms démonstratifs, les pronoms indéfinis — bien que ceux-ci soient rares dans notre corpus —, les pronoms corrélés ainsi que les pronoms relatifs modifiant un syntagme nominal.

Les déterminants possessifs sont considérés comme des expressions référant au possesseur et sont annotés comme appartenant à la chaine correspondant à ce référent. Ils peuvent par ailleurs faire partie du périmètre d'un syntagme nominal en tant que déterminants.

Enfin, les anaphores zéro signalent la présence d'un référent par l'intermédiaire du verbe. En français, elles apparaissent principalement dans des structures coordonnées ou avec des verbes à l'impératif. En italien, du fait des propriétés morphosyntaxiques de la langue, ces formes sont plus fréquentes et peuvent se rencontrer dans des contextes plus variés, généralement après une première mention explicite du référent.

4. Le rôle de l'adjudication dans la création d'un corpus *gold*

Une étape qui était importante pour la création de la ressource était l'étape d'adjudication, appelée parfois *curation* en anglais. Si dans certains cas la création d'un corpus *gold* consiste dans le choix des annotations qui ont obtenu le taux d'accord inter-annotateur plus élevé, dans d'autres cas, la conception de cette version du corpus annoté vise plutôt à obtenir un corpus composé d'annotations consensuelles. Cette étape permet de rassembler plusieurs annotations différentes, parfois divergentes, et à partir du désaccord, de réussir à créer une version considérée comme la référence, ou le *corpus gold*.

De manière similaire à la méthodologie d'annotation décrite par Gardent et Manuélian (2005), la constitution d'un corpus gold peut être réalisée par comparaison entre annotations différentes : « à partir d'un premier schéma, une phase d'annotation est entreprise au terme de laquelle les désaccords entre annotateurs sont discutés, résolus et les modifications nécessaires intégrées au schéma d'annotation » (Gardent & Manuélian, 2005, p. 4). L'objectif n'est plus réduit à la recherche du meilleur accord inter-annotateur selon le schéma imposé une fois pour toutes en début de campagne, mais d'affiner le schéma à travers la discussion entre annotateurs et d'aboutir à des annotations le plus possible consensuelles.

Cela peut être donc réalisé à travers des discussions faites par un « panel d'experts », possiblement déjà entrainés à la tâche ou qui ont déjà participé à la campagne d'annotation. L'adjudication peut être aussi réalisée par un expert qui se constitue comme « super adjudicateur », qui s'occupe de comparer les cas d'annotations divergentes et prend des décisions sur « la bonne annotation » à choisir. Dans ce cas, l'objectif visé est la cohérence avec la tâche d'annotation proposée, qui peut se réfléchir en deuxième lieu dans la cohérence au guide d'annotation proposé, si celui-ci a déjà passé les étapes d'affinement qui le rendent véritablement « à la hauteur de la tâche ». Si ce super adjudicateur devrait habituellement être une personne qui n'a pas été impliquée dans le processus d'annotation pour ne pas introduire des biais, souvent on se retrouve à devoir remplir ce rôle tout en ayant annoté précédemment, par manque de ressource ou d'experts de la tâche déjà formés.

Dans le cas de notre corpus, nous avons adopté une approche qui prévoit l'intervention d'un super-adjudicateur, *de facto* prenant des choix de manière ponctuelle sur les cas d'annotations divergentes. Néanmoins, nous avions mis en place un journal de bord partagé entre annotateurs lors du processus d'annotation, qui nous a permis de remettre en considération la manière d'annoter certains phénomènes. Ces discussions asynchrones ont fait office de discussions d'experts et ont contribué à l'amélioration du guide ou tout simplement au choix de l'annotation consensuelle. Cela parce que les annotateurs ont contribué à ce processus en marquant dans le journal de bord les difficultés rencontrés dans l'annotation, c'est-à-dire les cas de doute dans le processus d'annotation, soit dans le cas de délimitation des mentions, soit dans le rattachement de la mention à un référent donné. Un exemple de ces ajustements concerne le choix de ne pas annoter les pronoms présents dans les formules de clôture des histoires. Ce choix n'avait pas fait l'objet d'une définition explicite dans les premières versions du guide, mais a émergé à la suite de discussions entre annotateurs, documentées dans le journal de bord. En effet, sur le modèle du guide de *DEMOCRAT*, qui exclut de l'annotation les pronoms appartenant à des expressions figées, nous avons considéré les formules d'ouverture et de clôture comme des structures proches de ce type d'expressions. Dans ce cadre, les pronoms qu'elles contiennent ne sont pas interprétés comme renvoyant à des référents du texte et ne sont donc pas pris en compte dans l'annotation.

Cette étape nous a garanti un consensus plutôt général dans le cas de nos annotations, et donc une qualité que nous considérons fiable de notre corpus gold.

Chapitre 6. Campagnes d'annotation

Dans ce chapitre, nous décrivons le déroulement des campagnes d'annotation mises en place dans le cadre de ce travail. Nous présentons d'abord l'organisation de ce processus ainsi que les principes méthodologiques qui ont guidé ces choix, avant de détailler les différentes campagnes menées.

Le processus d'annotation constitue le cœur de notre travail de thèse et une étape centrale dans la constitution d'un corpus annoté exploitable pour l'analyse. Dans notre cas, l'objectif était de produire une ressource annotée en continuité référentielle suffisamment fiable pour permettre à la fois des analyses linguistiques qualitatives et des traitements quantitatifs comparables entre langues différentes. Pour atteindre cet objectif, nous avons adopté une méthodologie d'annotation itérative, inspirée des approches dites « agiles », appliquée à l'annotation de corpus (Fort, 2012; Voormann & Gut, 2008) dont les principes ont été décrits précédemment (cf. Chapitre 5, §2.1). L'approche que nous avons adoptée repose sur un principe d'ajustement progressif du guide d'annotation et du protocole méthodologique à partir de cycles successifs d'expérimentation, d'analyse des difficultés rencontrées et de révision du schéma d'annotation.

Chaque campagne constitue ainsi à la fois une phase de production de données annotées et un moment d'évaluation de la méthodologie adoptée (outil d'annotation et ses paramétrages, guide, accompagnement des annotateurs etc.). Nous distinguons en ce sens deux types d'étapes : des phases orientées vers la production de données, et des phases dédiées à la réflexion méthodologique et à l'amélioration du protocole d'annotation. Nous détaillons dans la Figure 41 les étapes méthodologiques faisant partie de cette organisation, qui seront approfondies par la suite. Le processus global d'annotation est structuré en plusieurs phases, caractérisées par des allers-retours entre le guide d'annotation et les retours des annotateurs ainsi que des enseignants-chercheurs impliqués dans le projet. Le guide d'annotation, entendu comme l'ensemble des consignes et des règles encadrant le processus d'annotation, a été régulièrement ajusté au cours des premières étapes du travail à partir des retours des annotateurs, afin de lever les ambiguïtés et de mieux refléter les pratiques d'annotation attendues au regard des objectifs du projet. Les enseignants-chercheurs ont notamment contribué à ce processus en questionnant certaines formulations du guide lorsqu'elles étaient jugées peu claires ou insuffisamment représentatives des réalités de l'annotation. Ce déroulement itératif permet aussi de vérifier de manière cyclique la qualité des annotations et surtout de corriger déjà dans les premières phases du projet les éventuelles incohérences internes au guide d'annotation (Guillaume & Fort, 2013; Voormann & Gut, 2008).

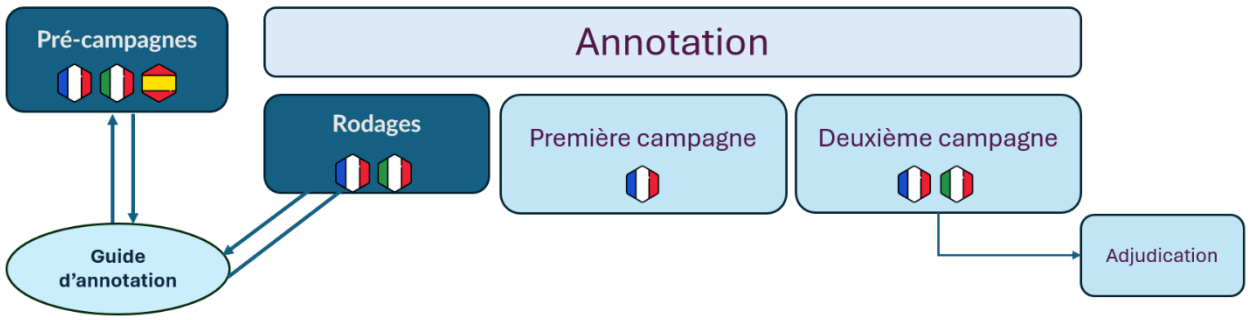


Figure 41. Schéma illustrant les étapes du processus d'annotation de notre corpus sur le plan méthodologique. Inspiré du schéma de Fort (2012).

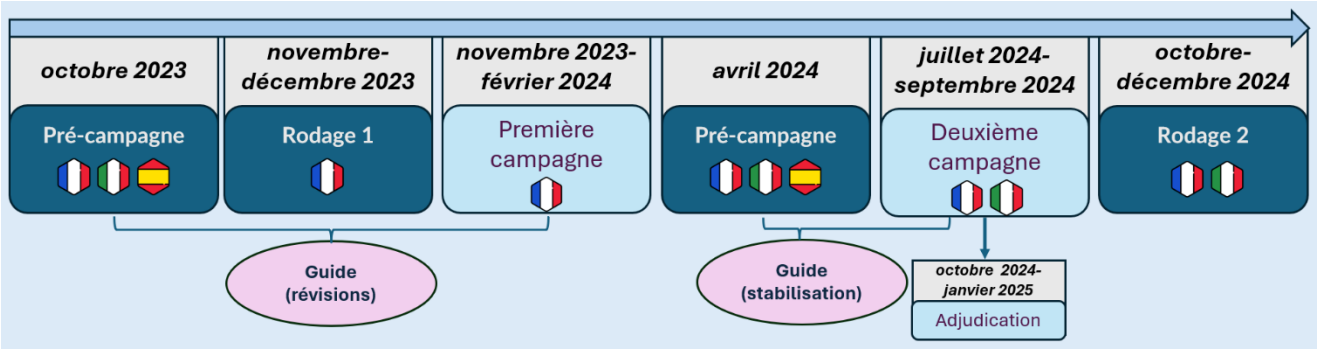


Figure 42. Frise chronologique détaillante l'organisation du processus d'annotation entre octobre 2023 et janvier 2025.

Dans la Figure 42 nous avons détaillé à travers une frise chronologique le déroulement temporel des différentes étapes du processus d'annotation.

Le processus d'annotation s'est organisé en plusieurs étapes successives, visant à tester, ajuster puis stabiliser progressivement le guide d'annotation avant et pendant les campagnes principales. Une phase de pré-campagne a d'abord été menée sous la forme de plusieurs tests d'annotation réalisés par les enseignants-chercheurs impliqués dans la partie française du projet. Ces sessions, conduites à la fois sur plateforme d'annotation et sur papier, ont porté sur 15 textes français et sur un échantillon réduit des textes italiens et espagnols. Elles ont permis d'identifier des difficultés d'interprétation du guide et d'apporter plusieurs corrections à ce document. Cette étape a été menée lorsqu'une grande partie des textes du CE1 au CM2 des corpus italien et espagnol étaient en cours de numérisation

Cette première phase a été suivie d'un temps de rodage, correspondant à l'expérimentation du guide dans le cadre de travaux pratiques en linguistique de corpus avec des étudiants de Master 1 en Sciences du Langage de l'Université Grenoble Alpes, entre novembre et décembre 2023. Ce dispositif a servi à éprouver le protocole d'annotation dans un contexte pédagogique et à observer son applicabilité par de nouveaux annotateurs, expérimentés en linguistique mais pas spécialistes de la tâche d'annotation en question.

La première campagne d'annotation proprement dite a ensuite concerné 111 textes du corpus français. Elle s'est déroulée entre novembre 2023 et février 2024. Nous allons détailler la

méthodologie de constitution du corpus à annoter adoptée lors de cette campagne et les résultats quantitatifs qui en découlent dans la section 3.

Une étape ultérieure de pré-campagne s'est déroulée en avril 2024. Dans cette occasion, nous avons pu réaliser une séance d'annotation collective qui a vu la présence de toutes les chercheuses et chercheurs impliqués dans le projet. Elle servait notamment à vérifier nouvellement la transférabilité du guide à l'italien et à l'espagnol.

Une deuxième campagne d'annotation a été organisée par la suite, avec deux volets menés en parallèle. Pour le français, l'annotation s'est étendue de juillet à fin septembre 2024. Pour l'italien, elle s'est déroulée entre juillet et octobre 2024. Cette phase d'annotation dédiée à l'italien a donné lieu à plusieurs allers-retours de modification du guide d'annotation et à des discussions régulières entre annotatrices, nourrissant la réflexion nécessaire à son adaptation au contexte linguistique italien. Enfin, les annotations issues de la deuxième campagne ont fait l'objet d'une phase d'adjudication, se déroulant entre la fin des volets respectifs et fin janvier 2025. Celle-ci a été principalement assurée par mes soins, tandis que les cas jugés ambigus ou problématiques ont été vérifiés avec deux enseignants-chercheurs responsables du recueil des données françaises et, pour l'italien, avec une enseignante-chercheuse spécialiste en linguistique italienne.

Après cette deuxième campagne, dont les résultats seront analysés par la suite (cf. Chapitre 7 et Chapitre 8), nous avons eu l'occasion de mener une ultérieure phase de rodage auprès du public des étudiants de Master 1 en Sciences du Langage. Celle-ci, n'étant plus ciblée à la stabilisation du guide, considéré suffisamment stable, nous a permis de tester sur une échelle plus large le fonctionnement du guide et, en particulier, de comprendre le fonctionnement de l'accord inter-annotateur sur un échantillon de textes comparable à celui annoté dans notre corpus *gold*.

En termes de ressources humaines, nous avons pu profiter des financements dédiés aux vacations recherche par le laboratoire LIDILEM, financements obtenus pour financer des actions liées aux données de cette thèse de doctorat. Entre ces actions, nous retrouvons les deux campagnes d'annotation en continuité référentielle des données du corpus, campagnes que nous allons évoquer par la suite. Ces vacations ont vu la participation des étudiantes du Master Sciences du Langage parcours Industries de la Langue en tant que participantes expertes grâce à leur double formation en linguistique et en Traitement Automatique de la Langue. Au total, nous avons pu profiter d'un volant d'environ cent heures de vacation.

1. Pré-campagnes : plusieurs allers-retours vers le guide et échanges entre enseignantes-chercheuses

Dans notre description du processus d'annotation, nous définissons comme *pré-campagnes* les phases exploratoires qui précèdent les campagnes d'annotation proprement dites. Contrairement à ces dernières, dont l'objectif est la production de données annotées, les pré-campagnes constituent des moments de réflexion collective entre chercheuses et chercheurs impliqués dans le projet. Leur finalité, dans le cadre de notre travail, était d'une part d'évaluer à quel point le guide d'annotation était clair, univoque et correspondant à la réalité de la tâche d'annotation, d'autre

part d'assurer la transférabilité d'un guide initialement conçu pour le français à l'annotation de textes en italien et en espagnol.

Ces étapes s'inscrivent dans la continuité des études réalisées sur l'état de l'art des corpus annotés en coréférence (cf. Chapitre 1, Chapitre 3). Avant la mise en place de la première pré-campagne collective, un travail préparatoire a été conduit afin de poser les bases méthodologiques du protocole d'annotation. Ce travail a consisté, d'une part, en une analyse approfondie des guides et modèles d'annotation existants et, d'autre part, en la sélection d'un outil d'annotation adapté à nos objectifs. En parallèle, j'ai mené des séances exploratoires d'annotation sur des échantillons du corpus français, dans le but de collecter des exemples à intégrer au guide et d'anticiper les difficultés susceptibles d'être rencontrées par les annotateurs. Ce processus a abouti à la rédaction d'une première version du guide d'annotation, achevée en octobre 2023. Ce travail a été réalisé sur le corpus français, déjà entièrement disponible (transcrit et normalisé), tandis que les transcriptions et normalisations des corpus italien et espagnol étaient encore en cours.

Dans un second temps, une fois la première version du guide finalisée, nous avons mené plusieurs séances de pré-campagne avec les chercheuses et chercheurs responsables du projet, en particulier avec l'équipe française. Ces annotations ont été effectuées sur 15 textes dans leur intégralité pour le français, mais elles ont concerné une quantité plus importante d'extraits de textes dont le nombre exact reste indéterminé. En effet, ces annotations ont été réalisées dans le but unique de réfléchir collectivement aux définitions constitutives du guide d'annotation, alors à sa toute première version, ainsi qu'aux exemples d'annotations qu'il incluait[82]. Lors de ces séances, les annotations ont été effectuées sur la plateforme d'annotation INCEpTION (Klie et al., 2018), avec le double objectif de présenter la plateforme aux chercheuses et chercheurs impliqués dans le projet et de faciliter le processus de réflexion grâce aux avantages de l'utilisation de la plateforme (possibilité de visualiser facilement les versions des différentes annotatrices et annotateurs en même temps, possibilité de modification rapide des annotations en cas d'erreur, etc.). Cependant, lors de ces réunions, un échantillon supplémentaire a été annoté dans une version papier, y compris sur les versions scannées des textes.

Nous avons pu élargir ce processus de réflexion aux corpus italien et espagnol en avril 2024 : lors d'une réunion avec les chercheuses italiennes et espagnoles du projet, nous avons soumis un échantillon composé de textes français, italiens et espagnols au même processus d'annotation et de discussion collective pour mettre à l'épreuve la validité et la transférabilité entre langues du corpus. Nous avons également exploité la plateforme d'annotation ainsi que laissé le choix d'effectuer des annotations sur des versions scans des textes proposés. Ces sessions n'avaient pas comme but de déterminer la validité de la tâche d'annotation elle-même, mais plutôt de questionner la description de cette tâche telle qu'elle était présentée dans le guide. Par conséquent, nous n'avons pas conservé ni les annotations obtenues durant ces séances ni le décompte exact des textes annotés. En particulier, notre préoccupation résidait dans la possibilité d'opérer un transfert entre langues à l'aide d'un guide initialement conçu pour annoter en langue

[82] Lors de la rédaction de ce manuscrit, le guide est à sa 8ème version.

française, et notre objectif était de vérifier si cela était possible à travers des campagnes d'annotation collectives.

Une des hypothèses sous-jacentes à l'écriture du guide était la possibilité d'effectuer une annotation de la continuité référentielle comparable entre les trois langues étudiées. En raison de la proximité syntaxique et morphologique entre ces trois langues romanes, il semblait possible d'appliquer un même cadre théorique et une même méthodologie d'annotation, de manière à garantir la comparabilité des annotations.

Dans une première étape, ce choix s'est traduit par la conception d'un guide reposant exclusivement sur des exemples en français. Cette décision se basait sur l'hypothèse d'une similarité suffisante dans la construction de la coréférence entre les trois langues. Toutefois, les séances de pré-campagnes ont rapidement mis en évidence plusieurs limites à cette approche. Elles ont soulevé des questionnements quant à l'applicabilité du guide dans les trois langues et ont conduit à des ajustements progressifs, en particulier pour mieux prendre en compte certaines spécificités linguistiques. Ces sessions ont ainsi permis d'enrichir le guide, notamment par l'ajout d'exemples en italien dans le texte du guide, afin de mieux accompagner les annotateurs italophones. Elles ont également contribué à préciser la définition de certains phénomènes, comme l'anaphore zéro. Peu fréquente en français, où elle apparaît principalement dans des structures coordonnées ou à l'impératif, elle est en revanche beaucoup plus répandue en italien, ce qui a nécessité des définitions plus explicites et un plus grand nombre d'exemples la concernant. Plus largement, ces allers-retours entre annotation collective, annotation individuelle et révision du guide ont permis de stabiliser progressivement le protocole d'annotation à adopter. À l'issue de ces phases, le guide pouvait être considéré comme suffisamment robuste pour être utilisé dans des campagnes d'annotation de plus grande ampleur impliquant plusieurs annotatrices et annotateurs.

2. Rodages : campagnes d'annotation en TP, ou le guide à l'épreuve des groupes d'annotateurs de Master

Lors de l'année universitaire 2023/2024, entre novembre et décembre 2023, nous avons mis à l'épreuve l'application du guide dans un contexte réel de campagne d'annotation, dans le cadre d'un cours de Linguistique de Corpus pour le Master Sciences du Langage de l'Université Grenoble Alpes. Après avoir introduit les notions de linguistique outillée, les principes d'annotation de corpus et de déroulement d'une campagne d'annotation, ainsi que les notions théoriques de coréférence et d'anaphore, nous avons illustré la consigne du travail pratique, centré sur l'annotation de la continuité référentielle dans des textes issus de notre corpus.

Les dix binômes participants ont annoté entre 11 et 18 textes français de niveau CE2 à l'aide de la plateforme INCEpTION, en s'appuyant sur le guide d'annotation fourni. Un binôme de locuteurs natifs italiens étant présent dans la classe, nous leur avons proposé d'annoter 20 courts textes de niveau CE1 du corpus italien, selon le même protocole. Au total, 140 textes français et 20 textes italiens ont été ainsi annotés.

Un rapport final était également demandé. Celui-ci visait à décrire le processus d'annotation : il était demandé d'expliciter dans une section de leur rapport les doutes, remarques et interrogations rencontrés lors de l'annotation et en lien avec le guide. Enfin la comparaison entre annotations dans le cas de divergence était aussi prévue dans l'absence de fonctionnalité d'adjudication sur la couche d'annotation utilisée dans INCEpTION. Pour cette raison, les binômes devaient comparer leurs annotations en cas de divergence et justifier leurs choix dans le rapport. Cette phase de comparaison a également servi d'évaluation qualitative de l'accord inter-annotateur. Il était par ailleurs précisé que l'objectif du travail ne portait pas sur l'évaluation des productions d'élèves, mais sur l'analyse réflexive du processus d'annotation lui-même. A noter ici que le TP s'appuyait sur la couche native d'annotation de la coréférence prévue par INCEpTION qui ne permettait pas d'effectuer l'adjudication sur la plateforme.

Les annotations produites par chaque binôme ont été analysées à la fois du point de vue qualitatif, à partir des observations empiriques détaillées dans les rapports finaux, et du point de vue quantitatif, afin d'évaluer l'accord inter-annotateur. Certaines de ces observations ont permis d'améliorer ultérieurement les critères d'annotation, notamment en enrichissant le guide avec des exemples annotés supplémentaires issus du corpus, en particulier pour la délimitation de certains types de mentions constituées de syntagmes nominaux. Cette itération a également été déterminante pour préciser le traitement des pronoms en discours direct, insuffisamment explicité dans les versions précédentes du guide, ainsi que pour proposer une reformulation de la définition des entités à annoter. Dans la version 5 du guide, utilisée lors de cette phase de rodage, la définition des référents à annoter était la suivante :

(64) Les cibles de l'annotation de la continuité référentielle dans ce projet sont tous les maillons relatifs aux **quatre personnages élicités par la consigne** (...) ainsi que **les maillons qui font référence à d'autres personnages animés qui agissent de manière telle que leur présence est importante pour l'intrigue du texte analysé**.
(Définition des référents à annoter, version 5 du guide d'annotation, novembre 2023)

Nous verrons par la suite comment cette définition a été revisitée, à la suite des remarques formulées par les annotateurs lors de cette phase de rodage et aux résultats obtenus lors de la première campagne d'annotation, qui ont mis en évidence le caractère ambigu et donc instable de cette définition.

En complément de cette analyse qualitative, une évaluation quantitative de l'accord inter-annotateur a été réalisée a posteriori à partir des exports des annotations au format CoNLL. Les mesures ont été calculées à l'aide de scripts Python, en mobilisant le coefficient kappa de Cohen ainsi que la F-mesure (F1-score), pour chaque texte annoté. Pour le corpus français, les résultats obtenus indiquent un accord moyen de 0,61 selon le kappa de Cohen et de 0,80 selon la F-mesure. Pour le corpus italien, les scores sont légèrement inférieurs, avec un kappa de 0,57 et une F-mesure de 0,72.

Ces résultats, bien que issus d'un contexte pédagogique et d'annotateurs non spécialistes de la tâche, témoignent d'un niveau d'accord globalement satisfaisant, en particulier pour l'échantillon

français. L'accord inter-annotateur un peu moins satisfaisant dans le cas de l'italien pourrait être lié à un guide encore insuffisant pour être appliqué à l'annotation des textes de cette langue, ou à la taille plus restreinte de l'échantillon.

Un an plus tard, lors de l'année universitaire 2024/2025, un second rodage a été mené selon un protocole similaire. Les annotations produites ont cette fois été exploitées pour mesurer l'accord inter-annotateur auprès de 27 annotateurs experts en linguistique, non spécialistes de la tâche. Ils ont annoté 132 textes de niveau CE2 en binôme ou en groupe (un seul groupe de trois étudiants).

A la différence du TP précédent, la fonctionnalité d'adjudication étant cette fois disponible, les étudiants devaient appliquer l'adjudication à chacun des textes. Outre les annotations, chaque binôme ou groupe devait ainsi fournir un rapport sur leur processus d'annotation, sur les choix effectués lors de l'adjudication pour la constitution d'un corpus *gold* et sur les points de divergence.

Cette deuxième phase de rodage expérimentale, tout en offrant la possibilité d'exploiter les outils d'adjudication et de calcul automatique du taux d'accord inter-annotateur, a contribué de manière moins significative à la réflexion autour de l'annotation et du guide. D'un côté, l'état plus avancé de notre travail justifie le peu de nécessité de contribution à cette étape. De l'autre côté, cette campagne nous a permis de vérifier la robustesse du protocole et de mesurer plus finement l'accord inter-annotateur. Les résultats indiquent un alpha de Krippendorff moyen de 0,80, avec un minimum de 0,66 et un maximum de 0,95. Selon les résultats fournis par les étudiants, cette phase de rodage obtient un taux de l'accord inter-annotateur moyen, calculé à l'aide de l'alpha de Krippendorff, de 0,8. L'accord minimum obtenu par un binôme ou groupe était de 0,66 et le maximum de 0,95.

A l'issue des différentes expériences que l'on vient de décrire, pré-campagnes et rodages, qui nous ont permis de valider le guide d'annotation proposé et d'améliorer au fur et à mesure les définitions des objets linguistiques à annoter, nous avons pu effectuer une première campagne d'annotation sur un échantillon de textes du corpus français.

3. Première campagne d'annotation : validation méthodologique de la constitution d'un corpus *gold*

Cette section présente les résultats issus de la première campagne d'annotation manuelle menée sur une portion restreinte du corpus français. Réalisée entre décembre 2023 et février 2024 par deux annotatrices expertes, cette campagne constitue une des premières étapes exploratoires dans l'analyse empirique de la continuité référentielle dans notre corpus. Elle poursuivait plusieurs objectifs : définir une méthodologie de pré-traitement et de sélection du corpus à annoter ; améliorer les critères de définition des entités à annoter ; tester la robustesse du guide d'annotation élaboré ; enfin, obtenir une première cartographie des chaines référentielles observables dans les productions écrites d'élèves de notre corpus.

3.1. Méthodologie de sélection des textes à partir du corpus longitudinal

Afin de garantir la représentativité du corpus annoté, nous avons fondé notre méthode de sélection sur la distribution des textes en fonction de leur longueur (mesurée en nombre de tokens) dans le corpus longitudinal français. Le Tableau 24 présente cette distribution en pourcentage pour les différents niveaux scolaires (CE1 à CM2).

Nombre de tokens par texte	Distribution des textes dans le corpus longitudinal - %			
	CE1 – 2	CE2 – 3	CM1 – 4	CM2 – 5
0-20	5,64	0,3	0	1
21-50	30,56	9,2	3,86	0,59
51-100	48,66	31,45	23,44	16,91
101-150	12,76	35,61	31,16	22,26
151-200	2,08	16,62	22,55	29,08
201-250	0,3	3,86	11,87	16,91
251-plus	0	2,97	7,12	13,95

Tableau 24. Distribution des textes par nombre de tokens sur le corpus longitudinal français.

Cette distribution met en évidence une forte variabilité selon les niveaux, avec une majorité de textes courts en CE1 et une proportion croissante de textes longs aux niveaux supérieurs. Sur cette base, comme nous l'avons décrit dans le chapitre précédent (cf. Chapitre 5, §2.2), nous avons défini une méthodologie d'échantillonnage visant à reproduire, autant que possible, cette distribution dans le corpus annoté, afin d'éviter un biais lié à la longueur des textes, susceptible d'influencer la structure des chaines référentielles observées.

La sélection du corpus annoté reproduit ainsi les tendances générales, avec des ajustements volontaires cohérents avec nos objectifs de recherche. Le Tableau 25 présente la distribution des textes dans le corpus annoté lors de la première campagne, permettant de comparer directement celle-ci à la distribution observée dans le corpus longitudinal.

Nombre de tokens par texte	Distribution des textes dans le corpus annoté - %			
	CE1 – 2	CE2 – 3	CM1 – 4	CM2 – 5
21-50	34,38	4	3,7	0
51-100	43,75	28	22,22	18,52
101-150	15,63	28	22,22	18,52
151-200	0	24	18,52	11,11
201-250	6,52	12	22,22	14,81
251-plus	0	4	11,11	37,04

Tableau 25. Distribution des textes par nombre de tokens sur le corpus annoté français - première campagne d'annotation.

3.2. Caractérisation du corpus annoté et premiers résultats

La taille du corpus à annoter étant initialement établie à 25 textes par niveau. Cette base a été par la suite ajustée pour mieux prendre en compte la distribution observée en termes de longueur sur le corpus longitudinal. En particulier, sur le niveau CE1, où les productions sont en moyenne plus courtes, le nombre de textes a été porté à 32, incluant notamment les textes les plus longs

disponibles pour ce niveau dans le corpus longitudinal. Pour les niveaux CM1 et CM2, deux textes supplémentaires ont été ajoutés par niveau, en privilégiant également les textes les plus longs, afin d'intégrer des textes contenant des chaines référentielles potentiellement plus complexes. Les textes très courts (moins de 20 tokens) ont été exclus, car jugés peu informatifs du point de vue de la cohérence textuelle. Le corpus sélectionné était ainsi constitué de 111 textes extraits de la partie longitudinale du corpus, produits par des élèves du CE1 au CM2, pour un total de 16 838 tokens. La caractérisation de ce corpus est illustrée dans le Tableau 26.

Niveaux scolaires	Nombre de textes	Nombre total de tokens	Nombre moyen de tokens par texte	Étendue de la longueur des textes
CE1 – 2	32	2 101	65,66	23-212
CE2 – 3	25	3 051	122,04	40-271
CM1 – 4	27	3 969	147	46-315
CM2 – 5	27	5 201	192,63	62-591
CE1-CM2 – 2-5	**111**	**14 322**	**129,03**	**23-591**

Tableau 26. Caractérisation de la longueur des textes du corpus annoté en nombre de tokens.

Il a été annoté par deux annotatrices expertes sur la plateforme INCEpTION, selon la version du guide d'annotation issue des phases de pré-campagne et de rodage précédentes, et à l'aide d'un schéma enrichi d'étiquettes spécifiques (ex. *cat, witch, ext1*, etc.) pour assurer le suivi des personnages issus de la consigne ou introduits par les élèves.

Contrairement aux campagnes successives, les fonctionnalités d'adjudication de la plateforme n'étaient pas disponibles pour le niveau d'annotation choisi. En effet, nous avons choisi d'utiliser la couche de coréférence native d'INCEpTION, qui permettait l'export des annotations au format CoNLL-2012, format indispensable pour la suite de nos analyses. Ce choix exclut la fonctionnalité d'adjudication, qui n'avait pas encore été intégrée à la plateforme pour ce niveau spécifique, en plus de la perte de la possibilité d'exploiter les fonctions de calcul automatique de l'accord inter-annotateur. Nous n'avons donc pas procédé à la réalisation d'une version de référence issue de l'adjudication entre les deux annotations disponibles. Par conséquent, les analyses présentées se basent sur une seule des deux annotations réalisées et non pas sur un *gold* obtenu à partir d'une étape d'adjudication (comme cela est le cas du *gold* obtenu à partir des annotations de la deuxième campagne, cf. ce chapitre, §4). Le Tableau 27 présente les résultats quantitatifs obtenus à partir de l'analyse de ces annotations.

Niveaux scolaires	Tot. mentions	3 mentions et plus	2 mentions	1 mention	Nombre moyen de référents par texte	Densité référentielle	Longueur moyenne des chaines (3 mentions+)
CE1 – 2	441	59	15	10	2,25	21%	6,8
CE2 – 3	622	63	12	11	2,76	20%	9,32
CM1 – 4	815	84	18	27	3,66	21%	8,95
CM2 – 5	1049	84	23	45	4,04	20%	12,07
CE1-CM2 – 2-5	**2 927**	**290**	**68**	**93**	**3,14**	**20%**	**9,3**

Tableau 27. Indices de description des chaines de référence dans le corpus annoté par niveau scolaire.

Grâce à cette étude, nous avions pu effectuer une première description empirique des chaines de référence dans notre corpus.

Les résultats mettent en évidence plusieurs tendances générales. Tout d'abord, le nombre moyen de tokens par texte et l'étendue augmentent avec le niveau scolaire, comme déjà observé dans le corpus longitudinal français (cf. Chapitre 4, §9.1). Par ailleurs, le nombre moyen de référents par texte augmente également, passant de 2,25 en CE1 à 4,04 en CM2, ce qui suggère plus d'aisance dans la gestion de plusieurs référents au sein d'un même texte avec l'augmentation du niveau scolaire. La longueur moyenne des chaines référentielles suit une progression similaire, passant de 6,8 mentions en moyenne en CE1 à 12,07 en CM2. En revanche, la densité référentielle, c'est-à-dire le ratio entre le nombre de mentions et le nombre de tokens par texte reste relativement stable (autour de 20 %), conformément aux attendus. Cet indice confirme d'ailleurs la conformité des textes de notre corpus au genre narratif. En effet, comme remarqué par Landragin et al. (2024), dans le corpus *DEMOCRAT* les textes narratifs présentent habituellement en moyenne une densité référentielle autour du 20%, plus importante que les textes appartenant à d'autres genres (par exemple, 13,81% dans le cas des textes juridiques et encyclopédiques du 21e siècle).

Ces premiers résultats tendent à confirmer que les textes d'élèves mobilisent une densité référentielle plutôt stable. Cela pourrait être interprété comme cohérent avec les attendus, c'est-à-dire qu'à une longueur plus importante des textes correspondent davantage de mentions référentielles. Cependant, cet indice pourrait aussi décrire l'absence, à des niveaux scolaires plus élevés, d'éléments autres de la narration, comme des descriptions du cadre de la narration. Nous verrons par la suite comment cet indice a été interprété dans des analyses ultérieures sur le corpus gold constitué successivement. Le nombre de référents augmente avec l'âge, tandis que la densité reste relativement constante, suggérant que les élèves parviennent à maintenir une gestion complexe des entités référentielles à des niveaux scolaires plus élevés, dans des textes plus longs et plus complexes. Malgré la richesse de cette première exploration, plusieurs limitations méthodologiques ont été mises en avant lors de cette campagne.

3.1. Limites méthodologiques de la première campagne

La première limitation que nous avions rencontrée était l'ambiguïté dans la définition des « personnages externes » à annoter. En effet, dans le guide proposé pour cette campagne d'annotation (le même adopté lors du premier rodage, décrit précédemment cf. ce chapitre, §2), nous avions fourni une définition initiale des personnages à annoter autres que ceux présents dans la consigne (*chat, robot, loup, sorcière*). Dans la version du guide utilisée pour cette campagne, telle était la définition des personnages à annoter :

(65) La cible de l'annotation de la continuité référentielle dans ce projet sont tous les maillons relatifs aux **4 personnages élicités par la consigne** (...), c'est-à-dire ces personnages qui sont identifiés en tant que chat, robot, sorcière et loup, ainsi que les maillons qui font référence à d'autres personnages animés qui agissent de manière telle que leur présence est importante pour l'intrigue du texte analysé.
(Guide d'annotation du corpus Scolinter, version 3).

Cette définition a donné lieu à des interprétations divergentes de ce qui pouvait être annoté en tant que personnage dans le texte. Dans le cas des textes d'enfants, il n'était en effet pas trivial de décider quand un personnage était actant actif dans la narration textuelle ou juste un personnage de deuxième plan ou encore part du décor. Selon une analyse empirique, cette définition nous semble avoir entrainé une sous-annotation des entités externes chez une des annotatrices et une sur-annotation de ces entités chez l'autre annotatrice, ce qui a engendré un désaccord sur le nombre d'entités annotées dans environ 50 % des textes.

Lors du premier rodage du guide, se déroulant en parallèle de cette campagne, les annotateurs avaient identifié cette même définition du guide comme un possible obstacle lors de l'annotation. Cela a confirmé la nécessité de revisiter et simplifier cette définition dans la version stabilisée du guide.

(66) L'objectif est d'annoter toutes les mentions qui se réfèrent à un personnage ou une entité animée spécifique dans l'univers du texte analysé. *(Définition des référents à annoter, version définitive du guide d'annotation)*

Cette nouvelle définition permet ainsi un déplacement d'une définition fondée sur le rôle narratif des personnages vers une définition plus opérationnelle, déterminée par les traits morphologiques des référents plutôt que d'une analyse narrative.

La deuxième limitation concerne l'annotation de l'anaphore zéro. Bien que ce type de mention soit central pour la comparaison entre langues (notamment avec l'italien), l'anaphore zéro a été annotée de manière inégale par les deux annotatrices (83 anaphores zéro annotés par annotatrice 1 contre 26 annotés par annotatrice 2). À la suite d'une exploration empirique des annotations, nous avons remarqué que les annotatrices ont souvent omis d'annoter des sujets implicites dans des phrases coordonnées ou subordonnées, ce qui a nécessité une meilleure explicitation de ce type de mention dans le guide.

Enfin, la troisième limitation concerne l'écart entre annotateurs : tout d'abord, les différences entre les deux annotateurs experts ont mis en lumière plusieurs différences quantitatives dans les annotations effectuées : 2 927 mentions étaient annotées dans un cas contre 2 691 mentions dans l'autre. L'écart moyen s'élève à 1,75 mention par texte, avec une variation de 0,55 entité annotée en plus ou en moins selon les versions. L'absence d'outils intégrés permettant le calcul automatique de l'accord inter-annotateur dans INCEpTION a par ailleurs limité la possibilité de suivre et de contrôler ces divergences au cours du processus d'annotation.

3.2. Vers l'amélioration du protocole d'annotation

À la lumière de ces résultats et des difficultés rencontrées, plusieurs pistes d'amélioration ont été envisagées pour la campagne d'annotation suivante. Tout d'abord, la définition des entités à annoter a été précisée. Plutôt que nous appuyer sur la notion de participation à l'histoire pour la définir (c'est-à-dire, le statut de personnage), nous avons décidé d'élargir la notion d'entité à l'annotation de toutes les entités animées dans le contexte du texte à annoter (y compris ces entités qui ne sont pas habituellement animées mais qui peuvent l'être dans le contexte de l'histoire analysée, par exemple le balai de la sorcière qui effectue des actions tel qu'un personnage autonome dans l'histoire et qui assume le statut d'animé à travers ces connotations).

Le guide a aussi été enrichi d'exemples issus du corpus annoté dans sa première itération pour clarifier des notions telles que l'anaphore zéro en français, peu représentée dans la version précédente du guide. Nous avons aussi inséré quelques exemples en italien dans le guide, pour accompagner les futurs annotateurs du corpus italien dans le travail d'annotation.

La dernière modification apportée concerne le développement d'une couche d'annotation personnalisée, associée à un jeu d'étiquettes fermé. Dans la version précédente, la saisie libre des étiquettes par les annotatrices avait en effet entraîné un certain nombre d'erreurs liées à des fautes de frappe, ce qui avait affecté les métriques d'accord inter-annotateur fondées sur ces étiquettes. La mise en place de cette nouvelle couche d'annotation, décrite dans la section consacrée à la deuxième campagne (cf. ce chapitre, §4.1), a également permis d'exploiter certaines fonctionnalités de la plateforme, notamment le calcul automatique de l'accord inter-annotateur ainsi que le support à l'adjudication semi-automatique.

3.3. Bilan de la première campagne d'annotation

Cette première campagne d'annotation a permis de poser les bases d'une analyse structurée de la continuité référentielle dans les écrits narratifs d'élèves de l'école primaire. Malgré les limites observées dans le guide d'annotation et les écarts entre annotateurs, les données issues du corpus de référence fournissent déjà des tendances sur la manière dont les élèves gèrent les référents dans leurs productions écrites.

Les résultats montrent une évolution progressive de la complexité des chaines référentielles avec l'âge des élèves, tant sur le plan de leur longueur que de leur densité. Toutefois, la densité référentielle reste relativement stable, ce qui suggère que l'augmentation de la longueur des textes ne s'accompagne pas nécessairement d'une complexification des textes décrits.

Les difficultés rencontrées dans l'identification des personnages « externes » ou dans l'annotation de l'anaphore zéro soulignent l'importance de clarifier les critères d'annotation et d'adapter les outils aux spécificités des textes produits par les enfants. Enfin, certaines des limites identifiées lors de cette campagne, tant du point de vue méthodologique qu'organisationnel, ont été exploitées afin d'améliorer l'organisation de la deuxième campagne d'annotation.

4. Deuxième campagne d'annotation : constitution d'un corpus *gold* par adjudication

Cette campagne d'annotation s'est déroulée entre juin et septembre 2024. Initialement elle portait sur 300 textes, 150 par langue (français et italien) et couvrant les niveaux CE1 et CE2. Deux textes ont toutefois été exclus du corpus français en raison de leur faible cohérence, portant ainsi le nombre total de textes annotés à 298. Un projet distinct était dédié à chaque langue sur la plateforme, ce qui a conduit à maintenir les deux corpus séparés tout au long de la campagne d'annotation. Parmi les annotatrices et annotateurs, l'une d'entre eux s'est occupée des deux langues, de l'adjudication et a coordonné la campagne d'annotation. Nous allons nous référer aux annotateurs comme Annotateur 1, Annotateur 2 et Annotateur 3, abrégés en A1, A2 et A3.

A1 était la doctorante qui a coordonné et travaillé sur les deux campagnes en parallèle, de langue maternelle italienne et avec une très bonne connaissance du français. A2 et A3, étaient deux étudiants du Master 2 de Sciences du Langage et Industries de l'Université Grenoble Alpes. A2 est de langue maternelle française et A3 de langue maternelle italienne. A2 et A3 ont été rémunérés pour leur activité d'annotation grâce à un financement dédié par le laboratoire LIDILEM et ont signé une autorisation de diffusion de leurs annotations. Le coût total des annotations de cette deuxième campagne d'annotation en termes d'heures équivaut à 100 heures (25 heures rémunérées pour chaque annotateur vacataire plus 50 heures faisant partie du travail de préparation de thèse pour la doctorante).

En plus de la plateforme contenant les textes à annoter et le guide d'annotation, les annotateurs avaient à disposition un dossier partagé sur un Cloud contenant une copie du guide d'annotation ainsi qu'un fichier de type tableur utilisé comme « journal de bord »[83]. Celui-ci nous a permis de noter des remarques sur les textes à annoter ou sur les difficultés rencontrées lors de l'annotation. Le fichier contenait plusieurs colonnes pour chaque texte, et il a servi à noter plusieurs catégories de problèmes rencontrés : problèmes de type technique (manque d'ergonomie de la plateforme d'annotation vis-à-vis de certains cas d'annotation) ; signalement à la coordinatrice de la campagne des erreurs rencontrées dans les textes avant d'effectuer l'annotation (des erreurs de segmentation ou d'orthographe encore présentes dans les normalisations) ; remarques sur des difficultés rencontrées pendant l'annotation, des doutes qui relèvent soit du manque de clarté du guide vis-à-vis du cas rencontré soit, plus souvent, à cause du peu de cohérence et cohésion du texte annoté. Une partie des remarques effectuées par le biais du journal de bord a ainsi contribuée à l'étape d'adjudication ainsi qu'à la modification ou à l'ajout d'éclaircissements ultérieurs dans le guide

[83] Nous remercions Lydia Mai Ho-Dac pour nous avoir suggéré l'adoption de cet outil.

d'annotation. L'étape d'adjudication (cf. Chapitre 5, §4) consiste dans la création d'un corpus de référence (ou corpus *gold*) à travers la comparaison et le choix, au cas par cas, des annotations les plus consensuelles en cas de divergences entre deux annotateurs.

4.1. Schéma et outil d'annotation adoptés

La plateforme d'annotation utilisée dispose d'une fonction spécifique facilitant le processus d'adjudication (voir Figure 43). Les annotations concordantes entre deux annotateurs (comme illustré par A1 et A2 dans l'exemple) sont automatiquement transférées vers la version *gold standard* (située en haut de l'écran dans la Figure 43), où elles peuvent être validées par le « super adjudicateur ». Les résultats que nous décrivons par la suite sont issus de cette version *gold* du corpus annoté.

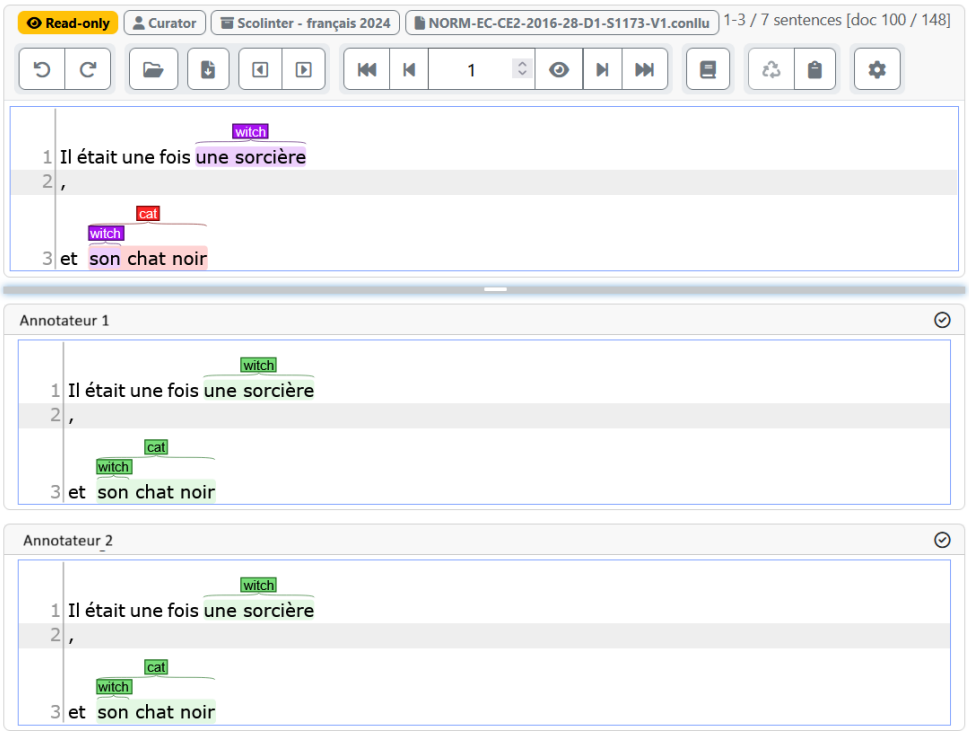


Figure 43. Ecran de la fonctionnalité d'adjudication sur la plateforme INCEpTION.
Les annotations identiques des deux annotateurs sont automatiquement reportées sur la version gold proposée.

Le schéma d'annotation utilisé lors de cette campagne sur le logiciel INCEpTION diffère de celui employé dans la première campagne. Plutôt que de recourir au niveau d'annotation de la coréférence proposé par défaut dans INCEpTION, nous avons développé une couche d'annotation personnalisée. Ce choix s'explique par les limitations du module natif : bien qu'il permette l'annotation et l'export des données au format CoNLL-2012, il ne prenait pas en charge, au moment de nos premières expérimentations, les fonctionnalités essentielles de calcul de l'accord inter-annotateur et de support à l'adjudication.

La couche d'annotation définie était composée de deux types d'annotation complémentaires : une annotation de type *Span*, permettant de délimiter les mentions, et une annotation de type *Link*, permettant de relier les mentions qui appartiennent à la même chaine référentielle.

L'annotation de type *Span* a été paramétrée avec une granularité au niveau du caractère, ce qui permet de délimiter des unités infra-token, telles que les pronoms clitiques en italien, que nous souhaitons inclure dans l'annotation. En effet, la tokenisation adoptée ne sépare pas ces éléments du token auquel ils sont rattachés (ni dans le cas du français, ni en italien), mais une annotation au niveau du caractère permet de délimiter cet élément au sein du token qui le contient. Nous retrouvons un exemple de cette nécessité et de l'annotation adoptée dans les Figure 44 et Figure 45.

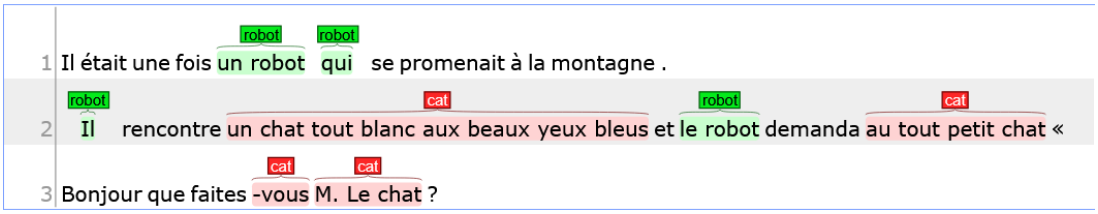


Figure 44. Annotation sur INCEpTION (corpus français, CE2, 1323) contenante un clitique, « vous » à la phrase 3.

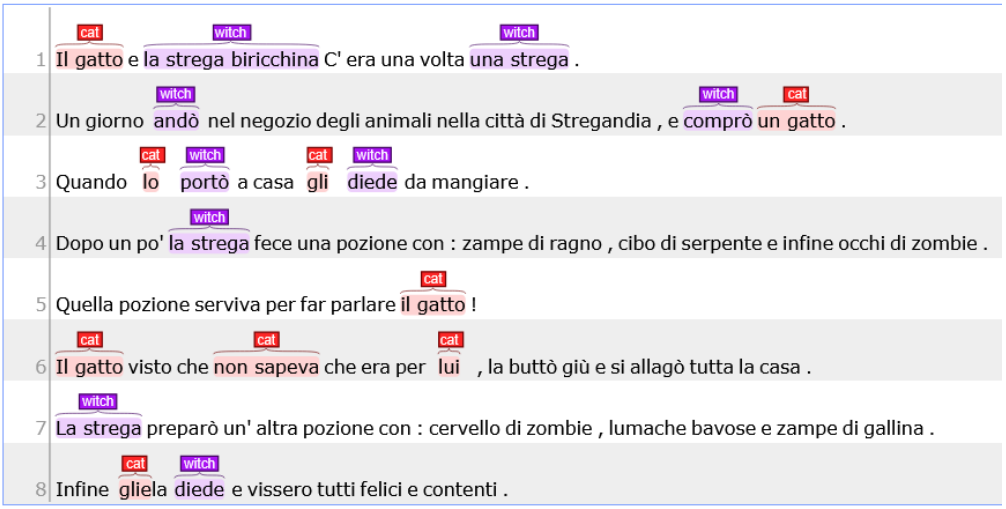


Figure 45. Annotation sur INCEpTION (corpus italien, CE2, 1427).
Un exemple d'annotation d'un clitique est présent dans la phrase 8, "glie".

En ce qui concerne le jeu d'étiquettes, il a été défini au préalable, de manière telle que les annotateurs n'avaient qu'à sélectionner l'étiquette souhaitée une fois la mention délimitée. Vu que la correspondance des étiquettes au caractère près est prise en considération dans le calcul de l'accord inter-annotateur ainsi que par la fonctionnalité de support à l'adjudication, imposer un jeu d'étiquettes fermé permet de réduire le taux d'erreur dû aux fautes de frappe dans la saisie des étiquettes que l'on avait pu observer dans les annotations issues de la première campagne d'annotation.

Ces choix ont permis d'exploiter au mieux les fonctionnalités d'adjudication semi-automatisée et de contrôler l'accord inter-annotateur au cours du processus d'annotation. Les textes ont été exportés au format UIMA-CAS 1.0 et traités à l'aide de la librairie *dkpro-cassis*[84]. Les liens de

[84] La documentation de la librairie est disponible au lien suivant : https://pypi.org/project/dkpro-cassis/

coréférence n'ont pas été exploités dans cette analyse, l'objectif étant ici de travailler à partir des étiquettes de référents seules, suffisantes pour reconstituer les chaines de continuité référentielle.

Nous avons en outre retenu cette plateforme d'annotation pour une fonctionnalité cruciale lors de l'annotation d'un corpus, à savoir la possibilité d'accéder à tout moment de la campagne au calcul du taux de l'accord inter-annotateur.

Celui-ci a été mesuré à l'aide de l'alpha de Krippendorff (Krippendorff, 1980) directement calculé par la plateforme. Il s'agit d'une mesure statistique de l'accord inter-annotateur, applicable à un nombre variable d'annotateurs, qui corrige le taux d'accord observé en tenant compte de l'accord attendu par hasard. Le choix s'est porté sur cette métrique en raison de sa disponibilité sur la plateforme d'annotation et de suivi des annotations. De plus, cette métrique permet de gérer les cas d'annotations manquantes, ce qui peut se produire dans le cas d'annotations discontinues comme celle de la coréférence, contrairement aux kappas de Cohen (Cohen, 1960) et de Fleiss (Fleiss, 1971), qui présupposent l'évaluation des mêmes items par le même nombre d'annotateurs.

Les résultats obtenus indiquent un accord moyen de 0,86 pour le français et de 0,75 pour l'italien. Bien qu'une légère baisse soit observée pour l'italien, ces valeurs restent globalement satisfaisantes au regard de la complexité de la tâche. Dans une perspective d'amélioration, l'utilisation de la métrique gamma (Mathet et al., 2015) est envisagée, celle-ci étant mieux adaptée à la prise en compte des frontières des mentions.

La phase d'adjudication a ensuite été réalisée par l'une des annotatrices, en collaboration avec deux enseignants-chercheurs responsables des corpus français et italien. Cette étape a conduit à la constitution d'un corpus de référence (*gold standard*). Deux textes du corpus français ont été exclus à ce stade en raison de leur faible cohérence, rendant leur annotation particulièrement difficile.

Enfin, contrairement à la première campagne, pour laquelle le calcul de l'accord inter-annotateur n'avait pu être réalisé de manière fiable, cette deuxième campagne a permis d'intégrer pleinement cette dimension dans l'évaluation de la qualité des annotations, grâce aux fonctionnalités offertes par la couche d'annotation personnalisée.

4.2. Synthèse : constitution et caractérisation du corpus annoté – deuxième campagne d'annotation

Dans la recherche de l'équilibre entre représentativité et exigences d'annotation, la solution proposée pour la sélection des textes à annoter était déjà de limiter nos annotations à deux niveaux scolaires, le CE1 et le CE2.

Au début de ce travail de thèse, le corpus français était déjà finalisé, tandis que les textes recueillis en Italie étaient encore en phase de traitement. En particulier, une grande majorité des textes de CP et une partie importante des textes ce CE1 avaient déjà été numérisés. En revanche, très peu des textes de CE avaient déjà fait l'objet des traitements de numérisation nécessaires à l'inclusion dans le corpus puis à leur analyse (décrits dans le Chapitre 4). Ainsi, si le corpus français pouvait déjà être exploité, le corpus italien était encore en cours de constitution. Il convient de noter que,

dans ce contexte, l'équipe responsables du recueil et du traitement des textes ne bénéficie pas d'un financement dédié à ces opérations. Le nombre restreint de chercheuses, étudiants et étudiantes travaillant sur ce projet explique les difficultés rencontrées dans le traitement des textes.

En ce qui concerne les étapes de pré-traitement des textes, elles restent globalement similaires à celles effectuées pour la première campagne d'annotation. Les textes sélectionnés étaient au nombre de 300, de niveau CE1 et CE2 en français et en italien, réduits à 298. Voyons maintenant plus en détail le processus de sélection des textes à annoter pour ces deux corpus.

Corpus français

Dans le cas du français, il a été possible de sélectionner les textes selon un critère de longueur cohérent avec celui adopté lors des campagnes précédentes. Les textes ont ainsi été choisis de manière à reproduire, autant que possible, la distribution du corpus longitudinal en termes de nombre de tokens pour les niveaux scolaires retenus.

Les textes de moins de 20 tokens ont été exclus de l'échantillon, car jugés trop courts pour permettre l'identification de chaines référentielles. Le Tableau 28 présente le détail de distribution pour le corpus longitudinal et comment cette distribution a été reproduite, avec des ajustements, sur le corpus à annoter. La Figure 46 présente la distribution des textes en nombre de tokens dans le corpus longitudinal français et dans le corpus sélectionné pour l'annotation. La comparaison de ces distributions montre que les proportions de textes répartis dans les différentes tranches de longueur restent globalement très similaires.

Nombre de tokens par texte	**Distribution des textes dans les corpus**			
	CE1 – 2		**CE2 – 3**	
	longitudinal	**gold**	**longitudinal**	**gold**
1-20	19	0	1	0
21-50	103	19	31	3
51-100	164	34	106	19
101-150	43	16	120	24
151-200	7	2	56	20
201-250	1	3	13	5
251-plus	0	0	10	3
Total	337	74	337	74

Tableau 28. Comparaison entre la distribution par longueur des textes dans le corpus longitudinal et dans le corpus à annoter du français.

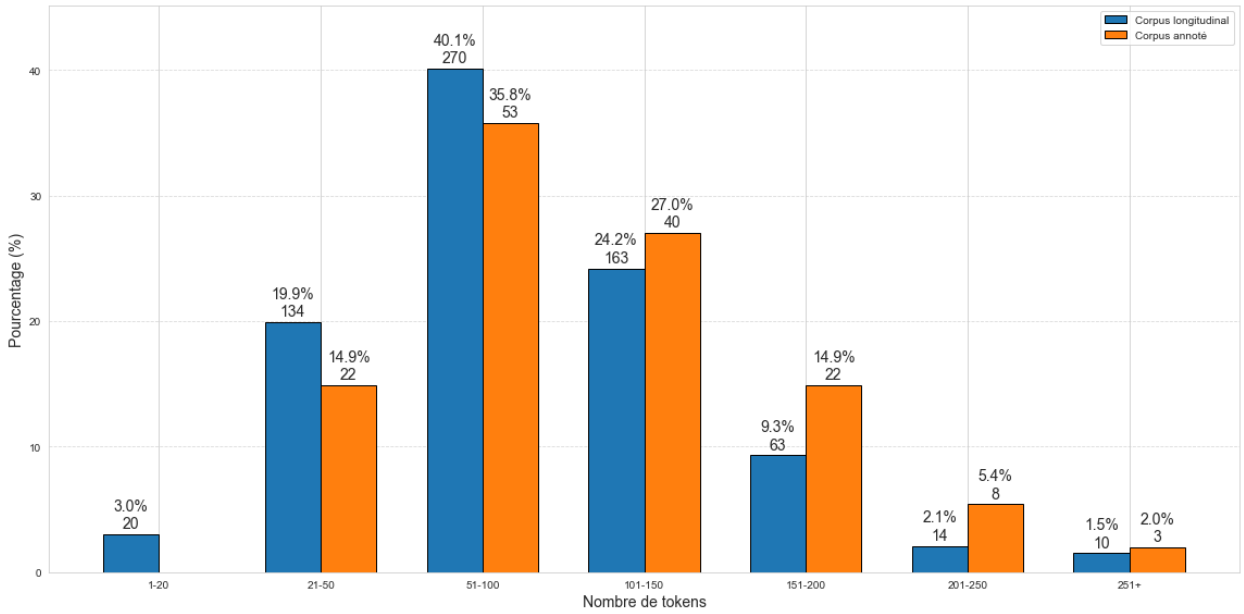


Figure 46. Distribution des textes en nombre de tokens - comparaison entre niveaux CE1 et CE2 du corpus longitudinal français et niveaux CE1 et CE2 du corpus sélectionné pour l'annotation de la deuxième campagne.

Corpus italien

Dans le cas du corpus italien, l'application de ce même critère de sélection n'a pas été possible en raison de l'état d'avancement du corpus au moment de l'annotation. Effectivement, une part importante des textes n'avait pas encore fait l'objet des étapes de numérisation, de transcription et de normalisation nécessaires à leur exploitation, ce qui n'a pas rendu possible la même caractérisation du corpus longitudinal effectuée pour le corpus français.

Pour obtenir un échantillon satisfaisant de textes de niveau CE1, dans le cadre de ce travail de thèse, nous avons contribué directement aux étapes de pré-traitement nécessaires à la numérisation des textes italiens. Nous avons ainsi participé à la numérisation, à la transcription et à la normalisation de textes de niveau CE2 (troisième année). Ce travail a été réalisé avec le soutien d'une étudiante de Master 1 en Industries de la Langue, locutrice italophone, rémunérée en vacations.

Vingt-cinq heures de vacations ont permis de transcrire et de vérifier 50 textes de CE2, tandis qu'environ 80 heures de travail supplémentaires ont été nécessaires pour compléter la transcription d'un second ensemble de 50 textes, ainsi que pour assurer la normalisation de l'ensemble des 100 textes et la vérification des transcriptions et normalisations.

Conclusion

Cette deuxième partie avait pour objectif de présenter le cadre méthodologique et les étapes de constitution du corpus annoté qui constitue la base empirique de ce travail de thèse. Nous avons décrit ic les différentes phases ayant mené à l'élaboration de la version annotée en continuité référentielle d'une partie du corpus *Scolinter*, depuis l'historique du recueil et de la constitution du corpus jusqu'à l'organisation des campagnes d'annotation, l'élaboration du guide et du schéma d'annotation, ainsi que les procédures d'adjudication nécessaires à l'obtention de notre *corpus gold*.

Le Chapitre 4 a d'abord retracé l'histoire et les conditions de constitution du corpus *Scolinter*, en montrant comment celui-ci s'inscrit dans la continuité de plusieurs projets successifs consacrés à l'étude des pratiques d'écriture à l'école primaire. La présentation des projets « Lire et écrire au CP » (Goigoux, 2016), *Scoledit* (Brissaud & Ponton, 2026; Wolfarth, 2019) et *Scolinter* (LIDILEM, 2025; Ponton et al., 2021) a permis de situer ce dernier corpus dans la prolongation des recherches longitudinales visant à étudier le développement des compétences d'écriture des élèves tout au long de l'école primaire. L'intégration de données recueillies en France, en Italie et en Espagne permet par ailleurs d'envisager des analyses contrastives entre langues et contextes éducatifs différents, tout en reposant sur un protocole de recueil commun garantissant la comparabilité des textes recueillis. Cette description a également mis en évidence les enjeux liés au recueil de corpus scolaires de grande ampleur, ainsi que les contraintes spécifiques associées à ce type de données, notamment en termes de ressources humaines et de temps nécessaires pour finaliser un tel corpus et en permettre l'exploitation et la diffusion.

Les différentes étapes de traitement et de numérisation des textes manuscrits du corpus ont ensuite été décrites. Ces étapes sont fondamentales pour permettre l'exploitation du corpus dans une perspective d'analyse linguistique. La numérisation, la transcription et la normalisation des textes constituent en effet des opérations essentielles pour rendre les données manipulables à l'aide d'outils informatiques et permettre leur annotation ultérieure. La présentation du protocole de transcription adopté, inspiré du modèle linéaire semi-diplomatique, a permis de montrer comment les choix de représentation des phénomènes graphiques, textuels et métatextuels ont été alignés avec les objectifs linguistiques du projet. Ces choix visent à préserver au mieux les caractéristiques des productions originales tout en assurant la lisibilité et l'exploitabilité du corpus. La phase de normalisation du corpus permet également de disposer d'une version plus facilement lisible des productions, tout en constituant une « interface entre la transcription de la production et les outils d'exploitation » (Wolfarth, 2019, p. 122). Ce sont ces versions normalisées des textes qui ont été retenues pour réaliser les annotations présentées dans ce travail. Au-delà de la description des aspects techniques de ces étapes, nous avons également souligné dans cette partie le caractère collectif et progressif de la constitution d'un corpus annoté d'écrits scolaires. Sa réalisation implique non seulement la coordination de plusieurs équipes de recherche, mais aussi la définition d'un protocole de recueil commun, la formalisation de guides de transcription et de

normalisation ainsi que la mise en place de procédures de vérification et de validation des données traitées.

L'ensemble de ces éléments méthodologiques prépare ainsi la mise en œuvre de l'annotation en continuité référentielle, qui constitue le cœur de ce travail de thèse.

Les Chapitres 5 et 6 ont ensuite été consacrés à la description de la méthodologie d'annotation adoptée et à la mise en œuvre des différentes campagnes d'annotation. Le Chapitre 5 a présenté les principes généraux qui ont guidé l'annotation du corpus ainsi que la méthodologie itérative adoptée pour élaborer un guide d'annotation applicable aux deux langues étudiées. Dans cette perspective, nous avons détaillé les critères retenus pour la sélection des textes à annoter au sein du corpus longitudinal, ainsi que les différentes étapes de pré-traitement nécessaires pour préparer les données à l'annotation. Dans ce chapitre ont également été exposés la structure du guide d'annotation développé dans le cadre de cette recherche, les phénomènes linguistiques qu'il vise à représenter et les choix théoriques et méthodologiques qui ont orienté sa conception.

Le Chapitre 6 a prolongé cette réflexion en présentant l'organisation concrète des campagnes d'annotation menées dans le cadre de ce travail. Nous y avons décrit les différentes phases du processus d'annotation, depuis les premières expérimentations exploratoires jusqu'aux campagnes d'annotation proprement dites, en passant par les phases de test et d'ajustement du guide. L'ensemble de ce processus s'inscrit dans une démarche progressive et itérative, fondée sur des allers-retours entre l'annotation effective des textes, l'analyse des difficultés rencontrées et la révision du guide d'annotation. Cette méthodologie a permis de stabiliser progressivement les critères d'annotation et de garantir la cohérence des décisions prises par les annotateurs.

Enfin, nous avons présenté les procédures mises en place pour assurer la qualité et la fiabilité des annotations produites, notamment à travers l'analyse des désaccords inter-annotateurs et la phase d'adjudication ayant conduit à l'établissement d'une version de référence des annotations. L'ensemble de ces étapes a permis de constituer une ressource annotée en continuité référentielle qui constitue l'un des apports méthodologiques de ce travail et qui servira de base aux analyses présentées dans la partie suivante.

Au-delà de la description du dispositif méthodologique, cette partie contribue ainsi à documenter les conditions de constitution d'un corpus d'écrits scolaires et son annotation en continuité référentielle. Nous proposons également ici une méthodologie d'annotation adaptée à des données non normées et comparables entre langues, fondée sur une démarche itérative d'élaboration et de validation du guide d'annotation. L'ensemble de ce travail méthodologique constitue le socle empirique et technique nécessaire à l'analyse des chaines de référence présentée dans la partie suivante.

La partie suivante sera consacrée à l'analyse des chaines de référence dans le corpus annoté, afin d'examiner la manière dont les élèves gèrent la présence et l'évolution des référents dans leurs textes narratifs et dans quelle mesure leurs choix linguistiques contribuent à la cohérence et à l'interprétabilité des récits produits.

Partie 3 – Résultats

Chapitre 7. *Approche quantitative de l'analyse des chaines de continuité référentielle* *189*
1. Description des indicateurs retenus et analyses quantitatives 191
2. Interdistance 200
3. Densité référentielle 204
4. Indice de stabilité ou d'instabilité 215
5. Distribution des catégories grammaticales de mentions 222
6. Conclusion 239

Chapitre 8. *Approche qualitative de l'analyse des chaines de continuité référentielle* *242*
1. Longueur des mentions 244
2. Composition des chaines 251
3. Nombre de personnages et modes de cohabitation des chaines 261
4. Différences de longueur des textes scolaires : entre choix pédagogiques et contraintes linguistiques. 289
5. Conclusion 296

Chapitre 9. *Conclusions* *299*
1. Quelles perspectives pour un travail multidisciplinaire ? 302
2. Perspectives en linguistique de corpus et en TAL 303
3. Perspectives en didactique 304

Chapitre 7. Approche quantitative de l'analyse des chaines de continuité référentielle

Dans ce chapitre nous allons d'abord décrire les indicateurs utilisés pour la description quantitative des chaines de référence dans le corpus annoté. Pour chaque indicateur, nous avons spécifié le cadre théorique dont il découle pour enfin définir les critères opérationnels adoptés pour le calcul de chaque indice. Ces définitions serviront à comprendre et interpréter les résultats issus de notre deuxième campagne d'annotation. Grâce à celle-ci, nous avons obtenu un corpus de référence annoté en français et en italien (la campagne est décrite dans le Chapitre 6). Son analyse nous permet d'identifier les spécificités de notre corpus et les variations éventuelles entre langues et entre niveaux scolaires. Pour ce travail de définition et de description des indicateurs utilisés, nous sommes redevables aux différents travaux que nous avons abordés dans nos chapitres théoriques sur la coréférence (cf. Chapitre 1 pour la théorie en linguistique et Chapitre 3 pour l'état de l'art en TAL). Entre autres, le volume de description de fin du projet *DEMOCRAT* (Landragin, 2021) a d'abord contribué à la réflexion théorique et opérationnelle autour de l'annotation de notre corpus, et a permis ensuite de rassembler les indices décrivant les chaines de référence en français.

Les travaux effectués sur le corpus *DEMOCRAT*, ainsi que la conception linéaire des chaines décrite par Schnedecker (2021), ont aussi le mérite de fournir un point de vue à la croisée entre linguistique et TAL sur les chaines de référence comme objet de recherche. Nous avons donc choisi des indices provenant des travaux sur ce corpus, ainsi que des travaux de description des chaines de Schnedecker (entre autres, Schnedecker, 1997, 2017, 2021). Nous mentionnons aussi le travail de thèse de Federzoni (2025). Ces indices nous fournissent ainsi des pistes pour poursuivre la description, quantitative puis qualitative, de nos annotations. Enfin, les indices retenus nous permettent de comparer notre corpus avec ceux existants en littérature (de moins les corpus francophones), de grande taille ou simplement décrits dans des travaux de recherche plus limités.

Dans leur article de description des chaines de référence dans les textes encyclopédiques du corpus *DEMOCRAT*, Oberlé et al. (2018, p. 92-93) décrivent trois groupes d'indicateurs, catégorisés selon leur fonction :

- Révéler les caractéristiques référentielles du texte, à travers le nombre de référents, des expressions référentielles (ER) et à travers le ratio entre nombre d'ER et nombre de tokens dans le texte (*densité référentielle*, approfondie dans la section 3 de ce chapitre) ;
- Décrire les caractéristiques des chaines de référence, à l'aide du nombre de chaines dans le texte, leur longueur (qui correspond au nombre de mentions qui la composent), le ratio entre chaines de référence et nombre de référents dans le texte, le coefficient de stabilité et la catégorie grammaticale ou type des mentions ;
- Montrer la relation entre les chaines et la structure du texte, grâce à l'analyse de la portée des chaines (leur permanence dans le texte), l'adéquation de la chaine au topique discursif, les modalités de cohabitation ainsi que l'incidence du découpage textuel et de la catégorie des mentions.

Schnedecker propose une liste qui comprend ces mêmes indices, auxquels s'ajoutent d'autres éléments d'analyse, en tant que possible « angle d'attaque systématique » (Schnedecker, 2017, p. 68) d'analyse des chaines. Dans sa contribution, qui examine les différences de genre textuel au prisme des chaines de référence entre deux corpus, de contes de fées et de faits divers, sont aussi mentionnés comme éléments d'analyse :

- L'ordre des mentions qui constituent les premières mentions des chaines ;
- La fonction grammaticale des mentions ;
- La sous-catégorie lexicale exprimée par les mentions ;
- Le rôle sémantique et le critère sémantique des mentions « (au sens d'identification des contenus généraux : changement climatique, appareil électro-ménager, etc.) » (Schnedecker, 2017, p. 68) ;
- La nature de la proposition d'occurrence de la mentions (principale, subordonnée etc.).

Bien que tous ces critères présentent un intérêt dans l'analyse et la représentation de la coréférence, nous avons retenu pour nos analyses seulement certains d'entre eux. Nous avons notamment sélectionné les indicateurs déjà utilisés dans l'état de l'art des corpus de coréférence, pour pouvoir comparer les spécificités de notre corpus avec des corpus annotés de manière similaire mais représentatifs de types et genres de textes différents. De plus, les indices retenus sont calculables à partir des niveaux d'annotation que nous avons réalisés, les autres niveaux (lexical, sémantique) nécessitant une annotation ad hoc pour être analysés. Cependant, d'autres niveaux pourraient être explorés à terme grâce à des techniques d'annotation automatique. Par exemple, en l'état, nous n'avons pas exploité l'analyse en dépendance déjà effectuée sur notre corpus de manière automatique pour définir le statut de la mention en tant que sujet ou objet dans la phrase, ni la nature de la proposition d'occurrence de la mention. Nous décrivons ici plus en détail certains des indices retenus dont la définition et le mode de calcul sont les moins transparents, tels qu'ils ont été adoptés dans la description de notre corpus annoté. En suivant la liste des propriétés structurelles des chaines de coréférence décrites précédemment (cf. Chapitre 1, §4), nous présentons ici certains choix de définition et de mode de calcul adoptés dans notre analyse.

Dans la description des analyses qui suivent, nous précisons au cas par cas la manière dont les mentions renvoyant à plusieurs référents ont été comptabilisées. Pour rappel, dans notre schéma d'annotation, nous avons choisi de superposer les étiquettes afin de représenter les cas où une même expression référentielle renvoie simultanément à plusieurs entités. Cette solution permet d'identifier les mentions plurielles tout en conservant l'information sur chacun des référents impliqués dans la chaine. Dans la plupart des cas, lorsque plusieurs personnages sont évoqués conjointement dans une même expression référentielle (par exemple *le chat et la sorcière*), nous avons considéré cette configuration comme une chaine collective distincte des chaines des deux référents singuliers. Ainsi, si un texte contient une chaine associée au personnage *cat* et une chaine associée au personnage *witch*, les mentions renvoyant conjointement à ces deux personnages sont regroupées dans une chaine supplémentaire *cat+witch*.

Du point de vue de l'annotation, ces mentions sont ainsi rattachées aux chaines individuelles correspondant à chacun des référents afin de conserver l'information sur les participants à l'action. Toutefois, dans les calculs quantitatifs présentés dans ce chapitre, ces occurrences sont comptabilisées uniquement dans la chaine collective correspondante, afin d'éviter de compter plusieurs fois une même expression référentielle. Ce choix méthodologique permet d'éviter le double comptage d'une même expression référentielle dans plusieurs chaines et garantit ainsi une représentation plus fidèle de la distribution des mentions dans le corpus lors des analyses quantitatives.

1. Description des indicateurs retenus et analyses quantitatives

Les analyses présentées dans cette partie ont été effectuées à partir des résultats de la dernière campagne d'annotation, pour plusieurs raisons. Tout d'abord, il s'agit de la campagne dans laquelle nous avons utilisé la version la plus aboutie du guide d'annotation, amélioré au fil de plusieurs itérations d'annotation et de révision, comme décrit précédemment (cf. Chapitre 6). De plus, à l'issue de cette campagne, nous avons pu réaliser une adjudication manuelle des annotations grâce à la plateforme d'annotation utilisée. Cette étape nous a permis d'obtenir un *corpus gold*, que nous décrivons ici d'abord de manière quantitative, puis qualitative dans le chapitre suivant. Cette comparaison permet de mettre en évidence certaines différences et similitudes quantitatives entre les deux langues et les deux niveaux scolaires. En effet, le corpus annoté comprend des textes produits par des élèves de CE1 et de CE2 dans les deux langues, ce qui permet d'observer à la fois les variations entre langues et les évolutions possibles entre niveaux scolaires. Bien que l'échantillon annoté demeure limité en taille à cause des contraintes de temps et de ressources disponibles, il nous semble néanmoins suffisant pour formuler des hypothèses exploratoires concernant les variations dans la constitution des chaines de référence selon la langue et le niveau scolaire.

Le Tableau 29 présente les principales caractéristiques des textes de ce *corpus gold*. Il détaille, pour chaque langue et pour chaque niveau, le nombre de textes, le nombre total de tokens hors ponctuation qui les composent, la longueur moyenne des productions ainsi que l'étendue de la longueur des textes. Comme indiqué précédemment dans la section consacrée à la constitution du corpus (cf. Chapitre 6, §4.2), la longueur moyenne des textes, mesurée en nombre de tokens, est plus élevée dans le corpus italien (143,66 tokens) que dans le corpus français (94,94). Dans les deux langues, les productions des élèves de CE2 sont plus longues que celles de CE1. L'ampleur de cette augmentation est toutefois comparable d'une langue à l'autre, les écarts observés étant très proches. En français, l'écart entre CE1 et CE2 est de +46,7 tokens, tandis qu'en italien il atteint +52,3 tokens. Il convient de noter que la longueur moyenne des textes de CE2 français est, à un token près, identique à celle des textes italiens de niveau CE1, ce qui suggère une quasi-équivalence quantitative entre ces deux sous-corpus.

Niveaux scolaires	Nombre de textes		Nombre total de tokens		Nombre moyen de tokens par texte		Étendue de la longueur des textes	
	fr	it	fr	it	fr	it	fr	it
CE1 – 2	74	75	5 298	8 812	71,59	117,49	22-245	34-676
CE2 – 3	74	75	8 753	12 737	118,28	169,83	43-364	45-365
CE1-CE2 – 2-3	**148**	**150**	**14 051**	**21 549**	**94,94**	**143,66**	**22-364**	**34-365**

Tableau 29. Caractérisation de la longueur des textes du corpus annoté en nombre de tokens.

Il convient ici de mentionner que nous nous sommes questionnée sur les possibles influences de la langue analysée sur la longueur des textes de notre corpus[85]. Nous avons ainsi cherché à déterminer si les différences observées dans la taille des textes entre français et italien pouvaient être liées à un phénomène de foisonnement linguistique.

Le taux ou coefficient de foisonnement est défini ici comme la variation de la longueur d'un texte à la suite de sa traduction. Cette variation peut être positive ou négative. Pour le calculer, nous avons appliqué la formule suivante :

$$taux\ de\ foisonnement = \frac{nb\ tokens\ texte\ cible - nb\ tokens\ texte\ source}{nb\ tokens\ texte\ source} * 100$$

Afin d'obtenir une estimation de ce coefficient entre le français et l'italien, nous avons d'abord tenté d'effectuer ce calcul à partir de certains corpus parallèles disponibles dans la collection Opus[86]. Nous nous sommes finalement tournés vers *Europarl* (Koehn, 2005), un corpus aligné constitué des procès-verbaux du Parlement européen et disponible en 11 langues européennes. Ce corpus contient des textes parallèles en français et en italien, il permet ainsi d'observer les différences de longueur entre les deux langues pour un même texte.

Les sous-corpus français et italien contiennent respectivement 53 095 838 et 49 021 839 tokens[87]. À partir de ces données, nous obtenons un taux de foisonnement d'environ -8% du français vers l'italien. Autrement dit, dans ce corpus, l'italien utilise en moyenne moins de mots que le français pour exprimer le même message. Un tel taux négatif n'est donc pas explicatif de la longueur supérieure des textes italiens observée dans notre corpus. Pour cette raison, nous n'avons pas pris ce facteur en compte dans l'interprétation des résultats et avons privilégié une présentation des données en valeurs absolues, notamment en termes de nombre de tokens ou de nombre d'annotations. Il convient toutefois de souligner que le corpus *Europarl* est représentatif d'un registre discursif spécifique, celui du langage politique et institutionnel. Les caractéristiques linguistiques de ce type de discours peuvent différer sensiblement de celles des textes narratifs produits par des élèves à l'école primaire. Les estimations issues de ce corpus doivent être interprétées avec prudence et ne peuvent constituer qu'un point de comparaison indicatif.

[85] Nous remercions ici les membres du CLLE de l'Université Jean Jaurès de Toulouse pour nous avoir proposé cette réflexion et pour nous avoir poussée à vérifier certaines hypothèses vis-à-vis des effets de langue dans ce travail de thèse.

[86] Ce site recueille plusieurs corpus parallèles. Il est disponible au lien suivant : https://opus.nlpl.eu/

[87] Le nombre de tokens provient d'Opus pour Europarl. Consulté le 23/11/2025.

Après avoir clarifié certains aspects liés à la nature du corpus objet de notre annotation, nous nous sommes ensuite penchée sur l'analyse quantitative des annotations effectuées. Le Tableau 30 décrit la répartition des annotations dans le corpus. En plus de détailler le nombre total d'expressions référentielles (ou mentions) annotées, par langue et par niveau scolaire, nous distinguons ici trois catégories : les *chaines*, c'est-à-dire les groupes de trois mentions ou plus qui indiquent le même référent, les *paires anaphoriques ou coréférentielles*, où deux mentions font référence au même individu, et les *singletons*, où le référent est nommé une seule fois dans le texte.

Niveaux scolaires	**Tot. mentions**		**3 mentions et plus**		**2 mentions**		**1 mention**	
	fr	it	fr	it	fr	it	fr	it
CE1 – 2	1 197	1 992	150	194	31	28	41	53
CE2 – 3	1 944	2 837	192	216	41	43	50	71
CE1-CE2 – 2-3	**3 141**	**4 829**	**342**	**410**	**72**	**71**	**91**	**124**

Tableau 30. Nombre d'expressions référentielles annotées et distribution des chaines (≥3 mentions, paires, singletons) par langue et par niveau scolaire.

En italien, nous pouvons observer une quantité de mentions annotées plus élevée qu'en français, avec une différence de 1 688 mentions en faveur de l'italien. De manière similaire, le corpus italien présente 410 chaines de coréférence, contre 342 pour le corpus français. Le nombre d'anaphores (paires anaphoriques ou coréférentielles) est pratiquement équivalent dans les deux langues, bien que l'italien se distingue par un plus grand nombre de singletons (124 contre 91 en français).

De manière similaire à ce que l'on observe pour la longueur des textes (voir Tableau 29), les valeurs obtenues pour certains indicateurs apparaissent proches entre le CE2 français et le CE1 italien. En particulier, le nombre total d'expressions référentielles annotées en CE2 français (1 944) est très proche de celui observé en CE1 italien (1 992), et l'on observe une tendance comparable pour le nombre de chaines de coréférence (192 en CE2 français contre 194 en CE1 italien). Cette proximité quantitative suggère que, pour ces indicateurs, les productions du CE2 français présentent des caractéristiques proches de celles observées dans les textes italiens de CE1.

1.1. Distribution des mentions par texte

Dans un deuxième temps, nous avons souhaité vérifier si ces différences sont effectivement dues à la différence de longueur des textes ou si d'autres facteurs peuvent l'expliquer. La distribution des mentions par texte semble en effet suivre la tendance observée pour la longueur des textes. Le Tableau 31 présente les valeurs de l'étendue et de la moyenne du nombre de mentions annotées dans les corpus français et italien, pour les niveaux CE1 et CE2. On observe que les textes italiens présentent une plus grande variabilité sur ce dernier indice que les textes français, notamment au niveau CE1 (étendue : 7–115 en italien, 4–44 en français). Les moyennes confirment également cette tendance : les textes italiens comportent en moyenne davantage de mentions que les textes français, aussi bien au niveau CE1 (26,54 en italien contre 16,17 en français) qu'au niveau CE2 (37,82 en italien contre 26,27 en français). Lorsque les deux niveaux sont considérés

conjointement, la moyenne italienne (32,18) reste supérieure à celle du français (21,22), tandis que l'étendue (7–115 en italien contre 4–75 en français) reflète une dispersion plus forte du nombre de mentions annotées en italien.

Niveaux scolaires	**Nombre de mentions par texte**					
	Moyenne		**Étendue**		**Écart type**	
	fr	it	fr	it	fr	it
CE1 – 2	16,18	26,55	4–44	7–115	9,28	16,02
CE2 – 3	26,27	37,83	6–75	8–74	12,78	15,06
CE1-CE2 – 2-3	**21,22**	**32,18**	**4–75**	**7–115**	**12,23**	**16,50**

Tableau 31. Distribution des mentions annotées par texte (singletons, anaphores et chaines confondues).

Comme dans le cas de la longueur des textes, nous pouvons nous demander si ces différences dans la distribution des mentions entre niveaux scolaires sont liées à l'augmentation de la longueur des textes précédemment évoquée, ou si elles peuvent être dues à une augmentation intrinsèque du nombre de mentions proportionnel au niveau scolaire, donc à une majeure complexité de la gestion référentielle. Nous allons approfondir cette question grâce à l'analyse de la densité référentielle, indice qui met en corrélation ces deux facteurs au sein des annotations (cf. ce chapitre, §3).

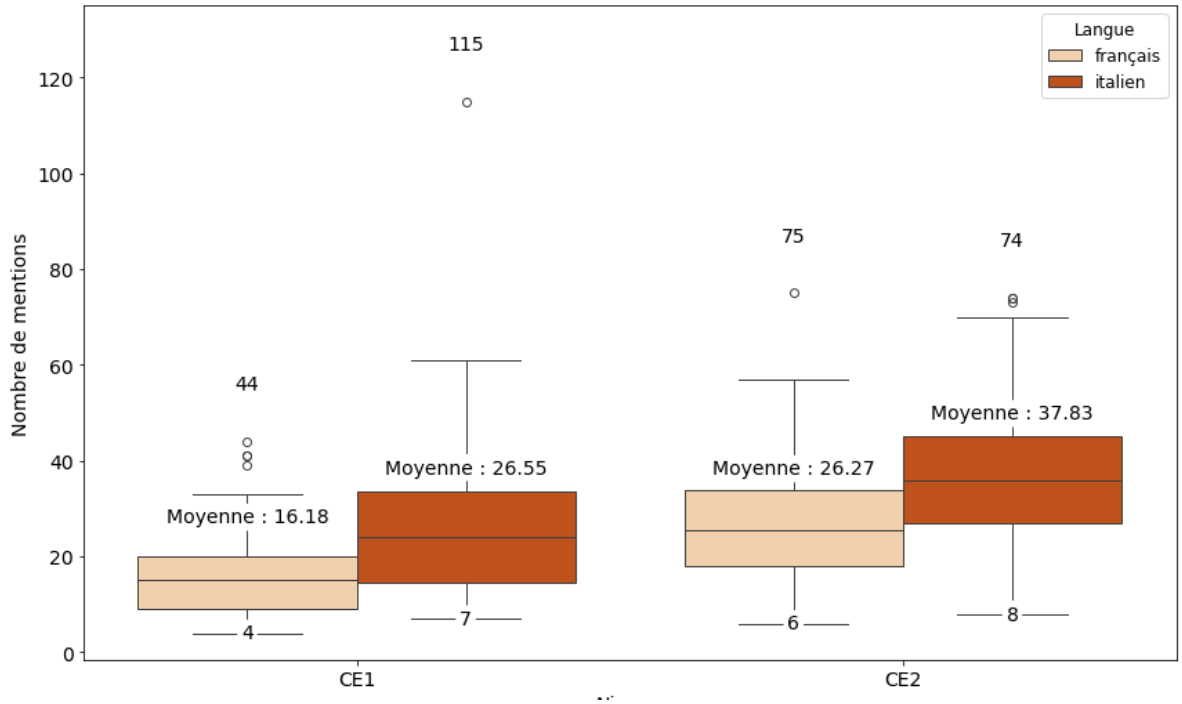


Figure 47. Distribution des mentions par texte dans le corpus annoté en français et en italien.

La Figure 47 précise la distribution des mentions dans les textes. En complément du Tableau 31, elle met en évidence le décalage entre les deux langues à parité de niveau scolaire, et la quasi-correspondance entre les valeurs des niveaux CE1 italien et CE2 français. Les résultats confirment également l'existence d'un effet du niveau scolaire dans les deux langues, les distributions de CE2 étant systématiquement situées au-dessus de celles de CE1. Enfin, la variabilité apparait plus

marquée dans le corpus italien, en particulier en CE1, comme en témoignent l'extension plus large des distributions et la présence de valeurs extrêmes élevées. Cette hétérogénéité interne est visible tant dans l'étendue que dans la dispersion interquartile. Nous pouvons aussi remarquer la similitude entre la distribution du CE2 français et le CE1 italien, qui présentent aussi pour cet indice une moyenne très similaire.

La comparaison entre niveaux scolaires met en évidence une progression quantitative du nombre de mentions par texte relativement limitée entre CE1 et CE2 dans chacune des deux langues. Les textes produits par les élèves de CE2 comportent en moyenne davantage de mentions que ceux de CE1, ce qui reflète en partie l'augmentation générale de la longueur des productions.

Un résultat particulièrement intéressant est la quasi-correspondance observée sur cet indice entre les valeurs du CE1 italien et celles du CE2 français. Cette convergence suggère que les différences observées entre les deux langues peuvent parfois être plus marquées que celles liées au niveau scolaire interlangue. Autrement dit, certaines propriétés référentielles des textes produits par les élèves italiens de CE1 apparaissent proches de celles observées dans les productions françaises de CE2.

1.2. Longueur des mentions

En ce qui concerne la longueur des mentions, il convient tout d'abord de rappeler que la taille minimale d'une mention correspond à un token. Dans certains cas, notamment en italien, ce token peut lui-même correspondre à un sous-token, comme dans le cas des pronoms clitiques. La longueur des mentions est ainsi calculée en nombre de tokens, hors ponctuation.

	Moyenne		**Étendue**	
Niveaux scolaires	fr	it	fr	it
CE1 – 2	1,54	1,66	1-9	1-15
CE2 – 3	1,59	1,66	1-11	1-10
CE1-CE2 – 2-3	**1,56**	**1,66**	**1-11**	**1-15**

Tableau 32. Longueur des mentions annotées (nombre de tokens) par langue et niveau scolaire.

Les principales statistiques descriptives de la longueur des mentions sont présentées dans le Tableau 32, qui indique pour chaque langue et chaque niveau scolaire la longueur moyenne des mentions ainsi que leur étendue. On observe que les mentions les plus longues en français sont composées de 11 tokens et apparaissent dans des textes de niveau CE2, tandis qu'en italien la mention la plus longue comporte 15 tokens et se rencontre dans un texte de niveau CE1. Les longueurs moyennes des mentions restent relativement similaires entre CE1 et CE2 dans les deux langues : respectivement 1,54 et 1,59 tokens en français, et 1,66 token pour les deux niveaux en italien.

La Figure 48 illustre plus précisément la distribution des mentions selon leur longueur dans les deux langues. Elle met en évidence la forte concentration des mentions de petite taille dans les deux corpus. Cette stabilité s'explique par la forte prédominance des mentions composées d'un seul token. Dans les deux langues, les mentions constituées d'un token représentent la grande

majorité des occurrences (56,3 % du total en français et 51,1 % en italien), suivies par celles qui en comportent deux (35,3 % en français et 37,8 % en italien). Les mentions comprenant trois tokens ou plus demeurent en revanche minoritaires : elles représentent 8,4 % dans le corpus français (264 mentions au total) et 11,1 % dans le corpus italien (536 mentions au total). Les longueurs moyennes des mentions restent très proches entre CE1 et CE2 dans les deux langues. Cette stabilité suggère que l'évolution entre ces deux niveaux scolaires ne concerne pas tant la complexité interne des expressions référentielles.

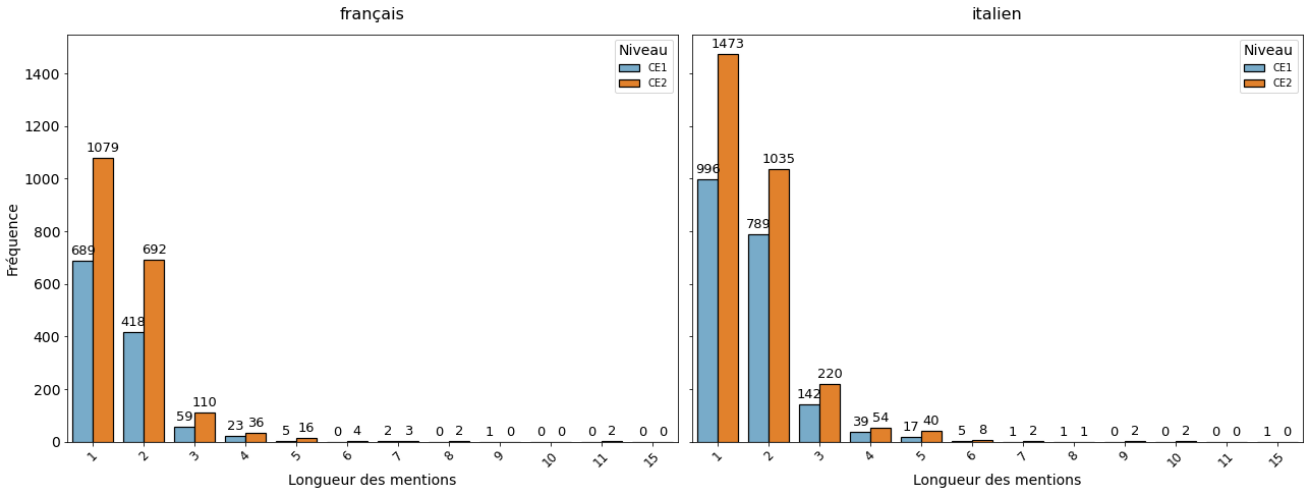


Figure 48. Distribution des mentions annotés par longueur (nombre de tokens). Comparaison entre niveaux par langue.

Dans la partie consacrée aux analyses qualitatives, nous reviendrons plus en détail sur la structure interne ainsi que sur les fonctions discursives possibles de ces mentions plus longues. Bien que leur présence demeure marginale dans le corpus, elles peuvent constituer des éléments particulièrement intéressants du point de vue de la description linguistique, mais aussi pour les retombées didactiques que ces observations peuvent suggérer (cf. Chapitre 8, §1).

1.3. Nombre de référents dans les textes

Le Tableau 33 synthétise plusieurs indicateurs relatifs à la distribution des référents dans les textes français et italiens aux niveaux CE1 et CE2 : le nombre moyen de référents par texte, ainsi que l'étendue (valeurs minimale et maximale) du nombre de référents observés par texte.

Niveaux scolaires	**Nombre total de référents annotés**		**Nombre moyen de référents/texte**		**Etendue**	
	fr	it	fr	it	fr	it
CE1 – 2	222	275	3	3,66	1-8	1-7
CE2 – 3	283	330	3,82	4,4	1-11	2-9
CE1-CE2 – 2-3	**505**	**605**	**3,41**	**4,03**	**1-11**	**1-9**

Tableau 33. Nombre de référents annotés par texte : total, moyenne et étendue selon la langue et le niveau scolaire.

On observe tout d'abord que le nombre moyen de référents par texte est légèrement plus élevé en italien qu'en français, aussi bien au niveau CE1 (3,66 contre 3) qu'au niveau CE2 (4,4 contre 3,82). En regardant plus en détail la distribution des textes par nombre de référents mobilisés (Tableau 34 et Figure 49), nous pouvons constater que la consigne d'origine, qui demandait de mentionner

seulement deux personnages dans son texte, a été respectée dans une grande partie des textes dans les deux langues (65 textes du corpus annoté français et 59 du corpus annoté italien). En français, le texte le plus « peuplé » comprend 11 personnages, alors qu'il est à 9 personnages en italien. Dans les deux langues, on retrouve quelques textes avec un seul protagoniste : huit au total en français (dont un seul en CE2) et trois en italien (en CE1). La gestion à deux référents apparait comme la configuration la plus fréquente dans les deux langues (43,92% du total des textes en français, 39,33% en italien), suivie par les textes contenant trois puis quatre personnages.

Nombre des référents par texte	**français**				**italien**			
	CE1		**CE2**		**CE1**		**CE2**	
		%		%		%		%
1	7	4,73	1	0,68	3	2	--	--
2	*36*	*24,32*	*29*	*19,59*	*28*	*18,67*	*31*	*20,67*
3	17	11,49	17	11,49	23	15,33	16	10,67
4	8	5,41	18	12,16	10	6,67	12	8
5	3	2,03	4	2,7	7	4,67	2	1,33
6	2	1,35	2	1,35	2	1,33	6	5
7	--	--	2	1,35	2	1,33	5	3,33
8	1	0,68	--	--	--	--	1	0,67
9	--	--	--	--	--	--	2	1,33
10	--	--	--	--	--	--	--	--
11	--	--	1	0,68	--	--	--	--

Tableau 34. Distribution des textes par nombre de référents - comparaison entre corpus.

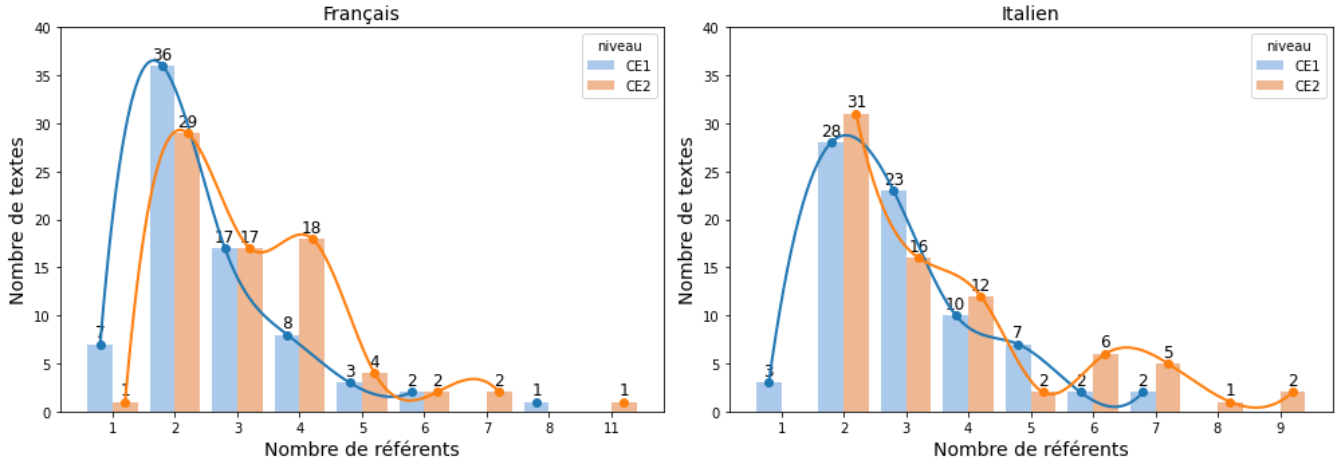


Figure 49. Distribution des référents dans le corpus annoté.

Nous nous sommes en outre intéressée aux textes ne contenant que deux personnages (ou deux référents) afin d'examiner quels binômes de personnages apparaissent plus fréquemment dans les textes. Une étude antérieure menée sur le corpus français par Ponton et al. (2019), fondée sur une exploration textométrique du corpus, a montré qu'au niveau lexical les lemmes *chat* et *sorcière* figuraient parmi les noms les plus fréquents dans le corpus. Les auteurs mettaient aussi bien en évidence un effet du genre du scripteur marqué dans le choix du lexique : *robot* et *loup* étaient les lemmes de type noms les plus spécifiques dans les écrits des garçons, tandis que les lemmes *sorcière* et *chaton* ce sont les plus spécifiques dans les écrits des filles. Cependant, nous n'avons pas à disposition le genre des élèves dans les métadonnées des textes italiens, et pour cela nous ne pouvons pas effectuer une comparaison avec cette étude sur l'influence du genre de l'élève dans le choix des personnages qui interviennent dans le texte. Néanmoins, nous pouvons vérifier si ces deux binômes, qui émergent à l'aide des analyses textométriques, sont aussi bien représentés au niveau du choix des binômes de référents dans les textes des deux corpus.

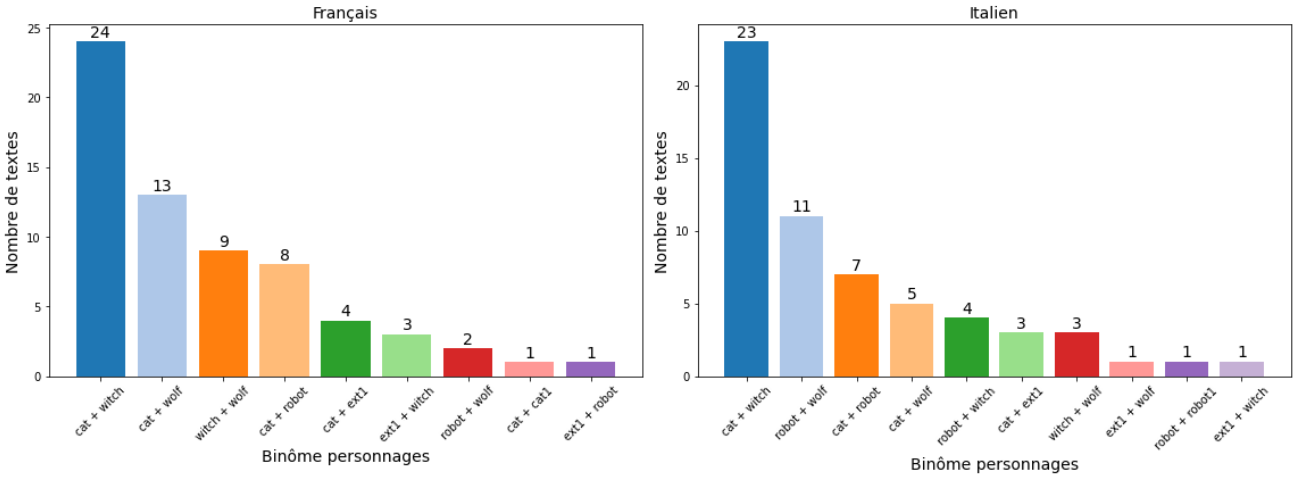


Figure 50. Distribution des binômes de personnages les plus fréquents dans les textes.

Effectivement, comme le montre la Figure 50, la distribution des binômes de personnages est fortement concentrée dans les deux corpus. Dans les deux langues, *chat/sorcière* constitue le binôme dominant (24 textes en français, 23 en italien). En français, il est suivi par *chat/loup* et *sorcière/loup*, tandis qu'en italien la deuxième position est occupée par *robot/loup*. Malgré ces différences de rang, les mêmes personnages principaux structurent l'ensemble des combinaisons, les autres binômes restant marginaux.

Cette analyse quantitative a servi de point de départ à une étude qualitative des différences de configuration des chaines de référence, selon que la narration met en jeu un seul personnage ou qu'elle implique la gestion de deux référents, de même genre ou de genres différents. Cette étude sera présentée dans une section du chapitre suivant, consacrée aux analyses qualitatives du corpus (cf. Chapitre 8, §3).

1.4. Longueur des chaines

Nous nous intéressons à présent plus spécifiquement à la longueur des chaines de référence (mesurée en nombre de mentions). La longueur des chaines permet de hiérarchiser l'importance d'un référent donné dans un texte, en partant du postulat que la chaine la plus longue correspond

habituellement au personnage « central » d'un texte, ou le plus saillant (Schnedecker, 2021, p. 89). Dans le Tableau 35, nous décrivons cet aspect des chaines dans une perspective fine, afin de proposer une caractérisation plus détaillée. Ces indices nous permettent de comparer les moyennes ainsi que la dispersion des données et la forme de leur distribution.

Niveaux scolaires	Max		Moyenne		Médiane		Écart type		Variance		Skewness		Kurtosis[88]	
	fr	it	fr	it	fr	it	fr	it	fr	it	fr	it	fr	it
CE1	30	58	7,29	9,71	6	8	4,17	7,06	17,39	49,76	2,02	2,67	6,58	12,75
CE2	35	35	9,44	12,41	8	11	6,04	7,84	36,43	61,49	1,35	0,85	2,16	0,07

Tableau 35. Indices de descriptions - longueur des chaines en termes de mentions.

Concernant la longueur moyenne et maximale des chaines, les textes italiens se distinguent par des valeurs supérieures. Les chaines y sont en moyenne plus longues : 9,71 contre 7,29 pour le français en CE1 ; 12,41 contre 9,44 en français en CE2. Elles peuvent également atteindre des longueurs maximales nettement supérieures, notamment au niveau CE1 (58 en italien, 30 en français). Ces résultats suggèrent, à première vue, une plus grande portée des chaines en italien. Il conviendra toutefois de vérifier ultérieurement à travers d'autres indices dans quelle mesure les chaines les plus longues correspondent effectivement à des référents particulièrement saillants ou fréquemment mobilisés dans le texte. La longueur maximale des chaines est systématiquement plus élevée en italien qu'en français, atteignant 58 mentions dans le corpus italien contre 30 à 35 mentions dans le corpus français selon le niveau.

Comme nous pouvons remarquer à travers la comparaison entre moyenne et médiane des longueurs des chaines, ainsi qu'à travers les distributions représentées dans les Figure 51 et Figure 52, globalement dans les deux corpus les distributions sont asymétriques à droite. Cela indique la présence d'une majorité de chaines courtes et d'un nombre plus restreint de chaines très longues. Cette distribution est d'ailleurs confirmée par les valeurs d'asymétrie positives[89] obtenues. Par ailleurs, la comparaison entre les niveaux scolaires met en évidence une augmentation de la longueur des chaines au niveau CE2 dans les deux langues, suggérant une tendance à la production de chaines plus étendues avec l'avancement de niveau scolaire. Enfin, les valeurs d'écart type relativement élevées, en particulier en italien (7,06 en CE1 et 7,84 en CE2, contre respectivement 4,17 et 6,04 en français), indiquent une dispersion importante des longueurs de chaines autour de la moyenne. Cette variabilité, plus marquée dans le corpus italien, reflète une plus grande hétérogénéité dans la longueur des chaines en italien, en cohérence avec la longueur des textes plus étendues observée dans cette langue.

[88] La *kurtosis* (ou coefficient d'aplatissement) mesure la forme de la distribution et le poids de ses queues. Une valeur proche de 0 correspond à une distribution proche de la loi normale, tandis que des valeurs positives ou négatives indiquent respectivement une distribution plus pointue ou plus aplatie que la normale (« Kurtosis », 2025).

[89] Le coefficient d'asymétrie, ou *skewness*, mesure le degré d'asymétrie d'une distribution d'une variable aléatoire réelle. L'asymétrie d'une distribution est généralement positive si la queue de droite, correspondante aux valeurs plus hautes, est plus longue ou grosse, et négative si la queue de gauche, des valeurs plus petites, est plus longue ou grosse (« Asymétrie (statistiques) », 2026).

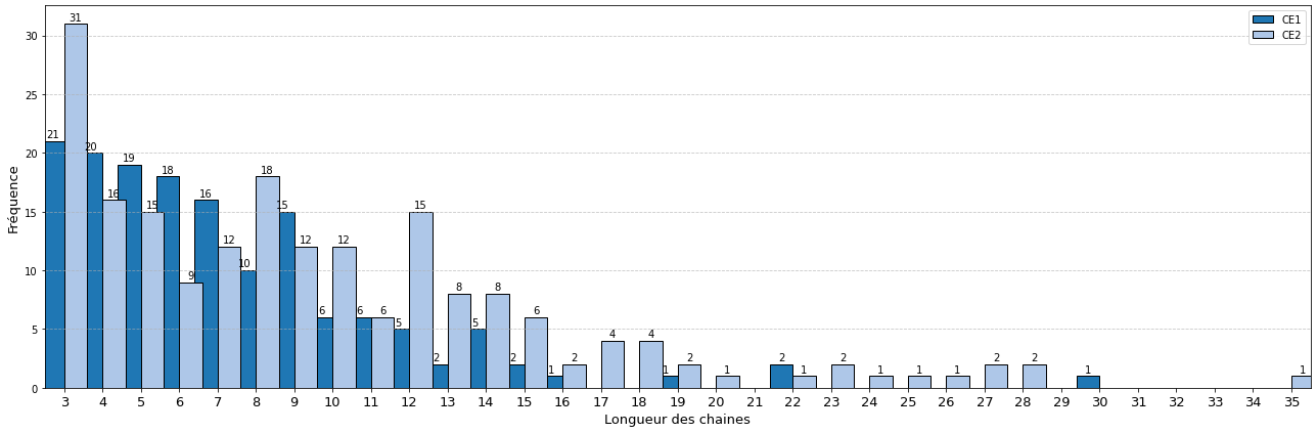


Figure 51. Distribution de la longueur des chaines dans le corpus français.

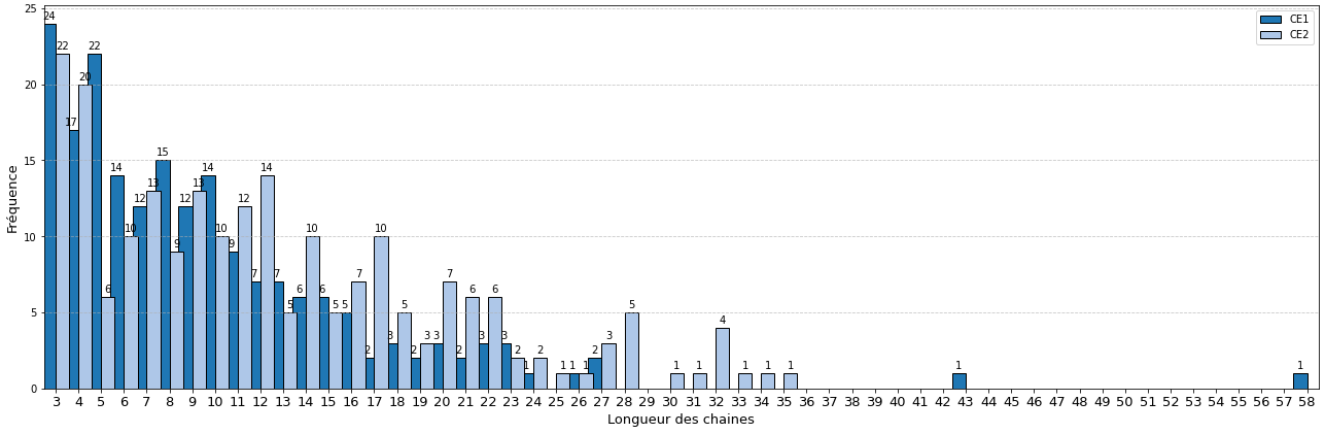


Figure 52. Distribution de la longueur des chaines dans le corpus italien.

Deux autres indices permettent d'approfondir l'analyse de certaines propriétés déjà mises en évidence. Il s'agit, d'une part, d'examiner si la longueur des textes est effectivement proportionnelle à leur contenu référentiel et, d'autre part, d'observer la manière dont les expressions référentielles se distribuent le long du texte. Ces deux questions peuvent être abordées respectivement à l'aide des indices de densité référentielle et d'interdistance.

2. Interdistance

Selon la théorie de l'accessibilité d'Ariel (1990), la distance entre une mention et son antécédent est un des facteurs qui déterminent l'accessibilité du référent. Ainsi, plus les deux expressions sont distantes, moins le référent sera considéré comme accessible, ce qui va influencer le choix d'un type de maillon plus informatif, moins rigide, moins atténué. Le rôle de la distance est aussi analysé par Schnedecker (1997), selon laquelle plus les mentions d'une chaine sont proches entre elles, plus la chaine semble solide, ce qui peut fournir des indices sur le degré de cohésion de la chaine ou encore sur la saillance du référent mentionné.

Cet indicateur, également désigné sous le terme de *distance inter-maillonnaire ou interdistance*, correspond à la somme des intervalles séparant toutes les mentions d'une chaine dans un texte, mesurée en nombre de tokens ou de phrases. Toutefois, le critère fondé sur le nombre de phrases n'a pas été retenu dans notre approche, compte tenu de la longueur limitée des textes analysés.

Nous avons également calculé le *ratio d'interdistance*, qui correspond au ratio entre la distance totale en nombre de mots (tokens hors ponctuation) et le nombre de mentions dans la chaine. Cette définition opérationnelle du taux d'interdistance provient de la publication de Rousier-Vercruyssen et Landragin (2019), où l'interdistance est ainsi définie à travers un exemple d'annotation.

Dans l'exemple (67), qui comprend 48 tokens au total, la chaine *cat* est composée de neuf mentions et la distance en termes de tokens entre ces neuf mentions est de 32 tokens[90].

(67) Il était une fois **[un chat]^{1}** **[qui]^{2}** avait oublié **[sa]^{3}** nourriture dans un voyage. Alors **[il]^{4}** prenait la nourriture des gens. 2 semaines plus tard **[le chat]^{5}** n'avait plus faim. Alors **[le chat]^{6}** va se coucher. Et après **[il]^{7}** avait encore faim et **[il]^{8}** mange enfin **[sa]^{9}** nourriture. (corpus français, CE2, 3055)

Afin de rendre plus explicite l'opérationnalisation de cet indice dans notre protocole d'annotation, nous proposons un second exemple appliqué directement à un des textes de notre corpus et à notre annotation. Le Tableau 36 illustre ainsi le calcul de l'interdistance pour les chaines *witch* et *cat* dans le texte (68). Il faut noter que dans ce calcul le référent *ext1* n'est pas pris en compte car il est représenté par une paire anaphorique. Nous avons donc calculé l'interdistance pour chaque chaine qui comporte trois mentions ou plus dans le texte ainsi que le ratio entre cette distance en termes de tokens hors ponctuation et de nombre de mentions présentes dans la chaine.

(68) **[La sorcière]**^{witch} et **[le chat]**^{cat} Il était une fois **[une sorcière]**^{witch} **[qui]**^{witch} se promenait dans une forêt. **[Elle]**^{witch} trouve **[un chat]**^{cat}, **[elle]**^{witch} **[lui]**^{cat} dit « **[je]**^{witch} vais **[te]**^{cat} prendre pour **[mon]**^{witch} petit déjeuner » **[elle]**^{witch} met **[le chat]**^{cat} dans la casserole et **[la sorcière]**^{witch} dit **[au chat]**^{cat} « bien fait pour **[toi]**^{cat} ! Fallait pas aller dans la forêt, fallait rester avec **[[ton]^{cat} maitre]**^{ext1} ou **[[ta]^{cat} maitresse]**^{ext1} » dit **[la sorcière]**^{witch} à voix haute ! (corpus français, CE1, 3012)

Chaine	Mentions	Interdistance (tokens)	Ratio interdistance (tokens/mentions)
witch	10	52	5,2
cat	9	46	5,11

Tableau 36. Calcul de la distance inter-maillonnaire sur les chaines du texte de l'exemple (68).

Dans le cas de ce texte, l'interdistance étant très similaire entre les deux chaines (les ratios sont respectivement de 5,2 pour la chaine *witch* et 5,11 pour *cat*), nous pouvons en déduire que la portée des chaines dans le texte est très similaire, et que les personnages auront donc une saillance similaire tout au long du texte. Cela est confirmé par l'analyse qualitative du texte, dans lequel on peut relever une cohabitation entre les deux chaines tout au long du récit.

Cet indice d'interdistance a déjà été étudié en lien avec la stabilité des chaines par Rousier-Vercruyssen et Landragin (2019). Dans leur étude, ils ont montré que dans les textes narratifs une

[90] La négation et le verbe «n'avait » comptent comme deux tokens séparés à cause des caractéristiques du tokenisateur utilisé.

interdistance élevée entre maillons correspond généralement à une stabilité réduite des chaines, c'est-à-dire à une « hétérogénéité lexicale » importante (Rousier-Vercruyssen & Landragin, 2019, p. 27). En revanche, cette corrélation apparait négative dans le cas des textes appartenant à des genres non narratifs, dans laquelle plus l'interdistance est élevée, plus la stabilité des chaines sera haute et marquée par une variation mineure dans la dénomination des référents. Ainsi, la distance entre maillons semble donc liée à la saillance des référents dans le texte, à leur accessibilité et permet de fournir une image, bien que partielle, des portées probables des chaines dans un texte. Dans le cas de notre recherche, l'interdistance a été calculée à la fois en termes absolus et sous forme de ratio, en rapportant la distance entre les mentions, exprimée en nombre de tokens, au nombre total de mentions dans la chaine. Les valeurs obtenues sont rapportées dans le Tableau 37.

Niveaux scolaires	**Min**		**Max**		**Moyenne**		**Médiane**		**Ecart type**		**Skewness**		**Kurtosis**	
	fr	it	fr	it	fr	it	fr	it	fr	it	fr	it	fr	it
CE1	1	1	172	460	44,29	70,24	40	61	32	65,66	1,45	3,22	2,79	15,71
CE2	0	1	258	308	71,23	104	60,5	98	48,74	66,57	1,11	0,59	1,62	0,18
CE1-CE2	**0**	**1**	**258**	**460**	**59,42**	**88,03**	**51**	**77**	**44,24**	**68,17**	**1,37**	**1,64**	**2,51**	**5,19**

Tableau 37. Caractérisation des valeurs absolues d'interdistance- comparaison entre français et italien.

La longueur supérieure des textes observés en italien se manifeste également dans cet indicateur. En effet, l'interdistance maximale est de 258 tokens en français contre 460 tokens en italien, et l'écart-type est de 44,24 en français contre 68,17 en italien, ce qui montre une forte dispersion de cette valeur pour l'italien. De manière générale, la moyenne d'interdistance est plus haute en italien (88,02 contre 59,41 en français), et la différence moyenne d'interdistance entre niveaux scolaires reste assez similaire entre langues, se situant à 26 tokens en français et à 33 tokens en italien. Ces statistiques semblent représenter la prédominance de textes plus longs en italien, et qui contiennent probablement des chaines dont la portée s'étale sur la plupart du texte.

Lorsque cet indice est calculé comme un ratio entre l'interdistance et le nombre de mentions dans la chaine considérée (cf. Tableau 38), les valeurs obtenues apparaissent beaucoup plus proches les unes des autres, et les différences observées précédemment sont nettement atténuées.

Niveaux scolaires	Min		Max		Moyenne		Médiane		Écart type		Skewness		Kurtosis	
	fr	it	fr	It	fr	It	fr	It	fr	it	fr	it	fr	It
CE1	0,33	0,17	45,67	58,67	6,34	7,96	5,21	6,21	4,95	7,78	4,18	3,75	27,68	17,83
CE2	0	0,25	40,67	54,50	8,28	9,47	6,46	7,29	6,03	7,46	2,57	2,57	8,2	9,54
CE1-CE2	**0**	**0,17**	**45,67**	**58,67**	**7,43**	**8,76**	**6**	**6,85**	**5,66**	**7,65**	**3,03**	**3,1**	**12,95**	**13,07**

Tableau 38. Caractérisation des valeurs du ratio d'interdistance- comparaison entre le corpus français et le corpus italien.

Comme le montre le Tableau 38, les valeurs de l'indice de distance intermaillonnaire normalisé sont globalement faibles et proches dans les deux langues et aux deux niveaux. Les moyennes sont légèrement plus élevées en italien qu'en français (respectivement 8,76 contre 7,43 pour le français), mais les écarts type restent limités (entre 5,66 et 7,65) et la variabilité demeure comparable entre les groupes. On observe en outre que, de manière générale, les valeurs de ce ratio entre distance et nombre de mentions sont légèrement plus élevées en italien qu'en français.

Les Figure 53 pour le français et Figure 54 pour l'italien permettent de visualiser la distribution des chaines en fonction de leur ratio de distance intermaillonnaire. Le Tableau 38 a montré que les valeurs moyennes de l'indice de distance intermaillonnaire normalisé sont globalement faibles et relativement proches entre les groupes, avec toutefois des valeurs systématiquement un peu plus élevées en italien qu'en français. Les Figure 53 et Figure 54 permettent de compléter cette description en montrant la distribution détaillée des chaines par valeur du ratio d'interdistance.

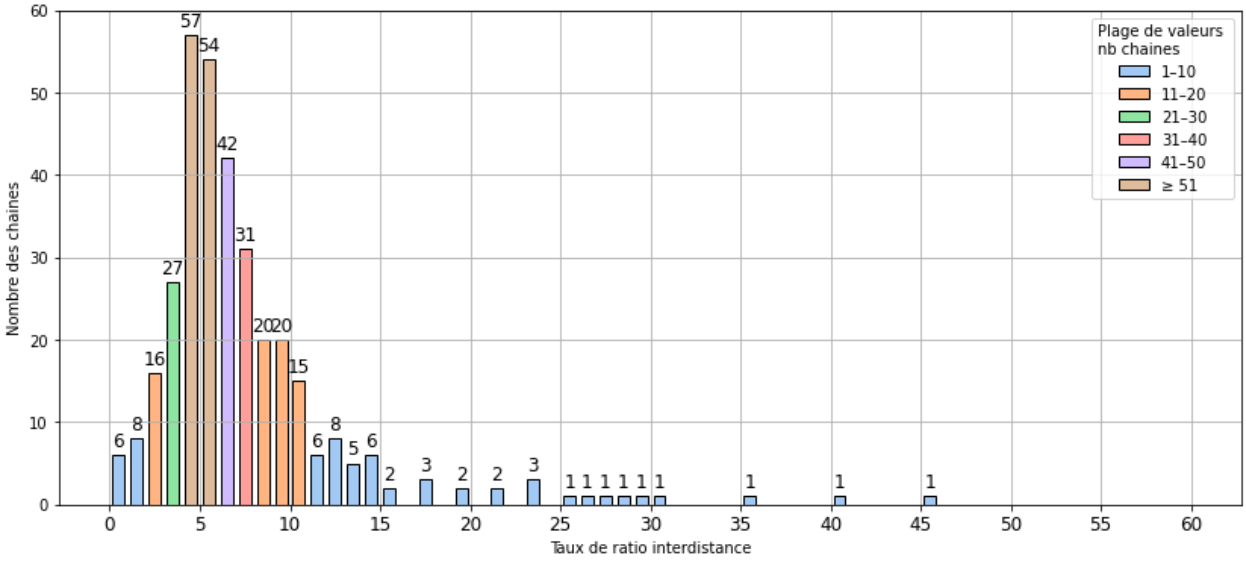


Figure 53. Distribution des chaines par taux de ratio d'interdistance dans le corpus français – tous niveaux confondus.

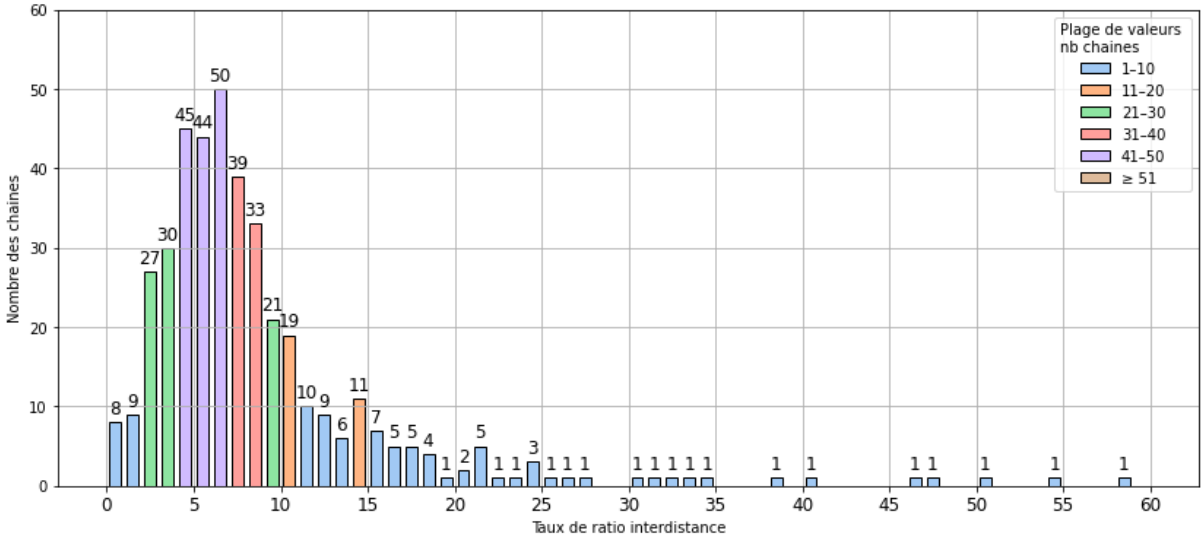


Figure 54. Distribution des chaines par taux de ratio d'interdistance dans le corpus italien – tous niveaux confondus.

On observe une forte concentration des chaines dans les valeurs basses de l'indice, principalement entre 3 et 11, ainsi qu'une distribution fortement asymétrique à droite, caractérisée par une longue traîne de valeurs élevées mais très peu fréquentes. Ces valeurs extrêmes, bien que rares, expliquent la différence importante entre les valeurs maximales et les moyennes observées dans le tableau. Les valeurs de l'italien sont aussi systématiquement plus hautes qu'en français dans les valeurs basses de l'indice.

Ces indices d'interdistance relativement courtes reflètent la taille des textes de notre échantillon, qui est réduite, et pourrait être un indicateur de la forte continuité narrative caractéristique des récits produits par les élèves, dans lesquels l'action se concentre généralement autour d'un nombre limité de personnages principaux et dans des successions rapides de mentions référentielles, jamais trop éloignées entre elles. Nous verrons si les analyses quantitatives successives confirment cette observation.

3. Densité référentielle

3.1. Définition classique et premiers résultats

Introduite par Boudreau et Kittredge (2005), la densité référentielle correspond au ratio entre le nombre de mentions et le nombre total de tokens (hors ponctuation) dans un texte, autrement dit à « la quantité d'éléments référentiels par rapport à la quantité de texte » (Boudreau & Kittredge, 2005, p. 43).

L'analyse descriptive des distributions de densité a mis en évidence des variations selon les niveaux scolaires et selon les langues. Toutefois, ces observations ne permettent pas, à elles seules, de déterminer si les écarts constatés sont significatifs ou non. C'est pourquoi nous avons cherché à compléter cette approche descriptive par une évaluation statistique de la significativité des différences observées entre groupes.

Pour éviter de fausser les analyses obtenues, nous avons calculé la densité référentielle pour chacun de nos textes en ne tenant pas compte des mentions imbriquées dans des mentions plus

élargies. Par exemple, dans l'exemple (69), les déterminants possessifs « sa » et « son » (surlignés dans le texte) n'ont pas été pris en compte pour le calcul de la densité référentielle car imbriqués dans des mentions plus larges, alors que les mentions qui les contiennent (« son chat » et « sa cousine la sorcière ») ont bien été prises en compte. Dans l'exemple (69) ci-dessous, la densité référentielle est obtenue en rapportant 35 mentions sur un total de 169 tokens, ce qui nous permet d'obtenir une densité référentielle égale à 22,8 (35/169).

> (69) Il était une fois dans une maisonnette, vivaient **[un chat]**^{cat} et **[[sa]**^{cat} **maîtresse]**^{ext1}. Mais un jour **[une vilaine sorcière]**^{witch} voulait un chat, et **[elle]**^{witch} savait où en trouver un, **[elle]**^{witch} alla au magasin des animaux, mais il n'y avait plus de chat noir. **[Elle]**^{witch} devint en colère et **[fouilla]**^{witch} toutes les maisons, et **[trouva]**^{witch} la maison d'**[un chat noir]**^{cat}, **[elle]**^{witch} fouilla et **[elle]**^{witch} trouva [**le chat]**^{cat} qu'il **[lui]**^{witch} fallait, **[elle]**^{witch} **[le]**^{cat} vola et **[l']**^{cat} **[emprisonna]**^{witch} jusqu'à ce que **[la dame]**^{ext1} se rendit compte qu'**[elle]**^{ext1} avait perdu **[[son]**^{ext1} **chat]**^{cat}. **[Elle]**^{ext1} alla chez **[[sa]**^{ext1} **cousine la sorcière]**^{witch} et **[se rendit compte]**^{ext1} que **[[son]**^{ext1} **chat]**^{cat} était là. […] **[Elle]**^{ext1} devint furieuse et **[elle]**^{ext1} dit « Alors c'est **[toi]**^{witch} qui as volé **[[mon]**^{ext1} **chat]**^{cat} ! » **[[Sa]**^{ext1} **cousine]**^{witch} **[lui]**^{ext1} répondit « C' est que **[j']**^{witch} ai besoin d' un chat comme les autres sorcières et il n' y en avait plus dans le magasin ». « D' accord, **[j']**^{ext1} ai compris, **[tu]**^{witch} veux vivre avec **[moi]**^{ext1} ». « Oui ». « **[On]**^{ext1,witch} peut le faire ». « D' accord ». Fin. *(corpus français, CE1, 841)*

Nous avons adopté cette définition de la densité car elle nous semble cohérente avec les descriptions proposées dans la littérature existante. Cela nous permet ainsi de comparer les résultats obtenus sur notre corpus avec ceux déjà rapportés dans des travaux antérieurs. Sur cette base, nous pouvons à présent examiner les valeurs de densité référentielle obtenues dans notre corpus à l'aide de cet indice.

Cet indice met en évidence une certaine stabilité de la densité référentielle entre les deux langues entre les deux niveaux (cf. Tableau 39). Pour rappel, nous avons analysé 148 textes pour le français (74 par niveau) et 150 pour l'italien (75 par niveau). En effet, cet indice présente des valeurs moyennes très proches dans les deux langues et aux deux niveaux : 22,12 en français et 22,10 en italien en CE1, puis 21,17 et 21,63 en CE2. On observe toutefois que les valeurs extrêmes ne suivent pas la même tendance, avec la valeur maximale plus élevée au CE1 qu'au CE2 en français (respectivement 33,33 et 28,20) alors qu'on observe le contraire en italien (la valeur maximale de 35,11 se trouve au CE2 en italien). Ces variations n'affectent pas la tendance globale, qui montre une légère diminution des valeurs de densité entre niveaux scolaires.

Niveaux scolaires	**Min**		**Max**		**Moyenne**		**Médiane**		**Écart type**	
	fr	it	fr	it	fr	it	fr	it	fr	it
CE1 – 2	12,94	10,48	33,33	32,69	22,12	22,10	21,17	22,02	5,25	4,04
CE2 – 3	13,33	12,93	28,20	35,11	21,17	21,63	21,53	22,03	3,94	4,24
CE1-CE2 – 2-3	**12,94**	**10,47**	**33,33**	**35,10**	**21,64**	**21,87**	**21,31**	**22,02**	**4,66**	**4,15**

Tableau 39. Calcul de la densité référentielle : ratio entre le nombre de maillons et le nombre total de tokens du texte.

En examinant ces moyennes, nous pouvons remarquer que le taux de densité obtenu est très proche de celui rapporté dans d'autres études qui portaient sur les textes narratifs, bien que notre annotation ait été limitée aux seuls personnages animés. Par exemple, Kim (2023) observe un taux du 19% dans le roman *Le Fantôme de l'Opéra* ; Obry et al. (2017) rapportent des taux compris entre 15,4% et 23,6% pour certains textes narratifs du corpus *DEMOCRAT*. Nous remarquons ici une différence importante entre notre résultat et celui de Schnedecker (2017), qui calcule un taux du 30% sur les contes de fées, qui se rapproche de nos valeurs maximales plutôt que moyennes. Il faut toutefois préciser que, dans notre annotation, seuls les personnages animés sont annotés dans le texte, alors que probablement dans le cas du travail de Schnedecker l'annotation porte sur toute entité présente dans les textes, y compris celles inanimées, ce qui pourrait expliquer cette différence du taux de densité référentielle, plus élevée dans le cas du calcul de Schnedecker car il comprend plus d'entités que notre annotation. Pour une vue d'ensemble, un tableau regroupant plusieurs de ces indices, comparables à ceux examinés ici, figure dans la première partie de ce travail (cf. Chapitre 1, §4.2, Tableau 1).

Pour ce faire, nous avons tout d'abord représenté la distribution des textes de notre corpus selon cet indice. L'inspection graphique des histogrammes montre que, dans les deux corpus, la distribution de la densité référentielle semble suivre une tendance approximativement normale et symétrique. Dans les deux langues, la majorité des textes se situe dans une zone centrale relativement resserrée, ce qui suggère une relative homogénéité de la densité référentielle d'un texte à l'autre, malgré la présence de quelques cas atypiques aux valeurs plus extrêmes. Ces propriétés de distribution, ainsi que les tailles de nos échantillons, qui sont modestes et qui varient d'une langue à l'autre, nous ont conduite à recourir à un test t de Welch afin de comparer les moyennes entre groupes indépendants. Ce test est adapté dans notre cas, car il permet de comparer les moyennes d'une variable continue entre deux groupes indépendants, y compris lorsque l'hypothèse d'égalité des variances ne peut pas être garantie.

Il convient de préciser que les comparaisons statistiques ont été effectuées en traitant les différents groupes (langues et niveaux scolaires) comme des échantillons indépendants. Ce choix est pleinement justifié dans le cas du corpus italien, où les groupes de CE1 et de CE2 ne partagent aucun élève en commun. En revanche, dans le corpus français, un nombre d'élèves très faible apparait dans les deux niveaux scolaires, 13 au total. Pour rendre les deux échantillons pleinement indépendants, nous avons donc retirer de l'analyse les 26 valeurs liées à ces 13 élèves, ce qui explique l'échantillon de 122 valeurs dans le cas du français. Cela explique aussi la légère variation dans les moyennes du corpus français par rapport à celles présentées dans le Tableau 39.

Pour décrire ces données, tout d'abord, nous avons vérifié la forme des distributions pour chaque langue afin de confirmer la pertinence de l'application de ce test.

Langue	N	Moyenne	Écart type	Variance	*Skewness*	*Kurtosis*
Français	122	21,63	4,63	21,52	0,14	-0,52
Italien	150	21,88	4,15	17,27	0,13	0,25

Tableau 40. Indices de description de la distribution des densités référentielles – comparaison entre langues.

Comme nous pouvons observer dans le Tableau 40, les distributions de la densité référentielle dans les deux corpus présentent une asymétrie très faible (*skewness* = 0,14 pour le français ; 0,13 pour l'italien) et une kurtosis proche de celle d'une loi normale (-0,52 et 0,25 respectivement). L'inspection graphique de l'histogramme (cf. Figure 56 et Figure 57) confirme une forme globalement symétrique pour les deux langues, avec de légères variations autour de cette forme idéale, qui ne constituent toutefois pas des écarts substantiels par rapport à une distribution normale. Des vérifications effectuées séparément sur chacun des sous-groupes confirment des distributions de même nature, caractérisées par une asymétrie et une kurtosis faibles, compatibles avec l'hypothèse d'une normalité approximative. Cette propriété justifie l'utilisation d'un test t de Welch pour les comparaisons entre groupes.

Dans le Tableau 41 ainsi que dans la Figure 55 nous pouvons observer les résultats obtenus en appliquant le test sur les trois différentes configurations des groupes que nous avons analysées.

	Français - Italien		**Français**		**Italien**	
	français	*italien*	*CE1*	*CE2*	*CE1*	*CE2*
N	122	150	61	61	75	75
Min	12,94	10,48	12,94	13,33	10,47	12,92
Max	33,33	35,11	33,33	28,21	32,69	35,11
Moyenne	21,63	21,88	22,04	21,22	22,12	21,63
Médiane	21,39	22,02	21,05	21,62	22,02	22,03
Écart type	4,67	4,16	5,16	4,05	4,08	4,24
Variance	21,52	17,27	26,65	16,41	16,63	18,01
Statistique t	−0,46		0,98		0,73	
p-value	0,65		0,33		0,47	
Degrés de liberté	245,52		113,57		147,76	

Tableau 41. Résultats obtenus à partir du t test de Welch en comparant les indices de densité référentielle.

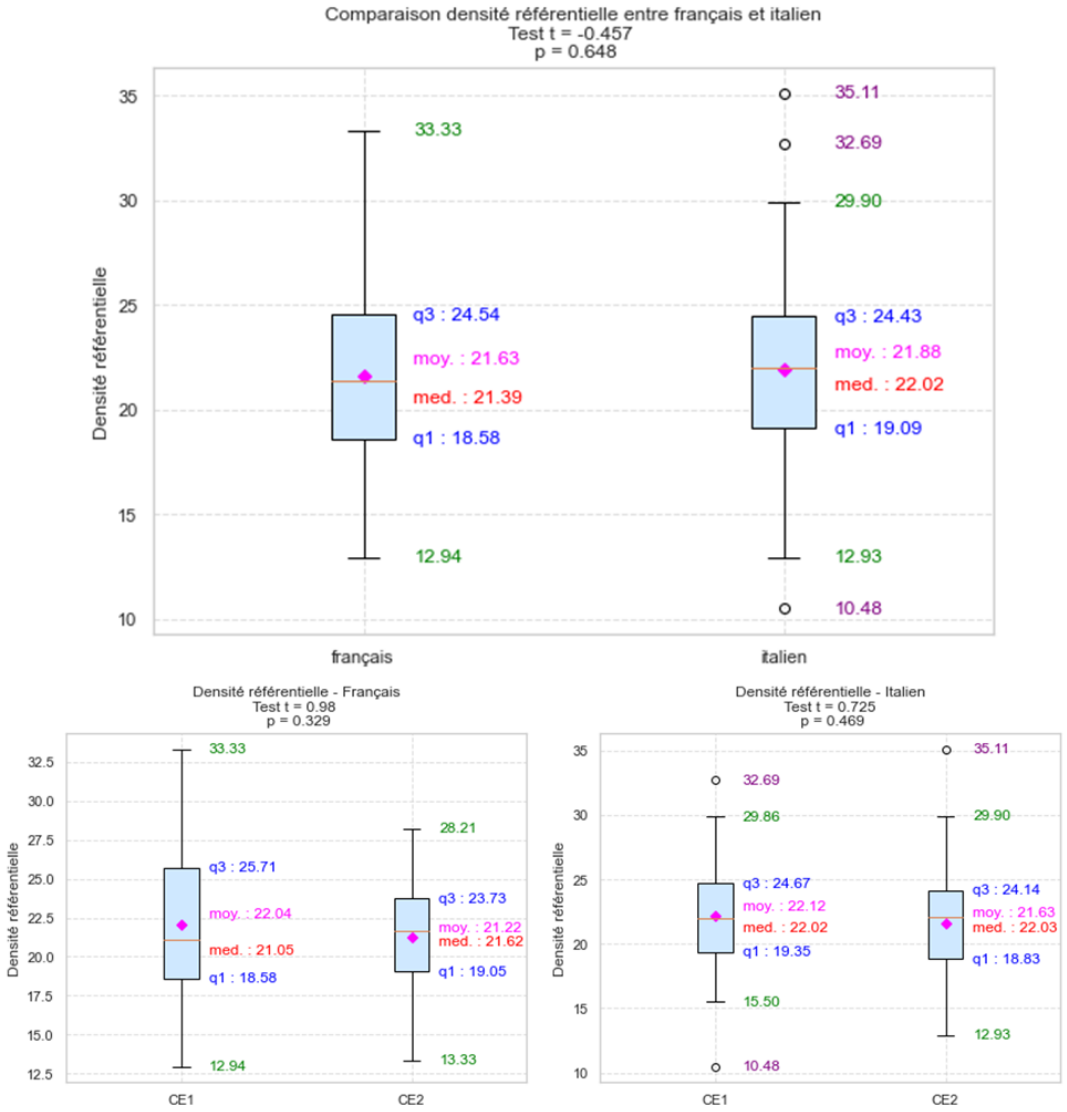


Figure 55. Comparaison des densités référentielle et p-values obtenues selon le *test t de Welch*. Comparaison entre langues et entre niveaux pour chaque langue.

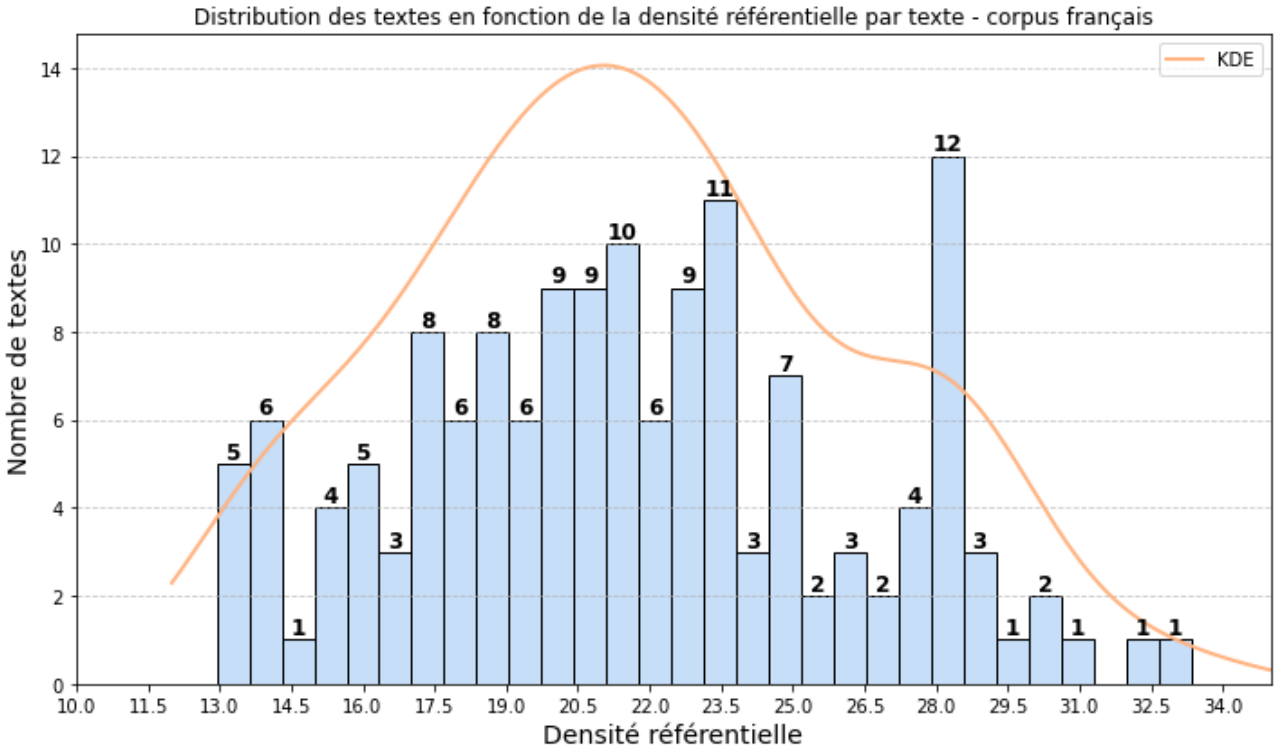


Figure 56. Distribution des textes en fonction de la densité référentielle - représentation de la distribution du corpus français.

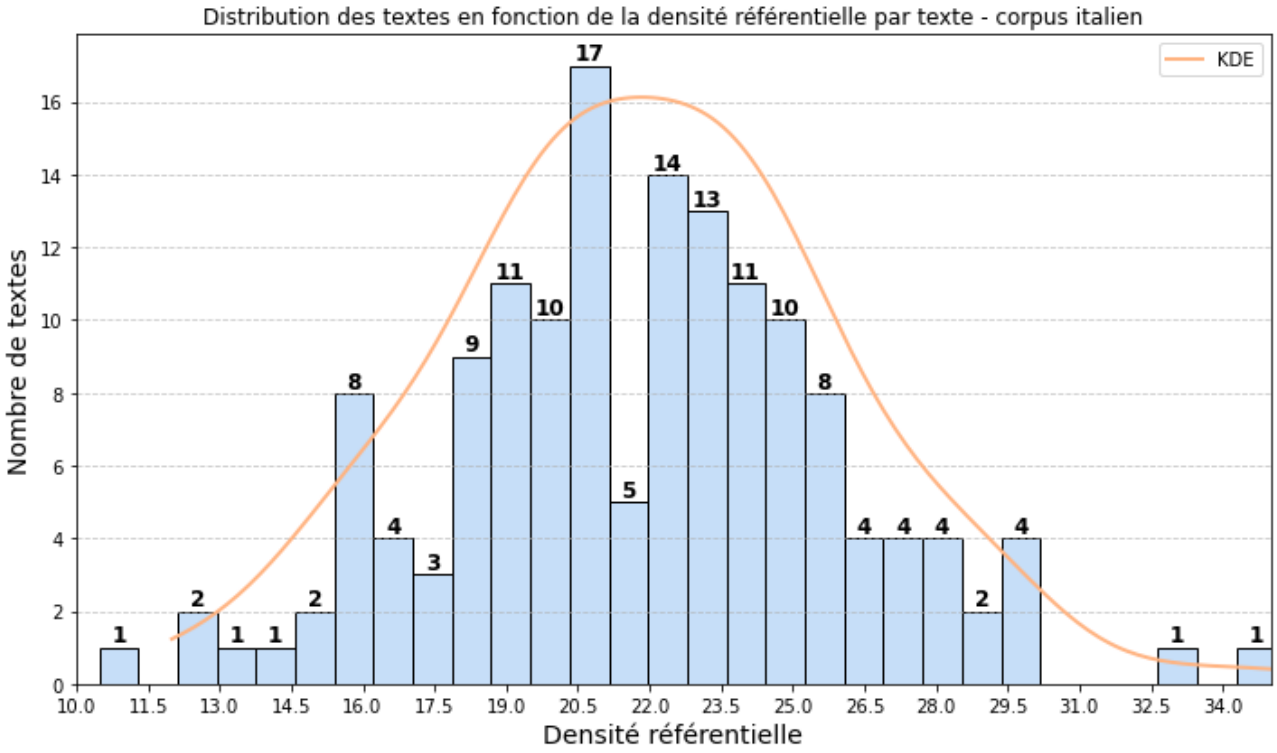


Figure 57. Distribution des textes en fonction de la densité référentielle - représentation de la distribution du corpus italien.

Bien que les indices obtenus entre niveaux scolaires présentent un écart d'un point environ, en appliquant le test t de Welch sur les densités référentielles de chaque texte, nous pouvons observer que la valeur de p n'est pas significative. En effet, la comparaison entre les textes français et italiens ne met pas en évidence de différence statistiquement significative entre les moyennes de densité

des deux langues, respectivement 21,63 pour le français et 21,88 pour l'italien (p = 0,65). Les deux distributions présentent en outre une dispersion comparable, avec un écart type de 4,67 pour le français et de 4,16 pour l'italien, ce qui indique une variabilité d'ampleur similaire dans les deux corpus. De même, les comparaisons entre niveaux scolaires au sein de chaque langue ne montrent pas d'effet significatif : en français, la différence entre le CE1 et le CE2 (22,04 contre 21,22) n'atteint pas le seuil de significativité (p = 0,33), avec des écarts types respectifs de 5,16 et 4,05. En italien, la moyenne est de 22,12 en CE1 contre 21,63 en CE2 (p = 0,47), et les écarts types correspondants sont de 4,08 et 4,24. Dans l'ensemble, ces résultats montrent que les différences observées entre groupes restent faibles au regard de la variabilité interne de chaque échantillon.

Aucune des comparaisons que nous avons effectuées ici ne permet de trouver de corrélation entre les différents niveaux scolaires ou entre les langues étudiées.

3.2. Limites de la définition classique et proposition d'un indice alternatif

La définition de la densité référentielle utilisée jusqu'à maintenant nous semble cependant présenter un défaut important. En effet, elle « aplatit » la présence de maillons comprenant deux tokens ou plus dans le texte. Si l'on analyse le texte (70), la densité référentielle y est égale à 20% (9 mentions sur 45 tokens). Or, dans ce texte, 6 mentions sur 9 comportent chacune deux tokens. Cela signifie que la densité référentielle de ce texte sera égale à celle d'un texte de même longueur contenant également 9 mentions, même si celles-ci ne comportent qu'un seul token chacune.

Cette définition n'est donc pas pleinement représentative de la quantité de matériel référentiel effectivement présent dans le texte, puisqu'elle ne prend pas en compte la présence de mentions composées de plusieurs tokens. Ces dernières, bien que moins fréquentes dans le corpus que les mentions composées d'un seul token (cf. §1.2) sont toutefois bien attestées dans le corpus, et pourraient influencer le taux de densité obtenu.

Nous proposons donc un autre mode de calcul de la densité référentielle, qui prend en considération le contenu des mentions en termes de nombre de tokens. Si l'on reprend l'exemple (70), son taux de densité référentielle est désormais de 35,55 (16/45), dès lors que l'on tient compte des 16 tokens qui font partie des chaines de référence présentes dans le texte.

(70) Il était une fois **[un chat persan]**^{cat}. **[Il]**^{cat} se promenait le long du trottoir. Un jour **[il]**^{cat} trouve **[un maitre]**^{robot}. C'était **[un robot]**^{robot}. **[Le robot]**^{robot} était gentil, très gentil. **[Il]**^{cat} s'est trouvé bien chanceux. Mais le lendemain matin **[le robot]**^{robot} abandonna **[le chat]**^{cat}. *(corpus français, CE1, 100)*

Effectivement, selon une première exploration empirique du corpus, il était possible de faire l'hypothèse d'une densité référentielle plus élevée dans les textes du corpus italien que dans ceux du corpus français. Les textes italiens semblaient en effet présenter une proportion de maillons plus importante que celle observée en français. Cependant, le premier test statistique effectué à partir de la définition classique de l'indice (cf. §3.1) ne permettait pas de confirmer cette hypothèse. Nous montrons ici comment le fait d'avoir repensé cet indice permet de faire apparaître une différence qui était auparavant masquée.

Si l'on applique ce nouveau mode de calcul de la densité référentielle, qui prend en compte la longueur des maillons, on obtient des résultats sensiblement différents de ceux présentés dans le Tableau 39. Ces nouveaux résultats sont illustrés dans le Tableau 42 : ici la moyenne varie d'environ 10 points entre les résultats précédemment obtenus et ceux calculés en tenant compte de la taille des maillons. Nous avons aussi un écart type bien plus élevé, avec 4 points d'écart avec le résultat précédemment obtenu en français et 3 points d'écart en italien. Pour comparer avec le Tableau de la section précédente, ici aussi nous avons mesuré les moyennes sur l'échantillon complet (148 textes en français, 150 en italien).

Niveaux scolaires	Min		Max		Moyenne		Médiane		Écart type	
	fr	it	fr	it	fr	it	fr	it	fr	it
CE1 – 2	14,71	19,48	57,14	53,23	34,51	36,93	34,75	37,50	9,45	7,36
CE2 – 3	16,66	20,21	52,12	50,53	34,49	36,37	34,59	37,60	6,82	6,83
CE1-CE2 – 2-3	**14,71**	**19,48**	**57,14**	**53,23**	**34,5**	**36,65**	**34,59**	**37,55**	**8,21**	**7,08**

Tableau 42. Calcul de la densité référentielle : ratio entre le nombre de tokens inclus dans des maillons et le nombre total de tokens du texte.

Comme pour la définition classique de la densité, nous avons d'abord examiné la forme des distributions avant de procéder aux comparaisons statistiques. Afin de compléter l'inspection graphique des distributions (cf. Figure 58) Tableau 44. Résultats obtenus à partir du t test de Welch en comparant les indices de densité référentielle (prise en compte de la longueur des maillons).

Le test t de Welch indique en effet une différence statistiquement significative entre le français et l'italien (p = 0,014) : les textes italiens présentent en moyenne une densité référentielle plus élevée (M = 36,65, écart type = 7,08) que les textes français Les résultats sont très proches de ceux obtenus à travers le calcul de l'indice qui suit la définition classique et confirment la distribution approximativement normale observée à l'aide des histogrammes. Ici (cf. Tableau 43) les distributions restent faiblement asymétriques et compatibles avec une normalité approximative (*skewness* respectivement de 0,18 et -0,33, et une *kurtosis* proche de 0 dans les deux langues).

Langue	N	Moyenne	Écart type	Variance	*Skewness*	Kurtosis
Français	122	34,31	8,26	68,25	0,18	0,18
Italien	150	36,65	7,08	50,21	-0,33	-0,26

Tableau 43. Indices de description de la distribution des densités référentielles (prise en compte de la longueur des maillons) - comparaison entre langues.

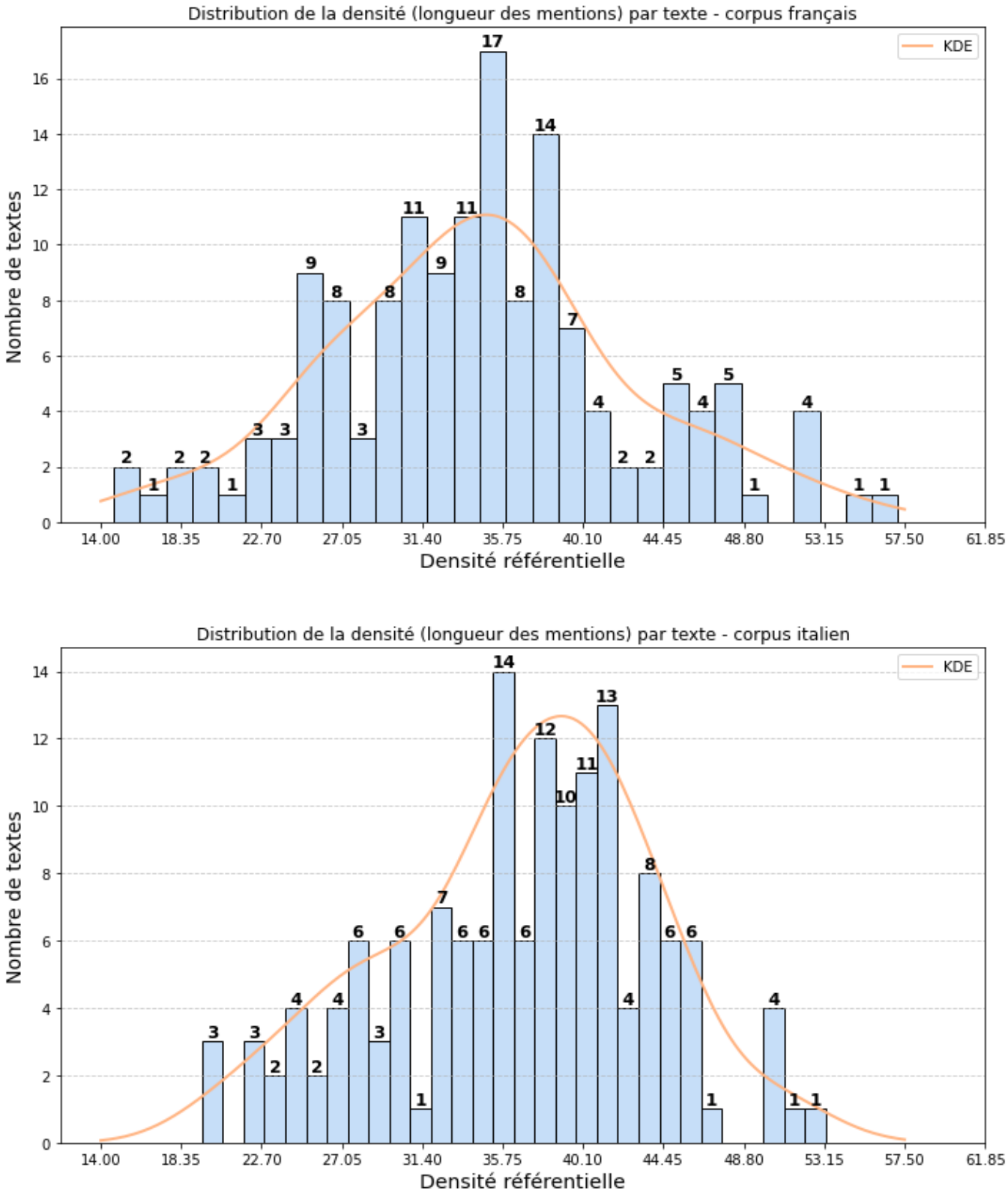


Figure 58. Distribution de la densité référentielle dans les textes prenant en compte la taille des maillons - comparaison entre français et italien.

Nous avons ainsi appliqué à nouveau le test t de Welch sur nos échantillons, et nous obtenons cette fois-ci une différence significative entre français et italien. Le fait que cette différence n'apparaisse pas avec la définition classique de l'indice confirme que celle-ci sous-estimait une partie du matériel référentiel effectivement présent dans les textes. Lorsque la densité référentielle est

calculée en tenant compte de la longueur des mentions, la comparaison entre les deux langues met en évidence une différence statistiquement significative.

	Français-Italien		**Français**		**Italien**	
	français	*Italien*	*CE1*	*CE2*	*CE1*	*CE2*
N	122	150	61	75	61	75
Min	14,7	19,48	14,7	16,67	19,48	20,22
Max	57,14	53,22	57,14	52,13	53,23	50,53
Moyenne	34,32	36,65	34,42	34,22	36,93	36,37
Médiane	34,59	37,55	34,21	34,61	37,5	37,6
Ecart type	8,26	7,08	9,54	6,83	7,36	6,84
Variance	68,25	50,21	91,03	46,59	54,17	46,76
Statistique t	−2,47		0,13		0,48	
p-value	0,014		0,89		0,63	
Degrés de liberté	239,48		108,67		147,21	

Tableau 44. Résultats obtenus à partir du t test de Welch en comparant les indices de densité référentielle (prise en compte de la longueur des maillons).

Le test t de Welch indique en effet une différence statistiquement significative entre le français et l'italien (p = 0,014) : les textes italiens présentent en moyenne une densité référentielle plus élevée (M = 36,65, écart type = 7,08) que les textes français (M = 34,50 ; écart type = 8,26) selon ce nouvel indice. En revanche, aucune différence significative n'est observée entre les niveaux scolaires au sein de chaque langue. En français, les moyennes sont très proches en CE1 (M = 34,51 ; écart type = 9,54) et en CE2 (M = 34,22 ; écart type = 6,83), et la comparaison ne révèle aucun effet (p = 0,89). De même, en italien, la moyenne est de 36,93 en CE1 (écart type = 7,36) contre 36,37 en CE2 (écart type = 6,84), et la différence n'est pas significative (p = 0,63). Ces résultats montrent que la différence mise en évidence par ce nouvel indice est principalement un effet inter-langue et non un effet lié au niveau scolaire.

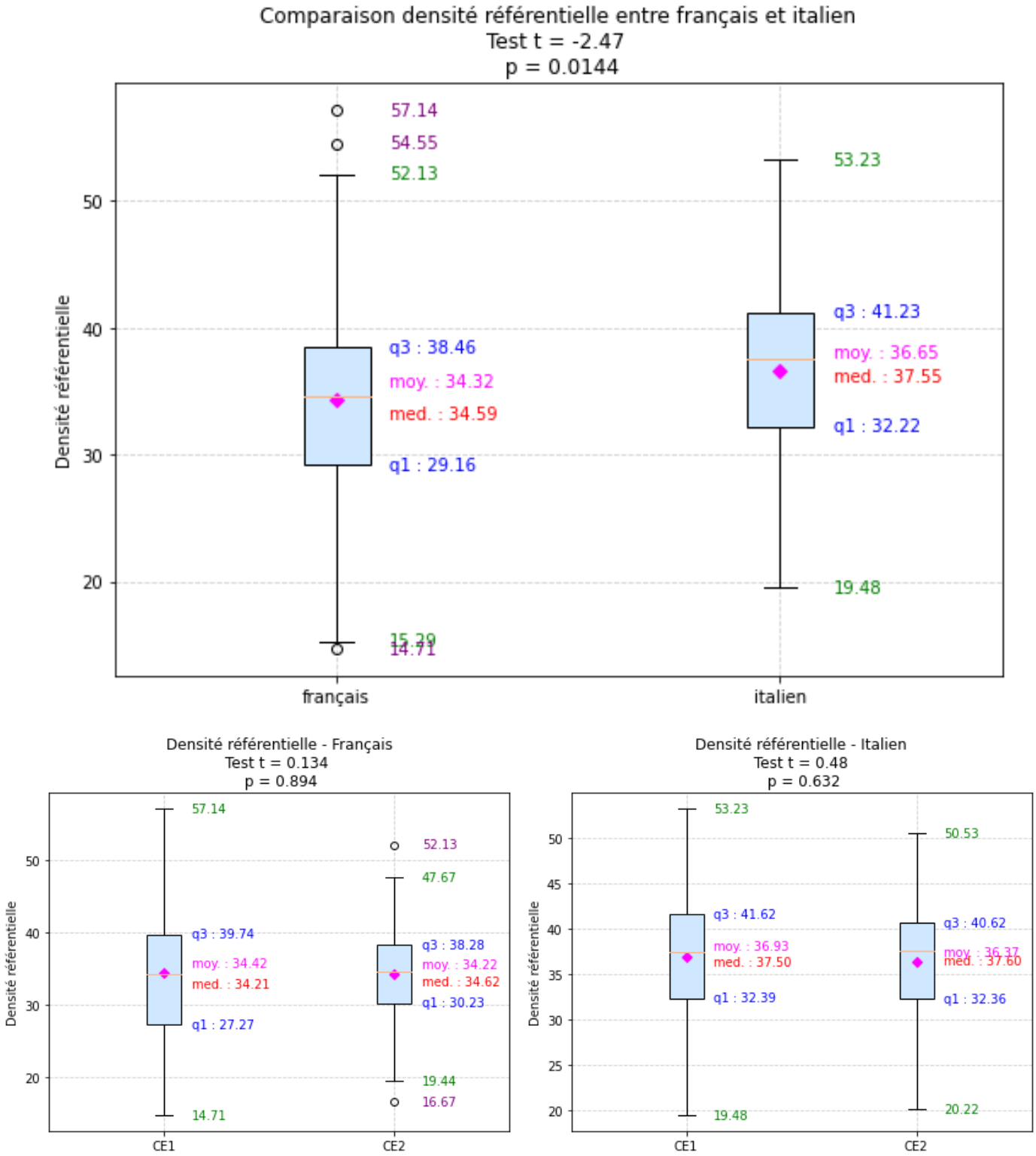


Figure 59. Calcul de la densité référentielle prenant en compte la longueur des mentions. Comparaison entre langues et entre niveaux scolaires pour chaque langue.

Ces résultats confirment que le choix de l'indice n'est pas neutre : lorsque l'on tient compte de la longueur des mentions, une différence entre les moyennes des deux langues apparait clairement, alors qu'elle restait invisible avec la définition classique de la densité référentielle. En revanche, aucune évolution n'est observée entre les deux niveaux scolaires, ce qui suggère que l'effet mis en évidence par ce nouvel indice est principalement lié à la langue plutôt qu'au niveau scolaire. Cela

confirme notre hypothèse initiale d'une majeure densité référentielle dans les textes de langue italienne, initialement constatée empiriquement.

4. Indice de stabilité ou d'instabilité

Après avoir examiné la densité référentielle, nous passons maintenant à l'analyse d'un autre indicateur utilisé pour décrire le fonctionnement des chaines de référence et en particulier la variation lexicale de leurs mentions, à savoir la stabilité référentielle. Comme nous l'avons montré dans la partie théorique (chapitre 1, §4.2), cet indice, tel qu'initialement défini par Perret (2000) permet d'évaluer la « stabilité désignationnelle » d'un référent. Cette notion a donné lieu à plusieurs définitions et à plusieurs propositions d'opérationnalisation dans la littérature (Oberlé, 2016; Obry et al., 2017; Perret, 2000; Rousier-Vercruyssen & Landragin, 2019). Ces différentes approches reposent sur des choix différents quant aux types de mentions pris en compte et à la manière de normaliser l'indice, et présentent chacune des avantages et des limites, et ont donné lieu aux définitions des indices de *stabilité*, *instabilité* et *CNSL* (*coefficient normalisé de stabilité lexicale*). Nous avons tiré profit de ces différentes propositions pour définir une procédure de calcul de ces trois indices adaptés à notre corpus et à notre méthodologie. Nous allons aussi examiner les limites de cet indice et proposer des perspectives pour une opérationnalisation différente.

Ces trois indices ont comme objectif de représenter tous les trois la variation des syntagmes nominaux présents dans une chaine donnée. Cependant, leurs caractéristiques de calcul ne les rendent pas directement comparables entre eux sur une même échelle. Alors que l'indice de *stabilité* et celui d'*instabilité* décrivent le rapport entre répétition et variation des syntagmes nominaux de la même chaine, rapport influencé par le nombre de mentions prises en compte, le *CNSL* tente de normaliser cette description, bornée entre les valeurs 0 (toutes les mentions nominales sont différentes) et 1 (toutes les expressions nominales sont égales). Nous allons voir dans cette section les choix d'opérationnalisation que nous avons effectués pour le calcul de ces indices ainsi que les descriptions des caractéristiques de notre corpus qui en découlent. Nous allons aussi retenir un seul indice, le *CNSL*, afin de comparer les langues et niveaux de notre corpus entre eux.

4.1. Sélection des chaines et des mentions pour le calcul de la stabilité

L'opérationnalisation de l'indice de stabilité repose sur une procédure de sélection et de filtrage progressive des chaines et des mentions, destinée à filtrer les chaines qui ne comportent pas un nombre suffisant de mentions nominales exploitables pour que le calcul des indices soit interprétable.

Dans un premier temps, nous partons de l'ensemble des chaines annotées dans le corpus. Parmi les mentions qui composent les chaines, nous avons retenu pour définir *n* (le nombre des mentions nominales dans la chaine) uniquement les mentions contenant au moins un nom, c'est-à-dire celles contenant au moins un token ayant la catégorie morphosyntaxique *NOUN*. Ce critère, bien que simpliste, permet d'obtenir un filtrage satisfaisant des mentions nominales. Toutes les formes des expressions référentielles retenues sont converties en minuscules afin de neutraliser les variations

de casse (par exemple en début de texte ou de phrase). Deux mentions sont considérées comme identiques si elles sont égales au niveau de la chaine de caractères. Cela nous a permis de définir la valeur de *x*, c'est-à-dire le nombre de dénominations différentes dans une chaine.

Conformément à la définition initiale de Perret, nous excluons ensuite du calcul la première mention de la chaine lorsqu'elle est introduite par un déterminant indéfini (POS-tag *DET* ou *ADP*, trait morphosyntaxique *Definite=Ind*), cette mention étant considérée comme jouant le rôle de « pivot référentiel ». En revanche, nous n'avons pas été en mesure d'exclure systématiquement du calcul les mentions contenues dans les titres des textes, comme le prévoit le calcul original de Perret.

À l'issue de ce double filtrage (sélection des mentions nominales, puis exclusion éventuelle de la première mention indéfinie), certaines chaines peuvent ne comporter qu'un très petit nombre de mentions nominales exploitables. Or, les indices étudiés ici visent à caractériser la variation lexicale à l'intérieur d'un ensemble de reprises, et ne sont donc interprétables que si cet ensemble comporte un minimum d'éléments. Pour cette raison, nous avons choisi de ne calculer les indices que pour les chaines contenant au moins trois mentions nominales. Les chaines ne satisfaisant pas ce critère sont exclues de cette analyse.

Pour rappel, dans la définition initiale donnée par Perret, cet indice sert à décrire la « stabilité désignationnelle » des reprises nominales. Il est « obtenu en divisant, pour un référent donné (ie. un personnage), le nombre total d'anaphores nominales par le nombre de désignations différentes. » (Perret, 2000, p. 17) Plus ce coefficient est élevé, plus les mêmes désignations sont répétées à l'intérieur des chaines. Plus il est bas, plus la chaine présente des variations lexicales à son intérieur.

4.2. Exemple du calcul de stabilité, instabilité et CNSL

Dans notre travail, nous avons choisi de calculer et comparer ces trois indices de description complémentaires, reposant sur les mêmes quantités pour chaque chaine : le nombre total de mentions nominales retenues (*n*) ainsi que le nombre de dénominations distinctes (*x*). À partir de ces deux valeurs, nous avons ainsi calculé l'indice de stabilité de Perret (2000) comme le ratio entre *n* et *x*, l'indice d'instabilité de Rousier-Vercruyssen et Landragin (2019) défini comme *x/n*100* et le coefficient normalisé de stabilité lexicale CNSL proposé par Oberlé (2016) défini comme $1 - \frac{x-1}{n-1}$. Cet indice mesure la stabilité référentielle en normalisant la proportion de variation lexicale dans une chaine, de sorte que la valeur est comprise entre 0 et 1, 1 correspondant à une stabilité maximale.

À titre d'exemple, pour mieux comprendre les critères de sélection des mentions appliqués, dans le texte (71) le calcul de la stabilité est effectué pour chacune des trois chaines présentes dans le texte, qui indiquent les référents *chat*, *witch* et *ext1*. Chaque chaine reçoit ainsi sa propre valeur de stabilité, calculée à partir du nombre total de mentions nominales et du nombre de désignations distinctes.

(71) **[Le chat]**^{cat} et **[la sorcière]**^{witch} Il était une fois **[un petit chat]**^{cat} et **[une sorcière]**^{witch}, **[la sorcière]**^{witch} est très méchante avec les chats. Un jour **[elle]**^{witch} voit **[le chat]**^{cat} **[dit]**^{witch} « Oh **[un chat]**^{cat}, **[il]**^{cat} **[me]**^{witch} tente bien pour faire une soupe ». **[Elle]**^{witch} dit **[au chat]**^{cat} « **[viens]**^{cat} **[[mon]**$^{\mathbf{witch}}$ **minou]**^{cat}. **[Je]**^{witch} ne vais pas **[te]**^{cat} manger. **[Entre]**^{cat} dans **[ma]**^{witch} maison. **[Je]**^{witch} vais **[te]**^{cat} donner à manger » **[le chat]**^{cat} entre dans la maison, **[il]**^{cat} voit des tas de bonnes choses à manger comme des gâteaux pour chat, des souris mais **[il]**^{cat} voit aussi un chaudron. **[Il]**^{cat} s'approcha et **[la sorcière]**^{witch} **[le]**^{cat} poussa et **[il]**^{cat} tombe dans le chaudron. **[La sorcière]**^{witch} alla chercher des crapauds, des yeux d'araignées mais **[elle]**^{witch} ne se doutait pas que **[le chat]**^{cat} était très très malin. **[Il]**^{cat} réussit à sortir du chaudron à temps, **[se cacha]**^{cat} et **[la sorcière]**^{witch} arriva. **[Elle]**^{witch} prend ce qu'**[elle]**^{witch} est allée chercher, **[elle]**^{witch} le jetait dans le chaudron, **[prend]**^{witch} **[son]**^{witch} livre de magie et **[lut]**^{witch} une potion magique : « Bygoudi Bygouda **[je]**^{witch} veux que **[ce chat]**^{cat} se transforme en soupe. » Mais **[elle]**^{witch} ne voit pas d'explosion, **[elle]**^{witch} ne vit plus **[le chat]**^{cat}, **[elle]**^{witch} regarda dans toute **[sa]**^{witch} maison et **[elle]**^{witch} ne retrouva pas **[le chat]**^{cat}. Mais **[le chat]**^{cat} appela **[la police]**^{ext1} et **[leur]**^{ext1} **[dit]**^{cat} : « **[Venez]**^{ext1} vite **[je]**^{cat} suis dans une maison, c'est la maison de **[Mme je suis méchant]**^{witch} en gros **[Madame sauvage]**^{witch} ». **[Le policier]**^{ext1} arrive et **[dit]**^{ext1} : « ah oui c'est **[la sorcière]**^{witch}. **[Ce chat]**^{cat} avait raison de **[m']**^{ext1}appeler, **[elle]**^{witch} est très dangereuse. » **[Il]**^{ext1} alla à la porte et **[toqua]**^{ext1} : toc toc toc. **[La sorcière]**^{witch} ouvrit et **[le policier]**^{ext1} **[lui]**^{witch} dit « **[vous]**^{witch} êtes en état d' arrestation pour violence sur chat sans défense. » **[Il]**^{ext1} sortit les menottes et **[lui]**^{witch} **[attacha]**^{ext1} aux poignets. **[Il]**^{ext1} **[la]**^{witch} met dans la voiture et **[le chat]**^{cat} peut se balader maintenant dans les rues de New York. Fin de l'histoire. *(corpus français, CE2, 3032)*

Le Tableau 45 illustre les éléments pris en compte pour le calcul et les différents indices de stabilité obtenus, ainsi que les valeurs obtenues pour chaque indice. Chaque ligne correspond ici à une chaine pour laquelle le calcul est applicable, identifiée par l'étiquette de son référent. Pour chaque chaine, on a listé les différents syntagmes nominaux, accompagnés de leur fréquence d'occurrence. Le tableau indique également la somme des mentions nominales dans la chaine ainsi que le nombre de dénominations distinctes, deux éléments fondamentaux pour tout calcul de la stabilité, indépendamment de l'opérationnalisation de l'indice.

Chaine	Mentions nominales uniques (fréq.)	Mentions nominales	Dénominations distinctes	Indice de stabilité	Indice d'instabilité (%)	CNSL
Cat	le chat (8) un petit chat (1) un chat (1) au chat (1) mon minou (1) ce chat (2)	14	6	2,33	42,86	0,62
Witch	la sorcière (7) une sorcière (1) Mme je suis méchant (1) Madame Sauvage (1)	10	4	2,5	40	0,67
Ext1	la police (1) le policier (2)	3	2	1,5	66,66	0,5

Tableau 45. Éléments pour le calcul des indices de stabilité, instabilité et CNSL pour les chaines du texte du corpus français, CE2, 3032.

L'indice de stabilité augmente avec la récurrence des mêmes désignations : plus sa valeur est élevée, plus la chaine est lexicalement stable, tandis qu'une valeur faible indique une plus grande variation des formes employées. À l'inverse, l'indice d'instabilité est à interpréter comme une mesure directe de la variation : plus il est élevé, plus la proportion de désignations différentes au sein de la chaine est importante. Le coefficient normalisé de stabilité lexicale (*CNSL*) normalise cette information sur une échelle comprise entre 0 et 1 : une valeur de 1 correspond à une stabilité maximale (toutes les mentions nominales sont identiques), tandis qu'une valeur de 0 correspond à une instabilité maximale (toutes les mentions sont différentes). Les valeurs intermédiaires représentent des degrés graduels de variation lexicale et permettent de comparer des chaines de longueurs différentes.

Dans l'exemple (71), les chaines des référents *cat* et *witch* présentent des indices de stabilité très proches (respectivement 2,33 et 2,5), malgré une différence du nombre de mentions entre les deux chaines (14 mentions nominales contre 10). Cette similarité s'explique par un rapport comparable entre le nombre de mentions et le nombre de dénominations uniques. La chaine *cat* compte quatre mentions de plus, mais seulement deux variantes supplémentaires par rapport à *witch*. Le degré de stabilité reste donc similaire, ce qui se traduit par des valeurs très proches entre les deux chaines, qui s'éloignent légèrement de la valeur associée à la chaine du personnage du policier *ext1* (de 1,5), qui selon cet indice représente la chaine la moins stable. L'indice d'instabilité est légèrement plus élevé pour *cat* (42,86%) que pour *witch* (40%), ce qui indique une plus forte instabilité de la chaine de *cat*, conformément à l'interprétation donnée de l'indice de stabilité. Le CNSL va dans le même sens : avec des valeurs respectives de 0,62 pour *cat* et 0,67 pour *witch*, les deux chaines présentent un niveau de stabilité globale similaire, *witch* apparaissant toutefois comme légèrement plus stable du point de vue de la normalisation.

À l'inverse, la chaine *ext1*, beaucoup plus courte, se distingue nettement par un indice de stabilité plus faible (1,5), un indice d'instabilité nettement plus élevé (66,66 %) et un CNSL de 0,5, ce qui

reflète une variation lexicale proportionnellement beaucoup plus importante, liée à l'alternance entre seulement deux désignations sur un nombre réduit de mentions (trois au total).
Cependant, comme nous l'avons anticipé, cet indice n'est pas représentatif de la variation dans les chaines qu'au niveau des mentions nominales, dans la mesure où il ne prend pas en considération les autres types de mentions, tels que les pronoms ou les anaphores zéro entre autres. Dans certaines approches, les noms propres sont également exclus du calcul : notamment, pour Schnedecker, « ce coefficient pose un certain nombre de questions : limité aux anaphores nominales, *il laisse de côté le nom propre (et ses sous-types)* qui n'est pas anaphorique, les pronoms personnels (et autres). » (Schnedecker, 2021, p. 97, notre italique). Ici nous avons retenu une approche qui exclut les syntagmes contenant exclusivement des noms propres mais qui comprend toutefois les syntagmes nominaux qui contiennent des noms propres, tels que les syntagmes nominaux de dénomination comme « un robot di nome Antenna di fuoco» (corpus italien, CE1, 432), ou encore de syntagmes nominaux accompagnés d'un nom propre tels que « le loup Cédric » (corpus français, CE2, 3060) ou « il gatto Thomas » (corpus italien, CE1, 534). Nous allons approfondir la structure, le contenu et la fonction de certaines de ces expressions nominales dans les analyses qualitatives qui portent sur les mentions longues annotées dans le corpus (cf. Chapitre 8, §1).

4.3. Comparaison des résultats

Le Tableau 46 présente le nombre total des chaines qui ont fait l'objet du calcul de cet indice, ainsi que les moyennes des indices de stabilité, d'instabilité et le CNSL.

Niveaux scolaires	**Chaines tot.**		**Stabilité moy.**		**Instabilité moy. (%)**		**CNSL moy.**	
	fr	it	fr	it	fr	it	fr	it
CE1 – 2	55	98	1,71	2,36	53,23	54	0,63	0,59
CE2 – 3	129	138	1,72	2,34	50,02	54,54	0,65	0,55
CE1-CE2 – 2-3	**184**	**236**	**1,71**	**2,35**	**50,98**	**54,32**	**0,64**	**0,57**

Tableau 46. Calcul des indices de stabilité, instabilité référentielle et coefficient normalisé de stabilité lexicale.

Dans le cas du français, sur les 342 chaines référentielles annotées, seules 184 répondent aux critères nécessaires pour être prises en compte dans le calcul de ces indices, soit environ 53,8% de l'ensemble des chaines. En ce qui concerne l'italien, sur un total de 410 chaines annotées, 236 remplissent les critères requis, ce qui correspond à environ 57,6% du total. Cela signifie que la mesure proposée ne porte pas sur l'intégralité des chaines annotées, mais sur un peu plus de la moitié de celles-ci, après exclusion des cas qui ne satisfont pas aux contraintes méthodologiques définies auparavant. À noter aussi que les chaines retenues au niveau CE1 sont sensiblement moins nombreuses que celles de niveau CE2. Cela pourrait être lié à une variété de facteurs : en CE2 les chaines sont en moyenne plus longues (cf. ce chapitre, §1.4) Les valeurs correspondantes des différents indices calculés sont présentées dans le Tableau 46.

On y observe que, dans les deux langues, la stabilité moyenne est globalement très proche entre les deux niveaux scolaires. En français, elle reste stable entre 1,71 en CE1 à 1,72 en CE2, tandis qu'en italien elle reste également très stable (2,36 en CE1 contre 2,34 en CE2). Cette tendance de l'indice de stabilité se reflète dans l'évolution de l'indice d'instabilité, qui demeure aussi très proche

entre niveaux, tant en français (de 53,23 à 50,02) qu'en italien (de 54 à 54,54). En ce qui concerne le coefficient normalisé de stabilité lexicale (CNSL), on observe en revanche une légère augmentation en français entre le CE1 et le CE2 (de 0,63 à 0,65), tandis qu'en italien la valeur moyenne diminue légèrement (de 0,59 à 0,55). Ces évolutions restent toutefois de faible amplitude. Globalement, les textes italiens semblent présenter une variation majeure dans les dénominations utilisées dans les chaines.

Parmi les trois indices de stabilité que nous avons calculés pour notre corpus, nous avons choisi d'utiliser le CNSL pour effectuer les comparaisons entre les deux corpus annotés, car sa normalisation permet de neutraliser partiellement l'effet de la longueur des chaines et de rendre les valeurs comparables d'un sous-corpus à l'autre. Le Tableau 47 résume la description de cet indice plus en détail. Dans les deux langues et pour les deux niveaux scolaires, on observe des valeurs moyennes relativement proches (entre 0,55 et 0,65), accompagnées d'écarts-types élevés (environ 0,29 à 0,35), ce qui indique une dispersion importante des valeurs décrites. Les distributions des chaines analysées par valeur du CNSL associées sont visibles dans les Figure 60 et Figure 61.

Niveaux scolaires	**Chaines tot.**		**Moyenne**		**Médiane**		**Écart type**	
	fr	it	fr	it	fr	it	fr	it
CE1 – 2	55	98	0,63	0,59	0,6	0,65	0,35	0,3
CE2 – 3	129	138	0,65	0,55	0,67	0,55	0,33	0,29
CE1-CE2 – 2-3	**184**	**236**	**0,64**	**0,57**	**0,67**	**0,59**	**0,33**	**0,29**

Tableau 47. Description des valeurs du CNSL - comparaison entre les deux corpus.

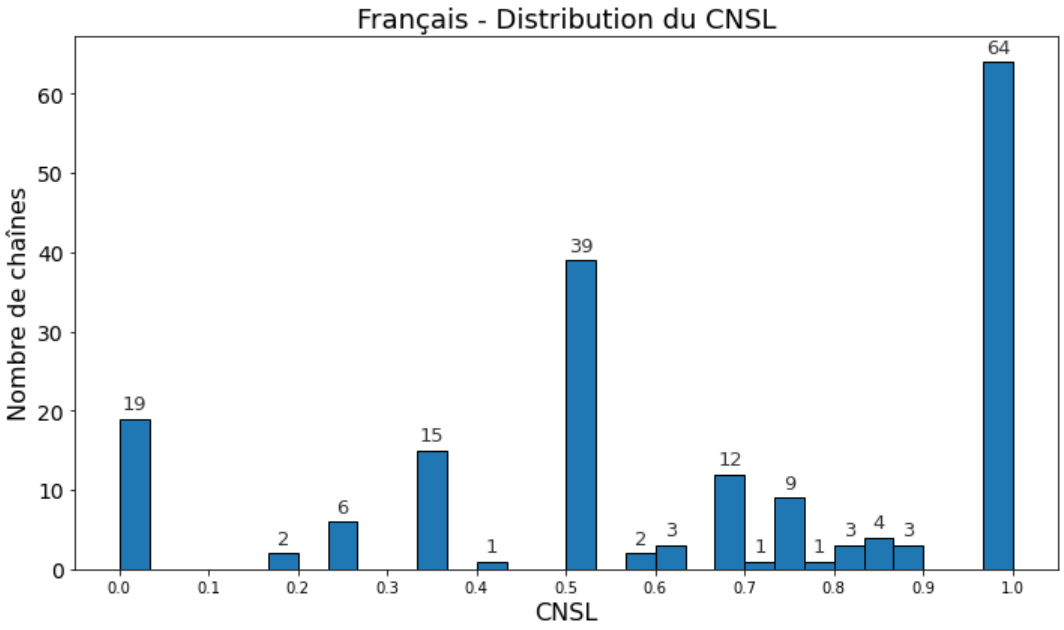


Figure 60. Distribution du CNSL sur les 184 chaines du corpus français.

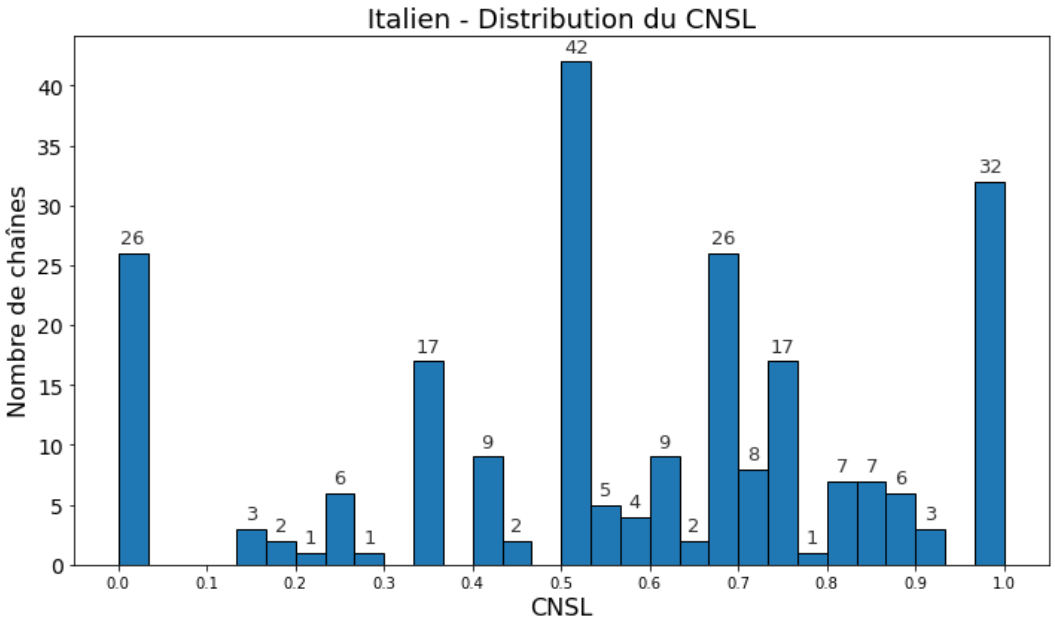


Figure 61. Distribution du CNSL sur les 236 chaines du corpus italien.

La distribution de cet indice en français (cf. ci-dessus) est fortement asymétrique. On observe une accumulation marquée des valeurs aux extrêmes de l'échelle, avec un nombre très élevé de chaines avec un indice de 1 (64 au total), donc qui sont totalement stables, et un second groupe de chaines ayant un indice de 0,5 (39 au total). Le troisième pic (19) est représenté par les chaines ayant un CNSL égale à 0, c'est-à-dire les chaines qui présentent de la variation sur tous les syntagmes nominaux. Les valeurs intermédiaires sont peu nombreuses et discontinues dans leur répartition. Dans le cas de l'italien (cf. Figure 61) la majorité des chaines présentent un indice de 0,5 (42 au total). De manière similaire au français, nous retrouvons ainsi des pics à l'indice 1 (32) de stabilité maximale et à l'indice 0 (26) d'instabilité maximale. Un pic se situe aussi à l'indice 0,67 (26 chaines). Très peu de chaines se situent entre 0 et 0,5 pour le français, alors qu'en italien la majorité des chaines se trouvent entre 0,5 et 1.

Les valeurs intermédiaires du CNSL, bien que peu nombreuses, correspondent néanmoins à des configurations des chaines qui présentent des valeurs intermédiaires de variations, donc assez stables à travers la présence de répétitions mais qui présentent néanmoins des variations lexicales.

À noter ici que, comme nous l'avons mentionné précédemment, cet indice a comme objectif de représenter la variation lexicale au sein des chaines. Cependant, la définition même de cet indice restreint cette mesure aux seuls syntagmes nominaux, en excluant toute expression référentielle d'un autre type. Dès une première analyse empirique, nous avions remarqué que cette contrainte de sélection aurait exclu une bonne partie du matériel référentiel que nous avions annoté, notamment au sein du corpus italien. Un exemple plutôt extrême, qui montre bien dans quels cas cet indice ne peut pas décrire la variation au sein des chaines et exclut même certains textes tout court, est le cas du texte (72).

(72) C'era una volta **[un robot]**[robot] **[che]**[robot] si chiamava **[Arbo]**[robot]. Poi c'era **[una strega]**[witch] **[che]**[witch] si chiamava **[Eva]**[witch]. Un giorno **[Arbo]**[robot] nel **[suo]**[robot] laboratorio segreto fece un'invenzione che scambia i corpi o li moltiplica. Allora **[Arbo]**[robot] provó a moltiplicare il **[suo]**[robot] corpo allora **[chiese]**[robot] a **[Eva]**[witch] di schiacciare il tasto di accensione e girare la manopola a due copie e **[Eva]**[witch] lo fece e la macchina fece quel che richiesto e l'esperimento è riuscito. *(corpus italien, CE1, 1376)*

Dans ce texte, les deux chaines suivent une structure similaire : la première mention est représentée par un syntagme nominal indéfini, qui n'est pas à prendre en compte dans le calcul de l'indice. Toutes les mentions qui suivent sont des noms propres (*Eva, Arbo*), des pronoms relatifs (*che*), des déterminants possessifs (*suo*) ou des anaphores zéro (*chiese*). Ces chaines, même si elles présentent des expressions référentielles variées, ne sont pas mesurables à l'aide de l'indice tel que décrit en littérature, ce qui mène à une perte de représentation des données à notre disposition lors qu'on effectue des descriptions sur la base de ces indices quantitatifs. Pour cette raison, cet indice ne nous semble pas totalement descriptif de la variété lexicale des chaines en tant que telles, car il se limite à la mesure de cette variété au niveau des syntagmes nominaux.

5. Distribution des catégories grammaticales de mentions

5.1. Cadre théorique et enjeux de la typologie des mentions

Ce critère d'analyse (dont les différentes catégories ont été détaillées dans le Chapitre 1, §6) représente « une clé d'interprétation décisive » (Schnedecker, 2021, p. 94) pour déterminer le contenu des chaines de référence et leur lien avec les différents facteurs de variation que nous avons identifiés, notamment la langue et le niveau scolaire, ou encore le genre textuel. Dans les approches de l'accessibilité du référent d'Ariel et de la continuité référentielle de Givón, présentées précédemment (cf. Chapitre 1, §5.2), le choix de la catégorie grammaticale d'un maillon est fortement conditionné par le statut du référent dans la représentation mentale du destinataire. Toutefois, ce choix ne dépend pas uniquement de ce paramètre. D'autres dimensions telles que la distance entre expressions référentielles, la saillance d'un référent, la compétition entre référents et l'unité discursive jouent également un rôle crucial. Ces principes théoriques impliquent qu'en présence de plusieurs référents potentiels, ceux-ci peuvent être hiérarchisés selon différents critères syntaxico-sémantiques, cotextuels et positionnels, permettant de formuler des hypothèses sur leur potentiel de reprise (Schnedecker, 2021, p. 130). Concrètement, un référent introduit par un syntagme nominal défini qui remplit la fonction de sujet aura un potentiel de réactivation plus élevé qu'un référent présenté par un syntagme nominal indéfini et qui a fonction d'objet.

Cette hiérarchisation des référents et son effet sur la désambiguïsation dans le rattachement d'un maillon à un référent donné peut être illustré par l'exemple (73), tiré du premier chapitre *d'Anna Karénine* de Tolstoï. Dans cet extrait, le maillon « La femme » (en gras dans le texte) pourrait renvoyer aussi bien à la princesse Daria Oblonsky, introduite par un syntagme nominal défini, qu'à l'institutrice française, introduite par un syntagme nominal indéfini. Or, c'est précisément la combinaison entre saillance du référent, choix du déterminant et contexte discursif qui permet de

désambigüiser le lien de coréférence et de conclure que « la femme » désignée ici est la princesse Daria Oblonsky.

(73) La maison Oblonsky était bouleversée. [***La princesse]***, ayant appris que son mari entretenait une liaison avec ***[une institutrice française]*** qui venait d'être congédiée, déclarait ne plus vouloir vivre sous le même toit que lui. Cette situation se prolongeait et se faisait cruellement sentir depuis trois jours aux deux époux, ainsi qu'à tous les membres de la famille, aux domestiques eux-mêmes. Chacun sentait qu'il existait plus de liens entre des personnes réunies par le hasard dans une auberge, qu'entre celles qui habitaient en ce moment la maison Oblonsky. **[La femme]** ne quittait pas ses appartements ; le mari ne rentrait pas de la journée ; les enfants couraient abandonnés de chambre en chambre ; l'Anglaise s'était querellée avec la femme de charge et venait d'écrire à une amie de lui chercher une autre place ; le cuisinier était sorti la veille sans permission à l'heure du dîner ; la fille de cuisine et le cocher demandaient leur compte. (incipit, *Anna Karénine*, Tolstoï)

Cet exemple, issu d'un texte narratif littéraire, illustre bien les prédictions que l'on peut formuler à l'aide des théories d'accessibilité et de continuité référentielle vis-à-vis de l'influence de la hiérarchisation des référents sur la sélection des types de mentions utilisés dans un discours donné. Toutefois, si ces théories cherchent à prévoir les différents types de reprise référentielle possibles selon les différentes caractéristiques des référents en jeu, comme remarqué par Schnedecker (Schnedecker, 2017), elles ne prennent pas en compte la variation que l'on peut rencontrer dans la formation des chaines selon le genre de texte représenté ou encore selon la maitrise de la langue d'un scripteur donné. Dans nos analyses, nous travaillons avec un corpus principalement narratif, défini par la consigne de production, et produit par des élèves d'école primaire. Puisque le genre textuel est constant, les différences observées entre niveaux scolaires et langues pourront être attribuées à ces facteurs de variation.

5.2. Annotation et catégories retenues

Afin d'opérationnaliser ces descriptions théoriques dans l'analyse de notre corpus, nous avons procédé à une annotation systématique et semi-automatisée de la catégorie morphosyntaxique de chaque mention annotée. Pour ce faire, nous nous sommes appuyée sur le *POS tag* défini par SpaCy (Honnibal et al., 2020). Nous avons ensuite utilisé un algorithme à base de règles qui nous a permis de reconstruire la catégorie d'appartenance de chaque mention (y compris des mentions imbriquées). Dans certains cas, néanmoins, nous avons effectué des corrections manuelles sur la sortie obtenue avant d'effectuer la caractérisation des typologies de mentions. Ces interventions ont porté soit sur des erreurs causées par notre algorithme de définition du *POS tag*, qui a parfois éliminé l'étiquette morphosyntaxique des résultats obtenus (dans ce cas, le *POS tag* était omis par exemple pour les pronoms clitiques en italien, pour lesquels nous avons déterminé la règle de *POS tagging* directement dans notre algorithme), soit sur des erreurs de *POS tag* évidentes (des noms propres « atypiques » identifiés comme noms communs, ex. « la strega Riunione » en italien).

Pour rappel, les catégories retenues pour déterminer la typologie de chaque maillon sont issues de travaux antérieurs (voir notamment Chapitre 1, §6, ou la taxonomie décrite par Federzoni et al., (2023)). Chaque mention a ainsi été classée dans l'une des catégories suivantes :

- **Syntagmes nominaux définis**, introduits par un déterminant défini (*le chat, la sorcière*). Rentrent dans cette catégorie aussi les syntagmes introduits par un numéral (*trois chatons*).
- **Syntagmes nominaux indéfinis**, introduits par un déterminant indéfini (*un chat, une sorcière*).
- **Syntagmes nominaux possessifs**, introduits par un déterminant possessif (*sa patronne, son balai*). Seuls les déterminants possessifs ont été inclus dans cette catégorie pour le décompte, y compris ceux imbriqués dans des mentions plus élargies.
- **Syntagmes nominaux démonstratifs** (*cette sorcière, ce chat*).
- **Syntagmes nominaux sans déterminant** (*reine, sorcière, chaton*).
- **Noms propres**, simples ou composés (*Victor, Baba Yaga*).
- **Pronoms**, regroupant les pronoms personnels sujet (*je, tu, il, elle, nous, vous, ils, elles*) et objet (*me, te, se, le, la, l', les, leur, moi, toi, soi, lui, eux, elles*), les pronoms réfléchis (*me* ou *-moi, te* ou *-toi, se, (-)nous, (-)vous, se*), le pronom relatif (*qui*) lorsque ceux-ci assurent une fonction référentielle, ainsi que les pronoms indéfinis (ex. *quelqu'un*).
- **Anaphore zéro**, lorsque la référence est assurée uniquement par un verbe, sans réalisation explicite d'un syntagme nominal ou pronominal (*le chat court et **se perd**)*.
- Nous avons aussi prévu la présence d'une **catégorie résiduelle** comprenant notamment certains SN introduits par un marqueur exclamatif en italien (*che bel gattino*). Cette catégorie a été finalement exploitée seulement dans le cas de l'italien et elle est restée vide en français.

Dans les catégories de syntagmes nominaux, la présence éventuelle d'adjectifs ou d'expansions du nom n'a pas été considérée comme discriminante : la catégorisation repose exclusivement sur le type de déterminant qui introduit le syntagme. Il faut noter ici que dans le cas de mentions contenant un nom propre ainsi qu'un syntagme nominal, celles-ci ont été catégorisées sous la catégorie du syntagme nominal. Rentrent dans cette catégorie aussi les syntagmes de dénomination tels que « un robot di nome Antenna di fuoco », *un robot nommé Antenne de feu*.

Comme cela a déjà été montré dans la littérature (voir Chapitre 1 , §6), l'un des descripteurs possibles des chaines de référence est la distribution des maillons selon leur type syntaxique, que l'on peut déterminer sur la base des catégories que nous venons d'illustrer. Cette distribution peut varier sensiblement d'une langue à l'autre en fonction des propriétés structurelles et des contraintes propres au système morphosyntaxique de chaque langue, qui tend en effet à organiser différemment l'usage des catégories grammaticales mobilisées dans les chaines. Elle peut également varier au sein d'une même langue, en fonction notamment du genre textuel considéré, comme plusieurs travaux l'ont souligné (cf. Chapitre 1, §4.3).

Ainsi, nous nous attendions à ce que les deux langues analysées dans notre travail présentent des distributions différentes des typologies de mentions, notamment en ce qui concerne les deux catégories les plus utilisées dans chacune d'elles, à savoir le pronom personnel en français (comme

observé par Federzoni et al., (2020) et l'anaphore zéro en italien, comme le suggèrent nos résultats empiriques. Enfin, nous mènerons une analyse exploratoire afin de déterminer si l'on observe, au sein d'une même langue, des différences dans la distribution sur les différentes catégories ou des variations en comparant les niveaux scolaires analysés.

5.3. Distribution globale des types de mentions dans le corpus

Pour tester ces hypothèses, qui portent sur la structure interne des chaines dans notre corpus, nous avons analysé la distribution des types grammaticaux des mentions en distinguant trois configurations : les chaines, correspondant à des groupes de trois mentions ou plus renvoyant à un même référent ; les paires anaphoriques ou coréférentielles[91], constituées de deux mentions renvoyant à une même entité ; et les singletons, à savoir les mentions liées à un référent n'apparaissant qu'une seule fois dans le texte.

Avant d'examiner plus spécifiquement la structuration linéaire des chaines de référence (cf. §5.5), nous présenterons dans un premier temps la distribution globale des types de mention observées dans l'ensemble de notre corpus annoté. Le Tableau 48 expose la distribution des mentions par type, en comparant le corpus français à l'italien, et la Figure 62 en propose une présentation visuelle, selon les catégories et les langues, en pourcentage.

Type de mention	français	français %	italien	italien %
Pronom	*1314*	*41,82*	942	19,51
SN défini	880	28,01	1267	26,24
SN indéfini	315	10,03	412	8,53
SN possessif	315	10,03	308	6,38
Anaphore Zéro	132	4,2	*1473*	*30,5*
Nom Propre	124	3,95	324	6,71
SN sans déterminant	42	1,34	60	1,24
SN démonstratif	20	0,64	36	0,75
Autre	0	0	7	0,14
Tot.	**3142**	**100**	**4829**	**100**

Tableau 48. Distributions des mentions par type grammatical (toutes mentions confondues).

[91] Nous pourrions utiliser le terme *anaphore*, de manière impropre, pour faire référence aux paires analysées pour un souci de simplification. Ici nous avons décidé d'utiliser le terme *paire* pour éviter cette erreur, pourtant très répandue en TAL.

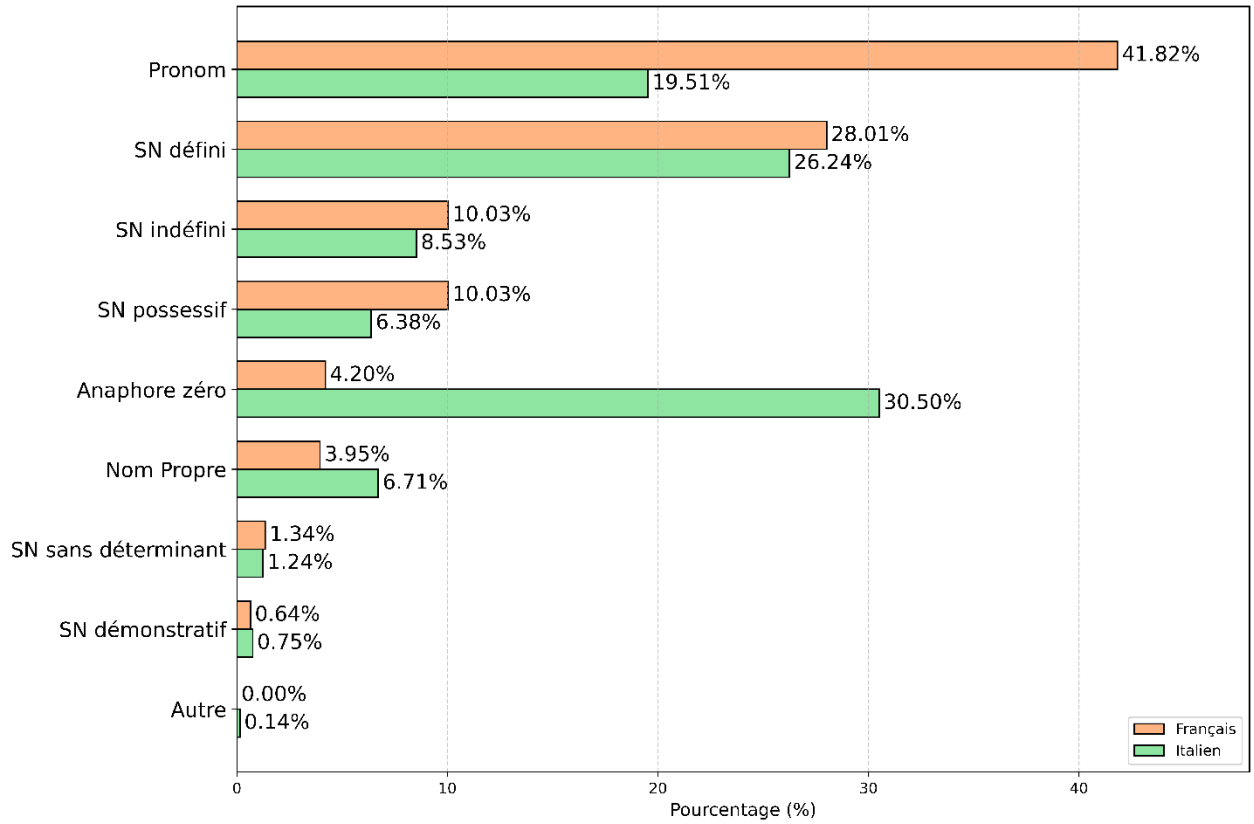


Figure 62. Distribution des mentions par type - comparaison entre corpus français et corpus italien.

Comme nous l'avions anticipé, en postulant une variation entre les deux langues analysées, la distribution des types de mention révèle une différence entre la catégorie majoritaire en français et en italien. En français, les pronoms représentent la catégorie la plus fréquente, 41,82% des mentions, suivis des syntagmes nominaux définis (28,01% des mentions annotées). Cette configuration correspond aux observations de Federzoni et al. (2020) dans leur description des chaines topicales du corpus *Annodis* (dans cette étude, sur un total de 3 456 mentions annotées, 1 026 étaient des pronoms, 1 198 des SN définis, suivis par 442 noms propres et toponymes). En italien, en revanche, les anaphores zéro apparaissent comme la catégorie la plus fréquemment employée (30,50% des mentions annotées), suivies également des syntagmes nominaux définis (26,24% des mentions annotées). Cette prédominance des anaphores zéro était effectivement attendue dans une langue à sujet nul telle que l'italien. Elle est par ailleurs confirmée par l'étude de Perugini et Zucchini (2025) sur les chaines de référence dans des écrits narratifs scolaires produits dans des collèges italiens (cf. Chapitre 2, §4.5), publiée peu après la publication de notre première analyse comparative du corpus annoté (Barletta & Ponton, 2025a), où il apparait également que l'anaphore zéro est la forme la plus fréquemment utilisée en italien. Ces tendances globales constituent un point de référence pour l'analyse plus fine de la distribution des types de mentions selon la langue et le niveau scolaire, en prenant en compte les configurations de chaine, paire anaphorique ou coréférentielle et singletons.

L'analyse de la distribution des types de mentions au sein de chaines de référence (définies ici comme des ensembles d'au moins trois expressions renvoyant au même référent), présentée dans le Tableau 49, vise à déterminer si les tendances observées dans l'ensemble des annotations se

retrouvent également dans la distribution des types dans les chaines de référence. La comparaison opérée entre langues et entre niveaux scolaires nous permet de tester l'hypothèse d'une variation dans l'utilisation des différents types de maillons selon ces deux facteurs. La mise en évidence de cette variation pourrait en effet contribuer à décrire de manière plus efficace les attentes vis-à-vis de la construction de la continuité référentielle selon le niveau scolaire.

Type ER	**français**				**italien**			
	CE1	***CE1%***	**CE2**	***CE2%***	**CE1**	***CE1%***	**CE2**	***CE2%***
Pronom	499	*45,57*	743	*41*	377	*20,02*	543	*20,26*
SN défini	271	*24,75*	550	*30,35*	494	*26,23*	709	*26,46*
SN indéfini	120	*10,96*	151	*8,33*	173	*9,19*	196	*7,31*
SN possessif	101	*9,22*	183	*10,10*	88	*4,67*	177	*6,60*
Anaphore Zéro	50	*4,57*	76	*4,19*	620	*32,93*	775	*28,92*
Nom Propre	36	*3,29%*	74	*4,08*	88	*4,67*	228	*8,51*
SN sans déterminant	16	*1,46*	20	*1,10*	28	*1,49*	25	*0,93*
SN démonstratif	2	*0,18*	15	*0,83*	10	*0,53*	25	*0,93*
Autre	-	-	-	-	5	*0,27*	2	*0,07*
Tot.	**1095**	***100***	**1812**	***100***	**1883**	***100***	**2680**	***100***

Tableau 49. Détail de la distribution des types de mentions dans les chaines - comparaison entre corpus et niveaux scolaires.

Le Tableau 49 décrit la distribution des différents types de mentions au sein des chaines, en valeurs absolues ainsi qu'en pourcentage par rapport au total des mentions pour chaque sous-partie du corpus (par exemple l'ensemble des mentions des chaines du CE1 français). Ces proportions sont utilisées ici afin de rendre les données comparables, dans la mesure où, comme nous avons montré dans la section dédiée à la densité référentielle (cf. ce chapitre, §3), le rapport entre le nombre de mentions et la longueur des textes est globalement stable dans le corpus et surtout pour chaque langue. Cette normalisation en pourcentage permet donc de comparer les profils de distribution sur une même échelle, du moins entre niveaux scolaires de la même langue.

À un niveau général, la distribution dans les chaines confirme les tendances déjà observées dans l'ensemble des annotations : en français, les pronoms constituent la catégorie la plus fréquente au sein des chaines (45,57% et 41%), tandis qu'en italien ce rôle est assumé par les anaphores zéro. Dans les deux langues, les syntagmes nominaux définis occupent la deuxième position en termes de fréquence, avec des valeurs entre 24,75% et 30,35% dans les différents niveaux.

SI l'on examine les variations entre niveaux scolaires pour chaque langue, on observe en français une diminution du pourcentage des pronoms (de 45,57% à 41%), et une augmentation des SN définis (de 24,75% à 30,35%). En italien, on a une diminution des anaphores zéro (de 32,93% au CE1 et 28,92% au CE2). On a aussi une augmentation des noms propres, qui passent de 4,67% à 8,51%, et des SN possessifs, qui varient de 4,67% à 6,60%. Ces évolutions quantitatives globales ne renseignent cependant pas encore sur la manière dont les différentes formes se répartissent à l'intérieur des chaines. Pour l'approfondir, il convient donc d'examiner plus précisément la position occupée par les maillons dans la chaine.

L'analyse des deux premières positions des chaines peut être menée à partir du Tableau 50, qui détaille leur distribution selon les niveaux scolaires, et de la Figure 63, qui présente la distribution des types de mentions sur les cinq premiers maillons des chaines, en comparant les deux langues.

Type ER	**français**				**italien**			
	Position 1		*Position 2*		*Position 1*		*Position 2*	
	CE1	**CE2**	**CE1**	**CE2**	**CE1**	**CE2**	**CE1**	**CE2**
Pronom	13	13	***83***	***87***	8	4	***68***	***49***
SN défini	**38**	**47**	***34***	***42***	**41**	**81**	***35***	***45***
SN indéfini	**91**	**111**	15	22	**123**	**90**	***25***	***66***
SN possessif	6	15	11	25	9	14	7	17
Anaphore Zéro	0	0	5	5	11	17	***48***	***27***
Nom Propre	2	3	1	6	2	9	4	7
SN sans déterminant	0	0	0	4	0	1	4	1
SN démonstratif	0	3	2	1	0	0	2	4
Autre							1	
Tot.	**150**	**192**	**151**	**192**	**194**	**216**	**194**	**216**

Tableau 50. Distribution des types de mentions en position 1 et 2 des chaines.

Comme nous pouvons observer à travers le Tableau 50, dans les deux langues la première position dans la chaine est majoritairement occupée par des syntagmes nominaux, indéfinis (202 au total en français, 213 en italien) puis définis (85 en français, 122 en italien), ce qui correspond à une stratégie attendue d'introduction du référent à l'aide de mentions explicites.

Comme nous allons approfondir par la suite (cf. Chapitre 8, §2.1), la présence de titres dans les textes, dont les mentions ont été inclues dans nos annotations, pourrait avoir influencé la présence de SN définis en position 1 des chaines. Le Tableau 51 met en relation le nombre de textes comportant un titre et le nombre de SN définis en position 1. Cette relation apparaît toutefois variable selon le niveau scolaire. En CE1, en français comme en italien, le nombre relativement faible de textes comportant un titre ne permet pas d'établir un lien clair avec la fréquence des SN définis en position initiale. En revanche, au niveau CE2, où les titres sont plus fréquents, notamment en italien, on observe une augmentation concomitante du nombre de SN définis en position 1, suggérant un effet plus marqué de la structure paratextuelle sur l'introduction des référents.

Niveaux scolaires	**français**		**italien**	
	Textes avec titres	*SN défini position 1*	*Textes avec titres*	*SN défini position 1*
CE1 – 2	7	34	7	40
CE2 – 3	22	45	42	79
CE1-CE2 – 2-3	**29**	**79**	**49**	**119**

Tableau 51. Distribution des titres et des SN définis en position 1 et 2 des chaines.

La deuxième position dans les chaines montre une organisation encore largement comparable entre les deux langues, la catégorie la plus fréquente étant celle du pronom dans les deux langues

(170 en français, 117 en italien), suivie par le syntagme nominal défini en français (76), tandis que les syntagmes nominaux indéfinis (91), les syntagmes définis (80) et les anaphores zéro (75) ont une fréquence comparable en italien. Dans la section dédiée à la composition des chaines (cf. Chapitre 8, §2) nous allons approfondir le rôle de certains de ces syntagmes selon leur position dans les chaines, dans le cas des types plus fréquents tels que les syntagmes définis en première position et les pronoms en deuxième, ou dans le cas de catégories bien moins représentées, telles que les SN possessifs en première position (cf. Chapitre 8, §2.3), qui sont peu fréquentes (seulement 21 et 23 cas en français et en italien) mais qui représentent des cas intéressant à étudier du point de vue linguistique, notamment pour approfondir comment l'ancrage à un référent existant est utilisé pour introduire un nouveau personnage. Nous pouvons maintenant analyser les distributions sur les cinq premières positions des chaines à partir de la Figure 63.

Nous pouvons à présent analyser les distributions sur les cinq premières positions des chaines à partir de la Figure 63. On observe par ailleurs que la majorité des mentions se concentre dans les premières positions des chaines : 1317 mentions sur 1571 en français et 1594 sur 1921 en italien apparaissent entre les positions 1 et 4.

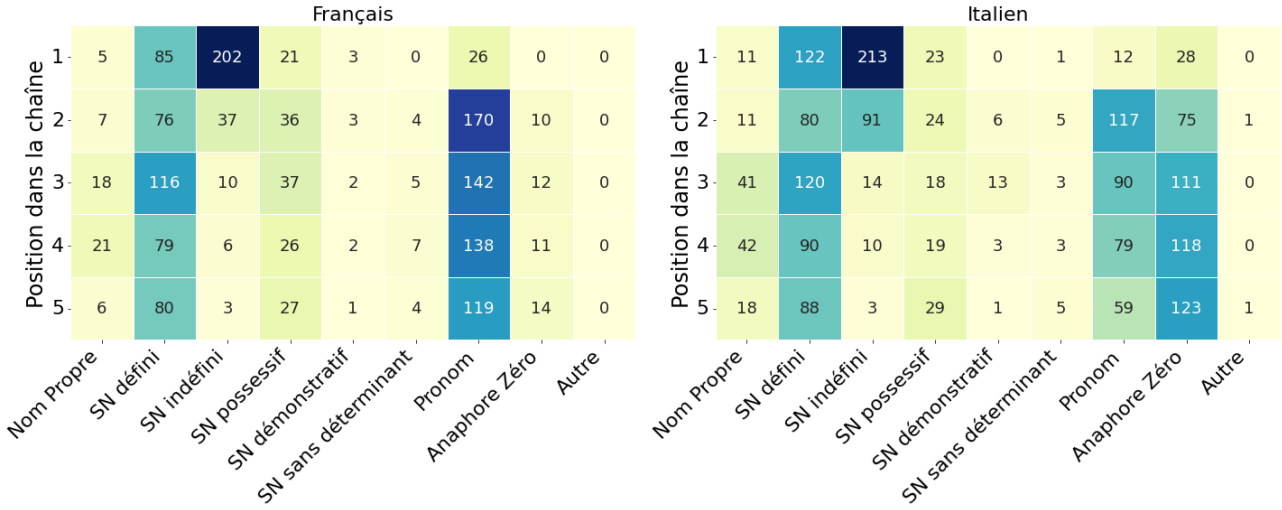


Figure 63. Distribution des types de mentions sur les 5 premiers maillons des chaines - comparaison entre langues.

En français, les positions médianes et finales sont dominées par les pronoms, qui restent la catégorie la plus nombreuse dans ces positions, entre 142 et 119, suivie par les SN définis, qui représentent la deuxième catégorie la plus fréquemment mobilisée dans le déroulement de la chaine, allant entre 116 et 79. Cela reflète une préférence marquée pour des reprises anaphoriques plus atténuées dans ces positions intermédiaires, mais aussi l'une des observations empiriques déjà formulées à propos de ce corpus, à savoir l'existence d'une division presque nette entre deux stratégies de gestion des chaines, l'une fondée principalement sur les SN définis, l'autre sur les pronoms. Ces deux stratégies seront décrites plus en détail dans une section ultérieure (cf. Chapitre 8, §3).

En revanche, dans les positions 3 à 5, l'italien se distingue par une présence particulièrement importante de SN définis (entre 120 et 90) et d'anaphores zéro (entre 123 et 111), devant les pronoms (de 90 à 59). Sans revenir sur la hiérarchie générale des catégories déjà observée dans la

distribution globale, cette configuration met surtout en évidence deux propriétés structurelles du corpus italien : d'une part, le recours massif au sujet nul dans le maintien référentiel, et d'autre part, l'existence de chaines construites de manière largement homogène, soit sur la base de SN définis, soit sur la base d'anaphores zéro. Cette organisation est ainsi pleinement comparable, dans son principe, aux stratégies homogènes observées en français, bien que les formes privilégiées diffèrent entre les deux langues.

On notera enfin quelques différences dans les catégories moins représentées dans les positions intermédiaires. Les SN possessifs sont légèrement plus fréquents en français qu'en italien (par exemple en position 3, 37 contre 18 ; en position 4, 26 contre 19), tandis que les noms propres sont en revanche plus nombreux en italien qu'en français (en position 3, 41 contre 18 ; en position 4, 42 contre 21 ; en position 5, 18 contre 6). Ces écarts suggèrent des préférences différentes dans l'exploitation des formes plus informatives pour le maintien ou la réactivation du référent. La comparaison entre les distributions observées en français en CE1 et en CE2, présentées dans la Figure 64, permet de préciser l'évolution de l'utilisation de types de mentions différents en français en fonction du niveau scolaire. Au total, cette distribution concerne 689 mentions en CE1 et 882 mentions en CE2.

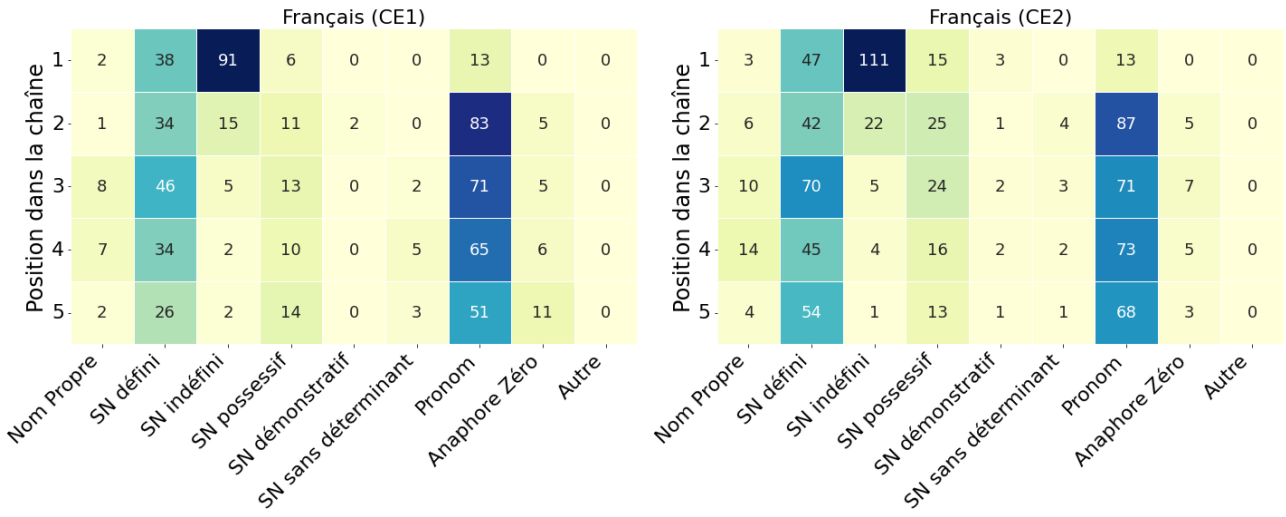


Figure 64. Distribution des types de mentions sur les cinq premiers maillons des chaines dans le corpus français - comparaison entre CE1 et CE2

Dans les deux niveaux, la distribution reste globalement similaire à celle observée sur l'ensemble du corpus français : les SN indéfinis, catégorie la plus fréquente en position 1, sont suivis majoritairement par des pronoms dans les positions 2 à 5, avec une certaine continuité dans l'utilisation des SN définis dans toutes les positions et dans les deux niveaux. On observe toutefois en CE2 une diversification plus marquée dans les positions 3 à 5, avec une augmentation notable du recours aux SN définis (70 en position 3, 45 en position 4 et 54 en position 5), aux SN possessifs (respectivement 24, 16 et 13) et aux noms propres (moins nombreux, 10, 14 puis seulement 4). Cette variation suggère une tendance, chez les élèves de CE2, à recourir aussi à des SN plus informatifs, tels que les SN définis ou les SN possessifs, pour réactiver ou maintenir le référent, plutôt qu'à s'appuyer majoritairement sur les formes pronominales.

Une comparaison analogue peut être menée pour le corpus italien, dont la Figure 65 présente la distribution des mentions par type grammatical sur les cinq premiers maillons des chaines en CE1 et en CE2. Au total, cette distribution repose sur 905 mentions en CE1 et 1016 mentions en CE2, ce qui confirme une augmentation globale du nombre de maillons annotés avec le niveau scolaire.

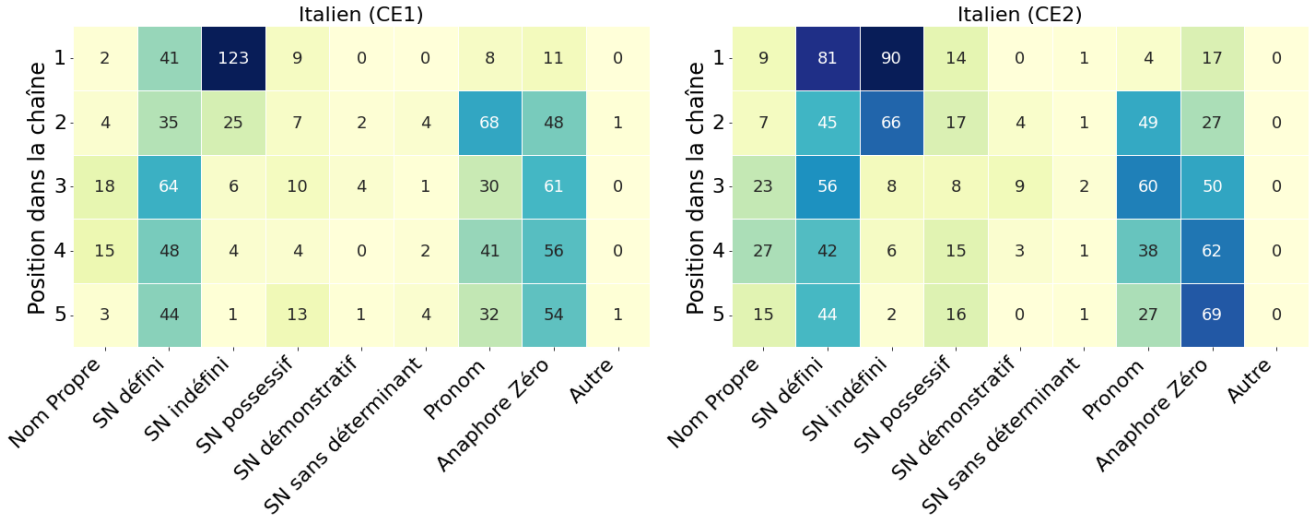


Figure 65. Distribution des mentions par type grammatical sur les cinq premiers maillons des chaines dans le corpus italien - comparaison entre CE1 et CE2

Tout comme en français, et selon les attentes liées au genre textuel analysé, les chaines s'ouvrent majoritairement à travers des SN indéfinis, qui réalisent le « pivot référentiel » déjà mentionné auparavant. Bien que les SN indéfinis représentent la catégorie dominante en position 1 au CE1, cela varie au CE2, où les SN définis et indéfinis partagent une fréquence comparable en position 1 des chaines. En position 2, les pronoms forment la catégorie la plus peuplée (68), suivie par les anaphores zéro (48). Cette distribution varie au CE2, où ce sont les SN indéfinis les plus nombreux (66), suivis par les pronoms (49) et les SN définis (45), et par les anaphores zéro en dernier (27).

Dans les positions 3 à 5 des chaines, les catégories des SN définis, pronoms et anaphores zéro semblent partager des fréquences similaires, avec une nette augmentation de l'utilisation de l'anaphore zéro en position 5 de la chaine. Cela semble suggérer, avec la progression scolaire, à la fois un recours plus systématique aux chaines reposant sur des formes verbales, et le maintien de référents dont la gestion s'appuie principalement sur des SN définis ou sur des pronoms, de manière comparable à ce que l'on observe dans le cas du français. Nous examinerons dans la suite, à travers une analyse qualitative, dans quelle mesure ces tendances quantitatives se traduisent effectivement par la mise en place de stratégies de gestion référentielle homogènes (cf. Chapitre 8).

5.4. Distribution dans les paires anaphoriques et dans les singletons – comparaison entre français et italien

Après avoir examiné ces distributions à l'intérieur des chaines, qui représentent la plus grande partie des mentions annotées dans notre corpus, il convient à présent d'analyser cette même distribution dans les mentions faisant partie de groupes moins nombreux de trois expressions

référentielles. Nous avons déjà précédemment mentionné ces configurations comme *paires anaphoriques* pour les groupes de deux mentions et *singletons* quand une seule mention est présente. Nous analysons à présent la distribution des mentions dans les paires anaphoriques, présentées dans le Tableau 52.

Paires anaphoriques				
Type ER	**français**		**italien**	
	CE1	**CE2**	**CE1**	**CE2**
Pronom	12	33	7	9
SN défini	20	16	13	15
SN indéfini	1	4	14	28
SN possessif	7	11	9	14
Anaphore Zéro	13	12	12	14
Nom Propre	5	1	0	4
SN sans déterminant	1	3	1	2
SN démonstratif	1	1	0	0
Autre	0	1	0	0
Tot.	**60**	**82**	**56**	**86**

Tableau 52. Distribution des mentions des paires anaphoriques/coréférentielles par type grammatical.

La distribution des typologies de mentions met en évidence des évolutions contrastées selon la langue. En français, l'augmentation entre le CE1 et le CE2 concerne principalement les pronoms, dont la fréquence progresse de manière notable, ainsi que, dans une moindre mesure, les syntagmes nominaux possessifs et indéfinis. Les anaphores zéro, en revanche, ne présentent pas d'augmentation entre les deux niveaux scolaires, ce qui confirme leur statut marginal dans le système référentiel du français. En italien, les variations les plus marquées dans les paires anaphoriques concernent les syntagmes nominaux indéfinis, qui doublent entre le CE1 et le CE2, ainsi que les syntagmes nominaux possessifs. Les anaphores zéro montrent également une légère augmentation, tandis que les autres catégories demeurent globalement stables. Cette distribution reflète une diversification modérée des ressources mobilisées dans les paires, sans remise en cause des préférences référentielles propres à chaque langue.

Au niveau des singletons, illustrés dans le Tableau 53, l'analyse révèle des distributions distinctes entre le français et l'italien.

Singletons				
Type ER	**français**		**italien**	
	CE1	***CE2***	***CE1***	***CE2***
Pronom	1	0	20	16
SN défini	12	9	10	24
SN indéfini	7	19	3	6
SN possessif	6	7	5	15
Anaphore Zéro	14	5	12	5
Nom Propre	1	7	0	4
SN sans déterminant	0	1	2	0
SN démonstratif	0	2	0	1
Autre	1	0	1	0
Tot.	**42**	**50**	**53**	**71**

Tableau 53. Distribution des singletons par type grammatical.

En français, les singletons sont majoritairement constitués de syntagmes nominaux indéfinis, suivis des syntagmes nominaux définis, tandis que les pronoms y sont très peu représentés. En italien, en revanche, les pronoms constituent la catégorie de singletons la plus fréquente, suivis des syntagmes nominaux définis, puis des syntagmes nominaux possessifs. Cette distribution peut paraître contre-intuitive dans la mesure où certaines catégories associées à une forte accessibilité référentielle, notamment les pronoms, apparaissent comme des maillons isolés. Ce phénomène s'explique par les choix méthodologiques opérés pour le comptage des chaines, qui considèrent comme nouvelle chaine tout cas où plusieurs référents sont indiqués par un même maillon, afin d'éviter la double comptabilisation d'une même mention dans plusieurs chaines. Ce choix a pu entraîner une sur-représentation de certaines catégories, en particulier des pronoms en italien, parmi les singletons.

5.5. Séquences de mentions – analyse des patterns plus fréquents

Après avoir décrit la distribution globale des types de mentions dans le corpus en fonction de leur position dans les chaines, et comparé les variations selon la langue et le niveau scolaire, nous abordons à présent les différentes séquences formées par les trois premières mentions des chaines et leur fréquence dans les corpus français et italien. L'examen des dix séquences les plus fréquentes constituées par les trois premiers maillons permet de préciser la structuration linéaire ou pattern des chaines de référence et de compléter l'analyse quantitative précédente par une approche plus exploratoire. Pour cette analyse, nous avons retenu uniquement les chaines correspondant à un seul référent.

La Figure 66 décrit les dix séquences les plus fréquentes du corpus français, représentant au total 200 chaines annotées.

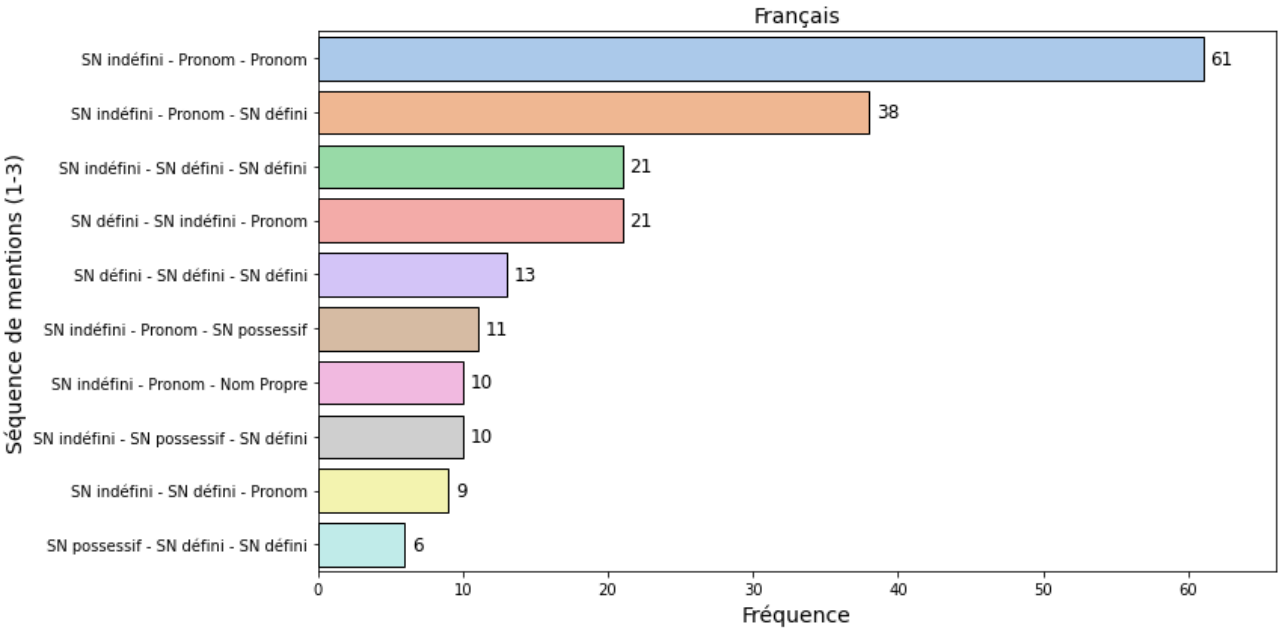


Figure 66. Distribution des *patterns* les plus fréquents des trois premières mentions dans 200 chaines de référence du corpus français.

Nous analyserons plus en détail les cinq premiers de ces patterns, notamment parce qu'ils figurent également parmi certaines des séquences les plus fréquentes en italien, comme on le verra par la suite.

On observe que la séquence la plus fréquente (61 occurrences) est constituée d'un SN indéfini, suivi de deux pronoms. Les chaines caractérisées par ce pattern vont de 3 à 23 mentions et leur longueur moyenne est de 9,1 mentions. Ce pattern correspond à une stratégie d'introduction explicite du référent, immédiatement suivie de reprises pronominales, plus atténuées et moins informatives. Ce schéma correspond à la chaine du robot du texte (74), composée par les expressions référentielles « un robot », « qui » et « il » dans les trois premières positions de la chaine.

(74) C'est **[un robot]**^{robot} **[qui]**^{robot} avait envie d'aller dans la forêt. Mais **[il]**^{robot} y voit **[un loup]**^{wolf} **[qui]**^{wolf} rêvait de manger **[le robot]**^{robot}. (…) (corpus français, CE1, 217)

La deuxième séquence la plus fréquente (38 occurrences) associe également un SN indéfini en première position à des reprises pronominales ou nominales définies (SN indéfini - pronom - SN défini), confirmant le rôle central du syntagme nominal indéfini dans l'amorce des chaines en français. Ce pattern est illustré par la chaine du *chat* dans le texte (75), constituée des syntagmes « un chat », « qui » est « le chat ».

(75) Il était une fois **[un chat]**^{cat} **[qui]**^{cat} se promenait dans un bois et **[il]**^{robot} arrive **[un robot]**^{robot} **[qui]**^{robot} s'appelait **[Roboti]**^{robot} et **[il]**^{robot} avait 6 ans, et **[le chat]**^{cat} était gris et blanc (…) (corpus français, CE1, 824)

Le pattern suivant débute également par un SN indéfini, repris par deux SN définis. Il s'inscrit dans la même stratégie homogène de constructions des chaines par des reprises nominales explicites que nous avons déjà mentionnée précédemment (cf. ce chapitre, §5.3), comme dans le cas de la chaine du *chat* du texte (76).

(76) Il était une fois **[un loup - garou]**^{wolf} **[qui]**^{wolf} voulait attaquer **[un chat]**^{cat}. **[Le chat]**^{cat} vivait dans une maison. **[Le chat]**^{cat} était tout seul. (…) (corpus français, CE1, 2983)

On observe toutefois, parmi les patterns les plus fréquents, une séquence qui pourrait apparaître comme atypique au regard des autres patterns analysés, à savoir SN défini - SN indéfini - pronom. Comme l'illustre la chaine de la *sorcière* de l'exemple (77), ce type de configuration peut s'expliquer par des productions comportant un titre, suivi de l'introduction narrative proprement dite du référent. Ce pattern ne constitue donc pas une anomalie référentielle, mais reflète des conventions textuelles spécifiques aux écrits narratifs ainsi que nos choix d'annotations.

(77) **[La sorcière]**^{witch} et **[le chat]**^{cat} Il était une fois **[une sorcière]**^{witch} **[qui]**^{witch} se promenait dans une forêt. (…) (corpus français, CE1, 3012)

Les autres patterns fréquents, avec des fréquences comprises entre 13 et 6 occurrences, présentent des configurations variées : la répétition du SN défini (13 occurrences), l'introduction par un SN indéfini suivie d'un pronom puis d'un SN ou d'un déterminant possessif (11 occurrences), ou encore l'alternance entre SN indéfini, pronom et nom propre (10 occurrences). Ce dernier pattern est illustré par la chaine du *loup* de l'exemple (78), et correspond à une introduction du référent du type « un N qui s'appelait X », où X est habituellement un nom propre. Cette séquence trouve sa correspondance sémantique dans les maillons de dénomination explicite du type « un loup nommé Ismaël », que nous allons approfondir dans une analyse qualitative ultérieure (cf. Chapitre 8, §1.2).

(78) Il y a **[une sorcière]**^{witch} **[qui]**^{witch} habitait dans une vieille maison. **[Elle]**^{witch} dormait. Il y avait **[un loup]**^{wolf} **[qui]**^{wolf} s'appelait **[Ismaël]**^{wolf}. **[Il]**^{wolf} volait des bijoux et des bonbons. (…) (corpus français, CE1, 1317)

La Figure 67 illustre les dix patterns les plus fréquents dans les chaines des référents singuliers du corpus italien. 198 chaines au total sont représentées dans cette figure.

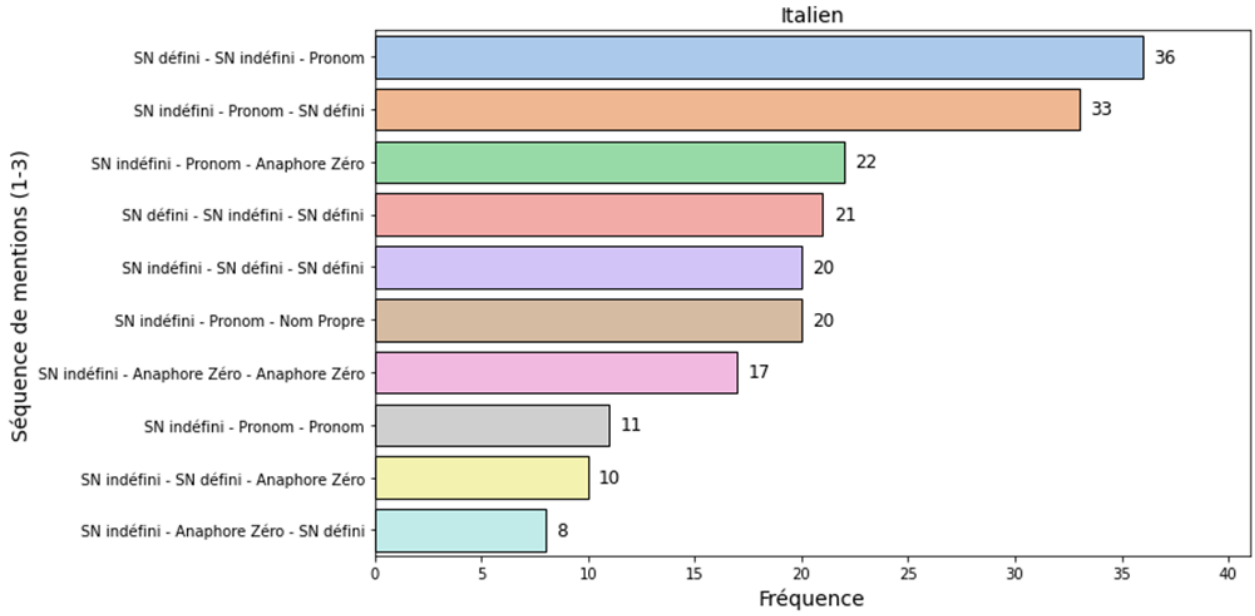


Figure 67. Distribution des patterns les plus fréquents des trois premières mentions dans 198 chaines de référence du corpus italien.

En italien, à différence du français, la première séquence est représentée par la succession de SN défini, SN indéfini et pronom. Celle-ci était la quatrième séquence la plus fréquente en français. Ce pattern caractérise 36 chaines au total, allant de 7 à 35 mentions, avec une longueur moyenne de 19 mentions. Dans 35 cas, ces chaines démarrent avec le syntagme nominal défini contenu ou qui constitue le titre du texte, comme dans l'exemple (79).

(79) **[Il gattino povero]**^{cat} C'era una volta **[un gattino selvaggio]**^{cat} **[che]**^{cat} viveva in un villaggio. (…) (corpus italien, CE2, 1663)

Le seul cas de ce pattern qui ne correspond pas à un titre on le retrouve dans le texte (80). Ici, la chaine du personnage *ext1* démarre par une définition du statut du personnage vis-à-vis du chat, déjà introduit. Le SN défini a fonction d'ancrage ici pour l'introduction du référent, successivement identifié comme « une princesse », à l'aide d'un SN indéfini. Cette expression référentielle démarre à son tour l'introduction du personnage à l'aide de la formule d'introduction du référent que nous avons identifiée précédemment, « un SN qui s'appelait X », où X est un nom propre (ici, la contrepartie italienne de Cendrillon).

(80) C'era una volta **[una gatta]**^{cat} **[il proprietario]**^{ext1} era **[una principessa]**^{ext1} **[che]**^{ext1} si chiamava **[Cenerentola]**^{ext1}. (…) (corpus italien, CE1, 707)

Le deuxième pattern le plus fréquent, caractérisé par la succession SN indéfini - pronom - SN défini, présente des propriétés intéressantes du point de vue de l'étude des modalités de cohabitation des chaines (cf. Chapitre 1, §4.4 où nous avions déjà abordé cette notion). Il est également le deuxième pattern le plus fréquent en français. À l'observation empirique, ce pattern de chaine semble apparaître fréquemment en alternance avec d'autres chaines. Ici, la chaine du *chat* du texte (81) et la chaine de la *sorcière* du texte (82) correspondent à cette description.

(81) C'era una volta **[una strega]**witch **[che]**witch viveva in una casetta in montagna e **[viveva]**witch da sola. Un giorno mentre **[volava]**witch sulla **[sua]**witch scopa **[vide]**witch **[un gattino]**cat. **[Lo]**cat **[guardò]**witch e **[vide]**witch che **[il micetto]**cat non aveva il collare quindi **[decise]**witch di portar**[lo]**cat con **[sé]**witch. (…) (corpus italien, CE1, 15)

(82) C'era una volta **[un gatto birichino]**cat **[che]**cat andava sempre in giro senza il permesso **[dei [suoi]cat genitori]**ext1. Un giorno **[cadde]**cat in una buca buia che C'era dentro **[una strega]**witch molto ma molto cattiva **[che]**witch **[lo]**cat rapì. **[Il gatto]**cat tutto impaurito cercò di scappare ma **[la strega]**witch glie**[lo]**cat impedì (…). (corpus italien, CE1, 186)

Cependant, nous retrouvons ce pattern aussi bien avec fonction d'introduction des référents, comme dans le texte (83), grâce à la présence du SN indéfini en début de chaine. Dans certains cas, la reprise pronominale est immédiate, et sert à introduire une phrase relative qui décrit la situation initiale dans laquelle se trouve le personnage en question, puis repris de manière explicite à l'aide du SN défini. Dans ce texte, les chaines de chat et de la sorcière présentent ces caractéristiques.

(83) C'era una volta **[un gatto]**cat **[che]**cat gironzolava nei boschi, poi comparve una casa in mezzo agli alberi, **[il gatto]**cat entrò e **[vide]**cat **[una strega]**witch **[che]**witch stava facendo la pozione per diventare tutto quello che esisteva. Poi il giorno dopo **[il gatto]**cat ritornò là **[dalla strega]**witch. (corpus italien, CE2, 773)

Le troisième pattern le plus fréquent partage les deux premiers maillons avec le pattern précédent : sa configuration est SN indéfini – pronom – anaphore zéro. En reprenant l'exemple déjà vu en (81), dans le texte (84) nous pouvons observer la chaine de la *sorcière* ayant cette structure.

(84) C'era una volta **[una strega]**witch **[che]**witch viveva in una casetta in montagna e **[viveva]**witch da sola. Un giorno mentre **[volava]**witch sulla **[sua]**witch scopa **[vide]**witch **[un gattino]**cat. **[Lo]**cat **[guardò]**witch e **[vide]**witch che **[il micetto]**cat non aveva il collare quindi **[decise]**witch di portar**[lo]**cat con **[sé]**witch. (…) (corpus italien, CE1, 15)

De manière similaire à la structure précédente, le SN indéfini d'introduction est repris par une relative, et l'anaphore zéro représente un ajout à l'information contextuelle contenue dans la relative, qui présente la situation initiale ou la condition initiale du personnage.

Ainsi, l'examen des séquences formées par les trois premiers maillons des chaines permet de préciser la structuration linéaire des chaines de référence dans les deux corpus. Malgré certaines différences dans la fréquence des formes privilégiées, les résultats mettent en évidence des configurations largement comparables dans les deux langues. Dans la majorité des cas, les chaines s'ouvrent par un syntagme nominal indéfini qui introduit le référent, conformément à la fonction discursive typiquement associée à ce type d'expression référentielle. Cette introduction est fréquemment suivie d'une reprise pronominale ou nominale définie, et dans certains cas d'une anaphore zéro, en particulier dans le corpus italien.

Les différentes séquences observées montrent également que l'introduction du référent s'inscrit souvent dans des structures syntaxiques plus larges, notamment à travers la présence d'une

proposition relative immédiatement associée au syntagme nominal indéfini. Cette structure permet d'enrichir dès l'introduction la représentation discursive du référent, en apportant une première caractérisation ou en décrivant la situation initiale dans laquelle se trouve le personnage. Dans ces configurations, les reprises ultérieures, qu'elles prennent la forme de pronoms, de syntagmes nominaux définis ou d'anaphores zéro, s'inscrivent dans une continuité informationnelle déjà amorcée par cette structure introductive.

La comparaison entre les deux langues met toutefois en évidence certaines différences dans les formes privilégiées pour assurer cette continuité référentielle. Alors que les chaines françaises reposent très fréquemment sur des reprises pronominales dans les positions suivantes, les chaines italiennes présentent plus souvent des configurations associant anaphores zéro et syntagmes nominaux définis. Ces différences reflètent en grande partie les propriétés morphosyntaxiques propres à chaque langue, notamment la possibilité du sujet nul en italien.

Dans l'ensemble, l'analyse des patterns confirme donc les tendances déjà observées dans les sections précédentes : l'introduction des référents repose majoritairement sur des syntagmes nominaux indéfinis, tandis que leur maintien dans le discours mobilise des formes plus atténuées et moins informatives, dont la distribution varie selon la langue. Ces résultats permettent ainsi de mieux caractériser les stratégies de construction des chaines de référence dans les récits produits par les élèves et constituent un point d'appui pour les analyses qualitatives présentées dans le chapitre suivant.

6. Conclusion

Dans ce chapitre, nous avons proposé une première caractérisation quantitative des chaines de continuité référentielle observées dans le corpus *Scolinter*. Les analyses ont été réalisées sur un *corpus gold* obtenu par adjudication manuelle à partir des annotations de deux annotateurs pour chaque texte. Le corpus comprend 148 textes en français (74 de niveau CE1, 74 de niveau CE2) et 150 textes en italien (75 de niveau CE1, 75 de niveau CE2). À partir de ces données, nous avons examiné plusieurs indicateurs décrits dans le cadre théorique (cf. Chapitre 1, §4.2), permettant de caractériser les caractéristiques référentielles des textes et de décrire les chaines qu'elles forment. Nous avons ainsi analysé le nombre de mentions, de chaines, de paires référentielles et de singletons, et leurs distributions dans les textes ; la longueur des mentions, pour en saisir l'éventuelle complexité ; le nombre de référents dans les textes et les binômes fréquents de personnages ; la longueur des chaines, la distance entre mentions d'une chaine ou interdistance, la densité référentielle, la stabilité référentielle, et enfin la distribution des types de mentions selon leur position dans la chaine.

Le choix de ces indicateurs répond à un double objectif. D'une part, ils permettent de proposer une première description quantitative du corpus annoté et de caractériser certaines propriétés générales de la gestion référentielle dans les textes produits par les élèves. D'autre part, ces indices ont été retenus parce qu'ils sont fréquemment utilisés dans les travaux portant sur les corpus annotés en coréférence et permettent ainsi de mettre en perspective les propriétés de notre corpus avec celles observées dans d'autres ressources comparables.

Tous les indicateurs sélectionnés sont calculables à partir des niveaux d'annotation réalisés dans ce travail. Certains d'entre eux ne sont pas directement comparables aux indices issus d'autres corpus, tels que la longueur des textes ou des chaines, mais ils permettent néanmoins de rendre compte de manière fine du contenu référentiel des textes et d'articuler différentes propriétés quantitatives. Certains servent ainsi de base au calcul de mesures dérivées, essentielles pour notre analyse, comme la longueur des chaines, utilisée dans le calcul de l'interdistance, ou le nombre de mentions, mobilisé pour le calcul de la densité référentielle.

Dans cette perspective, l'analyse du nombre de mentions et la distribution entre chaines, paire référentielles et anaphores, et leur rapport à la longueur des textes, constitue un premier point d'entrée pour comparer les deux langues. Cela met tout d'abord en évidence certaines différences globales entre les deux langues étudiées. Les textes du corpus italien apparaissent en moyenne plus longs et comportent un nombre plus élevé de mentions et de chaines que les textes du corpus français. Cette différence semble toutefois largement liée à la longueur globale des productions plutôt qu'à des variations dans la gestion référentielle : les textes italiens comportant davantage de tokens (143,66 tokens par texte en moyenne) que les textes français (94,94 tokens par texte en moyenne), ils offrent davantage d'occasions de développement des chaines de référence. Dans ce contexte, il est intéressant de noter que les textes de CE2 en français présentent une longueur moyenne très proche de celle des textes de CE1 en italien, ce qui se reflète également dans le nombre total de mentions observées dans ces deux sous-corpus. Ce résultat suggère que certaines

différences observées entre les deux langues pourraient être en partie liées à la longueur des productions plutôt qu'à des différences dans l'organisation référentielle elle-même.

Afin de vérifier plus systématiquement cette relation entre longueur des textes et nombre de mentions, nous avons examiné la densité référentielle des textes. Cette analyse a été motivée par une première observation empirique, selon laquelle les textes italiens semblaient présenter une densité référentielle plus élevée que les textes français. Le calcul de cet indice permet en effet d'évaluer la proportion d'expressions référentielles relativement à la longueur des textes. Les résultats montrent que, lorsque la densité est calculée selon sa mesure classique (c'est-à-dire nombre de mentions sur nombre de tokens du texte), les différences observées entre les deux langues ne sont pas statistiquement significatives. En revanche, lorsque la densité est calculée en prenant explicitement en compte le nombre total de tokens contenus dans les mentions d'un texte, la différence entre les deux langues devient significative, bien que pas importante (la densité du français étant à 34,32 et celle de l'italien à 36,65). Ces résultats suggèrent que la relation entre longueur des textes et nombre d'expressions référentielles dépend en partie de la manière dont cet indice est opérationnalisé.

L'analyse du nombre de référents par texte montre par ailleurs que les récits produits par les élèves mobilisent un nombre relativement limité d'entités. En moyenne, les textes comportent entre deux et quatre référents. Toutefois, la majorité des productions se concentre autour de deux personnages principaux, ce qui correspond globalement à la consigne de production narrative donnée aux élèves. Les textes ne comportant qu'un seul référent restent relativement rares dans le corpus, mais ils présentent un intérêt particulier pour l'analyse qualitative dans la mesure où ils mettent en évidence certaines stratégies spécifiques de construction du récit. Ces cas seront examinés plus en détail dans le chapitre suivant. Nous analyserons également des textes présentant des configurations à deux référents, afin d'approfondir les stratégies de cohabitation des chaines mises en place par les élèves dans ces cas.

L'analyse de la longueur des expressions référentielles met en évidence une forte prédominance des mentions très courtes dans l'ensemble du corpus. La majorité des expressions référentielles est constituée d'un seul token, suivie par les mentions composées de deux tokens, tandis que les expressions plus longues apparaissent de manière nettement plus marginale. Ces mentions longues, bien que peu fréquentes, présentent un intérêt particulier pour l'analyse linguistique et didactique, dans la mesure où elles correspondent souvent à des syntagmes nominaux enrichis permettant de caractériser les personnages ou d'introduire des informations descriptives supplémentaires.

L'étude de la stabilité référentielle a permis d'examiner la variation lexicale à l'intérieur des chaines de référence. Toutefois, l'application de cet indice à notre corpus met en évidence certaines limites méthodologiques. En effet, les mesures de stabilité reposent uniquement sur les syntagmes nominaux et excluent par définition les autres types d'expressions référentielles, tels que les pronoms ou les anaphores zéro. Or, ces formes jouent un rôle particulièrement important dans la structuration des chaines dans les textes du corpus, en particulier dans le cas de l'italien. Cette restriction conduit donc à exclure une partie significative du matériel référentiel annoté et limite

la capacité de cet indice à rendre compte de la diversité réelle des stratégies référentielles observées dans les textes. Cependant, les résultats des indices de stabilité obtenus montrent une stabilité similaire entre niveaux pour chaque langue et une majeure stabilité des mentions en français, l'italien présentant ainsi une plus grande variation dans les dénominations utilisées.

Enfin, l'analyse de la distribution des catégories grammaticales de mentions permet de dégager plusieurs tendances générales dans la structuration des chaines de référence. Les résultats montrent que les deux langues présentent des préférences différentes dans le choix des types grammaticales des mentions. En français, les pronoms constituent la catégorie la plus fréquente dans les chaines, tandis qu'en italien ce rôle est principalement assuré par les anaphores zéro, ce qui reflète les propriétés morphosyntaxiques propres à une langue à sujet nul.

L'examen de la position des mentions dans les chaines montre par ailleurs que celles-ci s'ouvrent majoritairement par des syntagmes nominaux indéfinis, qui remplissent leur fonction typique d'introduction de nouveaux référents dans le discours. Les positions suivantes sont plus fréquemment occupées par des pronoms, des syntagmes nominaux définis ou, dans le cas de l'italien, aussi par des anaphores zéro. L'analyse des séquences formées par les premiers maillons des chaines met en évidence certaines configurations récurrentes dans les deux langues, notamment des séquences associant un syntagme nominal indéfini initial à des reprises pronominales ou nominales définies, ce qui reflet la présence de la formule d'ouverture typique des contes de fée « il était une fois ».

Dans l'ensemble, ces analyses quantitatives permettent ainsi de dégager plusieurs tendances générales dans la structuration des chaines de référence dans les récits produits par les élèves. Elles mettent en évidence à la fois des régularités communes aux deux langues et certaines variations liées aux propriétés morphosyntaxiques des systèmes linguistiques considérés. Toutefois, ces observations restent encore globales et ne permettent pas de rendre pleinement compte de la diversité des configurations référentielles observées dans les textes. Pour approfondir ces résultats, le chapitre suivant proposera une analyse qualitative des chaines de référence, afin d'examiner plus finement les stratégies linguistiques mobilisées par les élèves pour introduire, maintenir et organiser les référents dans leurs récits.

Chapitre 8. Approche qualitative de l'analyse des chaines de continuité référentielle

En complément de l'analyse quantitative approfondie dans le chapitre précédent, il nous a semblé nécessaire de mener également une analyse qualitative sur les textes de notre corpus, afin d'obtenir une compréhension plus fine des modalités concrètes de constitution des chaines dans les textes narratifs. Cette démarche vise à éclairer les stratégies effectivement mises en œuvre dans la gestion des chaines référentielles et à préciser le rôle que ces chaines jouent dans l'écriture d'une narration de fiction par des scripteurs en phase d'apprentissage.

D'un point de vue méthodologique, cette analyse qualitative s'inscrit dans le prolongement des résultats statistiques précédemment exposés (cf. Chapitre 7). Ici, certaines des tendances dégagées par l'analyse quantitative orientent les choix des phénomènes examinés. Il s'agit, d'une part, d'explorer certains items correspondant à des valeurs extrêmes ou atypiques au regard des distributions observées ; d'autre part, d'examiner des propriétés qui ne peuvent être appréhendées de manière satisfaisante par le seul traitement quantitatif et qui exigent un retour au texte dans son intégralité, notamment pour analyser les enchainements discursifs, la progression narrative et la cohérence globale.

Ce chapitre s'attachera d'abord à l'analyse des mentions dont la longueur dépasse six tokens, qui constituent les formes les moins fréquentes du corpus. L'objectif sera d'en décrire la structure interne ainsi que le rôle sémantique rempli dans la construction des chaines de référence. Nous approfondirons ensuite l'examen de certains éléments particulièrement saillants au regard des analyses statistiques précédentes, notamment en ce qui concerne le type des mentions et leurs positions dans les chaines. Nous analyserons ainsi les syntagmes nominaux définis, les syntagmes nominaux possessifs et les syntagmes démonstratifs en position 1, ainsi que les pronoms apparaissant en position 2 dans la chaine. Il s'agira d'en proposer une interprétation fonctionnelle et discursive. Le chapitre se poursuivra par une analyse des stratégies de gestion référentielle en fonction de la configuration des personnages dans le texte, en distinguant les récits centrés sur un seul personnage de ceux qui en mobilisent deux. Nous examinerons en particulier si la présence de deux personnages du même genre ou de genre différent peut influencer les choix référentiels. Enfin, une étude empirique sera consacrée aux facteurs susceptibles d'expliquer les variations observées dans la longueur des textes entre les corpus français et italien.

Afin d'articuler plus étroitement l'analyse référentielle et la structuration narrative, nous mobiliserons également des apports issus des recherches en psychologie du développement sur la narration chez les enfants (Longobardi et al., 2014; Losh & Capps, 2003; Stein & Glenn, 1975) ainsi que les travaux réalisés sur *Scolinter* (Gutiérrez-Caceres et al., 2023). Ces recherches abordent la narration du point de vue du développement de ressources linguistiques et cognitives nécessaires à la construction d'un récit cohérent. Si l'on adopte cette perspective, en plus des analyses effectuées sur nos annotations, les chaines en tant que mentions directes des actants du texte ne sont pas une facette isolée de ce dernier, mais se constituent comme partie intégrante de la cohésion narrative.

Dans cette dernière étude (Gutiérrez-Caceres et al., 2023), les auteurs analysent et comparent les stratégies narratives mobilisées par des élèves de CP (première année) dans les trois pays représentés au sein du corpus, dans le but de mettre en évidence d'éventuelles différences liées aux contextes scolaires. Leur démarche repose sur plusieurs postulats : d'une part, la production d'un texte narratif à l'écrit implique la mise en œuvre de stratégies langagières spécifiques au genre narratif, susceptibles d'assurer la cohérence du texte et de maintenir l'attention du lecteur ; d'autre part, la compétence narrative est envisagée comme une capacité à développer dans un contexte de littéracie socialement situé, l'école constituant le lieu privilégié de l'appropriation des stratégies communicationnelles, en particulier à l'écrit. Dans cette perspective, l'évaluation du développement de ces compétences narratives s'appuie sur un ensemble d'indicateurs variés reconnus dans la littérature, notamment en psychologie du développement, pour apprécier l'évolution de la narrativité chez les enfants (Longobardi et al., 2014; Losh & Capps, 2003; Stein & Glenn, 1975, entre autres).

En nous appuyant sur ces travaux, nous mobilisons également les huit composantes de la structure narrative proposées par Stein et Glenn (1975, cités par Longobardi et al., 2014) afin d'évaluer dans quelle mesure les textes du corpus correspondent aux schémas prototypiques du récit. Ces catégories, fréquemment observées dans les contes et les fables, permettent de décrire l'organisation interne des narrations produites par les élèves (Stein & Glenn, 1975). Elles sont les suivantes :

1) *Introduction :* présentation du ou des personnages principaux de l'histoire (ex. « Il était une fois un petit chat qui s'appelait François ») ;
2) *Setting* : la description du contexte de l'histoire, la mise en place du cadre spatio-temporel et situationnel de l'histoire (« Il habitait dans un très grand château et n'avait pas de maître. ») ;
3) *Initiating event* : l'évènement déclencheur de la narration, qui introduit le conflit et met en mouvement la narration (« Un jour, des pirates attaquent son château ; François doit abandonner sa maison. ») ;
4) *Internal response* ; réaction interne du personnage, en termes de pensée et émotions, à l'évènement déclencheur (« Il est triste d'abandonner son château…») ;
5) *Planning* : les objectifs du personnage et ses plans d'action (« … mais c'est nécessaire, il doit partir. »);
6) *Attempt* : les actions entreprises par le personnage pour atteindre ses objectifs (« Il s'embarque alors sur un navire de chatons pirates pour faire le tour du monde. ») ;
7)*Direct Consequences :* conséquences immédiates des actions du personnage (« Au cours de ses voyages, il rencontre un garçon nommé Thomas qui accepte de devenir son maître. ») ;
8) *Resolution :* résolution du conflit et clôture du récit (« Thomas et le chaton François rentrent chez eux et vivent heureux, fin ! »).

L'adoption de cette grille d'analyse permet d'examiner la structuration narrative des textes du corpus de manière systématique et de mettre en relation l'organisation macro-structurelle du récit avec les choix linguistiques observés, notamment en matière de gestion référentielle. Il ne s'agit

pas ici d'évaluer les propriétés littéraires des productions, mais plutôt de décrire la présence de ces composantes structurelles de la narration.

Par ailleurs, Gutiérrez-Caceres et al. (2023) prennent également en considération d'autres indicateurs de complexité narrative : la présence de formules d'ouverture et de clôture des narrations ; l'utilisation de formes verbales au passé ou au présent ; la présence de représentations des états mentaux, tels que les liens de causalité entre événements ou comportements et des inférences sur les états affectifs et émotionnels des personnages ; la présence de marqueurs destinés à capter l'attention du lecteur, tels que le discours direct ou indirect, des intensificateurs ou des procédés d'interpellation.

Dans notre analyse, une attention particulière sera accordée maintenant à la possible correspondance entre certaines mentions longues et des fonctions descriptives, qu'il s'agisse de la caractérisation visuelle des personnages ou de la représentation de leurs états mentaux. La présence de ces éléments peut en effet être envisagée comme un indice du développement de la complexité narrative chez les jeunes scripteurs, dans la mesure où elle témoigne d'une capacité accrue à enrichir le récit au-delà de la simple succession événementielle.

1. Longueur des mentions

Nous avons précédemment mentionné les longueurs maximales des mentions relevées dans les deux langues (cf. Chapitre 7, §1.2). À partir de cette analyse quantitative, nous observons que la grande majorité des mentions sont constituées d'un seul token, suivies par celles de deux tokens, tandis que les mentions de taille supérieure à 6 tokens restent marginales dans notre corpus annoté. Nous nous sommes penchée sur la composition et le contenu de ces mentions plus longues en raison de leur rareté dans le corpus, pour en analyser la position dans la chaine, la fonction référentielle (d'introduction ou de maintien d'un référent) ainsi que la structure interne, afin d'identifier les types d'expansion du syntagme nominal les plus utilisés selon la langue. Selon Garcia-Debanc et al. (2022), ces expansions peuvent être de trois types : 1) les adjectifs, qui s'accordent en genre et nombre avec le nom ; 2) les compléments prépositionnels (aussi appelés compléments du nom) : introduits par une préposition, spécifient la référence ou permettent de caractériser davantage le nom ; 3) la proposition subordonnée relative, fréquemment utilisée par les élèves sans être dans les programmes scolaires. Ici nous allons davantage analyser les deux premiers types d'expansion, le troisième type n'étant pas inclus dans la limite des mentions annotées selon les choix effectués dans notre guide d'annotation.

L'analyse de certains de ces éléments nous semble constituer ainsi une piste de réflexion pertinente pour la linguistique et pour la didactique. Du point de vue linguistique, nous pouvons étudier les éventuelles constantes en termes de fonction et de structure interne de la mention pour les décrire. Comme remarqué par Garcia-Debanc et al. (2022, p. 211), le travail sur les expansions du nom à l'école primaire devrait inciter à leur utilisation à l'écrit, en particulier dans les descriptions. Du point de vue didactique, porter l'attention de l'élève sur ces éléments permettrait de mobiliser sa connaissance épilinguistique, c'est-à-dire sa connaissance intuitive des phénomènes discursifs.

Dans notre corpus, les longueurs maximales des mentions observées atteignent onze tokens en français et quinze tokens en italien. Nous avons ainsi choisi de concentrer notre analyse qualitative sur les mentions dont la longueur est comprise entre six tokens et le maximum attesté dans chaque langue. En effet, en dessous de six tokens, les mentions plus courtes sont nettement plus fréquentes. Ce choix permet également de se concentrer sur les formes les plus rares du corpus, dont l'examen est particulièrement éclairant pour en comprendre le fonctionnement : il s'agit de déterminer si ces occurrences relèvent d'anomalies ou, au contraire, si elles constituent des formes plus complexes et potentiellement plus riches, susceptibles d'ouvrir des pistes d'évolution et des perspectives en didactique. Ce sous-corpus comprend ainsi trente-sept maillons au total : quatorze en français et vingt en italien. Le sous-corpus des mentions analysées dans cette partie est disponible en Annexe K[92].

Dans la variabilité de structure interne de la mention que nous avons pu tracer, il était possible de dégager des patterns syntaxiques fréquents, dans la même langue ou similaires entre les deux langues. Comme préconisé par les théories de Ariel (1990) ou par Charolles (1987), selon lesquels la position de la mention dans la chaine peut influencer son type grammatical, ces mentions longues se positionnent principalement en position 1 dans les chaines analysées. Nous avons principalement identifié deux catégories de mentions longues ayant fonction introductive : des mentions descriptives du type SN indéfini comprenant des adjectifs ou des compléments du noms, et des mentions de dénomination explicite du personnage contenant un SN indéfini ainsi qu'un nom propre.

1.1. Mentions longues à fonction introductive : SN indéfini + adjectifs ou compléments du nom

Un premier ensemble des mentions longues analysées correspond aux syntagmes nominaux indéfinis, complétés par des adjectifs, qui réalisent des descriptions des personnages. Ce type de mention apparait majoritairement en début de texte, en position 1 dans la chaine ou du moins parmi les trois premières expressions référentielles. Ainsi elle remplit habituellement une fonction d'introduction et de description physique (des caractéristiques visuelles) du personnage nouvellement introduit. Dans le texte de l'exemple (85), le *chat* est introduit par un tel syntagme, « un chat tout blanc aux beaux yeux bleus » (surligné dans le texte), lequel en décrit les caractéristiques physiques. La particularité de ce texte réside également dans la présence d'un syntagme expansé similaire, « au magnifique robot jaune, vert et bleu » (surligné dans le texte) : ce syntagme défini et prépositionnel réinstancie et décrit le *robot* bien plus tard par rapport à son introduction, et survient en position 10 dans la chaine. Les adjectifs portent sur la caractérisation visuelle du robot, à travers un adjectif qualificatif à valeur positive (« magnifique ») et plusieurs adjectifs de couleur (« jaune, vert et bleu »).

(85) Il était une fois **[un robot]**^{robot} **[qui]**^{robot} se promenait à la montagne. **[Il]**^{robot} rencontre **[un chat tout blanc aux beaux yeux bleus]**^{cat} et **[le robot]**^{robot} demanda **[au**

[92] Trois maillons issus du corpus italien, bien que leur longueur varie incluse entre six et quinze tokens, n'ont pas été pris en compte dans cette analyse qualitative car leur longueur est le résultat d'une erreur dans l'adjudication.

tout petit chat][cat] « Bonjour que faites**[-vous]**[cat] **[M. Le chat]**[cat] ? » « **[Je]**[cat] joue à cache-cache avec **[[mes]**[cat] **copains]**[ext1], et **[vous]**[robot] que faites**[-vous]**[robot] **[M. Le robot]**[robot] ? » « **[Je]**[robot] me balade » s' exclama **[le robot]**[robot]. « Voulez***[-vous]***[robot] bien être **[mon]**[cat] ami ? » demanda **[le chat]**[cat] **[au magnifique robot jaune, vert et bleu]**[robot]. (…) *(corpus français, CE2, 1323)*

Des structures similaires apparaissent dans le corpus italien. Dans un texte, une description très proche de celle du texte (85) du *robot* est employée, cette fois-ci en tant que maillon introductif. Dans l'exemple (86) la première mention de la chaine du personnage du *robot* est constituée par un syntagme nominal indéfini « un simpatico robottino azzurro, giallo, rosso e blu ». Ce syntagme inclut un adjectif qualificatif descriptif du caractère du robot (« simpatico » ; gentil) ainsi que des adjectifs de couleur, similaires à ceux de l'exemple analysé précédemment dans le corpus français (« azzurro, giallo, rosso e blu » ; bleu clair, jaune, rouge et bleu).

(86) C'era una volta **[un gatto** solo nella neve e molto infreddolito**]**[cat]. **[Quel gatto]**[cat] era tutto nero. Un giorno arrivò **[un simpatico robottino, azzurro, giallo, rosso e blu]**[robot]. **[Il gatto]**[cat] **[gli]**[robot] chiese « chi sei **[tu]**[robot] ? » « **[un robottino simpatico e carino]**[robot] » **[rispose]**[robot]. « E **[tu]**[cat] che ci fai qui tutto solo nella neve? » **[chiese]**[robot]. « **[Mi]**[cat] hanno abbandonato » **[rispose]**[cat]. « **[Ti]**[cat] **[capisco]**[robot], ***[la donna]***[ext1] ***[che]***[ext1] ***[mi]***[robot] *ha costruito* **[mi]**[robot] ha rimpiazzato per **[un robot dell'ultimo modello]**[ext2]. » (…) *(corpus italien, CE1, 200)*

Dans l'exemple (87), également issu du corpus italien, le personnage du *robot* est décrit dès sa première mention dans le corps du texte (en excluant le titre, indiqué dans le texte) par le biais d'un syntagme nominal indéfini qui inclut des adverbes ainsi que des adjectifs. Ces adjectifs relèvent de la description de qualités mentales et comportementales du personnage (« intelligente, giocherellone, divertente e studioso » ; intelligent, joueur, amusant et studieux), plutôt que des caractéristiques physiques, comme cela était le cas dans les exemples précédents. La liste d'adjectifs est introduite par une expression intensificatrice, deux adverbes superlatifs absolus répétés indiquant un degré élevé, unis par la conjonction adversative « ma » avec valeur d'emphase (« molto ma molto intelligente »; très mais très intelligent). L'accumulation d'adjectifs en plus des adverbes contribuent à la représentation du caractère du personnage dès son introduction dans le texte.

(87) **{[Il robot]**[robot] e **[il lupo]**[wolf]**}**
C'era una volta **[un robot molto ma molto intelligente, giocherellone, divertente e studioso]**[robot]. (…) (corpus italien, CE2, 1665)

L'exemple (88) illustre un exemple similaire d'utilisation d'adverbes de degré pour le référent *strega* (*sorcière*). Le personnage est introduit par un syntagme nominal indéfini, qui comprend des adverbes superlatifs dans la même configuration que celle identifiée dans l'exemple précédent

suivie d'un adjectif évaluatif («una strega molto ma molto strana », *une sorcière très mais très bizarre*). Ce premier syntagme est immédiatement complété par une reformulation contrastive, introduite par une négation, qui oppose explicitement le personnage à la catégorie stéréotypée des sorcières (« non è [una strega come le altre cattiva], ma [era] buona e generosa *», ce n'était pas une sorcière méchante comme les autres, mais elle était gentille et généreuse*). Ce procédé combine description et prise de position axiologique, et contribue à individualiser le référent dès son introduction dans le récit. L'utilisation d'une négation est d'ailleurs indicative du point de vue du narrateur ou de la narratrice, car il permet d'indiquer des événements ou comportement contraire aux attentes du lecteur (Losh & Capps, 2003).

(88) **{[La strega]**^{witch} e **[il lupo]**^{wolf}}
C'era una volta **[una strega molto ma molto strana]**^{witch}, non è **[una strega come le altre cattiva]**^{witch}, ma **[era]**^{witch} buona e generosa. (…) *(corpus italien, CE2, 1438)*

Ces syntagmes nominaux ayant fonction d'introduction et de description sont souvent composés d'une tête de type syntagme nominal indéfini et incluent des adjectifs qui permettent de caractériser le personnage introduit, au niveau physique, visuel ou perceptif, ou aussi relationnel. Dans l'exemple (89), c'est une description physique visuelle du chat qui le caractérise dès sa première mention, « un chat blanc avec des tâches noires ».

(89) Il était une fois **[une sorcière]**^{witch} qui cherchait un chat. **[Elle]**^{witch} cherchait dans tous les magasins. Un jour **[elle]**^{witch} trouva **[un chat blanc avec des taches noires]**^{cat}. **[Elle]**^{witch} jouait tout le temps avec **[lui]**^{cat}. Et **[ils]**^{cat,witch} devinrent meilleurs amis pour toujours. Et **[le chat]**^{cat} eut **[des chatons]**^{cat1} avec **[une chatte blanche]**^{cat2}. *(corpus français, CE1, 1325)*

Cette description physique du pelage du chat est présente aussi dans un autre exemple d'introduction du personnage *chat*, le texte (90) : ici le référent est introduit à l'aide d'un SN indéfini qui associe des traits perceptifs de type tactile au référent en question, « au pelage très doux ».

(90) Il était une fois **[un petit chat au pelage très doux]**^{cat} **[qui]**^{cat} vivait avec **[un robot grand]**^{robot}, **[il]**^{robot} faisait 1 mètre 50. *(corpus français, CE2, 1544)*

Dans le cas de l'exemple (91), la description du personnage *loup* est relationnelle et situationnelle plutôt que physique ou caractérielle : le loup est caractérisé par son statut social (« seul et sans famille »), qui motive la suite de la narration, ainsi que le choix du titre de l'histoire.

(91) {**[Le loup Cédric]**^{wolf} veut se marier}
Il était une fois **[un loup seul et sans famille]**^{wolf}, **[il]**^{wolf} voulait absolument une famille vu qu'**[il]**^{wolf} était **[le seul loup de la forêt]**^{wolf}. *(corpus français, CE2, 3060)*

L'analyse des maillons longs à fonction introductive montre ainsi que les élèves peuvent recourir à des syntagmes nominaux indéfinis enrichis par des adjectifs ou des compléments du nom pour introduire et caractériser les personnages dès leur première mention. Ces expansions remplissent une fonction descriptive, permettant de construire la représentation du référent sur les plans physique ou visuel, caractériel, social ou relationnel. Cette caractérisation initiale contribue à établir la présence du référent et participe à la richesse lexicale et narrative des productions.

En revanche, l'analyse qualitative du corpus met en évidence une pénurie de descriptions explicites des états mentaux des personnages lors de leurs introductions, au moins observables à travers l'analyse des mentions plus longues. Ce constat ouvre des perspectives didactiques importantes. Il invite en effet à interroger la place accordée, dans l'enseignement de la production écrite, à l'utilisation des adjectifs comme outils de description des personnages, et à l'analyse des différents types de descriptions possibles pour en caractériser les états internes.

1.2. Mentions longues à fonction introductive : dénomination explicite

Dès le début de notre travail d'annotation sur le corpus français, nous avons identifié un autre type récurrent de maillon introductif, explicitement pris en compte dans le guide d'annotation comme structure devant être annotée en tant que maillon. Il s'agit des mentions constituées d'un syntagme nominal, généralement indéfini, intégrant une structure de dénomination explicite du référent, réalisée au moyen du participe *nommé/e* en français ou de l'expression *di nome* en italien, et suivie d'un nom propre. Ce syntagme nominal expansé d'un complément prépositionnel est largement représenté dans le corpus italien, avec 38 occurrences au total dans l'entièreté du corpus annoté, et seulement 5 occurrences dans le corpus français annoté. Parmi les occurrences italiennes, cinq correspondent à des maillons dont la longueur varie entre six et neuf tokens. La variabilité observée porte principalement sur l'expansion du noyau nominal, notamment par l'ajout de déterminants possessifs (« suo » ; le sien) et des noms de relation (« amico » ; ami) comme dans l'exemple (95) ou d'adjectifs qualifiant la tête nominale (« vecchia », vieille dans (92) et « piccolo » (petit) dans (94)), ou sur la compositionnalité du nom propre inventé par l'élève (« Antenna di fuoco » dans (93) ; Antenne de feu).

(92) una vecchia strega di nome Riunione *(corpus italien, CE1, 11)*
(93) un robot di nome Antenna di fuoco *(corpus italien, CE1, 432)*
(94) un piccolo gattino di nome Bibi *(corpus italien, CE2, 844)*
(95) il suo amico gatto di nome Caddy *(corpus italien, CE2, 1440)*

D'autres exemples montrent que ces syntagmes peuvent introduire les personnages et rajouter des détails additionnels, comme l'âge ou la situation sociale et familiale, comme dans le cas de l'exemple (96). Le troisième personnage est introduit à l'aide d'un SN indéfini « un bambino » (un enfant) complété de deux compléments prépositionnels, l'un introduisant son nom propre (« di nome Giacomo » ; nommé Giacomo), le deuxième son âge (« di 2 anni » ; de deux ans), et unis par la conjonction e (pas nécessaire dans le contexte de cette phrase, qui pourrait d'ailleurs être interprétée comme erronée ici).

(96) C'era una volta **[un uomo]**^{ext1} **[che]**^{ext1} aveva **[una moglie]**^{ext2} e **[un bambino di nome Giacomo e di 2 anni]**^{ext3} (…). (*corpus italien, CE2, 1104)*

Nous retrouvons toujours dans le corpus italien (texte (97)) un maillon similaire du point de vue syntaxique et sémantique, mais qui indique deux personnages simultanément. Ce maillon comprend 10 tokens, il apparait également en début de texte et il introduit deux personnages tout en établissant leurs relations avec le premier référent, ici la sorcière nommée Orazia (« una strega di nome Orazia »). Ce syntagme nominal initial constitue le tout premier maillon de toutes les chaines du texte et le centre du réseau de personnages principaux, auquel s'ajoutent son mari, « Lupo Lucio »[93], et leurs deux enfants, introduits par un syntagme nominal défini (« i due figli ») et par un seul complétement prépositionnel incluant leur deux noms propres, « Strega Dina » et « Lupo Nello ».

(97) C'era una volta **[una strega di nome Orazia]**^{witch}, **[[suo]**^{witch} **marito Lupo Lucio]**^{wolf} e **[i due figli di nome Strega Dina e Lupo Nello]**^{ext1,ext2}. (…) *(corpus italien, CE2, 1108)*

Suite à notre analyse, nous pouvons donc remarquer que ce type de maillon apparait majoritairement en début de texte ou parmi les premiers maillons de la chaine, et remplit une fonction descriptive et introductive à la fois. La présence du nom propre permet d'individualiser immédiatement le référent, tout en autorisant par la suite des reprises sous des formes moins rigides (pronoms ou syntagmes définis).

1.3. Maillons les plus longs : corpus français

Dans le corpus français, le maillon le plus long est composé de 11 mots. Celui-ci, en jaune dans l'exemple (98), est un syntagme nominal défini « la sorcière » dont l'expansion est constituée par un superlatif relatif « la plus âgée de toutes les sorcières du monde », qui nous semble avoir une occurrence assez rare dans notre corpus annoté. Il constitue le troisième maillon d'une chaine qui introduit une troisième sorcière dans le texte : déjà au premier maillon, elle est qualifiée de « très très vieille » à l'aide d'un SN indéfini, repris à l'aide d'un pronom personnel, puis elle est décrite en relation à son âge à sa troisième reprise comme « la sorcière la plus âgée de toutes les sorcières du monde » (surligné dans le texte).

(98) **[Une vieille sorcière des bois]**^{witch} et **[[son]**$^{\mathbf{witch}}$ **chat]**^{cat} s'appelaient **[Marie-Memèr]**^{witch} et **[[son]**$^{\mathbf{witch}}$ **chat Catoumane]**^{cat}. **[Il]**^{cat} était très méchant avec les gens qui **[les]**^{cat,witch} entouraient. Mais un jour **[la sorcière]**^{witch} **[qui]**^{witch} faisait le mal **[une autre sorcière]**^{witch1} arriva en ville et **[essaya]**^{witch1} de **[l']**^{witch}attraper mais **[Marie-Memèr]**^{witch} ne se laissa pas faire et **[elles]**^{witch,witch1} se lancèrent un combat de magie. Le lendemain le combat commença mais **[une très très vieille sorcière]**^{witch2} arriva,

[93] Le nom de ce personnage est très probablement inspiré du personnage homonyme du programme télévisé italien pour enfant, la *Melevisione*. Plus d'information sur ce célèbre programme sont disponible à la page Wikipedia suivante : https://it.wikipedia.org/wiki/Melevisione (Consultée le 16/02/2026).

[elle]witch2 avait 364 ans, c' était **[la sorcière la plus âgée de toutes les sorcières du monde]**witch2. Donc **[elle]**witch2 menotta **[les deux sorcières]**witch,witch1 et **[les]**witch,witch1 **[jeta]**witch2 en prison ! *(corpus français, CE2, 1991)*

À noter dans ce texte qu'une mention présente dans la deuxième phrase, « il », nous interroge sur les choix de normalisation effectués sur ce texte, à cause de son contexte droit qui présente un pronom pluriel, discordant avec ce pronom au singulier en tant que sujet de la phrase.

L'autre maillon composé de 11 tokens est en réalité constitué par une mention imbriquée dans une autre mention. On peut la retrouver dans l'exemple (99), en jaune dans le texte. La mention « un gros chat tendre et dodu » est suivie par un complément prépositionnel mentionnant le *loup* « aussi affamé » protagoniste de la narration.

(99) **{[Le loup affamé]**wolf}

Il était une fois, dans la forêt, **[un loup affamé]**wolf. **[Il]**wolf cherchait de quoi se nourrir. **[Il]**wolf avait faim, très faim. **[Il]**wolf cherchait **[sa]**wolf nourriture de partout. Dans les ruelles, les prairies, les forêts, etc. Mais **[il]**wolf ne trouvait rien. Pourtant un jour, **[il]**wolf vit dans les herbes hautes d'une prairie, **[un chat]**cat, **[un gros chat tendre et dodu pour [un loup aussi affamé]wolf]cat**. Vite ! **[Le chat]**cat **[le]**wolf regardait, **[le loup]**wolf se cachait dans les hautes herbes. **[Il]**wolf avançait jusqu'**[au chat]**cat lentement mais sûrement. Et puis, **[il]**wolf bondit sur **[le chat]**cat **[qui]**cat n'avait pas le temps de s'enfuir et **[le]**cat **[mangea]**wolf. Et, depuis ce jour, **[le loup]**wolf cherchait un coin pour digérer **[le chat]**cat car **[il]**wolf avait aussi avalé les poils **[du chat]**cat. Fin *(corpus français, CE2 , 3026)*

1.4. Maillon plus long : corpus italien

L'exemple de maillons plus longs en italien (jusqu'à 15 tokens ponctuation exclue) est particulièrement intéressant car il s'insère dans une chaine fortement cataphorique, que l'on retrouve dans l'exemple (100) : le personnage du *loup* est introduit progressivement dans le texte. Le premier maillon de la chaine lui faisant référence est constitué par le pronom indéfini « qualcuno » (quelqu'un), suivi immédiatement dans la phrase suivante par sa description (la mention la plus longue occupe ici la deuxième position dans la chaine, elle est surlignée en jaune dans le texte). Cette description est reprise par un syntagme nominal démonstratif (« quell'ombra sconosciuta », cette ombre inconnue) qui reprend la tête nominale du maillon précédent (« ombra »). Ce n'est qu'à la quatrième mention que l'auteur ou autrice nous révèle que ce mystérieux personnage est en fait « un lupo » (un loup).

Le maillon descriptif est ici composé par un noyau indéfini complété d'un adjectif qualificatif, « buia » (sombre) et complété par un complément prépositionnel introduit par « con » (avec) qui rattache trois nouveaux syntagmes nominaux, « occhi gialli » (yeux jaunes), « muso a punta » (museau pointu) et « denti che spuntano dalla bocca » (dents qui dépassent de sa bouche), ce dernier groupe nominal incluant une relative à son intérieur.

(100) Una sera buia nel castello di **[un robot di nome Antenna di fuoco]**[robot], **[qualcuno]**[wolf] bussò alla porta, **[Antenna di fuoco]**[robot] terrorizzato aprì la porta, si vedeva [**un'ombra buia, con occhi gialli, muso a punta e denti che spuntano dalla bocca]**[wolf]. **[Quell' ombra sconosciuta]**[wolf] entrò nel castello, era **[un lupo]**[wolf]! **[Antenna di fuoco]**[robot] scappò con tutte le forze e **[si nasconde]**[robot] dietro il divano, **[il robot]**[robot] lanciò: bicchieri, scatole e matite, il rumore era forte PINC ZAC BOOM ma non c'era niente da fare, allora **[Antenna di fuoco]**[robot] lanciò un coltello e **[il lupo]**[wolf] morì e da quel giorno i lupi non esistettero più. *(corpus italien, CE1, 432)*

En conclusion, l'analyse des mentions les plus longues et les moins fréquents dans les corpus français et italien montre que, malgré la diversité apparente des syntagmes nominaux, les mentions « complexes » se répartissent principalement en deux types : mentions descriptives et de dénomination explicite. Les mentions descriptives consistent en un syntagme nominal indéfini ou défini qui inclut un ou plusieurs adjectifs. Ces expressions référentielles servent à caractériser physiquement, mentalement ou socialement le personnage au moment de son introduction. Les maillons de dénomination explicite, quant à eux, incluent une structure de type *di nome* en italien ou *nommé(e)* en français, suivie d'un nom propre, permettant d'identifier précisément le référent. Ces deux types de maillons apparaissent dans les deux corpus, avec une distribution légèrement différente selon la langue mais avec une fonction similaire d'introduction et de description.

Comme nous l'avions pressenti, l'analyse des syntagmes nominaux et des expansions que l'on peut retrouver dans le cas des maillons longs dans le corpus peut fournir des pistes didactiques intéressantes. Celles-ci peuvent porter sur plusieurs niveaux textuels et sur plusieurs procédés : sur l'introduction et la description des personnages pour en établir la présence dans le texte, enrichir la variété lexicale, mieux le distinguer par ses caractéristiques des autres personnages et permettre ainsi par la suite de le réinstancier de manière non ambigüe à travers la mention de ces mêmes caractéristiques. Ce travail peut ainsi contribuer à la construction d'un répertoire adjectival portant sur ces descriptions physiques, relationnelles et des états mentaux des personnages, et pousser les élèves à les utiliser pour instaurer la présence des personnages à travers des descriptions en début de texte.

2. Composition des chaines

Après avoir décrit la composition des chaines du point de vue des types de maillons qui les composent (cf. Chapitre 7, §5 pour une vue quantitative globale), nous abordons à présent leur dimension qualitative, en nous concentrant sur l'interaction entre les types de mentions et certains éléments de textualité. Dans une conception linéaire de la chaine référentielle, il s'avère pertinent d'étudier l'interaction du contenu des mentions avec l'organisation textuelle et avec le niveau sémantique du texte, au-delà des analyses quantitatives portant sur la distribution des types de maillons et leurs variations selon la position dans la chaine. Nous avons ainsi analysé plusieurs éléments issus de nos analyses quantitatives : la présence de SN nominaux définis dans les titres, de SN possessifs et des SN démonstratifs en tête de chaine et leur rôle dans l'introduction d'un référent donné, ainsi que la typologie des pronoms en deuxième position de chaine et comment

cela est influencé par la présence des formules d'incipit des contes de fées typiques de chaque langue.

Nous avons ainsi analysé qualitativement les types de mentions les plus fréquents à une position donnée ainsi que les types appartenant aux catégories moins nombreuses. Cette analyse exploratoire devrait permettre de vérifier, dans le deuxième cas, quelles sont les raisons pour lesquelles ces catégories sont peu mobilisées : à travers l'étude de leurs cas d'utilisation, on peut décrire une trajectoire d'utilisation entre les deux niveaux scolaires observés et éventuellement dégager des pistes didactiques pour leur inclusion dans l'écriture des textes.

2.1. SN définis en position 1

Pour ce qui concerne les syntagmes nominaux définis en position 1 des chaines, nous faisons l'hypothèse que leur distribution est en partie conditionnée par la présence des titres en début de texte, ainsi que par les choix d'annotation adoptés dans notre corpus, où les titres ont été considérés comme les premières mentions des chaines de référence. Le traitement des chaines plurielles constitue également un facteur explicatif : lorsque plusieurs référents distincts sont impliqués, celles-ci sont comptabilisées comme des chaines séparées, ce qui peut influer sur la distribution des mentions en position initiale

Dans notre corpus, les SN définis apparaissent en effet très fréquemment en tête de chaine. Dans les deux langues, ils constituent la deuxième catégorie la plus nombreuse en première position, juste après les SN nominaux indéfinis (cf. Figure 68). Ceux-ci sont toutefois moins nombreux en français (79 occurrences au total, dont 33 au CE1 et 45 au CE2) qu'en italien, où l'on en recense 118 (39 au CE1, et 79 au CE2).

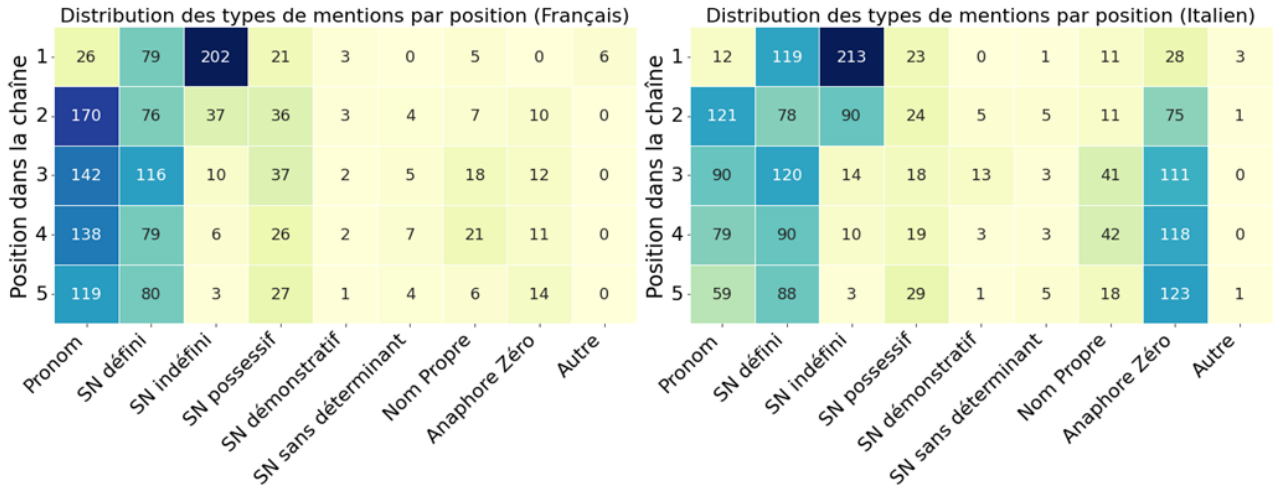


Figure 68. Distribution des types de mentions sur les 5 premiers maillons des chaines - comparaison entre langues (reprise de la Figure 63).

Niveaux scolaires	**français**			**italien**		
	Textes avec titres	*SN défini position 1*	*SN défini position 2*	*Textes avec titres*	*SN défini position 1*	*SN défini position 2*
CE1 – 2	7	34	34	7	40	35
CE2 – 3	22	45	42	42	79	43
CE1-CE2 – 2-3	**29**	**79**	**76**	**49**	**119**	**78**

Tableau 54. Distribution des titres et des SN définis dans les chaines annotées.

Ces résultats peuvent être mis en relation avec la présence des titres dans les textes, selon la langue et le niveau scolaire, comme le synthétise le Tableau 54. Ici on compare la distribution des titres dans les textes annotés de notre corpus avec la distribution des SN définis en position 1 et 2 dans les chaines.

Dans le corpus français, on compte 29 textes contenants des titres : parmi ceux-ci, 27 présentent un titre en début de texte, tandis que deux textes contiennent chacun deux titres, signalant la présence de deux récits différents au sein de la même production. Un seul texte, exclu de ce décompte, contient un titre à la fin de la narration[94]. Dans le corpus italien, on recense au total 55 textes dotés d'un titre. Parmi ceux-ci, 6 textes présentent l'insertion du titre de l'histoire à la fin du texte[95], eux aussi exclus du décompte présenté dans le Tableau 54. Les exemples (101) et (102) illustrent respectivement les cas où un seul SN défini constitue le titre du texte, et qui peut indiquer un ou plusieurs personnages, et le cas où deux SN définis composent le titre et démarrent deux chaines différentes.

(101) **[Les deux amis]**^{cat,robot} Il était une fois **[une robot]**^{robot} **[qui]**^{robot} s'appelle **[Roti]**^{robot}. (…) *(corpus français, CE1, 1969)*

(102) **[La strega]**^{witch} e **[il gatto]**^{cat} C'erano una volta **[una strega]**^{witch} e **[un gatto]**^{cat} **[che]**^{cat,witch} non andavano molto d'accordo insieme. (…) *(corpus italien, CE1, 536)*

Si l'on croise cette quantification de la présence des titres avec celle des SN définis en position initiale de chaine, on observe que le nombre des titres est très faible en CE1, avec seulement 7 occurrences en italien et en français. Cette fréquence augmente fortement au niveau CE2 : dans le corpus français, le nombre de titres (22) représente alors près de la moitié des SN définis en position initiale (45), tandis que dans le corpus italien, la hausse est encore plus marquée, avec les titres qui représentent toujours quasiment la moitié (42) des SN définis en position 1 des chaines (79). Ces observations étayent l'hypothèse selon laquelle la présence de titres en début de texte, combinée aux choix d'annotation retenus, contribue de manière significative à la fréquence élevée des SN définis en première position de chaine.

Restent toutefois à examiner les cas où un SN défini apparait en tête de chaine sans correspondre à un titre. En effet, le début d'une chaine référentielle ne coïncide pas nécessairement avec le

[94] C'est le texte de niveau CE1, 2820, disponible en Annexe.
[95] Ce sont les textes de niveau CE1 81, 84, 91,

début du texte, en particulier lorsqu'il s'agit de l'introduction de personnages secondaires. Afin d'analyser ce phénomène, nous avons identifié les maillons de type SN défini apparaissant en position 1 des chaines à partir des sorties intermédiaires de notre algorithme de classification des mentions par typologie. Les textes contenant ces occurrences ont ensuite été examinés afin de sélectionner des exemples illustrant les cas les plus représentatifs de ce fonctionnement dans le corpus.

L'exemple (103), illustre ce phénomène : le personnage secondaire du *chien* y est introduit au moyen d'un SN défini, motivé par une caractérisation contextuelle (« le chien d'à côté », surligné dans le texte). De même, un autre personnage secondaire est introduit par un SN défini comportant une préposition (« au grand maître des chats », surligné dans le texte), montrant que le recours au SN défini peut également s'expliquer par des relations contextuelles ou inférentielles, indépendamment de la structure titre–texte.

(103) Il était une fois **[un petit chat]**^{cat} **[qui]**^{cat} s'appelait **[Mimi]**^{cat}. **[Il]**^{cat} vivait dans une maison avec **[[son]^{cat} frère]**^{cat1}. Un jour **[il]**^{cat} fut vendu au magasin puis **[il]**^{cat} ne savait pas comment faire sans **[[son]^{cat} frère]**^{cat1} mais **[il]**^{cat} ne se fit pas adopter alors **[il]**^{cat} préférait rester chez **[lui]**^{cat} en toute sécurité dehors dans le jardin mais **[le chien *d'à côté]***^{ext1} n'arrêtait pas de **[l']**^{cat}embêter. « Mais alors comment faire ? **[le]**^{ext1} tuer, **[le]**^{ext1} faire vendre au marché, ou demander à **[l']**^{ext1}adopter ? » « **[Je]**^{cat} le sais comment faire pour sortir de cette maison ! » dit **[le chat]**^{cat} à **[[son]^{cat} frère]**^{cat1}. « Pourquoi ne pas **[le]**^{ext1} faire fuir ? » « Comment ? » **[lui]**^{cat} demande **[[son]^{cat} frère]**^{cat1}. « Et bien en **[lui]**^{ext1} faisant peur » dit **[Mimi]**^{cat} « **[tu]**^{cat} n'as qu'à demander **[au grand maître des chats]**^{cat2}, qu'en dis-**[tu]**^{cat} ? » « Non ! **[Je]**^{cat} ne veux pas ! **[Je]**^{cat} ne veux pas ! **[J']**^{cat}ai trop peur » « Ah ! Ah ! Ah. **[T']**^{cat} as peur. Pfuh **[t']**^{cat} es bête, c'est **[un chat]**^{cat2} **[qui]**^{cat2} **[nous]**^{cat,cat1} a fait vivre alors **[tu]**^{cat} disais que **[tu]**^{cat} avais peur. » « Bien sûr que non » dit **[le chat]**^{cat} « bon est-ce que **[tu]**^{cat} comprends l'histoire maintenant ? » « oui » « alors **[tu]**^{cat} vois. » Fin. *(corpus français, CE1, 93)*

Un autre cas à considérer est celui des chaines plurielles. Dans notre protocole d'annotation, celles-ci ont été traitées comme des démarrages de chaines distincts lorsqu'une mention regroupait des référents auparavant instanciés individuellement. Ainsi, un SN défini pluriel peut apparaître en position initiale de chaine tout en renvoyant à des entités déjà introduites dans le texte.

Toutefois, ce type de syntagme ne donne pas systématiquement lieu à l'ouverture d'une nouvelle chaine référentielle. Dans certains cas, le regroupement des référents s'effectue progressivement au fil du texte et la référence plurielle se construit de manière cumulative. L'exemple (104) illustre ce phénomène : le groupe désigné par « les deux amis », est constitué à partir de référents introduits antérieurement, le *chat* et la *girafe*, et s'enrichit ensuite de nouveaux membres au cours de la narration. Dans ce cas, le syntagme nominal défini pluriel ne marque pas un nouveau démarrage de chaine autonome, mais s'inscrit dans un processus d'extension référentielle au sein de la chaine existante.

(104) **[Un chat tout seul]**^{cat} mais un jour **[il]**^{cat} rencontra **[un loup]**^{wolf} alors **[il]**^{cat} courut puis **[se reperdit]**^{cat}. **[Il]**^{cat} trouva **[une girafe]**^{ext1}. **[Il]**^{cat} était plus soulagé. Puis **[les deux amis]**^{cat,ext1} repartirent puis **[ils]**^{cat,ext1} trouvent **[un lapin]**^{ext2}. Puis **[ils]**^{cat,ext1,ext2} repartirent, **[ils]**^{cat,ext1,ext2} trouvèrent **[un lion]**^{ext3}. **[Tous les animaux]**^{cat,ext1,ext2} disent « au secours, **[le tigre]**^{ext4} va **[nous]**^{cat,ext1,ext2} manger » puis **[ils]**^{cat,ext1,ext2} ont réussi à s'échapper. **[Ils]**^{cat,ext1,ext2} retrouvèrent **[le loup]**^{wolf} au même endroit puis la faim arrive. *(corpus français, CE1, 2984)*

Dans l'exemple (105), la chaine plurielle composée du *chat* et du *robot* émerge de manière progressive à partir de la seconde moitié du texte, après une série de mentions de types variés qui désignent séparément les deux référents. Tout au long du récit, ceux-ci demeurent majoritairement traités comme des entités distinctes, clairement identifiées au moyen des noms propres qui les désignent, la référence plurielle s'inscrivant ainsi dans une organisation référentielle diffuse plutôt que comme un regroupement stable et central dès son apparition.

(105) C'era una volta **[un gatto di nome Malic]**^{cat}. **[Viveva]**^{cat} a Parigi. **[Non aveva]**^{cat} padrone. Mentre **[camminava]**^{cat} **[vide]**^{cat} un volantino che rappresentava un negozio di giocattoli. **[Malic]**^{cat} si incuriosisce e **[va]**^{cat} nel negozio misterioso. **[Vide]**^{cat} molte cose che **[non aveva mai visto]**^{cat} ma C'era **[una cosa]**^{robot} che **[lo]**^{cat} **[attirava]**^{robot}. Era **[un robot]**^{robot}. **[Decise]**^{cat} di prender**[lo]**^{robot}. **[Uscì]**^{cat} dal negozio e **[lo]**^{robot} **[portò]**^{cat} nel **[suo]**^{cat} posticino preferito vicino a un cantiere. A un certo punto **[il robot]**^{robot} si svegliò! E **[disse]**^{robot} "ciao. **[Tu]**^{cat} chi sei?" **[il gatto]**^{cat} rispose "chi. **[Io]**^{cat}?" "sì" disse **[il robot]**^{robot} "io sono **[Malic]**^{cat}" **[rispose]**^{cat} "ciao **[Malic]**^{cat}. **[Io]**^{robot} mi chiamo **[Sfasac]**^{robot}". **[I due]**^{cat,robot} decisero di fare una passeggiata. Purtroppo **[Sfasac]**^{robot} non conosceva molto le strade di Parigi. **[Malic]**^{cat} invece vagava per tutto il mondo. **[Lo]**^{robot} **[porta]**^{cat} nei posti più famosi. **[Si divertivano]**^{cat,robot} molto insieme. ***Passarono*** un po' di mesi ma un giorno **[Malic]**^{cat} riconobbe **[un uomo]**^{ext1}. Era **[il [suo]^{cat} padrone]**^{ext1}! **[Malic]**^{cat} corse ad abbracciar**[lo]**^{ext1}. **[L' uomo]**^{ext1} **[lo]**^{ext1} riconobbe **[Malic]**^{cat}. **[Sfasac]**^{robot} ci mise un po' per prender**[ci]**^{ext1} confidenza ma alla fine non era poi così male. **[L' uomo]**^{ext1} **[li]**^{cat,robot} portò in casa **[sua]**^{ext1} e vissero tutti felici e contenti. *(corpus italien, CE1, 386)*

L'ensemble de ces analyses met en évidence la diversité des fonctions assumées par les SN définis en position initiale de chaine, qu'il s'agisse de titres, d'introductions contextuelles ou de regroupements référentiels. Cette diversité invite à examiner, dans la section suivante, le fonctionnement des syntagmes nominaux possessifs en position initiale de chaine et leur emploi dans le corpus annoté.

2.2. SN possessifs en position 1

Les syntagmes nominaux possessifs font partie des maillons typiquement attendus dans le maintien d'un référent donné. Pour cette raison nous avons analysés les cas, peu fréquents mais néanmoins présents, où ce type de maillon est présent en tête de chaine ayant fonction d'introduction d'un référent donné. Ces configurations, qui peuvent sembler atypiques,

s'observent principalement dans des contextes d'introduction par ancrage ou de mobilisation d'un savoir partagé entre locuteur et destinataire. Cette notion d'ancrage a été définie par Prince (1981) comme l'introduction d'un nouveau référent dans le discours à l'aide des relations établies avec des entités déjà actives dans le discours. (cf. Chapitre 1, §6.5 sur les déterminants possessifs et sur la notion d'ancrage qui leur est liée).

Dans l'exemple (106), le référent *chat* est introduit à travers son lien avec le référent déjà instancié du *petit garçon*, introduit déjà depuis deux mentions en tant que sujet (bien qu'introduit dans une phrase qui présente un dysfonctionnement apparent au niveau de l'utilisation de la relative). Le type de cette première mention du chat est acceptable dans cette position dans la chaine, ce référent étant inférable à partir du cadre relationnel activé par la possession.

Ce fonctionnement met en évidence la manière dont l'ancrage initial autorise l'emploi de formes typiquement associées au maintien du référent dès la première occurrence, et comment ces formes contribuent à la stabilité de la chaine tout au long du récit (dans les cas des mentions « ton maitre » et « son maitre », surlignés dans le texte).

(106) Il était une fois **[un très gentil loup]**^{wolf} **[qui]**^{wolf} vivait dans la forêt. Et un beau jour **[un petit garçon]**^{ext1} **[qui]**^{ext1} se promenait dans la forêt avec **[[son]**^{ext1} **chat]**^{cat}. Mais **[le chat]**^{cat} avait très peur et puis **[il]**^{cat} s'enfuit. **[Le loup]**^{wolf} vit **[le chat]**^{cat} et **[dit]**^{wolf} « **[n'aie pas]**^{cat} peur, **[je]**^{wolf} **[te]**^{cat} protégerai ». « Mais à une condition, que **[[ton]**^{cat} **maitre]**^{ext1} **[me]**^{wolf} donne de la viande » « C'est d'accord mais comment vais**[-je]**^{cat} prendre la viande ? » « Eh bien **[tu]**^{cat} dis à **[[ton]**^{cat} **maitre]**^{ext1} que **[tu]**^{cat} as faim et **[il]**^{ext1} **[t']**^{cat} en donnera. » Et tous les jours **[le chat]**^{cat} alla chercher de la viande chez **[[son]**^{cat} **maitre]**^{ext1} et **[le loup]**^{wolf} était bien content. Fin *(corpus français, CE1, 2004)*

De manière comparable, dans l'exemple (107), le référent *robot* est introduit par un SN défini accompagné d'un possessif (« *son* fidèle serviteur le robot »), l'existence du robot étant inférée à partir du référent déjà établi (le *loup*) qui en est le patron.

(107) Il était une fois **[un loup]**^{wolf} et **[[son]**^{wolf} **fidèle serviteur le robot]**^{robot}. **[Le robot]**^{robot} faisait tout ce que **[le loup]**^{wolf} demandait mais un jour **[le robot]**^{robot} dit **[au loup]**^{wolf} : « **[Je]**^{robot} m'en vais. » **[Le loup]**^{wolf} dit « : - Pourquoi pars - **[tu]**^{robot} ? » **[Le robot]**^{robot} dit : « **[Je]**^{robot} prends grève, **[je]**^{robot} suis fatigué » dit **[le robot]**^{robot}. « Mais un robot est fait de fer, il n'est jamais fatigué, » dit **[le loup]**^{wolf} « Mais **[je]**^{robot} prends grève quand même. **[Je]**^{robot} prends grève de 100 ans. » **[Le loup]**^{wolf} reste bouche bée devant **[le robot]**^{robot} **[qui]**^{robot} s'en va. Fin. *(corpus français, CE2, 562)*

Enfin, l'exemple (108) met en évidence un cas plus complexe, impliquant plusieurs référents de type *robot*. Bien que le rattachement des différentes mentions à une même chaine référentielle puisse apparaitre partiellement ambigu, l'introduction repose ici sur des relations internes à un groupe déjà constitué (« i suoi amici robot », ses amis robots ; « gli altri robot », les autres robots). La mobilisation de ce cadre référentiel partagé permet l'emploi de formes définies ou collectives

dès la première position de chaine, confirmant que l'apparition de marques de maintien en position initiale s'inscrit dans une continuité avec les mécanismes d'introduction analysés précédemment.

(108) Tanto tempo fa C'era **[un robot]**[robot] **[che]**[robot] cammina e **[vede]**[robot] **[una befana]**[witch], meno male **[la befana]**[witch] non **[l']**[robot] aveva visto, **[lui]**[robot] non sapeva che esistono le befane, lo **[andò]**[robot] a dire **[ai [suoi]**[robot] **amici robot]**[robot1]
"ehi **[robot]**[robot1], lo sapete che esistono le befane, **[volete venire]**[robot1] con **[me]**[robot]?"
"sì" **[dicono]**[robot1]. Lo disse **[agli altri robot]**[robot1] e **[gli altri]**[robot1] hanno detto sì. **[Si sono avventurati]**[robot,robot1] nel bosco per rivedere **[quella befana]**[witch] " e **[gli altri]**[robot1] dove sono? " disse **[il robot]**[robot] che ha visto **[la befana]**[witch]
"prima **[erano]**[robot1] 140 adesso sono 30" **[un altro robot]**[robot2] ha detto
"**[si saranno persi]**[robot1]"
"no, perché **[eravamo agganciati]**[robot,robot1] a una catena" disse **[il robot]**[robot] **[che]**[robot] aveva visto **[la befana]**[witch]
"oh! **[Eccola]**[witch] lì! **[La befana]**[witch] e proprio lì!"
"È vero, **[eccola]**[witch] lì, **[abbiamo staccato]**[robot1] la catena, **[siamo corsi]**[robot1] **[dalla befana]**[witch] e **[abbiamo chiamato]**[robot1] **[i soccorsi]**[ext1]. **[I soccorsi]**[ext1] avevano un jet che sparava reti, pum!!! **[Hanno sparato]**[ext1] la rete e **[l']**[witch] **[abbiamo presa]**[ext1]".
(corpus italien, CE1, 284)[96]

L'analyse des syntagmes nominaux introduits par un possessif en position de premier maillon de chaine montre que leur emploi, bien que peu fréquent, est tout à fait acceptable lorsque l'introduction du référent s'effectue par ancrage, c'est-à-dire en s'appuyant sur un référent déjà introduit dans le texte. Dans ces cas, le possessif permet d'inférer l'existence du nouveau référent à partir d'une relation de possession, d'appartenance ou d'inclusion au sein d'un cadre référentiel partagé. L'utilisation de formes typiquement associées au maintien du référent dès la première occurrence apparait ainsi comme motivée et cohérente, à condition que la chaine concernée ne soit pas la première à apparaitre dans le texte. Ces observations confirment que l'introduction par possessif est surtout possible à partir de la deuxième chaine référentielle, ou des chaines suivantes, lorsque le contexte discursif offre déjà les points d'ancrage nécessaires à l'identification du référent.

2.3. SN démonstratifs en position 1

Comme exposé dans le chapitre théorique (cf. Chapitre 1, §6.4), le syntagme démonstratif renvoie à une entité déjà introduite dans le discours ou encore saillante dans le discours. Il fonctionne comme une expression focalisante, en ce qu'il isole un référent particulier et en réactive la présence au centre de l'attention du lecteur. Conformément à ces propriétés fonctionnelles, et comme le confirment les résultats de l'analyse quantitative (cf. Chapitre 7, §0), les syntagmes

[96] Pour améliorer la lisibilité du texte, nous avons ici disposé le texte en plaçant chaque replique sur une ligne différente, alors que le texte d'origine ne présente pas cette structure. La transcription annotée est disponible en Annexe.

nominaux démonstratifs sont effectivement très peu attestés en position initiale de chaine en français et ne sont pratiquement pas mobilisés en position 1 en italien.

On ne relève en effet que trois occurrences de syntagmes nominaux démonstratifs en position initiale de chaine, toutes attestées en français, dont deux au sein d'un même texte, l'exemple (109). Cette faible fréquence confirme que l'introduction d'un référent nouveau à l'aide d'un démonstratif constitue une stratégie marginale dans les productions écrites scolaires analysées, conformément à la fonction généralement attribuée à ce type de syntagme dans la littérature. Le texte (109) se distingue par ailleurs du reste du corpus en ce qu'il relève davantage d'un type de discours plutôt descriptif que narratif (pour une définition du texte descriptif, nous renvoyons ici à Adam & Petitjean (1982)). Les deux personnages y sont introduits à l'aide de syntagmes nominaux démonstratifs, et les chaines se présentent en succession dans le texte : l'introduction du nouveau référent *sorcière* correspond à un changement de thème et le référent *chat* représenté dans la chaine précédente n'est plus récupéré. En outre, ce texte ne présente pas de déroulement d'actions comparable à celui que l'on observe dans les textes narratifs, mais se limite à une succession de descriptions, organisées autour des propriétés attribuées aux référents (« il est mignon, elle a l'air gentille plus ou moins ») ou par l'énoncé d'actions génériques (« elle peut voler, elle peut faire des potions »).

(109) **[Ce chat]**^{cat} veut des caresses, **[il]**^{cat} est mignon. On voit qu'**[il]**^{cat} veut plein de libertés. **[Cette sorcière]**^{witch} a l'air gentille plus ou moins. **[Elle]**^{witch} peut voler sur **[son]**^{witch} balai magique et **[elle]**^{witch} peut faire des potions : sort de gel, potion de lenteur, de trou noir, de téléportation ou de double. **[Elle]**^{witch} peut faire bonne cuisinière dans les restaurants. *(corpus français, CE2, 1959)*

Si, dans un texte de type majoritairement descriptif, l'emploi d'un tel syntagme peut être considéré comme acceptable, le second texte comportant un démarrage de chaine à l'aide d'un syntagme nominal démonstratif, à savoir le texte (110), présente une configuration différente L'introduction du référent *robot* à l'aide de ce type de syntagme peut y être interprétée comme un dysfonctionnement, dans la mesure où ce référent n'a pas été préalablement introduit dans le discours. La reprise démonstrative du référent *chatte*, en revanche, apparait tout à fait acceptable du point de vue de l'accessibilité du référent, car elle n'intervient qu'en troisième occurrence dans la chaine du référent *chat*. Le démonstratif contribue ici à renforcer la saillance du référent déjà activé, tout en apportant des informations supplémentaires sur ses propriétés.

(110) Il était une fois **[une chatte abandonnée]**^{cat}, **[elle]**^{cat} était très belle pourtant, mais un jour **[ce robot nommé Hubert]**^{robot} trouva **[cette chatte magnifique nommée Moumoune]**^{cat}, puis après avoir réfléchi **[Hubert]**^{robot} dit tout fort « oui, **[je]**^{robot} **[te]**^{cat} veux » peu après **[il]**^{robot} **[la]**^{cat} prit dans **[ses]**^{robot} bras et **[l']**^{cat} **[amena]**^{robot} dans **[sa]**^{robot} maison. **[Ils]**^{cat,robot} s'aiment tellement fort qu'**[ils]**^{cat,robot} ne se séparaient plus. Fin *(corpus français, CE2, 834)*

En italien, la seule occurrence de syntagme démonstratif en position initiale est constituée par un singleton, « quei bambini » (ces enfants), attesté dans l'extrait (111). Dans ce cas, le démonstratif

introduit un maillon de type singleton, qui ne renvoie pas à un référent préalablement introduit dans le discours, mais assume une valeur descriptive et contextuelle. Il permet de faire émerger un groupe de référents génériques, dont la fonction principale n'est pas d'être développé ultérieurement dans le texte, mais de servir de point de comparaison implicite. Le syntagme « quei bambini », et le fait qu'ils soient en train de fêter l'arrivée des bonbons, sert ici à créer un contraste avec la frustration du loup, privé de friandises, en donnant ainsi plus de profondeur à la description des états mentaux du personnage en question.

(111) C'era una volta **[un lupo]**^{wolf} **[che]**^{wolf} ogni giorno doveva alzarsi e correre per vedere se **[la strega Gegi]**^{witch} aveva portato dei dolcetti ma ogni giorno **[continuava]**^{wolf} a guardare se **[la strega Gegi]**^{witch} aveva portato qualcosa ma niente, **[non trovò]**^{wolf} neanche una mosca. Allora un giorno **[disse]**^{wolf} piagnucolando "**[io]**^{wolf} non ho fatto niente di male, **[io]**^{wolf} sono bravo, guarda **[quei bambini]**^{ext1} come festeggiano per i dolcetti" (…). *(corpus italien, CE1, 319)*

L'analyse de ces syntagmes confirme le caractère marginal et potentiellement dysfonctionnel des syntagmes démonstratifs. Son emploi devrait être lié à la réactivation d'un référent donné, plutôt qu'à son introduction. Si dans les textes à dominante descriptive une introduction de référent à travers ce type de syntagme peut sembler acceptable, cela n'est pas le cas dans un texte narratif, où l'introduction d'un référent à l'aide de ce type de syntagme donne lieu à une incohérence de la chaine. Cela confirme que le syntagme démonstratif peut constituer un marqueur de saillance et de focalisation d'un référent donné, dont l'usage en tête de chaine est contraint par le type de discours et par les conditions d'accessibilité du référent.

2.4. Pronoms en position 2 : la présence du pronom relatif

En français, une étude précédente (décrit dans Barletta et Ponton, (2025b), cf. Chapitre 5, §1.1) avait déjà mis en évidence la fréquence des pronoms relatifs omis dans les textes d'élèves. Ce constat est cohérent avec les descriptions proposées par Roubaud et Garcia-Debanc (2014), selon lesquelles les introductions de contes, utilisées sous la forme globale « il était une fois… qui… », constituent des structures relatives complexes, dont l'emploi peut parfois donner lieu à des dysfonctionnements. L'utilisation de ces structures à l'écrit demeure fragile chez les élèves du primaire. Leur enseignement n'est pas prévu à l'école primaire, bien qu'elles soient présentes très précocement dans les usages oraux et écrits des enfants. Du point de vue linguistique, une analyse de Schnedecker avait déjà remarqué la structure de la formule d'introduction des contes de fées ; celle-ci « introduit un référent au moyen d'un SN indéfini, qui est expansé, dans une grande majorité des cas, par une relative » (2021, p. 106). La formule peut ainsi introduire dans la relative soit un nouveau référent soit fournir davantage de détails sur le référent de la principale. Sur la base d'une observation empirique de notre corpus, ce deuxième cas nous semble le plus fréquent que l'on a observé. Voyons maintenant si effectivement cette correspondance entre la formule d'introduction et le pronom relatif se réalise aussi dans notre corpus.

Dans les chaines du corpus français, les pronoms qui paraissent en position 2 des chaines appartiennent à des catégories différentes : sur les 201 pronoms dans cette position, nous

retrouvons des pronoms personnel sujet (*je, il, elle, vous, ils, elles*), des pronoms objet (*moi, toi, lui*) ou le pronom relatif « qui », avec 93 occurrences dans cette position sur le total de 143 occurrences au total dans les chaines. Nous faisons l'hypothèse que cette forte proportion est due en grande partie à la fréquence élevée des formules d'ouverture narratives stéréotypées, et plus particulièrement à la formule « il était une fois », qui a 99 occurrences dans le corpus français annoté. D'autres formules introductives, moins fréquentes, présentent un fonctionnement comparable, telles que « c'est l'histoire de » (6 occurrences) ou encore des constructions en « il y avait », parfois précédées ou suivies d'un syntagme nominal.

Dans ces structures, la deuxième position dans la chaine d'un référent est très majoritairement occupée par un pronom relatif, qui assure la liaison immédiate entre l'introduction du référent et sa première description. Les exemples (112)à (116) illustrent cette structure récurrente, dans laquelle nous retrouvons une formule d'ouverture plus ou moins stéréotypée, suivie d'un syntagme nominal indéfini qui introduit un nouveau référent, et qui est immédiatement suivi d'une proposition relative introduite par *qui* :

(112) Il était une fois **[un petit chat blanc]**^{cat} **[qui]**^{cat} vivait dans la forêt.
(corpus français, CE1, 212)

(113) Il était une fois **[un chat]**^{cat} et **[une sorcière]**^{witch} **[qui]**^{cat,witch} voulaient tellement avoir un balai, mais **[ils]**^{cat,witch} n'avaient pas assez d'argent.
(corpus français, CE1, 1850)

(114) C'est l'histoire d'**[un chat]**^{cat} **[qui]**^{cat} avait une belle maison.
(corpus français, CE1, 1829)

(115) Un jour, il y avait **[une sorcière]**^{witch} **[qui]**^{witch} vivait avec **[un chat]**^{cat}.
(corpus français, CE1, 2530)

(116) Il y a **[une sorcière]**^{witch} **[qui]**^{witch} habitait dans une vieille maison.
(corpus français, CE1, 1317)

Dans ces configurations, la proposition relative remplit une fonction centrale dans l'introduction référentielle : elle permet d'associer immédiatement à l'apparition du référent une première caractérisation, qu'il s'agisse d'un état, d'une propriété ou d'une action initiale, tout en assurant la continuité syntaxique et référentielle entre l'introduction du syntagme nominal et le développement du récit. La fréquence élevée des pronoms en position 2 s'explique ainsi par la présence de ces formules d'ouvertures stéréotypées, typiques des textes narratifs.

Un fonctionnement comparable est observable dans le corpus italien. Sur un total de 121 occurrences de pronoms en deuxième position dans les chaines, 90 correspondent au pronom relatif *che*. Comme en français, ces occurrences apparaissent majoritairement dans des structures introductives stéréotypées, en particulier dans la formule « *C'era una volta »*, comme l'illustre l'exemple (117).

(117) C'era una volta **[un robot]**^{robot} **[che]**^{robot} si chiamava **[Arbo]**^{robot}. Poi c'era **[una strega]**^{witch} **[che]**^{witch} si chiamava **[Eva]**^{witch}. *(corpus italien, CE1, 1376)*

3. Nombre de personnages et modes de cohabitation des chaines

Il apparait maintenant pertinent d'analyser les stratégies de gestion des chaines référentielles en tenant compte de deux facteurs principaux : le nombre des référents annotés dans le texte et le genre grammatical de ces référents. C'est effectivement reconnu que « la concurrence discursive entre les différents référents partageant des traits communs (animéité, genre grammatical notamment) (Arnold & Griffin, 2007; Fukumura & van Gompel, 2011) constitue le point central des problèmes de clarté référentielle dans les récits produits par les enfants. Ces derniers ne parviennent pas toujours à choisir une forme qui soit suffisamment explicite ou descriptive pour les différencier les uns des autres » (Dupret, 2022, p. 10).

L'hypothèse sous-jacente à cette analyse est donc que, dans l'étape d'apprentissage de l'écrit représentée dans notre corpus, les élèves mettent en œuvre des stratégies de gestion référentielle visant à éviter les ambiguïtés. Cette hypothèse s'inscrit dans la continuité de travaux montrant que « ils·elles sont sensibles à la présence d'un autre référent dans l'évitement du recours à des formes pronominales qui pourraient s'avérer ambiguës » (Serratrice, 2013, citée dans Dupret, 2022, p. 11). En outre, la variation entre différentes stratégies devrait être d'autant plus observable en fonction du nombre et du genre grammatical des référents.

Pour rappel, l'ambiguïté référentielle se manifeste dans les cas où une expression référentielle se trouve à avoir potentiellement deux sources (cf. Chapitre 1, §3.2). Plusieurs facteurs peuvent provoquer cette confusion dans le rattachement : l'ordre des mentions, la longueur des chaines, la catégorie grammaticale et la fonction syntaxique des mentions sont prises en compte dans la résolution des ambiguïtés référentielles, ou encore le sémantisme des verbes, ainsi que le nombre des référents saillants impliqués dans le récit et leur genre grammatical (Schnedecker, 2021, p. 104). En effet, lorsque deux ou plusieurs personnages partagent le même genre grammatical, le risque de confusion lors du rattachement de mentions peu informatives, telles que les pronoms, est potentiellement plus élevé. À l'inverse, lorsque les référents relèvent de genres grammaticaux distincts, cette confusion de rattachement devrait être moins probable.

Dans le cas de textes écrits par des scripteurs en phase d'apprentissage, le nombre de personnages ainsi que leur genre devrait plus fortement influencer les stratégies de construction des chaines. En effet, lorsqu'un seul personnage est présent, la reprise peut s'effectuer davantage à l'aide de maillons peu informatifs, plus faibles, alors que s'il y a plusieurs référents de genres similaires, leur gestion demandera la sélection de mentions plus rigides et plus informatives pour éviter le plus possible les interférences entre chaines. Ici, nous allons nous focaliser d'abord sur les textes qui présentent un seul référent annoté, pour ensuite analyser les textes où deux référents sont présents, du même genre puis contenant deux référents de genres différents. Ainsi, nous pourrons analyser les caractéristiques des chaines pour en décrire la variation, les stratégies fréquemment utilisées, ou décrire les cas de dysfonctionnement que l'on rencontre dans nos analyses.

Pour mener cette analyse, l'ensemble des textes annotés du corpus a été pris en compte. Le recensement du nombre de personnages par texte, que nous avons déjà mentionné précédemment (cf. Chapitre 7, §1.3), et qui est illustré dans la Figure 69, met en évidence une

distribution globalement comparable dans les deux corpus étudiés. Dans le corpus français, huit textes ne comportent qu'un seul référent (sept produits au niveau CE1 et un au niveau CE2) ; en italien, cette configuration est plus marginale et concerne uniquement trois textes du niveau CE1. La grande majorité des productions, dans les deux langues, met en jeu deux personnages référentiels centraux, conformément à la consigne initiale donnée aux élèves.

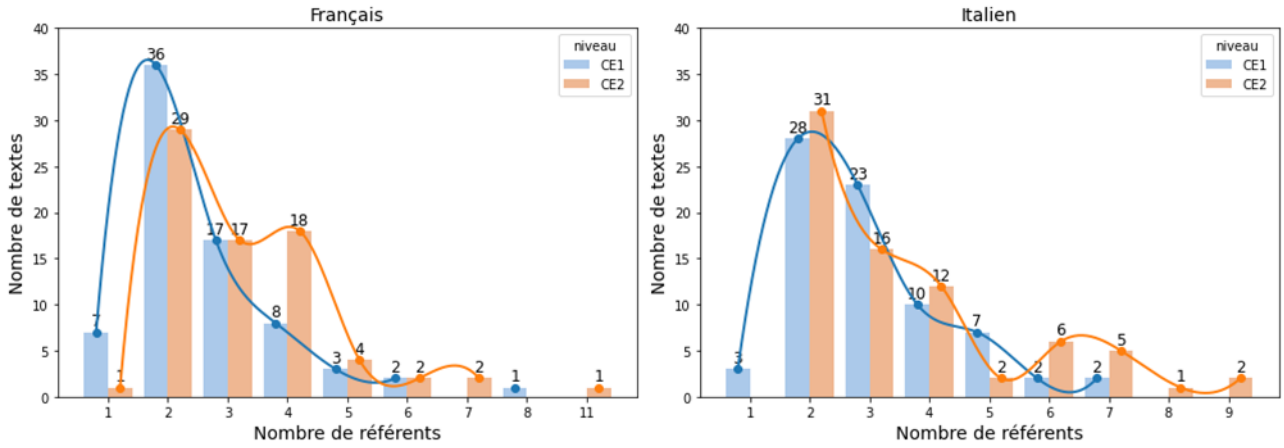


Figure 69. Distribution des textes par nombre de référents dans le corpus annoté. Reprise de la Figure 49.

Dans cette perspective, nous avons ainsi sélectionné dans notre échantillon certains des textes présentant trois configurations différentes : 1) les textes contenant un seul personnage ; 2) les textes comportant deux personnages du même genre grammatical ; 3) les textes comportant deux personnages de genre grammatical différent. En nous appuyant sur les étiquettes relatives aux personnages principaux possibles et imposés par la consigne d'écriture (*chat, sorcière, loup, robot*), nous distinguons ainsi les textes présentant deux personnages du même genre grammatical (par exemple le *chat* et le *robot*) de ceux qui comprennent deux personnages de genres grammaticaux distincts (le *chat* et la *sorcière*). Nous avons par la suite vérifié si le genre du personnage attendu correspondait au genre employé par l'élève. Nous décrirons ainsi la composition des chaines pour chaque référent annoté dans le texte en termes de types de mention pour ensuite analyser la modalité de cohabitation utilisée (si applicable), pour enfin spécifier davantage quelles composantes narratives sont présentes dans le texte. Cela devrait permettre de fournir un cadre plus complet de la « qualité narrative » intrinsèque aux textes analysés et de la mettre en perspective avec la gestion référentielle au sein du texte.

Nous rajoutons ici qu'en conduisant ces analyses nous avons remarqué une des possibles limites de la typologie des mentions que nous avions adoptée, et de nos annotations : dans plusieurs cas nous retrouvons ainsi des mentions annotées qui contiennent deux types différents, surtout des SN définis et indéfinis accompagnés de noms propres, ce qui revient régulièrement dans nos analyses.

Afin de rendre plus lisible la composition des chaines en termes de types de mentions, chaque texte annoté est accompagné d'un schéma représentant graphiquement le déroulement des chaines référentielles et la nature des mentions qui les composent (Figure 70 à Figure 91).

Dans ces schémas, chaque chaine est représentée sur une ligne distincte, identifiée à gauche par l'étiquette du référent correspondant. Chaque carré figurant sur la ligne correspond à une mention annotée dans l'ordre d'apparition dans le texte ; l'intérieur du carré indique le type de mention (SN indéfini, SN défini, pronom personnel ou relatif, nom propre, anaphore zéro, déterminant possessif, etc.). Les mentions plurielles renvoyant simultanément à deux référents sont représentées par un même carré placé sur les deux lignes concernées. Les mentions imbriquées (par exemple un syntagme nominal contenant un possessif annoté séparément) sont figurées par une représentation emboîtée des éléments correspondants, afin de rendre visible la structure interne de la mention. Enfin, pour des raisons de lisibilité graphique, certaines chaines ont été réparties sur plusieurs segments de ligne : lorsque le nombre de mentions excédait l'espace disponible dans le cadre du schéma, la représentation a été poursuivie à la ligne suivante. Cette division ne correspond pas à une rupture dans la chaine référentielle, mais constitue uniquement un choix de mise en page destiné à faciliter la visualisation.

3.1. Textes contenant un seul référent – stratégie de gestion à baisse informativité

Après analyse, les productions écrites mettant en scène un seul personnage (11 au total dans notre corpus annoté, 8 en français et 3 en italien) semblent présenter des stratégies de gestion référentielle qui privilégient davantage des types de mention moins informatives, moins rigides et plus atténuées. Cela devrait être lié à la présence d'un seul référent saillant dans le texte, et donc à la possibilité moindre d'ambiguïté référentielle pour ces mentions moins sémantiquement informatives.

Corpus français : textes avec un référent annoté

Par exemple, dans les textes (118) et (119) du corpus français de niveau CE1, la gestion référentielle de maintien du référent est quasi exclusivement pronominale. La seule variation dans la chaine est l'introduction des déterminants possessifs pour introduire d'objets potentiellement fonctionnels à la description du personnage.

Dans le texte (118), après l'introduction du personnage du *robot* effectuée à l'aide d'un SN indéfini, « un robot », ce qui correspond au type de mention attendu pour l'introduction d'un nouveau personnage, nous retrouvons trois mentions pronominales, suivies par une mention de type déterminant possessif « sa », introduisant un autre référent, un instrument utilisé par le robot, non animé et donc pas pris en compte par notre annotation, suivie de quatre mentions pronominales sujet. La stratégie de gestion référentielle est ici quasi exclusivement pronominale sur les 9 mentions de la chaine.

(118) Il était une fois **[un robot]**^{robot}. **[Il]**^{robot} vivait dans des endroits sous terre. **[Il]**^{robot} était très grand. **[Il]**^{robot} ne lâchait jamais **[sa]**^{robot} pelle. **[Il]**^{robot} voulait aller rejoindre le noyau de la terre pour voir ce qu'il y avait à l'intérieur. Mais **[il]**^{robot} n'y arrivait pas alors **[il]**^{robot} était obligé de remonter mais cela prenait des années et des années mais **[il]**^{robot} n'y arrivait pas. *(corpus français, CE1, 3045)*



Figure 70. Schéma de la chaine dans le texte du corpus français, CE1, 3045.

Dans cette production, la dimension descriptive domine la narration : même lorsque des verbes sont présents, ceux-ci servent à décrire le personnage (« il était très grand »), sa condition (« il vivait dans des endroits sous terre ») ou ses états mentaux (« il voulait aller rejoindre le noyau de la terre »). On observe aussi ici que à la description d'une situation initiale du robot ne fait pas suite une résolution de celle-ci mais que la résolution de l'histoire correspond à un échec des plans du personnage (« il n'y arrivait pas »). En ce sens, il manque certains éléments qui caractérisent habituellement la narration : si ce texte présente une introduction, la description d'un contexte, un événement déclencheur (l'envie du robot de rejoindre le centre de la terre), la planification des actions du personnage et une vraie résolution de l'intrigue sont absentes.

Dans le texte (119), nous retrouvons un exemple similaire de stratégie de gestion référentielle, dans un texte encore plus descriptif que le précédent. Cette production relève davantage de la description que de la narration : les caractéristiques physiques et les habitudes d'un *chat*, dont la chaine comporte 7 mentions, y sont présentées à travers des phrases courtes. La première mention est un syntagme nominal indéfini, « un chat », qui représente les premiers mots du récit. Elle est suivie de deux pronoms relatifs, puis d'un enchainement de quatre pronoms personnels. Les deux pronoms relatifs introduisent des éléments de description : une caractéristique physique (« un chat qui a une queue noire ») et une action habituelle du personnage principal du texte (« qui chasse tout le temps les souris »). Cette description condensée du référent et l'absence d'un verbe dans la phrase principale lui font assumer une fonction quasi de titre du texte. La suite du texte décrit quatre états mentaux du référent : le premier est une répétition de la situation déjà introduite dans la relative (« il joue beaucoup à chasser les souris »). Les trois phrases suivantes contiennent toutes les trois le verbe « aimer » en tant que verbe principal, verbe qui supporte ici des descriptions des états mentaux du référent (« il n'aime pas les chiens », « il n'aime pas se battre », « le seul truc qu'il aime c'est manger »). Comme déjà mentionné, ici les éléments typiques de la narration sont pratiquement absents.

(119) **[Un chat]**^{cat} **[qui]**^{cat} a une queue noire et **[qui]**^{cat} chasse tout le temps les souris. **[Il]**^{cat} joue beaucoup à chasser les souris. Mais **[il]**^{cat} n'aime pas les chiens et **[il]**^{cat} n'aime pas se battre. Le seul truc qu'**[il]**^{cat} aime c'est manger. *(corpus français, CE1, 3046)*

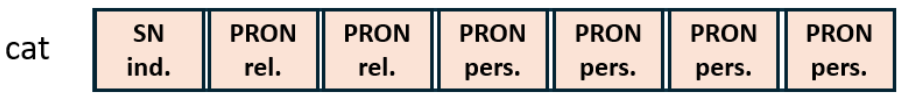


Figure 71. Schéma de la chaine dans le texte du corpus français, CE1, 3046.

Les deux textes montrent ainsi que, lorsque la narration ne met en scène qu'un seul protagoniste, le maintien du référent peut s'effectuer majoritairement à l'aide de reprises pronominales : aucun référent concurrent n'est présent, et cela permet ainsi de rattacher de manière non ambigüe les pronoms au seul référent saillant décrit. En revanche, ces deux textes ne peuvent pas se caractériser comme « prototypiquement narratifs » au regard des critères retenus précédemment : la gestion référentielle y semble maitrisée, mais elle s'inscrit dans des productions où la structure narrative est peu ou pas élaborée.

Dans la même veine, le texte (120) montre une stratégie de gestion homogène et largement pronominale. La chaine comporte ici 11 mentions. Le référent *sorcière* est introduit à l'aide d'un SN indéfini, puis maintenu tout au long du texte par une alternance entre pronoms personnels et déterminants possessifs. Ces derniers apparaissent intégrés à des syntagmes nominaux désignant des éléments associés au personnage (« sa maison imaginaire dans l'espace », « sa potion magique »).

Cette alternance est interrompue par une succession de quatre pronoms personnels sujets au centre du texte. La dernière mention de la chaine prend la forme d'un syntagme nominal défini (« La sorcière ne sort pas »), qui intervient après une description météorologique (« C'est l'hiver, il fait froid et il y a du vent »). Cette mention opère une refocalisation explicite sur le personnage à la suite d'un segment descriptif.

(120) **[Une sorcière]**^{witch} se promène dans la rue et tout d'un coup **[elle]**^{witch} s'envole vers **[sa]**^{witch} maison imaginaire dans l'espace. **[Elle]**^{witch} est arrivée à **[sa]**^{witch} maison, **[elle]**^{witch} rentre et **[elle]**^{witch} se couche. Le matin **[elle]**^{witch} va chercher le pain et des croissants pourris et **[elle]**^{witch} prépare **[sa]**^{witch} potion magique. C'est l'hiver, il fait froid et il y a du vent. **[La sorcière]**^{witch} ne sort pas. Enfin l'été il y a un gros soleil dehors et il fait très chaud dehors. *(corpus français, CE1, 2901)*

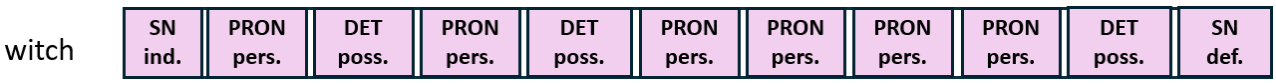


Figure 72. Schéma de la chaine du texte du corpus français, CE1, 2901.

Contrairement au texte précédent, ici le personnage en question réalise plusieurs actions (« se promène dans la rue », « s'envole vers sa maison imaginaire dans l'espace », « va chercher le pain et des croissants pourris », « elle prépare sa potion magique » etc.). Néanmoins, ici nous n'avons pas non plus un texte narratif que l'on pourrait catégoriser comme typique, c'est-à-dire comprenant une situation de début, un obstacle rencontré par les personnages principaux et la résolution de cet obstacle comme fin de l'histoire. Le cadre est posé en fin d'histoire à travers des représentations de la météo (« il fait froid et il y a du vent », « il fait très chaud dehors »), dont la première description est indirectement mise en lien avec l'action de la sorcière, qui « ne sort pas ».

Le seul texte du corpus français de niveau CE2 mettant en scène un personnage unique (exemple (121)) présente toutefois une variété légèrement plus grande dans la gestion de la chaine, qui comporte ici 9 mentions.

(121) Il était une fois **[un chat]**[cat] **[qui]**[cat] avait oublié **[sa]**[cat] nourriture dans un voyage. Alors **[il]**[cat] prenait la nourriture des gens. 2 semaines plus tard **[le chat]**[cat] n'avait plus faim. Alors **[le chat]**[cat] va se coucher. Et après **[il]**[cat] avait encore faim et **[il]**[cat] mange enfin **[sa]**[cat] nourriture. *(corpus français, CE2, 3055)*

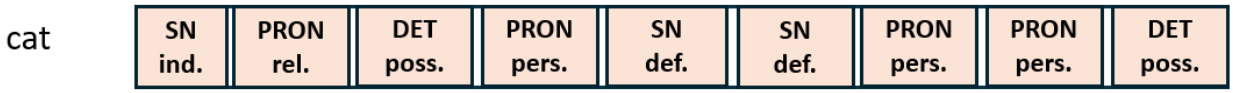


Figure 73. Schéma de la chaine du texte du corpus français, CE2, 3055.

Dans ce texte relativement court, des syntagmes nominaux définis s'alternent avec des pronoms personnels et des déterminants possessifs, notamment pour introduire ou maintenir des référents non animés. Cette variation suggère une maîtrise plus avancée des ressources référentielles disponibles, bien que le texte soit peu cohérent, mettant en scène un enchainement d'actions qui semble relever de la situation initiale (le manque de nourriture pour le chat). Il faut noter ici qu'après une succession de pronoms ou de déterminants, le personnage est refocalisé à l'aide d'un SN défini car un nouveau référent animé, « les gens », venait d'être introduit dans la narration, bien qu'introduit en tant que complément du nom dans le syntagme « la nourriture des gens ». Par ailleurs, bien que ce cas de figure n'ait pas été inclus dans notre guide d'annotation, nous nous posons *a posteriori* la question de la validité du fait de ne pas annoter ce référent, qui est animé mais qui ne semble pas réaliser des actions au sein du texte et qui représente ici un singleton. Ce texte semble lui aussi décrire une suite d'actions plutôt que correspondre à une structure narrative proprement dite, bien qu'ici certaines des actions décrites semblent liées par des liens de causalité (le chat oublie sa nourriture, vole de la nourriture autrui pour se nourrir).

Corpus italien : textes avec un référent annoté

Après l'analyse des textes du corpus français mettant en scène un seul référent, nous souhaitons aborder à présent les textes du corpus italien correspondant à cette même configuration. En français, les pronoms constituent la catégorie de mention la plus fréquente dans les chaines référentielles (cf. Chapitre 7, §5 pour l'analyse quantitative des distributions des types de mention), et ils assurent également une fonction centrale dans le cas de construction uniforme des chaines en absence de concurrence référentielle. Sur la base de cette correspondance entre fréquence quantitative et rôle discursif des types de mention, nous faisons l'hypothèse que, dans le corpus italien, ce même rôle discursif est majoritairement pris en charge par l'anaphore zéro, comme le suggérait déjà la distribution des types de mention analysée précédemment (cf. . Chapitre 7). Bien que les données relatives à la gestion d'un seul référent soient très limitées en italien (pour rappel, nous avons seulement 3 textes ayant cette configuration en italien), il reste néanmoins possible d'observer ces productions et d'analyser les stratégies de gestion référentielle qu'elles mobilisent, afin de les comparer aux textes français et d'évaluer la validité de cette hypothèse, qui reste à être confirmée sur une plus grande quantité de textes.

Effectivement, dans l'exemple (122), nous retrouvons 15 mentions annotées au total. L'introduction du référent est effectuée par une mention de dénomination explicite « Un gatto di

nome Alex » (un chat nommé Alex). En dehors de cette première mention, seulement deux autres mentions sous forme de nom propre simple « Alex » sont présentes dans le texte, tandis que les 12 mentions restantes correspondent à des anaphores zéro. La chaine est ainsi maintenue presque exclusivement par la flexion verbale, qui suffit à assurer la continuité référentielle en l'absence de tout autre référent concurrent.

Tout comme l'exemple précédent en français, ce texte présente une succession d'actions effectuées par le personnage, et il présente deux particularités : d'une part, le chat subit un reproche par un référent qui n'est pas introduit de manière explicite dans le texte (« si fa sgridare ») ; d'autre part, nous retrouvons en fin de texte une accumulation de 9 adverbes consécutifs avec fonction d'intensificateurs « distrugge tutta, tutta, tutta, tutta, tutta, tutta, tutta, tutta, tutta, tutta l'università » (il détruit toute (…) toute l'université), amplifiant l'effet de narration.

(122) **[Un gatto di nome Alex]**^{cat} gioca con un gomitolo e **[rompe]**^{cat} i letti, la tv, i pupazzi, il divano ecc. E **[butta]**^{cat} giù il frigo ed **[esce]**^{cat} di casa e **[rompe]**^{cat} un'auto pure la scatola **[ha rotto]**^{cat}. **[Alex]**^{cat} ritorna e **[si fa sgridare]**^{cat}. **[Ha fatto]**^{cat} la cacca fuori e pure la pipì. **[Va]**^{cat} anche al mercato. Insomma **[Alex]**^{cat} è passeggione: **[è andato]**^{cat} in Cina, Giappone, Africa, America, Sud America, Australia ecc. **[Va]**^{cat} in una scuola due giorni dopo **[va]**^{cat} all'università e poi **[distrugge]**^{cat} tutta tutta tutta tutta tutta tutta tutta tutta tutta tutta tutta l'università. *(corpus italien, CE1, 599)*

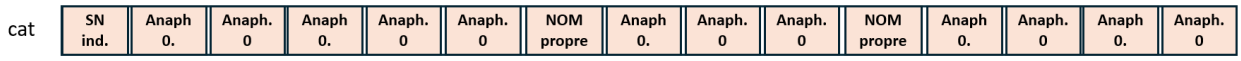


Figure 74. Schéma de la chaine du texte du corpus italien, CE1, 599.

La narration demeure ainsi centrée sur des actions effectuées par le personnage, et toute représentation de ses état mentaux internes est absente. Le seul adjectif de description du personnage d'Alex intervient à la moitié du texte en tant que justification des actions du personnage, de plus à travers un adjectif qualificatif qui n'est pas attesté en italien standard mais transparent sur le plan sémantique : « passeggione » (« Insomma Alex è passeggione » ; Bref, Alex est un grand flâneur).

Le référent *chat* dans l'exemple (123) présente une oscillation de genre grammatical, phénomène néanmoins interprétable et acceptable dans le contexte du texte, dans la mesure où il s'agit d'un animal, catégorie dont le genre grammatical peut être parfois en décalage avec le genre du référent attribué par l'élève. Cette variation de genre ne peut pas entrainer de difficulté majeure de rattachement référentiel, le personnage étant unique et fortement saillant tout au long de la production. Le genre réel du référent est ainsi défini dès la troisième mention à l'aide du nom propre choisi, « Ghiacciolina », puis confirmé plus tard à travers un pronom personnel féminin, « lei » (elle), avant de revenir à la dénomination au masculin (probablement perçue comme neutre) dans un syntagme démonstratif « questo gatto » (ce chat) en fin de chaine. Dans ce texte demeurent prévalentes les descriptions physiques du référent « un gatto carino » (un chat mignon) ; « lei era anche bellissima quando si vestiva », (elle était aussi très belle quand elle

s'habillait) ; « e questo gatto è bianco con pochissimi pallini marroni », (et ce chat est blanc avec très peu de petits points marrons).

Ces descriptions physiques alternent avec des descriptions du caractère du personnage, initialement stéréotypées « era bella e buona », (elle était belle et gentille), puis justifiées par des actions du personnage, « era buona con tutti e ascoltava tutti », (elle était gentille avec tout le monde et elle écoutait tout le monde).

Tout comme les textes du français, celui-ci relève davantage de la description que de la narration, bien que le texte démarre avec la formule « c'era una volta », toute autre caractéristique d'une narration est effectivement absente.

(123) C'era una volta **[un gatto carino]**^{cat} **[che]**^{cat} si chiamava **[Ghiacciolina]**^{cat}, **[era]**^{cat} bella e buona, **[era]**^{cat} buona con tutti e **[ascoltava]**^{cat} tutti, e **[Ghiacciolina]**^{cat} era una mangiona e il **[suo]**^{cat} mangiare preferito erano i croccantini e il frutto preferito è la fragola.
[Lei]^{cat} era anche bellissima quando si vestiva.
E **[questo gatto]**^{cat} è bianco con pochissimi pallini marroni. *(corpus italien, CE1, 627)*

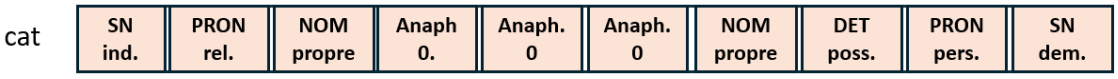


Figure 75. Schéma de la chaine du texte du corpus italien, CE1, 627.

En résumé, après analyse, les productions écrites mettant en scène un seul personnage (11 au total dans notre corpus annoté, 7 en français et 3 en italien) présentent des stratégies de gestion référentielle qui privilégient davantage des types de mention moins informatives, moins rigides et plus atténuées. Cela correspond à une utilisation majeure des anaphores pronominales dans le cas du français et d'anaphores zéro en italien. En effet, en l'absence de compétition entre plusieurs référents saillants, le risque d'ambiguïté de rattachement à une chaine est très réduit ou inexistant. Cela permet de maintenir la présence d'un même référent dans le texte à l'aide de formes référentielles moins chargées sémantiquement et à travers une sélection des types de mention plus homogène. Bien que d'autres référents puissent apparaitre ponctuellement dans ces textes, nos annotations ciblant exclusivement les personnages animés des textes, nous considérons que les autres référents qui apparaissent seront non animés donc automatiquement moins saillants, mois persistants et devraient donc jouer un rôle marginal au niveau discursif. Bien que quantitativement peu représentés, ces textes offrent un terrain d'observation intéressant pour comparer les stratégies de gestion en absence de compétition référentielle dans les deux langues. En outre, la présence d'un référent seul permet de mieux observer le rôle discursif de ces chaines homogènes en isolement dans le texte.

Ces résultats mettent en évidence que la gestion référentielle dans les textes à personnage unique repose avant tout sur des stratégies de maintien à faible informativité, rendues possibles par l'absence de concurrence entre référents saillants. Si le français privilégie les anaphores pronominales et l'italien l'anaphore zéro, ces différences formelles relèvent davantage de

contraintes typologiques que de divergences fonctionnelles. Cette convergence invite à examiner, dans les sections suivantes, comment ces stratégies évoluent lorsque plusieurs référents entrent en compétition au sein du même texte.

3.2. Textes contenant deux référents de genres grammaticaux différents

Lorsque deux référents de genres grammaticaux différents sont présents au sein d'un même texte, la gestion référentielle est susceptible de se complexifier à cause de la nécessité de gérer des éventuelles concurrences entre référents saillants. Contrairement aux textes à référent unique, le maintien référentiel ne peut plus reposer sur des stratégies d'utilisation de mentions à basse informativité, mais exige la présence d'indices suffisants pour garantir des rattachements aux chaines correctes. Toutefois, la différence de genre grammatical entre les deux référents d'un texte pourrait constituer une facilitation dans cette gestion de deux référents : dès lors que les mentions encodent le trait de genre syntaxique, cette information contribue à restreindre l'éventail des interprétations possibles et à diminuer le risque de confusion entre référents.

On peut ainsi avancer que, même dans une configuration à deux référents, lorsque ceux-ci relèvent de genres grammaticaux distincts, le maintien peut être effectué au moyen de formes peu informatives et atténuées, à condition que celles-ci indiquent de manière explicite le genre grammatical. À défaut d'un marquage morphologique direct, la désambiguïsation peut également être assurée par la cohérence sémantique et/ou contextuelle de la mention avec l'un des deux personnages.

L'observation empirique des textes du corpus met en évidence deux modes principaux de gestion de ces configurations à deux personnages : l'une qui tend à privilégier une gestion référentielle à base de syntagmes nominaux, plus informatifs, l'autre qui s'appuie davantage sur des mentions moins informatives, telles que les pronoms et les anaphores zéro. Il convient toutefois de souligner que ces deux stratégies correspondent aux pôles d'un continuum plutôt qu'à des catégories strictement étanches. Dans les productions analysées, elles coexistent fréquemment au sein d'un même texte. La gestion référentielle se caractérise alors par des ajustements progressifs en fonction des besoins de clarification ou, au contraire, des possibilités offertes par le contexte discursif et l'exploration des limites de ces possibilités dans cette étape d'apprentissage de l'écriture représentée dans notre corpus.

La présente section vise ainsi à mettre à l'épreuve l'hypothèse formulée ci-dessus à partir de l'analyse d'un binôme particulièrement représenté dans le corpus (45 fois dans le corpus français et 53 dans l'italien) : le binôme *chat/sorcière*. Ces deux figures sont étroitement associées dans l'imaginaire culturel et narratif, où le chat occupe fréquemment la fonction de familier[97] de la sorcière. Dans les textes du corpus, nous avons sélectionné les textes contenant seulement au moins trois mentions annotées pour chacun des deux référents *cat* et *witch* annotés.

[97] Le familier (génie familier, démon familier ou tout simplement familier) est « une entité, animal ou esprit, parfois imaginaire et invisible, à laquelle les hommes s'adressent pour demander des conseils ou obtenir des services, en particulier liés à la sorcellerie. (« Familier (esprit) », 2024)

Corpus français : textes avec deux référents annotés

Dans le corpus français, 56 textes comportent ce binôme en cooccurrence avec d'autres référents. Parmi eux, 19 textes présentent au moins trois mentions annotées pour chacun des deux référents (*cat* et *witch*). Sept de ces textes relèvent du niveau CE1 et se caractérisent par une brièveté marquée (entre 15 et 73 tokens), ce qui limite le développement des chaines référentielles et, par conséquent, les possibilités d'observer des phénomènes de concurrence ou de désambiguïsation plus complexes.

Afin d'analyser des configurations dans lesquelles les chaines sont suffisamment développées pour faire apparaitre des stratégies de maintien, nous avons ainsi choisi de concentrer notre analyse sur les textes de niveau CE2 contenant ces deux référents. Plus étendus (entre 75 et 142 tokens), ces textes présentent des structures narratives plus élaborées et des chaines de référence plus longues, permettant d'observer de manière plus fine les mécanismes de gestion et d'éventuelles ambiguïtés dans la réalisation des chaines de continuité référentielle des deux personnages. Ce choix méthodologique se justifie également par la portée didactique de l'analyse, ces productions offrant un matériau particulièrement pertinent pour réfléchir aux enjeux de l'enseignement de la cohésion référentielle à ce stade de l'apprentissage de l'écriture.

Sur le plan référentiel, le texte (124), de longueur réduite, comporte 12 mentions annotées, dont 7 constituées par des formes pronominales (sujets ou objets). La chaine du *chat*, introduite à l'aide d'un SN indéfini, est dominante dans la première partie du texte et bénéficie d'un maintien pronominal continu. La chaine de la *sorcière*, bien qu'introduite dès l'incipit, n'est véritablement réactivée qu'après cette partie centrée sur le *chat*, à travers le syntagme nominal défini « la sorcière ». Cette partie présente donc une gestion par succession des deux chaines, qui rend possible l'emploi pronominal. La brève cohabitation des deux chaines dans la seconde moitié du texte est majoritairement gérée par des pronoms (« elle », « le »), sans ambiguïté démontrable grâce à la différence de genre grammatical. Le seul syntagme nominal lexical nouveau correspond à « chat vert », qui ne sert pas à réintroduire le référent, mais à signaler sa transformation en termes d'attributs.

Le texte présente certains éléments de la grille d'analyse de la narration retenue : les personnages principaux et le contexte de l'histoire sont décrits de manière succincte et selon certains stéréotypes dans la première phrase du texte (« Une sorcière avait une maison noire et un chat noir »). L'événement déclencheur pourrait être identifié dans le fait que le chat se cache dans la maison, ce qui provoque l'insatisfaction de la sorcière, qui constitue une réponse interne ou représentation d'état mental (« Un jour la sorcière elle en avait marre ») et la pousse à changer la couleur du chat « en chat vert », ce qui est implicitement lié à la possibilité de le voir maintenant, par contraste, dans la maison toute noire, cette dernière transformation constituant la résolution et le final de l'histoire.

Globalement, ce texte illustre une gestion majoritairement pronominale des référents, conforme à l'hypothèse initiale selon laquelle, en présence de deux personnages de genres distincts et dans un texte court à faible concurrence référentielle, le maintien peut s'effectuer efficacement au

moyen de formes peu informatives, dès lors que le marquage morphologique suffit à garantir la désambiguïsation.

(124) **[Une sorcière]**witch avait une maison noire et **[un chat noir]**cat. **[Il]**cat se cachait dans la maison pour ne pas qu'on **[le]**cat voie. **[Il]**cat ne sortait que la nuit. Un jour **[la sorcière]**witch en avait marre. Avec **[sa]**witch baguette magique **[elle]**witch **[l']**cata transformé en **[chat vert]**cat comme ça **[elle]**witch **[le]**cat voyait tout le temps. (*corpus français, CE2, 207)*

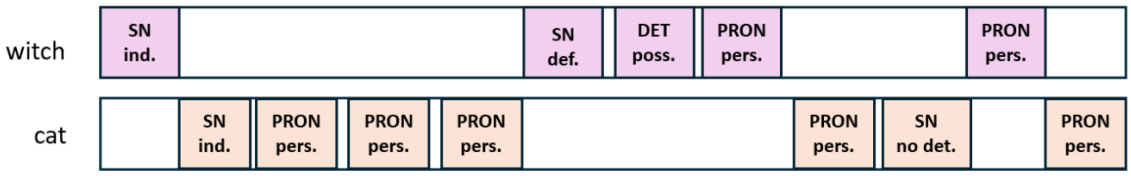


Figure 76. Schéma illustrant les chaines du texte du corpus français, CE2, 207.

Le texte (125) semble présenter des stratégies de gestion plus variées. Il comporte 20 mentions au total, dont 8 pronoms personnels. Les deux référents sont introduits dès l'incipit par des formules contenant le verbe *s'appeler* et le nom propre du référent (où les noms propres « Craboudine » et « Grifdure » sont très expressifs).

La chaine de la sorcière est activée puis maintenue en premier, et le chat est introduit en fonction de sa présence (« Elle avait un chat »). Après son introduction, la chaine du *chat* est maintenue à travers un pronom personnel introduisant une présentation stéréotypée (« il était très intelligent ») avant de donner lieu à deux mentions fusionnées, décrivant le rapport entre les deux personnages (« Ils restaient toujours ensemble, rien ne pouvait les séparer »). Ici l'événement déclencheur, le renversement de la potion par la part du chat, provoque une réaction émotionnelle de la sorcière (« elle était si furieuse… ») qui cause la résolution de l'histoire (« …qu'elle abandonna son chat »). Dans ce texte, les pronoms personnels sujets au singulier indiquent toujours le référent le plus proche.

(125) {**[Le chat]**cat et **[la sorcière]**witch}
Il était une fois **[une sorcière]**witch **[qui]**witch s'appelait **[Craboudine]**witch. **[Elle]**witch avait **[un chat]**cat, **[il]**cat s'appelait **[Grifdure]**cat, **[il]**cat était très intelligent. **[Ils]**cat,witch restaient toujours ensemble, rien ne pouvait **[les]**cat,witch séparer. Un jour catastrophe **[le chat]**cat avait fait une grosse bêtise en jouant, **[il]**cat avait renversé la potion de **[[sa]cat maîtresse]**witch, **[elle]**witch était si furieuse qu'**[elle]**witch abandonna **[[son]witch chat]**cat. *(corpus français, CE2, 1352)*

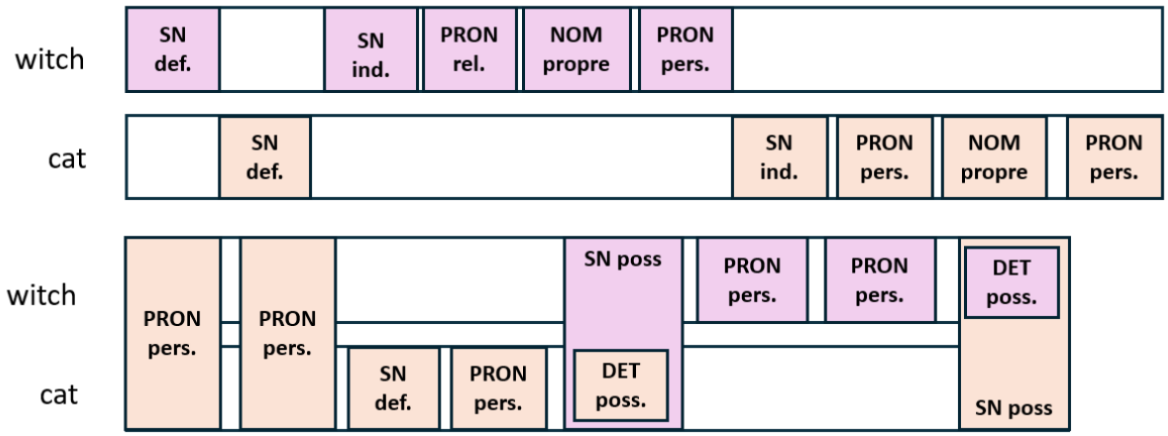


Figure 77. Schéma illustrant les chaines du texte du corpus français, CE2, 1352.

Dans le texte (126), nous avons 12 mentions pour la chaine du *chat* et 14 pour la *sorcière*.

Sur le plan narratif, le texte présente une structuration relativement lâche, caractérisée par une succession de segments : l'introduction du *chat* abandonné ; la présentation stéréotypée de la *sorcière* (méchante avec les humains, gentille avec les chats) ; une première action conjointe (la sorcière ramène le chat chez elle) suivie d'un dialogue explicitant son objectif (tuer les humains) ; une reprise finale d'éléments déjà décrits dans le texte (la méchanceté envers les humains, l'abandon du chat, puis son adoption de la part de la sorcière). Cette organisation produit une impression de circularité. Cette progression partiellement redondante suggère une planification narrative encore en cours d'élaboration.

Sur le plan référentiel, le premier référent introduit est le *chat,* à travers un SN indéfini (« un chat abandonné »), conformément aux attendus de l'ouverture de chaine après la formule « il était une fois ». Toutefois, dès la phrase suivante, la narration se focalise sur la sorcière, qui devient rapidement le personnage saillant de l'histoire. Cette centralité se manifeste par la densité des mentions la concernant et par son rôle d'agent principal (ramener le chat, demander un poil, préparer la soupe, etc.). Le nom propre de la sorcière « Océane » n'est introduit que dans la seconde moitié du texte, alors que la chaine référentielle de la sorcière est déjà solidement installée, et sa description de méchante réitérée, contribuant ainsi à l'impression de circularité que l'on vient de décrire. La gestion référentielle repose ici majoritairement sur des syntagmes nominaux, les formes pronominales apparaissent ainsi surtout dans les dialogues directs. Ce texte se place ainsi du côté nominal du continuum des stratégies de gestion que nous avions évoqué, et il est caractérisé par une recherche de la non-ambiguïté de la part de son auteur ou autrice.

(126) Il était une fois **[un chat abandonné]**^{cat}. Un jour il y avait **[une sorcière très méchante]**^{witch} mais avec les chats **[elle]**^{witch} était trop gentille. Mais avec les humains non. **[Elle]**^{witch} rentra chez **[elle]**^{witch} avec **[le chat]**^{cat} et **[elle]**^{witch} demanda **[au chat]**^{cat} « veux-**[tu]**^{cat} **[me]**^{witch} donner un poil de **[toi]**^{cat} ? C'est pour les méchants humains » **[le chat]**^{cat} répond « bon ben allez, si **[tu]**^{witch} veux » et **[la sorcière]**^{witch} cracha **[son]**^{cat} poil pour le mettre dans la grande soupe pour tuer les humains. Et **[la sorcière]**^{witch} s'appelait **[Océane]**^{witch}, **[la sorcière]**^{witch} était très maléfique mais seulement qu'avec

les humains. Et il y avait **[le chat]**^{cat} que **[la sorcière]**^{witch} avait adopté, **[le chat]**^{cat} car un jour **[il]**^{cat} était abandonné et maintenant **[le chat]**^{cat} est à **[la sorcière]**^{witch} et **[la sorcière]**^{witch} était méchante avec **[le chat]**^{cat}. *(corpus français, CE2, 3012)*

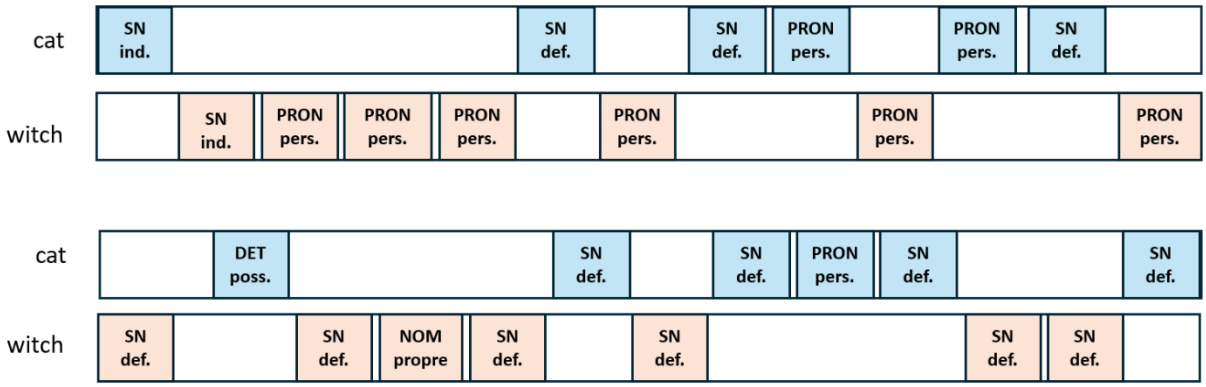


Figure 78. Schéma illustrant les chaines du texte du corpus français, CE2, 3012.

Autre extrême de la tendance observée précédemment, le texte (127) se caractérise par une présence quasi systématique de syntagmes nominaux définis pour assurer le maintien des référents. Ici la chaine du chat est composée de 22 mentions et celle de la sorcière par 14 mentions.

Sur le plan narratif, le texte présente une structure linéaire et chronologique, centrée sur une succession d'actions inscrites dans un cadre domestique immédiatement identifiable dès l'introduction des personnages (« le chat était sur son tapis pendant que la sorcière est en train de travailler »). Globalement, ce texte s'organise autour des descriptions d'épisodes domestiques plutôt que comme un récit proprement dit.

Du point de vue de la gestion référentielle, le texte présente une stratégie fortement nominale. Les chaines du *chat* et de la *sorcière* sont maintenues presque exclusivement par la répétition des syntagmes définis (« le chat », « la sorcière »), y compris dans des contextes où un maintien pronominal serait possible sans risque d'ambiguïté. Les formes pronominales sont rares et essentiellement limitées à des pronoms objets (« lui ») ou à une occurrence de pronom sujet pour la sorcière en fin de texte (« elle »). Cette répétition systématique produit un effet de sur-spécification référentielle.

Ce texte se situe ainsi à l'extrémité « nominale » du continuum décrit précédemment : la gestion référentielle ne s'appuie pas prioritairement sur des formes atténuées, malgré la configuration favorable que constitue la différence de genre grammatical. Il constitue, de ce point de vue, un contrepoint intéressant au texte (124), où le maintien pronominal dominait dans un contexte de concurrence référentielle plus limité.

(127) Il était une fois **[un chat]**^{cat} et **[une sorcière]**^{witch}. **[Le chat]**^{cat} était sur **[son]**^{cat} tapis pendant que **[la sorcière]**^{witch} est en train de travailler, et d'un coup **[le chat]**^{cat} miaule, **[la sorcière]**^{witch} **[lui]**^{cat} donne de la pâté de chat. **[Le chat]**^{cat} miaule toujours donc **[la sorcière]**^{witch} met **[le chat]**^{cat} dehors. **[Le chat]**^{cat} miaule, **[saute]**^{cat} par la

fenêtre et d'un coup **[la sorcière]**witch crie et **[le chat]**cat miaule encore plus fort donc **[la sorcière]**witch refait rerentrer **[le chat]**cat et **[le chat]**cat ne miaule plus et **[le chat]**cat s'endort et quand **[le chat]**cat se réveille **[la sorcière]**witch était avec **[lui]**cat. Pour le petit-déjeuner **[le chat]**cat mange des croquettes et **[la sorcière]**witch mange un bol de lait et deux tartines de beurre et confiture. **[La sorcière]**witch et **[le chat]**cat sortent dehors. **[La sorcière]**witch donne la balle **[au chat]**cat et **[le chat]**cat fait la passe à **[la sorcière]**witch et ainsi de suite, **[la sorcière]**witch dit **[au chat]**cat « c'est l'heure de manger **[[mon]witch petit chat]**cat » **[elle]**witch **[lui]**cat a préparé des croquettes*. (corpus français, CE2, 1596)*

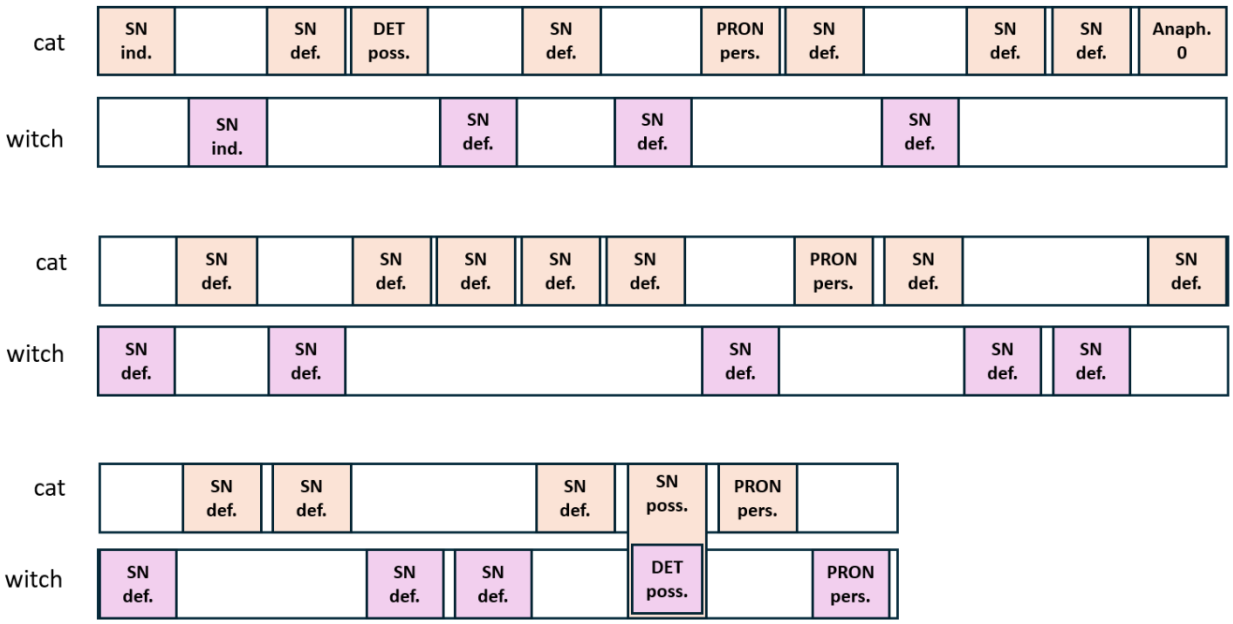


Figure 79. Schéma illustrant les chaines du texte du corpus français, CE2, 1596

Le texte (128) présente une structure narrative plus canonique et orientée vers un schéma conflictuel que les précédents. Ici la chaine du *chat* contient 13 mentions, celle de la *sorcière* 9.

La situation initiale installe un seul protagoniste (« un petit chat ») dans un cadre spatial relativement neutre et commun dans les fables (la forêt). L'événement déclencheur correspond à la rencontre avec la sorcière « très fâchée », qui l'introduit en tant qu'antagoniste. Leur rencontre constitue ainsi l'événement déclencheur, qui provoque une réponse de la sorcière, qui attaque le chat. La scène de combat entre les deux personnages, en plus de représenter l'état émotionnel du chat (« le chat était énervé »), contient aussi du lexique typique du registre familier destiné aux enfants (« faire bobo »). La narration trouve sa résolution ici dans la mort de la sorcière.

Sur le plan référentiel, la chaine du *chat* est installée seule au début du texte et bénéficie d'un maintien pronominal (« il »), rendu possible par l'absence initiale de concurrence. Après l'introduction de la *sorcière*, l'alternance entre les chaines s'intensifie, notamment dans la scène de combat. On observe alors un retour aux syntagmes nominaux définis (« le chat », « la sorcière

») afin de sécuriser les rattachements dans un contexte d'actions croisées. Les pronoms objets référés à la sorcière (« lui », « la ») sont néanmoins employés sans générer d'ambiguïté.

Comparé aux textes (124) et (127), ce prochain texte occupe une position intermédiaire sur le continuum des stratégies. Le chat, qui semble constituer l'héro positif de l'histoire, est susceptible d'être repris à l'aide des pronoms personnel sujet, alors que la sorcière est indiquée à travers des syntagmes nominaux, des pronoms personnel objet et un seul pronom personnel sujet, qui réinstancie immédiatement un syntagme nominal défini en fin de texte, dans une reprise redondante (« la sorcière elle est morte »).

(128) **[Un petit chat]**^{cat} se promène dans la forêt et tout à coup **[le chat]**^{cat} trouve une maison au loin, et **[le chat]**^{cat} s'approche doucement vers la maison. Quand **[il]**^{cat} est en face **[il]**^{cat} a peur et <u>après</u> un bon moment **[il]**^{cat} est rentré. Une fois dedans **[il]**^{cat} rencontre **[une sorcière]**^{witch} **[qui]**^{witch} est très fâchée, quand **[il]**^{cat} est rentré **[la sorcière]**^{witch} a fermé la porte, et **[la sorcière]**^{witch} jetait des sorts mais **[le chat]**^{cat} esquivait tous les sorts et un long moment **[le chat]**^{cat} était énervé alors **[le chat]**^{cat} **[lui]**^{witch} saute à la gorge, **[il]**^{cat} **[la]**^{witch} mord et un long moment **[le chat]**^{cat} continue à faire bobo à **[la sorcière]**^{witch} et encore et un long moment **[la sorcière]**^{witch} **[elle]**^{witch} est morte. *(corpus français, CE2, 2893)*

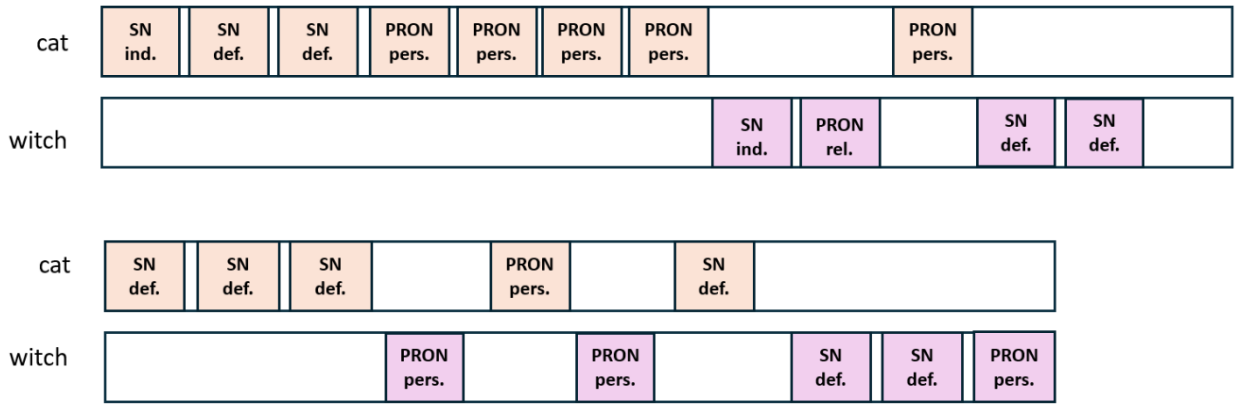


Figure 80. Schéma illustrant les chaines du texte du corpus français, CE2, 2893.

Le texte (129) présente une organisation narrative relativement développée par rapport aux textes analysés précédemment. La situation initiale installe un protagoniste unique (« un chat ») dans un cadre extérieur (la forêt), animé par une motivation simple, à savoir chercher à manger. Après un événement apparemment incohérent avec le reste de la narration (« il trouva une balle devant une maison », balle qui ne sera plus nommée dans le texte), qui sert à introduire la présence d'une maison dans la narration, nous retrouvons une réponse interne à la découverte de cette maison (« il voulait voir ce qu'il y avait dans cette maison vieille »). Cette réponse interne met en mouvement une série de planifications et tentatives de rentrer dans le bâtiment. L'accès à la maison introduit ainsi la présence de la *sorcière* et sa description, correspondant à plusieurs stéréotypes typiques des sorcières « une sorcière avec un balai magique, un chapeau de sorcière, une robe noire » et qui a « une voix pointue »). Après un échange entre les deux personnages sous

forme de discours direct, la sorcière transforme le chat en crapaud. Cependant, la résolution finale n'est pas conflictuelle car « les deux sont amis ».

Du point de vue référentiel, la chaine du *chat* est largement dominante dans la première moitié du texte et elle est maintenue de manière principalement pronominale (« il »). À partir de l'introduction de la *sorcière*, les deux chaines s'alternent, notamment dans le dialogue. On observe alors une combinaison de reprises nominales (« la sorcière ») et pronominales (« elle », « le », « lui »), avec les reprises pronominales toujours majoritaires.

Comparé aux textes précédents, ce texte occupe une position intermédiaire sur le continuum des stratégies. La première partie illustre une gestion largement pronominale, facilitée par la focalisation sur un seul référent. La seconde partie, marquée par la cohabitation de deux personnages et par une forte densité de dialogue direct, combine maintien pronominal et réactivation nominale, sans tomber dans la répétition systématique observée dans le texte (127) et en évitant la sur ou sous-spécification lors des reprises.

(129) **{[Le chat]**^{cat} et **[la sorcière]**^{witch}}

Il était une fois **[un chat]**^{cat} **[qui]**^{cat} était dans la forêt, **[il]**^{cat} cherchait à manger : des souris, des crapauds. **[Il]**^{cat} trouva une balle devant une maison, **[il]**^{cat} voulait voir ce qu'il y avait dans cette maison vieille. **[Le petit chat]**^{cat} essayait de pousser la porte. Mais **[il]**^{cat} était trop petit. **[Il]**^{cat} essaya de trouver quelque chose pour aller dans la maison. **[Il]**^{cat} trouva une balle de tennis et une corde épaisse et longue. **[Il]**^{cat} prit la corde et la **[mit]**^{cat} dans **[sa]**^{cat} bouche, puis **[il]**^{cat} l'accrocha à la poignée de la porte et **[tira]**^{cat}. La porte s'ouvrit et **[il]**^{cat} voit **[une sorcière]**^{witch} avec un balai magique, un chapeau de sorcière, une robe noire. **[Il]**^{cat} entra dans la maison et **[la sorcière]**^{witch} **[le]**^{cat} voit, **[elle]**^{witch} **[lui]**^{cat} a dit avec une voix pointue « pourquoi **[tu]**^{cat} es venu dans **[ma]**^{witch} maison **[petit rat]**^{cat} ? » « Mais non **[je]**^{cat} ne suis pas un rat, **[je]**^{cat} suis **[un chat]**^{cat}. » « **[je]**^{witch} vais **[te]**^{cat} transformer en crapaud. » **[Elle]**^{witch} prononçait **[sa]**^{witch} formule et **[le]**^{cat} **[transforma]**^{witch} en crapaud et **[les deux]**^{cat,witch} sont amis. *(corpus français, CE2, 1866)*

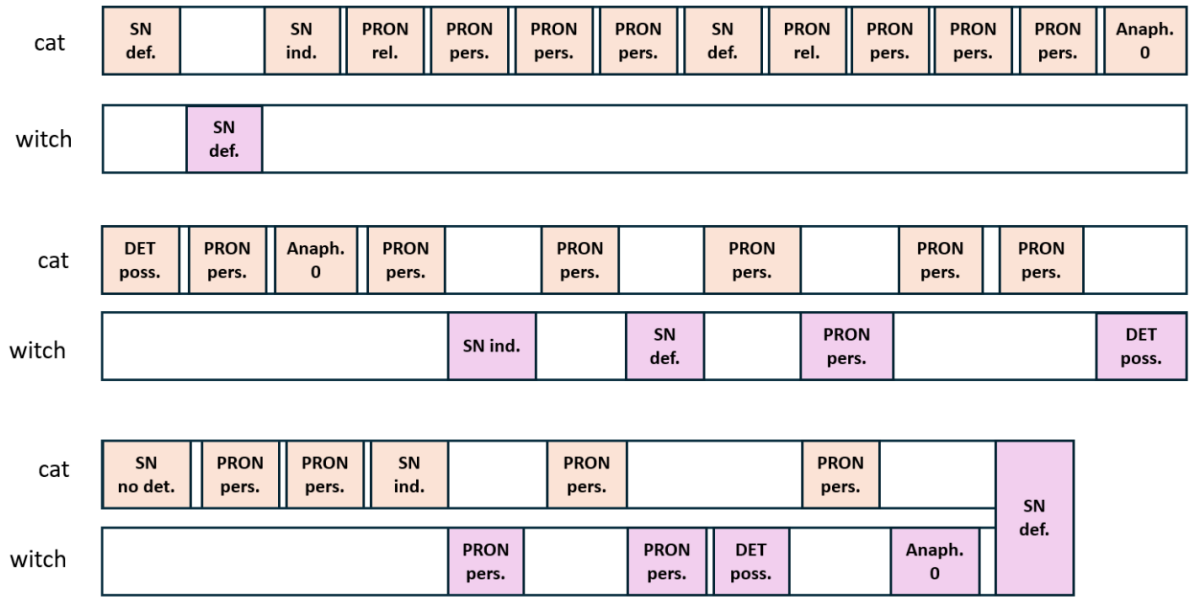


Figure 81. Schéma illustrant les chaines du texte du corpus français, CE2, 1866

Corpus italien : textes avec deux référents annotés

Dans le corpus italien, 53 textes comportent le binôme sorcière/chat, dont 20 mettent en scène exclusivement ces deux référents. Nous en avons sélectionné trois pour notre analyse qualitative.

Dans le texte (130), la chaine du *chat* comporte 9 mentions et celle de la *sorcière* 10. Sur le plan narratif, l'action ici s'organise autour de la métamorphose du chat, qui est l'événement principal du récit. Malgré son changement de chat à humain, explicité dans le texte, sa désignation reste stable en tant que « chat Thomas », ce qui garantit la continuité référentielle tout en présence d'une transformation. Une brève description physique de la sorcière est présente en début de texte (« aveva i capelli marroni, gli occhi azzurri », [elle] avait les cheveux bruns, les yeux bleus). La narration établit un lien de causalité entre la transformation du chat et son état mental (il gatto Thomas era molto ma molto arrabbiato perché non voleva essere un umano », le chat Thomas était vraiment mais vraiment vénère car il ne voulait pas être un humain).

Dans ce texte, le premier référent, la *sorcière*, est introduit de manière explicite à travers la formule SN indéfini + « qui s'appelait » + nom propre. Le deuxième référent, le *chat* Thomas (« il gatto Thomas »), ne fait pas l'objet d'une introduction explicite selon les attendus, sa première mention étant déjà sous forme de SN défini + nom propre. Si les mentions de la chaine de la *sorcière* semblent plus variées en termes de type, celle liées à la chaine du *chat* ne varient pas autant : nous retrouvons 4 occurrences du SN déf + nom propre dans les deux ordres possibles « il gatto Thomas » ou « Thomas il gatto ». Les anaphores zéro sont aussi présentes dans ce texte, 4 au total dont trois dans la chaine de la sorcière. On retrouve aussi un syntagme nominal démonstratif + nom propre qui réinstancie la sorcière après son introduction « questa strega Elena ».

(130) C'era una volta **[una strega]**^{witch} **[che]**^{witch} si chiamava **[Elena]**^{witch}, **[aveva]**^{witch} i capelli marroni, gli occhi azzurri. Un giorno a **[questa strega Elena]**^{witch} **[le]**^{witch} viene una idea di trasformare **[il gatto Thomas]**^{cat} in un umano allora **[preparò]**^{witch} una

pozione magica che era fatta di rospi, uccelli magici e galline. Il giorno dopo **[il gatto Thomas]**[cat] diventò **[un umano]**[cat] e la pozione **[lo]**[cat] aveva fatto anche parlare, ma **[il gatto Thomas]**[cat] era molto ma molto arrabbiato perché **[non voleva essere]**[cat] un uomo e allora **[la strega Elena]**[witch] il giorno dopo preparò un'altra pozione magica che **[Thomas il gatto]**[cat] era tornato alla normalità ma **[la strega]**[witch] era molto felice perché **[aveva fatto diventare]**[witch] **[un gatto]**[cat] in **[un umano]**[cat] e vissero felici e contenti. *(corpus italien, CE1, 534)*

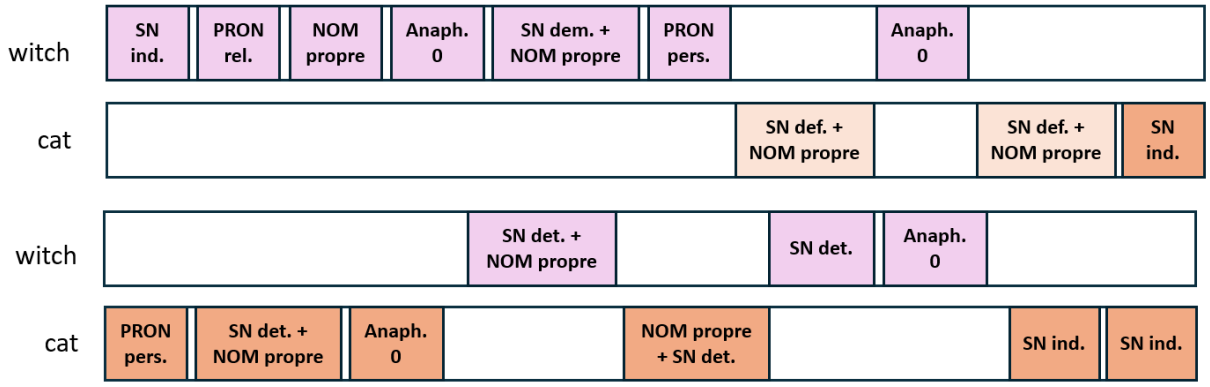


Figure 82. Schéma illustrant les chaines du texte du corpus français, CE1, 534.

Le texte (131) présente une configuration légèrement différente. La chaine du *chat* comporte 13 mentions et celle de la sorcière 16. Bien que le titre mentionne « la gatta misteriosa » (la chatte mystérieuse, au féminin), le texte introduit le référent par un SN masculin (« un gatto », un chat), et toutes les mentions ultérieures, nominales ou pronominales, sont également au masculin. Le trait morphosyntaxique de genre féminin annoncé dans le titre n'est donc pas repris dans le texte.

La structure narrative est relativement développée : introduction du chat, présentation de la sorcière (avec nom propre « Malefica »), séquence d'actions domestiques (la préparation d'une tarte aux myrtilles), rencontre avec un petit chaton (« gattino »), description détaillée des caractéristiques physiques et affectives), et l'adoption finale. La résolution est positive et s'inscrit dans un schéma narratif classique de rencontre et adoption du chaton.

Sur le plan référentiel, la gestion repose essentiellement sur des syntagmes nominaux définis ou des noms propres (« Malefica », « il gattino », « Pilù ») et sur des anaphores zéro pour la sorcière (« decise », « incominciò », « mise »). La chaine du chat présente une variation dénominative intéressante : « un gattino », puis attribution d'un nom propre (« Pilù »), qui devient ensuite la forme dominante. Cette variation marque une évolution statutaire du personnage (d'animal inconnu à individu nommé et adopté), tout en assurant la continuité référentielle. L'incohérence initiale entre le genre du titre et celui du texte souligne toutefois certaines limites dans la planification globale des chaines référentielles.

(131) {**[La strega]**witch e **[la gatta misteriosa]**cat}

C'era una volta **[un gatto]**cat **[che]**cat viveva nel bosco davanti a una casa abbandonata. Dove c'era dentro **[una strega di nome Malefica]**witch.

Un giorno **[Malefica]**witch decise di fare una torta ai mirtilli. Quindi **[decise]**witch di andare nel bosco vicino a casa sua a raccogliere un po' di frutta.

[Arrivati]witch alla radura, **[Malefica]**witch incominciò a raccogliere le bacche per fare la torta. Dopo un po' **[ritorna]**witch a casa e inizia a preparare la torta. Una volta finita la **[mise]**witch sopra il davanzale della finestra per farla raffreddare. Poi **[incominciò]**witch a pulire e **[sentì]**witch un rumore. **[Si girò]**witch e **[vide]**witch **[un gattino]**cat **[che]**cat stava mangiando la torta. **[Era]**cat piccolo. Con due occhi blu enormi. Il pelo nero e **[era]**cat dolce come un coniglietto.

[Malefica]witch **[lo]**cat prese e **[lo]**cat **[chiamo]**witch **[Pilù]**cat.

[Pilù]cat era freddo e piccolo quindi **[Malefica]**witch decise di adottar**[lo]**cat e da quel giorno **[Malefica]**witch si prenderà cura di **[Pilù]**cat. *(corpus italien, CE2, 847)*

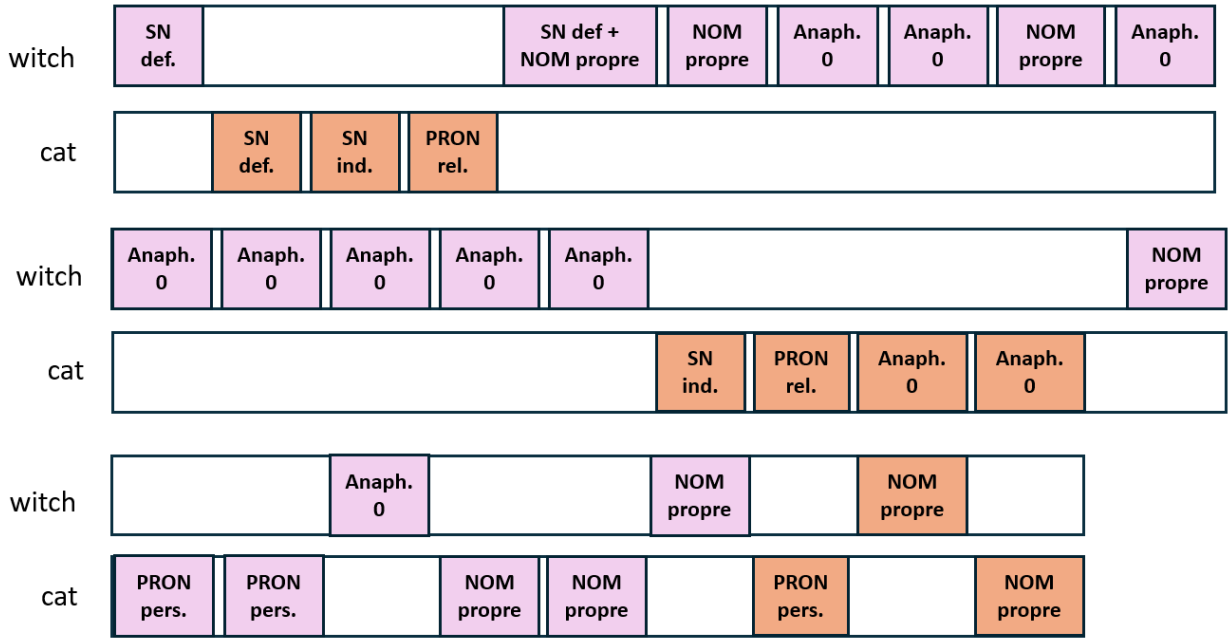


Figure 83. Schéma illustrant les chaines du texte du corpus français, CE2, 847.

Le texte (132) illustre une gestion combinant syntagmes définis et anaphores zéro. Les deux référents (« una strega » et « un gatto ») sont introduits conjointement dès l'incipit. La narration repose sur une alternance soutenue entre les chaines, notamment dans les séquences de dialogue direct.

On observe de nombreuses anaphores zéro pour la sorcière (« fece », « cerca cerca e cerca », « disse », « aveva sbagliato ») ainsi que pour le chat (« si nascose », « restò »).

La séquence de métamorphose subie par le chat (« lo fece diventare una persona ») introduit un nouveau statut pour ce personnage, qui devient humain (« la persona »). Sa nouvelle forme devient ensuite la forme référentielle dominante. Contrairement au texte (130), la transformation

entraîne ici une modification de la désignation lexicale, reflétant le changement ontologique du personnage. Néanmoins, la continuité référentielle est maintenue sans ambiguïté grâce à la cohérence du fil narratif.

(132) **{[La strega]**witch e **[il gatto]**cat}

C'erano una volta **[una strega]**witch e **[un gatto]**cat **[che]**cat,witch non andavano molto d'accordo insieme. Un giorno **[la strega]**witch fece una pozione da usare contro **[il gatto]**cat, ma **[il gatto]**cat era molto furbo e **[si nascose]**cat sotto il letto **[della strega]**witch, **[la strega]**witch cerca cerca e cerca ma **[non trovò]**witch niente. **[La strega]**witch disse "forse **[so]**witch dove si è nascosto **[il [mio]witch gatto]**cat, sotto il **[mio]**witch letto." **[Il gatto]**cat sentì le parole **[della strega]**witch e **[disse]**cat tra sé e sé "aiuto, aiuto dove **[mi posso nascondere]**cat? **[la strega]**witch sta arrivando!" **[il gatto]**cat andò nella **[sua]**cat cuccia a far finta di dormire. Però **[il gatto]**cat era molto ingenuo. E così **[restò]**cat lì, **[la strega]**witch nel passare vide **[il gatto]**cat e **[gli]**cat **[versò]**witch la pozione. Ma però **[la strega]**witch aveva sbagliato e **[lo]**cat **[fece diventare]**witch **[una persona]**cat. **[La persona]**cat si svegliò e **[disse]**cat "grazie **[strega]**witch per **[aver[mi]cat liberato]**witch dall'incantesimo" **[la strega]**witch disse "**[ho sbagliato]**witch a fare la pozione perché ci **[ho messo]**witch delle trecce di persone", **[la persona]**cat disse "fa niente, anzi **[mi]**cat **[hai fatto]**witch una sorpresa". **[La strega]**witch disse "**[facciamo pace]**cat,witch" e **[la persona]**cat disse "certo che **[facciamo pace]**cat,witch!" e vissero tutti felici e contenti. *(corpus italien, CE1, 536)*

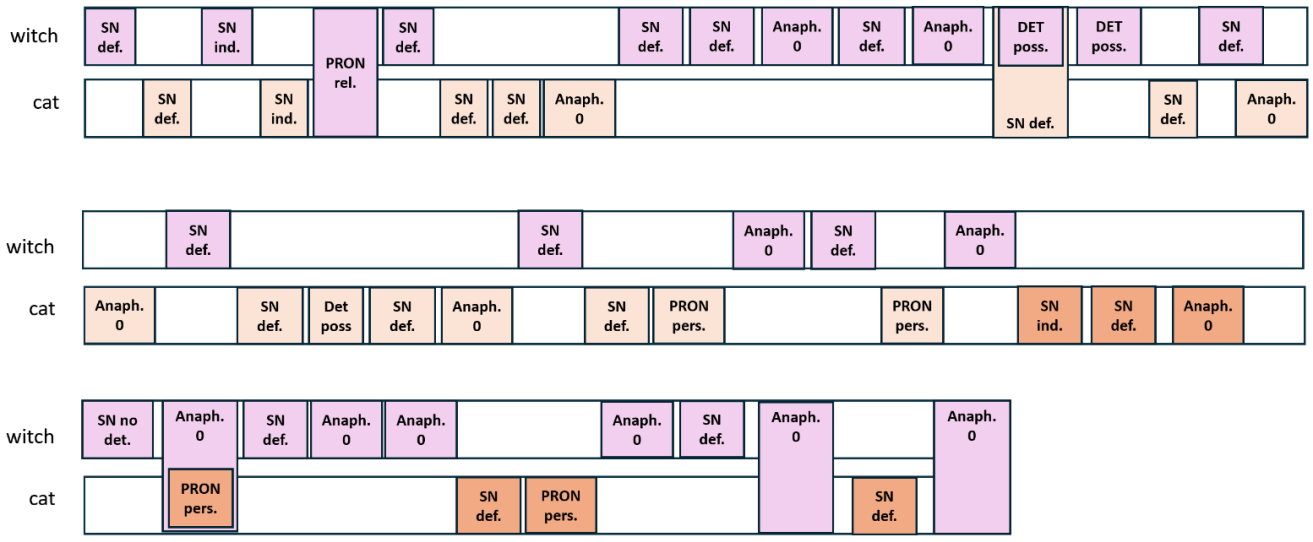


Figure 84. Schéma illustrant les chaines du texte du corpus français, CE2, 536.

En conclusion, dans l'ensemble des textes analysés dans cette section, la différence de genre grammatical entre les deux référents semble effectivement constituer un facteur facilitateur dans la gestion de la référence. Elle autorise, dans de nombreux cas, un maintien pronominal sans ambiguïté, notamment lorsque l'un des référents demeure temporairement dominant ou lorsque l'alternance entre les chaines reste localement contrôlée. Les textes les plus économiques exploitent pleinement cette ressource morphologique, en s'appuyant sur des formes peu

informatives (pronoms personnels en français, anaphores zéro en italien) dont le trait de genre suffit à contraindre l'interprétation, en plus de la sémantique du texte.

Toutefois, cette possibilité n'est pas systématiquement exploitée. Plusieurs productions se situent du côté nominal du continuum, privilégiant la répétition de syntagmes définis même en l'absence de risque réel d'ambiguïté. D'autres adoptent des stratégies intermédiaires, ajustant le degré d'informativité des mentions en fonction de la densité interactionnelle, de la complexité narrative ou de la saillance momentanée des personnages. Ainsi, la différence de genre n'entraîne pas automatiquement une gestion pronominale : elle ouvre un espace de possibilités que les élèves exploitent de manière variable.

Ces observations permettent de mieux circonscrire le rôle du marquage morphologique dans la désambiguïsation référentielle. Elles invitent également à s'interroger sur les stratégies mises en place lorsque cette ressource disparaît. La section suivante examine précisément cette configuration plus contraignante, dans laquelle les deux référents partagent le même genre grammatical, et où les indices morphologiques ne suffisent plus à distinguer les chaines concurrentes.

3.3. Textes contenant deux référents du même genre grammatical

Dans cette section, l'analyse qualitative de la construction des chaines référentielles permet de dégager les stratégies de gestion mises en œuvre lorsque l'élève choisit de faire coexister, au sein d'un même texte, deux personnages principaux de même genre grammatical. Ce type de configuration constitue un contexte particulièrement propice à l'émergence d'ambiguïtés de rattachement, et oblige les élèves à mobiliser des types de mentions plus rigides et plus informatives afin d'assurer la compréhension du discours, ou à se reposer davantage sur la sémantique des verbes utilisés dans la narration et sur la gestion de la saillance des référents.

Corpus français : textes avec deux référents annotés

Dans le texte (133), le *chat* semble représenter le personnage principal du texte (13 mentions contre les 4 du *loup*) à cause de la portée de sa chaine dans le texte, bien qu'il soit introduit comme deuxième référent. Après une phase initiale marquée par des syntagmes nominaux définis, le maintien du *chat* devient majoritairement pronominal dès que sa saillance est installée. La chaine du *loup-garou*, plus courte, est principalement maintenue par des SN définis. L'absence d'ambiguïté tient au fait que le chat soit le personnage principal du texte : les pronoms sont interprétables en fonction de la structure du récit.

Sur le plan narratif, ce texte contient plusieurs éléments caractéristiques, bien que synthétiques. L'introduction des personnages et la description du cadre sont simples. Ici la planification des actions du loup est présente dès la phrase d'introduction des deux personnages (« Il était une fois un loup-garou qui voulait attaquer un chat »), ce qui est lié à l'événement déclencheur (le loup-garou veut détruire la maison du chat). La réponse interne du chat est représentée à travers ses actions, « il se réveilla. Il sortit pour voir ce qui se passait ». La description de la tentative de résolution de l'histoire, le chat qui pose un piège, a comme conséquence directe la résolution de l'intrigue, le loup-garou qui tombe dans le piège. Le texte se conclut avec une réaction montrant

l'état émotionnel du chat, qui dit « enfin » avant de rentrer dans sa maison. Tout en étant simple, cette narration contient cependant plusieurs éléments narratifs. Les états mentaux sont représentés de manière indirecte, et les liens causaux restent cependant implicites.

(133) Il était une fois **[un loup-garou]**^{wolf} **[qui]**^{wolf} voulait attaquer **[un chat]**^{cat}. **[Le chat]**^{cat} vivait dans une maison. **[Le chat]**^{cat} était tout seul. Un jour **[le loup-garou]**^{wolf} essayait de casser la maison **[du chat]**^{cat}. Il faisait nuit et **[le chat]**^{cat} **[il]**^{cat} se réveilla. **[Il]**^{cat} sortit pour voir ce qui se passait. **[Il]**^{cat} ne voit rien, **[il]**^{cat} va se recoucher. Le lendemain matin **[il]**^{cat} posa un piège. Et un soir **[le loup-garou]**^{wolf} tombe dans le piège. Le lendemain matin **[le chat]**^{cat} dit « enfin » et **[il]**^{cat} va dans **[sa]**^{cat} maison. *(corpus français, CE1, 2983)*

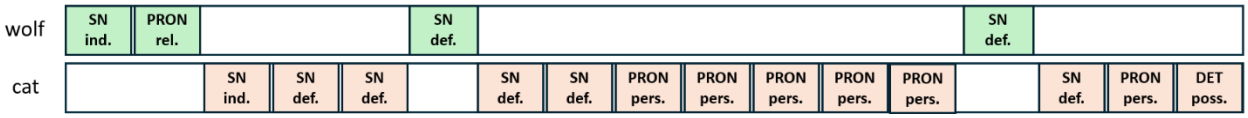


Figure 85. Schéma illustrant les chaines du texte du corpus français, CE1, 2983.

Nous retrouvons dans le texte (134) une gestion globalement comparable à celle observée dans le texte (133), c'est-à-dire une chaine où s'installe une gestion pronominale quasiment ininterrompue lorsque le référent est bien établi dans la narration. Il faut remarquer ici que les deux textes (133) et (134) appartiennent au même auteur ou autrice. Dans ce dernier, écrit en CE2, la gestion des référents semble cependant bien plus variée que dans le texte précédent, au niveau de la variété lexicale et au niveau des « patrons » des chaines employés.

Après l'introduction des deux référents, le *chat* et le *loup*, et la présentation de la situation initiale qui est conflictuelle, dans le déroulement de la situation centrale à la narration, la séquence d'actions effectuées par le *loup* donne lieu à une succession de pronoms personnels de troisième personne (« il »). Cette succession intervient après une mention de type nom propre, « Maléficore », très rigide et très informative, bien qu'indirectement liée au loup par exclusion, étant le seul autre nom propre du texte en plus de celui explicitement lié au chat, « Roqui ». En fin de texte, la refocalisation sur le personnage du *chat* est effectuée à l'aide d'un syntagme nominal défini, qui déclenche à nouveau la cohabitation des deux chaines à travers une stratégie de gestion variée comme celle présente dans le début du texte, enrichie davantage grâce à la présence d'un court échange entre les deux personnages sous forme de dialogue, contenant des pronoms déictiques et une anaphore zéro.

Sur le plan narratif, ce texte est très similaire au précédent, où le chat est la victime d'un loup, qu'ici embête le chat. Le développement narratif, riche en actions, est ici plus dynamique : les personnages sont caractérisés par des adjectifs, bien que stéréotypés (ce méchant loup ; ce pauvre animal). La situation de conflit présentée initialement provoque l'événement déclencheur et les plans successifs du chat qui « décida de mettre des pièges de partout ».

La tentative se déploie à travers la mise en place par le *chat* de pièges dignes d'un dessin animé (tapettes, trous recouverts de feuilles), suivie de l'arrivée du *loup*. La progression narrative repose sur une chaine d'actions rapides, majoritairement attribuées au *loup*, qui est à son tour maintenant

victime du *chat* : il marche sur la tapette à souris, il tombe dans le trou creusé par le *chat* et y reste coincé. La réaction émotionnelle du *loup* est ici représentée par l'onomatopée « aie aie aie ». L'épisode se clôt par une négociation dialoguée et une réconciliation entre les deux personnages, qui remplit la fonction de conséquence finale et de clôture de l'épisode.

Dans ce texte, l'utilisation des pronoms personnels est réussie : la série d'actions effectuées par le loup et reliées à sa chaine à travers une succession de pronoms personnels n'est pas référentiellement ambiguë, tout comme les déictiques présents dans le dialogue direct en fin de texte, dont les répliques sont sémantiquement cohérentes avec les actions réalisées dans le texte. Bien que la structure narrative ne soit pas très développée et soit plus similaire à la description d'un dessin animé, la cohérence référentielle est respectée ici.

(134) C'est l'histoire d'**[un chat]**^{cat} **[qui]**^{cat} se faisait toujours embêter par **[un loup]**^{wolf}. **[Ce méchant loup]**^{wolf} n'arrêtait pas d'aller chez **[le chat]**^{cat} et de mettre le bazar. **[Ce pauvre animal]**^{cat} *dor* mais un jour **[Rouqui le chat]**^{cat} décida de mettre des pièges de partout. Pour que **[Maléficore]**^{wolf} ne rentre plus, **[Rouqui]**^{cat} met des tapettes à souris. **[Il]**^{cat} mit des trous avec des feuilles dessus tout autour de la maison **[du chat]**^{cat}. Un soir **[Maléficore]**^{wolf} vient, **[il]**^{wolf} voit les feuilles donc **[il]**^{wolf} rentre vite dans la maison **[du chat]**^{cat}, **[il]**^{wolf} croit que c'était un piège les feuilles. **[Il]**^{wolf} avança et **[il]**^{wolf} marcha sur une tapette à souris, **[il]**^{wolf} cria « aie aie aie » **[il]**^{wolf} sort vite de la maison mais **[il]**^{wolf} tomba dans un trou et **[il]**^{wolf} resta coincé. **[Le chat]**^{cat} vient voir **[le loup]**^{wolf}, **[Maléficore]**^{wolf} dit « **[sors]**^{cat}**[-moi]**^{wolf} d'ici s'il **[te]**^{cat} plait ». **[Le chat]**^{cat} dit « oui mais **[tu]**^{wolf} arrêtes de **[m']**^{cat}embêter » **[le loup]**^{wolf} dit « oui ». **[Le chat]**^{cat} **[le]**^{wolf} sort et **[ils]**^{cat,wolf} deviennent amis. *(corpus français, CE2, 2983)*[98]

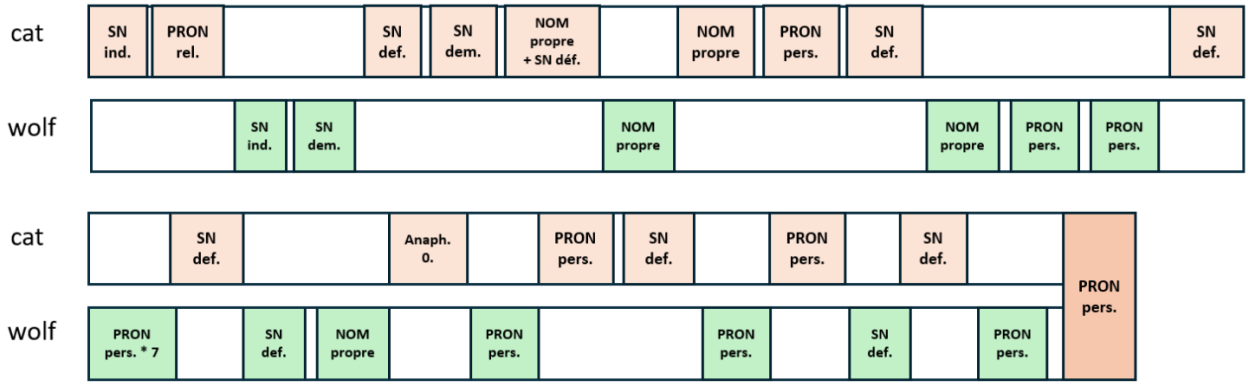


Figure 86. Schéma illustrant les chaines du texte du corpus français, CE2, 2983.

Une gestion différente est présente dans le texte (135) : les chaine du *chat* et du *loup* se caractérisent par une utilisation majoritaire de SN définis. Cependant, ce texte présentait plusieurs ambiguïtés référentielles, qui étaient déjà observées lors de l'étape d'annotation.

[98] Dans cette exemple le mot en vert, « dor » est fruit de la reconstruction des omissions effectuée automatiquement, dont la méthodologie est décrite dans le Chapitre 5, §1.1.

En plus du pronom relatif lié à la présentation du chat et d'un déterminant possessif lié au loup, on y relève pour chacune des deux chaines un pronom personnel et une anaphore zéro. Cette gestion entrecroisée apparait, à première vue, plus équilibrée en raison de la variété des types de mentions mobilisés. Elle donne toutefois lieu, en pratique, à des ambiguïtés de rattachement référentiel. Ainsi, le déterminant possessif « son » et la phrase « Il se retourne » (surlignés dans le texte) ont constitué deux points de difficulté lors de l'annotation, dans la mesure où il n'était pas possible de déterminer avec certitude à quel référent le déterminant et le pronom devraient être rattachés, à cause de l'absence de différenciation morphologique et de l'impossibilité de s'appuyer sur la sémantique des verbes pour désambiguïser. Le rattachement à la chaine du *chat* a été retenu pour « il se retourne », alors que « [son] chemin » a été rattaché au loup, étant plus logique qu'il utilise les cailloux pour distinguer son chemin depuis celui du chat. Cette interprétation demeure cependant incertaine.

Sur le plan narratif, cette production contient une présentation des deux personnages et de la situation initiale, où le chat se promène dans la forêt. Cependant, le reste des événements décrits est complexe à démêler et ne présente pas des liens de causalité directe. On peut déduire la présence d'un conflit qui se résout entre les deux personnages principaux mais qui traverse des actions qui ne sont pas directement liées entre elles (le chat griffe le loup, le loup tombe, le loup se déguise en chat). La résolution du conflit est aussi abrupte. La structure globale tente de correspondre aux requis de la narration mais l'absence de causalité entre événements limite fortement sa réussite.

(135) Il était une fois **[un chat]**^{cat} **[qui]**^{cat} se promenait dans la forêt mais soudain **[un loup]**^{wolf} a mangé des cailloux. C'est pour reconnaitre **[son]**^{wolf} chemin **[du chat]**^{cat}. **[Il]**^{cat} se retourne. **[Le chat]**^{cat} a griffé **[le loup]**^{wolf} et **[tombe]**^{wolf} par terre boum. Donc **[le chat]**^{cat} est perdu mais **[le loup]**^{wolf} a fait semblant de tomber alors **[il]**^{wolf} se déguise en **[chat]**^{wolf}. **[Le chat]**^{cat} se retourne et **[voit]**^{cat} **[un énorme chat]**^{wolf} donc **[le loup]**^{wolf} et **[le chat]**^{cat} sont devenus amis. Fin de l'histoire. *(corpus français, CE1, 950)*

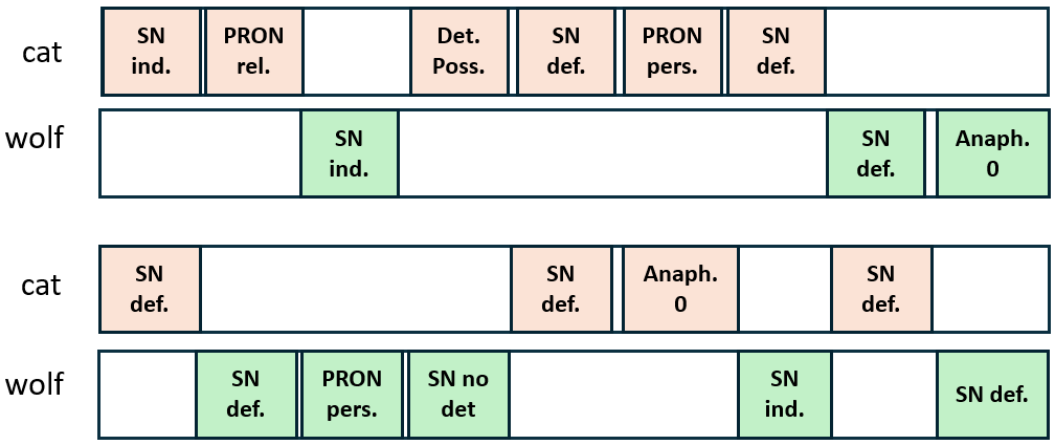


Figure 87. Schéma illustrant les chaines du texte du corpus français, CE1, 950.

Dans le texte (136), sensiblement plus court que les exemples précédemment illustrés, la distinction entre personnage principal et personnage secondaire apparait plus clairement, bien qu'au niveau numérique les deux chaines soient quasiment équivalentes (5 mentions dans la chaine du *chat*, 4 dans celle du *robot*).

L'introduction du chat suit le schéma prototypique d'ouverture narrative par un SN indéfini (« un chat persan »), immédiatement suivi d'un maintien pronominal (« il »). La chaine est ensuite principalement assurée par des pronoms personnels sujets, ce qui confère au référent une forte continuité discursive. Elle se clôt par un SN défini (« le chat ») en position d'objet direct dans la phrase finale. La chaine du robot est introduite de manière progressive : d'abord par le SN indéfini relationnel (« un maître »), puis par une reformulation (« un robot »). Le maintien ultérieur repose sur des SN définis (« le robot »), sans recours au pronom sujet.

Le seul point potentiellement ambigu concerne le pronom de la phrase « Il s'est trouvé bien chanceux ». D'un point de vue strictement syntaxique, le dernier sujet mentionné est « le robot », ce qui pourrait orienter l'interprétation vers ce référent. Toutefois, la cohérence sémantique et narrative favorise clairement le rattachement au chat : c'est lui qui vient de « trouver un maître » qualifié de « gentil ». L'interprétation repose donc ici sur la sémantique du texte et sur la logique du contexte. La saillance du chat tient surtout à la modalité de reprise, effectuée à travers le recours répété au pronom personnel. Le robot, en revanche, est majoritairement maintenu par des syntagmes nominaux définis, ce qui semble lui conférer une position de personnage secondaire dans le texte.

Sur le plan narratif, la production est minimaliste : la situation initiale et la description de la rencontre entre personnages n'a pas de lien de causalité avec l'abandon du chat en fin d'histoire, ce qui laisse l'impression d'un épisode incohérent plutôt qu'un texte narratif.

(136) Il était une fois **[un chat persan]**^{cat}. **[Il]**^{cat} se promenait le long du trottoir. Un jour **[il]**^{cat} trouve **[un maître]**^{robot}. C'était **[un robot]**^{robot}. **[Le robot]**^{robot} était gentil, très gentil. **[Il]**^{cat} s'est trouvé bien <u>*chanceux*</u>. Mais le lendemain matin **[le robot]**^{robot} abandonna **[le chat]**^{cat}. *(corpus français, CE1, 100)*[99]

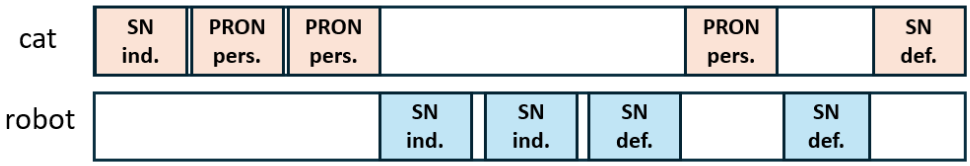


Figure 88. Schéma illustrant les chaines du texte du corpus français, CE1, 100.

[99] Le mot surligné ici, « chanceux » a été indiqué par les balises <unsure> lors de sa transcription, ce qui indique un doute du transcripteur dans la reconstruction du mot.

Corpus italien : textes avec deux référents annotés

En italien des stratégies de gestion légèrement différentes ont été observées. Dans le texte (137), nous retrouvons deux chaines, celle du *robot* et du *chat*, gérées majoritairement à l'aide des noms propres. Il faut noter ici que la troisième mention, « il Mattia », constitue une déviation de la norme en italien standard. Cependant, cette forme a été gardée lors de la normalisation car elle correspond à un usage dans l'oral spontané de certaines variétés d'italien du Nord, où le nom propre peut être précédé d'un déterminant défini. La chaine du robot contient 7 mentions, celle du chat 8, et 1 mention plurielle est aussi présente.

Le robot est introduit par un maillon de dénomination explicite (cf. ce chapitre, §1.2), « un robot chiamato Mattia » (un robot nommé Mattia). La deuxième mention est représentée par un SN défini, « il robot » (le robot), puis toutes les reprises successives sont des noms propres ou des anaphores zéro. La chaine du *chat* est également introduite par un SN indéfini, puis reprise à l'aide de SN définis ou de pronoms jusqu'à l'attribution d'un nom propre, qui se réalise au sein du texte (« ti chiamerò Silvestro », je vais t'appeler Silvestro). Une fois le chat nommé, le nom propre est utilisé pour le maintien du référent. Ici la gestion des chaines est effectuée à l'aide de formes très rigides et informatives comme les noms propres, et les cas d'utilisation des pronoms personnels ou des anaphores ne provoquent aucune ambiguïté référentielle.

Sur le plan narratif, comme dans plusieurs textes que nous avions analysés, la production porte sur la narration d'un épisode : le robot Mattia va dans la forêt, voit un chat sur un arbre, le fait descendre de l'arbre, il lui donne le nom Silvestro, ils rentrent à la maison ensemble, le robot donne du lait au chat avant de jouer ensemble. C'est un récit linéaire où on pourrait retrouver des éléments de la narration tels que la formule d'introduction initiale ou le discours direct, mais où ils manquent un réel événement déclencheur ou une résolution de l'intrigue.

(137) C'era una volta **[un robot chiamato Mattia]**^{robot} e **[il robot]**^{robot} va nel bosco e sull'albero c'era **[un gatto]**^{cat} e **[*il* Mattia]**^{robot} disse "**[vieni]**^{cat} giù" **[al gatto]**^{cat} e **[il gatto]**^{cat} veniva giù. **[Mattia]**^{robot} disse "**[ti]**^{cat} **[chiamerò]**^{robot} **[Silvestro]**^{cat}". **[Mattia]**^{robot} e **[Silvestro]**^{cat} vanno a casa. **[Mattia]**^{robot} dava il latte a **[Silvestro]**^{cat} e dopo **[giocavano]**^{cat,robot}. *(corpus italien, CE1, 325)*

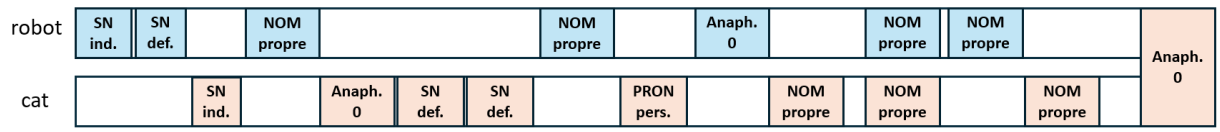


Figure 89. Schéma illustrant les chaines du texte du corpus italien, CE1, 325.

Dans le texte (138), nous retrouvons deux référents de genre féminin, une *chatte* et la *sorcière*. Ici les deux chaines ont un nombre similaire de mentions (6 le chat, 4 la sorcière) et ont une persistance similaire au sein du texte. Les deux personnages sont introduits à l'aide de mentions contenant des noms propres bien que dans deux configurations différentes, l'une constituée de nom propre + SN indéfini, l'autre SN indéfini + « di nome » + nom propre. La cohérence référentielle

ici est réussie car majoritairement gérée par des SN nominaux ou comprenant des noms propres, seulement deux formes pronominales et une anaphore zéro sont présentes ici.

Sur le plan narratif, tout comme le texte précédent, suite à l'introduction d'une situation initiale (la sorcière veut lancer un sort au chaton Valentina pour la rendre méchante), la réaction du chat est de tenter de s'échapper, et sa peur est explicitement formulée ici. La résolution de l'intrigue est ici partielle : après un an, la sorcière parvient à capturer la chatte et le récit s'interrompt sur une action de la sorcière, qui « correva di qua e di là », probablement une réaction émotionnelle implicite.

Bien que la cohérence référentielle soit bien gérée dans ce texte, sa structure narrative manque d'éléments.

(138) C'era una volta **[Valentina una gattina]**^{cat}. E un giorno **[una strega di nome Emma]**^{witch} **[gli]**^{cat} voleva fare un incantesimo per far**[la]**^{cat} diventare cattiva. **[La gattina poverina]**^{cat} correva sempre perché **[aveva paura]**^{cat} **[della strega Emma]**^{witch}. Dopo 1 anno **[la strega]**^{witch} è riuscita a catturare **[la gattina]**^{cat} proprio il 31 dicembre. Mentre i fuochi d'artificio suonavano **[la strega]**^{witch} correva di qua e di là. *(corpus italien, CE1, 1380)*

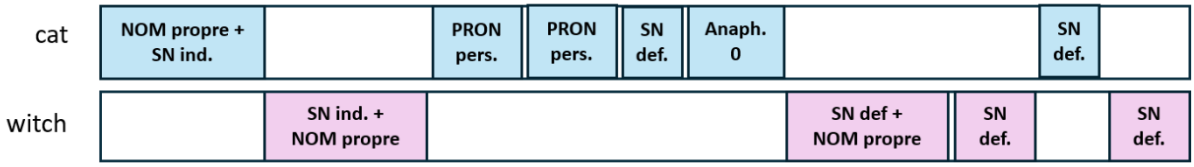


Figure 90. Schéma illustrant les chaines du texte du corpus italien, E1, 1380

Dans le texte (139), nous retrouvons également le binôme *chatte/sorcière*. La chaine de la *sorcière* comporte 13 mentions, celle du chat 18. Les deux référents sont introduits par des SN indéfinis, et la gestion référentielle est très variée ici, dans un texte qui se configure comme référentiellement dense. La première partie de l'histoire introduit les deux personnages, une chatte sauvage et une sorcière. L'événement déclencheur est la demande de la chatte à la sorcière de la « faire devenir magique », demande qui est satisfaite par la sorcière. Le « conflit » ici devrait être dans le fait que la chatte, après avoir quitté la maison de la sorcière, ne retrouve plus sa maison ni se rappelle d'être magique, avant d'être sauvé par la sorcière, sauvetage souligné par un échange sous forme de discours direct de « Grazie » et « Prego ! » entre les deux personnages. La résolution de l'histoire est ici aussi (de manière similaire à d'autres textes déjà analysés) l'adoption du chat de la part de la sorcière, ce qu'on retrouve aussi dans la mention « l'animale della strega ». Cette résolution est suivie par des descriptions de vie quotidienne, où les deux sont toujours ensemble et la sorcière donne à manger au chat. L'histoire se conclut ici avec la formule de clôture typique des narrations en italien, « e vissero tutti felici e contenti » (ils vécurent heureux).

Sur le plan langagier, nous retrouvons une légère incohérence dans l'utilisation de la conjonction adversative « ma » (mais) pour introduire la dernière phrase du texte. Cette utilisation nous semble ici influencée par l'emphase de l'oral.

Cette production, tout comme les précédentes analyses, est effectivement un bon exemple de gestion de la continuité référentielle. Au niveau narratif, elle semble plus achevée que les histoires précédemment analysées dans le cas de l'italien, et nous pourrons ici faire l'hypothèse que la « meilleure qualité » de ce texte pourrait aussi dériver d'une gestion apparemment plus variée de la continuité référentielle.

(139) C'era una volta **[una gatta selvaggia]**^{cat} e un giorno **[incontrò]**^{cat} **[una strega]**^{witch}. **[La gatta]**^{cat} **[gli]**^{witch} disse: «**[Mi]**^{cat} **[fai]**^{witch} diventare magica?» e **[la strega]**^{witch} accettò e **[l']**^{cat}**[ha fatta diventare]**^{witch} magica e **[la gatta]**^{cat} se ne andò. **[La gatta]**^{cat} dopo non trovò più casa **[sua]**^{cat}, **[non si ricordava]**^{cat} più che aveva la magia ma **[la strega]**^{witch} **[la]**^{cat} trovò e **[la]**^{cat} **[portò]**^{witch} a casa **[sua]**^{witch} e **[la gatta]**^{cat} disse: «Grazie!» e **[la strega]**^{witch} disse: «Prego!» e **[la gatta]**^{cat} diventò **[l'animale [della strega]**^{witch}**]**^{cat} e **[il gatto]**^{cat} andava sempre in giro con **[la [sua]**^{cat} **padrona]**^{witch} **[che]**^{witch} era **[la strega]**^{witch} ma un giorno **[la strega]**^{witch} preparò la pappa per **[la gatta]**^{cat} che era il tonno e **[la gatta]**^{cat} lo mangiò e vissero tutti felici e contenti... *(corpus italien, CE1, 28)*

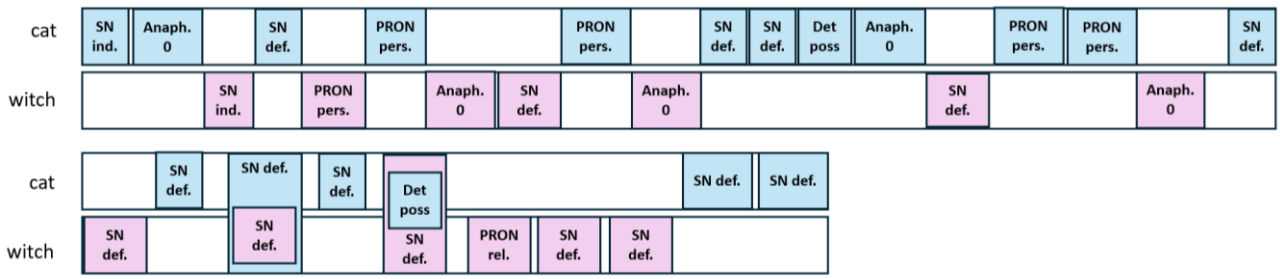


Figure 91. Schéma illustrant les chaines du texte du corpus italien, CE1, 28.

En conclusion, l'analyse conduite dans cette section avait pour objectif d'examiner dans quelle mesure le nombre de référents et leur genre grammatical influencent les stratégies de gestion des chaines référentielles dans les productions de notre corpus. L'hypothèse de départ postulait que la variation de ces paramètres structurels (un ou eux référents, du même genre ou de genres différents) devait se répercuter sur le degré d'informativité des mentions mobilisées et sur le risque d'ambiguïté référentielle.

Les textes mettant en scène un seul référent confirment clairement cette hypothèse. En l'absence de concurrence entre chaines saillantes, le maintien référentiel peut s'effectuer au moyen de formes peu informatives. En français, cela se traduit par un recours massif aux pronoms personnels ; en italien, par l'exploitation de l'anaphore zéro, rendue possible par la richesse morphologique du système verbal. Malgré la différence formelle entre les deux langues, la fonction discursive demeure comparable. Dans cette configuration, la gestion référentielle apparait généralement maîtrisée, même lorsque la structuration narrative reste peu élaborée.

Lorsque deux référents sont présents et relèvent de genres grammaticaux distincts, la situation se complexifie sans devenir pour autant instable. Le marquage morphologique du genre constitue alors un indice décisif de désambiguïsation dans les deux langues. Les textes analysés montrent

que cette ressource peut être pleinement exploitée : le maintien pronominal reste possible, dès lors que l'alternance entre chaines demeure localement contrôlée et qu'alternent des moments de succession et des moments d'entrecroisement. Toutefois, tous les scripteurs n'investissent pas cette possibilité de la même manière. Certains élèves privilégient des stratégies nominales plus sécurisantes, répétant les syntagmes définis même en l'absence de risque réel d'ambiguïté. On observe ainsi un continuum entre gestion fortement pronominale et gestion majoritairement nominale, la différence de genre grammatical ouvrant un espace de possibilités plutôt qu'imposant un mode unique de maintien.

La configuration la plus contraignante apparait lorsque deux référents partagent le même genre grammatical. Dans ce cas, le marquage morphologique des mentions peu informatives ne suffit plus à assurer la désambiguïsation. Les textes révèlent alors une diversité de stratégies compensatoires : recours accru aux syntagmes nominaux définis ou aux noms propres, gestion séquentielle des chaines pour éviter les « zones de risque », appui sur la cohérence sémantique des verbes, ou encore structuration narrative plus marquée permettant de hiérarchiser les rôles actantiels. Les cas d'ambiguïté observés, notamment lors de l'annotation, montrent que l'absence d'indices morphologiques explicites peut fragiliser le rattachement lorsque la planification narrative demeure lâche ou lorsque les actions ne sont pas clairement articulées causalement.

L'ensemble des analyses met ainsi en évidence l'interdépendance étroite entre cohésion référentielle et organisation narrative. Une structuration causale claire, une hiérarchisation des personnages ou une progression linéaire contribuent à stabiliser les chaines, même en contexte potentiellement ambigu. Inversement, lorsque la planification narrative est peu développée, la gestion référentielle tend à s'appuyer davantage sur des formes nominales explicites ou à générer des hésitations interprétatives. Enfin, la comparaison entre le français et l'italien suggère que les différences observées relèvent principalement de contraintes typologiques (pronominalisation explicite en français, utilisation de l'anaphore zéro en italien) plutôt que de divergences fonctionnelles.

Ainsi, le nombre de référents et leur genre grammatical apparaissent comme des variables qui pourraient être exploitées sur le plan didactique. Ces résultats ouvrent des perspectives didactiques, en particulier quant au rôle que peut jouer la variation contrôlée du nombre et du genre des personnages sur les stratégies référentielles déployées lors de l'apprentissage de la gestion de la cohésion textuelle.

4. Différences de longueur des textes scolaires : entre choix pédagogiques et contraintes linguistiques.

Dans nos analyses quantitatives, nous avons mis en évidence une différence remarquable entre la longueur moyenne des textes du corpus en langue italienne et celle des textes rédigés par les élèves français. Comme nous l'avons pu observer précédemment, les textes du corpus italien ont une longueur moyenne de 143,66 token et les productions du corpus français de 94,94 token. Ainsi, les productions d'élèves italiens présentent une longueur moyenne supérieure de 48,72 tokens à

celle des textes français (cf. Tableau 55). A parité de niveau scolaire, nous nous interrogeons sur les facteurs de variation qui entrent en jeu pour expliquer cette différence de longueur entre les textes des deux langues.

Niveaux scolaires	**Nombre de textes**		**Nombre moyen de tokens par texte**	
	fr	it	fr	it
CE1 – 2	74	75	71,59	117,49
CE2 – 3	74	75	118,28	169,83
CE1-CE2 – 2–3	**148**	**150**	**94,94**	**143,66**

Tableau 55. Caractérisation du corpus annoté - reprise d'une partie du Tableau 29.

Ayant partiellement exclu l'hypothèse d'une influence du taux de foisonnement susceptible d'expliquer cet écart, plusieurs hypothèses descriptives peuvent être avancées. Elles relèvent à la fois du style d'écriture utilisé à l'école dans chaque pays, et comment celui-ci est préconisé dans les programmes d'enseignement en vigueur, ainsi que des caractéristiques propres aux langues analysées, telles que l'orthographe ou leur degré de transparence.

La première de ces hypothèses porte ainsi sur le style d'écriture adoptée dans les productions de notre corpus. Alors que dans le corpus français le style d'écriture demeure uniforme en raison de l'adoption généralisée de la cursive à l'école en France, le corpus italien se caractérise par une plus grande variabilité des styles utilisés. Dans le corpus annoté italien, en effet, nous avons recensé 64 textes écrits en script majuscule (voir Figure 92), 16 en script (Figure 93) et 70 textes rédigés en cursive (Figure 94).

Cette variabilité observée dans le corpus italien, et la longueur majeure des textes que nous retrouvons dans cette même partie du corpus, nous interroge ainsi sur les orientations institutionnelles et pédagogiques qui encadrent l'enseignement de l'écriture dans les systèmes scolaires des deux pays et sur les retombées de ces orientations sur l'écriture des élèves. Nous abordons ici surtout le développement de l'écriture au sens du développement des aspects graphomoteurs de l'écriture, aspect que nous avons sélectionné pour l'analyse entre les différents facettes présentes dans la polysémie de la notion d'écriture (Brissaud et al., 2016).

Depuis les premiers travaux sur l'apprentissage de l'écriture chez les enfants, menés au début du 20ème siècle, il a été montré que les enfants écrivent plus rapidement et de manière plus lisible en script qu'en cursive. Ces études semblent également indiquer que le script soutient mieux l'apprentissage de l'écriture que l'écriture cursive (Bara et al., 2011 pour les recherches citées). Les différents types d'écriture demandent en effet un effort différent de la part de l'élève, tant au niveau des règles de production motrice que dans l'exécution des mouvements nécessaires à la formation des lettres (Bara et al., 2011, p. 42). L'écriture script, plus proche du dessin, apparait ainsi comme plus spontanément accessible aux jeunes enfants, tandis que l'écriture cursive favoriserait, à plus long terme, une plus grande rapidité et une meilleure fluidité du geste graphique. Par ailleurs, les performances orthographiques ne reposent pas uniquement sur les connaissances lexicales et orthographiques, mais dépendent également du niveau de maitrise graphique de

l'élève (Fayol & Miret, 2005, cités par Bara et al., 2011). Autrement dit, la gestualité acquise lors de l'apprentissage de l'écriture influence directement la performance orthographique de l'élève (Bara et al., 2011, p. 42).

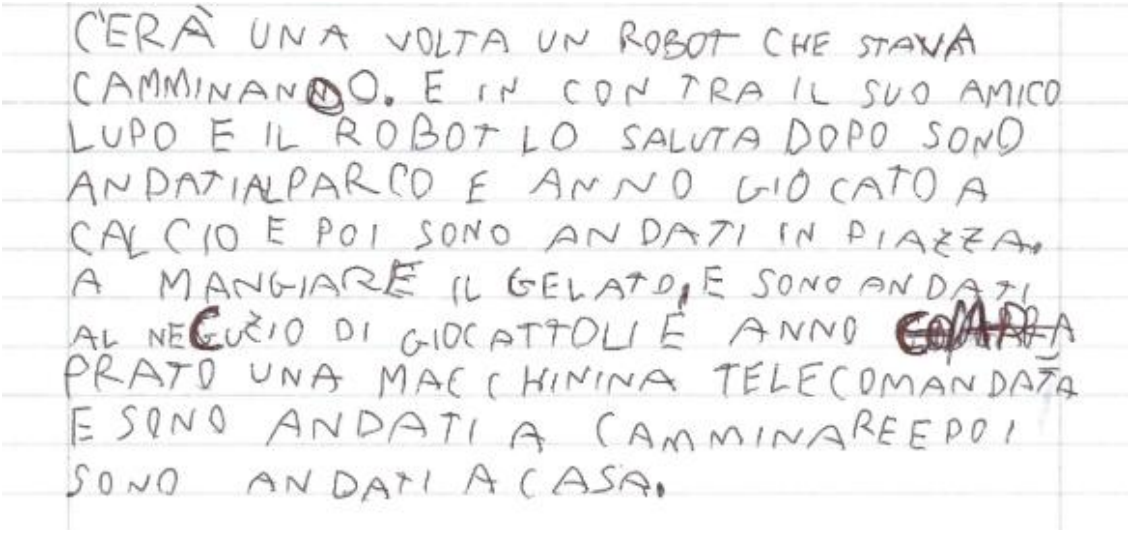

C'ERÀ UNA VOLTA UN ROBOT CHE STAVA CAMMINANDO. E IN CONTRA IL SUO AMICO LUPO E IL ROBOT LO SALUTA DOPO SONO ANDATI AL PARCO E ANNO GIOCATO A CALCIO E POI SONO ANDATI IN PIAZZA A MANGIARE IL GELATO, E SONO ANDATI AL NEGOZIO DI GIOCATTOLI E ANNO COMPRATO UNA MACCHININA TELECOMANDATA E SONO ANDATI A CAMMINARE E POI SONO ANDATI A CASA.

Figure 92. Texte du corpus italien (CE1, 821). Exemple d'écriture en script majuscule.

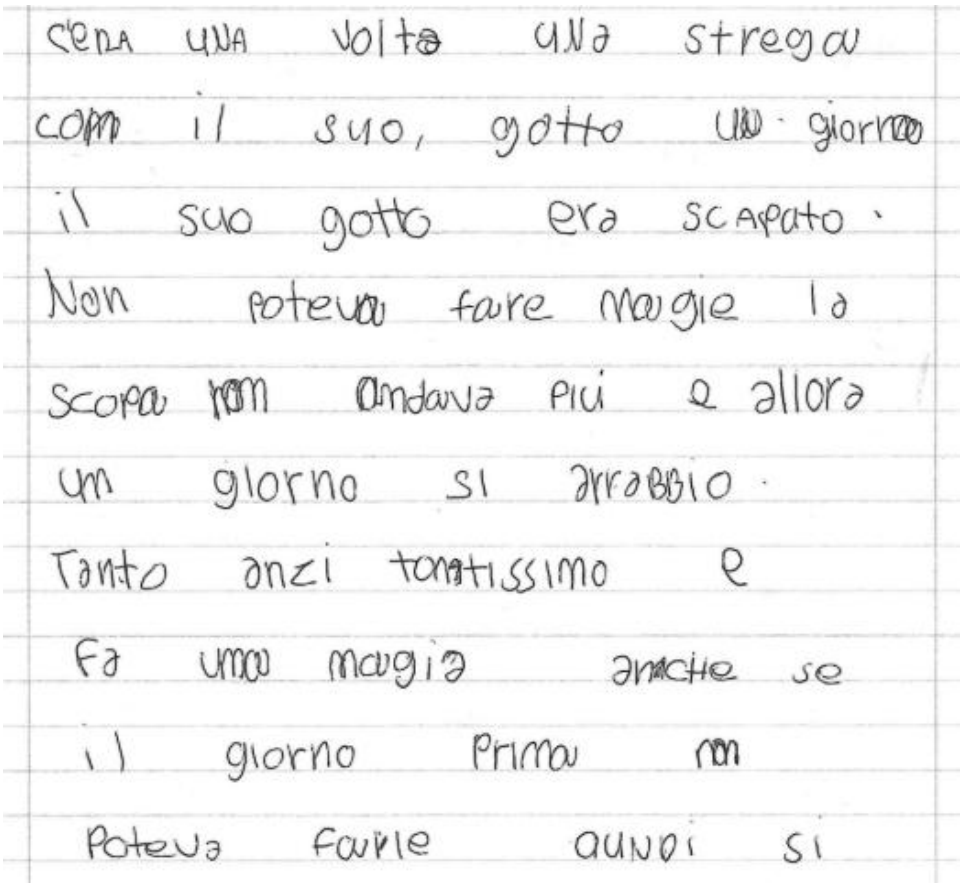
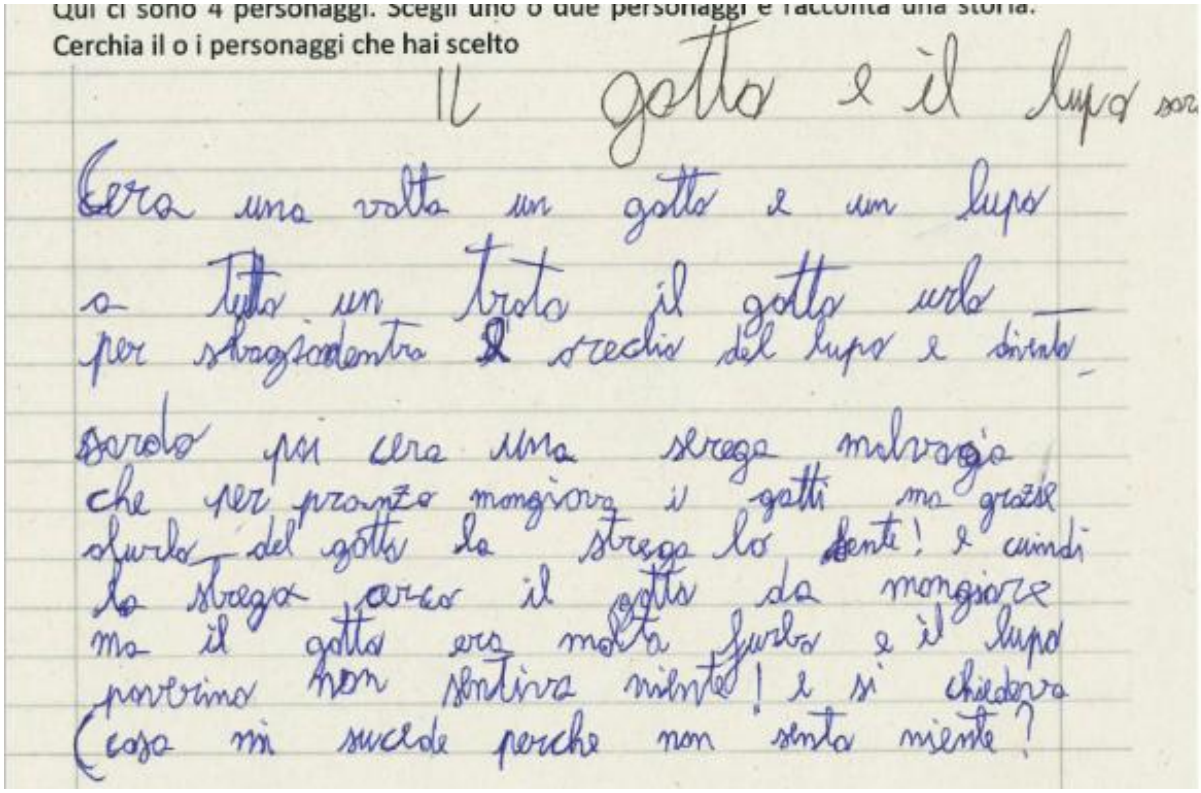

C'era una volta una strega con il suo gatto un giorno il suo gatto era scapato. Non poteva fare magie la scopa non andava più e allora un giorno si arrabbio. Tanto anzi tantissimo e fa una magia anche se il giorno prima non poteva farle quindi si

Figure 93. Extrait du texte du corpus italien (CE2, 761). Exemple d'écriture en script.


Qui ci sono 4 personaggi. Scegli uno o due personaggi e racconta una storia.
Cerchia il o i personaggi che hai scelto

Il gatto e il lupo

C'era una volta un gatto e un lupo
a tutto un tratto il gatto urla
per sbaglio dentro l'orecchio del lupo e diventa
sordo poi c'era una strega malvagia
che per pranzo mangiava i gatti ma grazie
sordo del gatto la strega lo sente! e quindi
la strega cerca il gatto da mangiare
ma il gatto era molto furbo e il lupo
poverino non sentiva niente! e si chiedeva
(cosa mi succede perche non sento niente?

Figure 94. Extrait du texte du corpus italien (CE2, 1423). Exemple d'écriture cursive.

La deuxième hypothèse porte sur les différences des préconisations quant au style d'écriture majoritairement enseigné ou à enseigner dans les deux systèmes scolaires. Le style d'écriture à adopter en primaire fait effectivement débat depuis plusieurs années et dans différents pays (Ediger, 2002). Dans certains pays, le script est l'écriture la plus enseignée à l'école (au Mexique, au Canada, aux Etats-Unis), alors que dans certains autres, comme la France, c'est la seule cursive qui est enseignée à l'école. Si, en Italie, l'utilisation de l'écriture cursive n'était pas systématiquement imposée à l'école lors de l'apprentissage de l'écriture (du moins, elle n'était pas explicitement indiquée en tant que telle), c'est en revanche le cas en France, où une attention particulière est portée à l'acquisition de l'écriture cursive dès les premières étapes de l'enseignement de l'écriture.

En France, les programmes scolaires et les documents d'accompagnement pour les enseignants mentionnent explicitement l'écriture cursive en tant qu'habileté qu'il faut maitriser et exécuter « avec une vitesse et une sureté croissantes » (Ministère de l'Education Nationale & Direction générale de l'enseignement scolaire, 2015) durant les cycles 2 et 3. L'objectif est double ici : consolider le geste graphique et ainsi maitriser les codes de l'écrit (Ministère de l'Education Nationale de la Jeunesse et des Sports, 2025). L'analyse des documents officiels produits entre 2015 et 2025 met en évidence une attention constante portée à l'écriture, envisagée d'abord comme action et habileté physique, avant d'être mobilisée comme activité intellectuelle. Dans le Bulletin de 2015, bien que l'usage d'outils numériques facilitant la rédaction et la relecture soit pris en compte, l'écriture manuscrite cursive reste considérée comme le point d'entrée dans l'apprentissage de l'écriture. Les exercices de copie, le travail de tracé des lettres, et la maitrise de la mise en page y sont particulièrement valorisés, en articulation étroite avec l'apprentissage

parallèle de la lecture. (Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche, 2015).

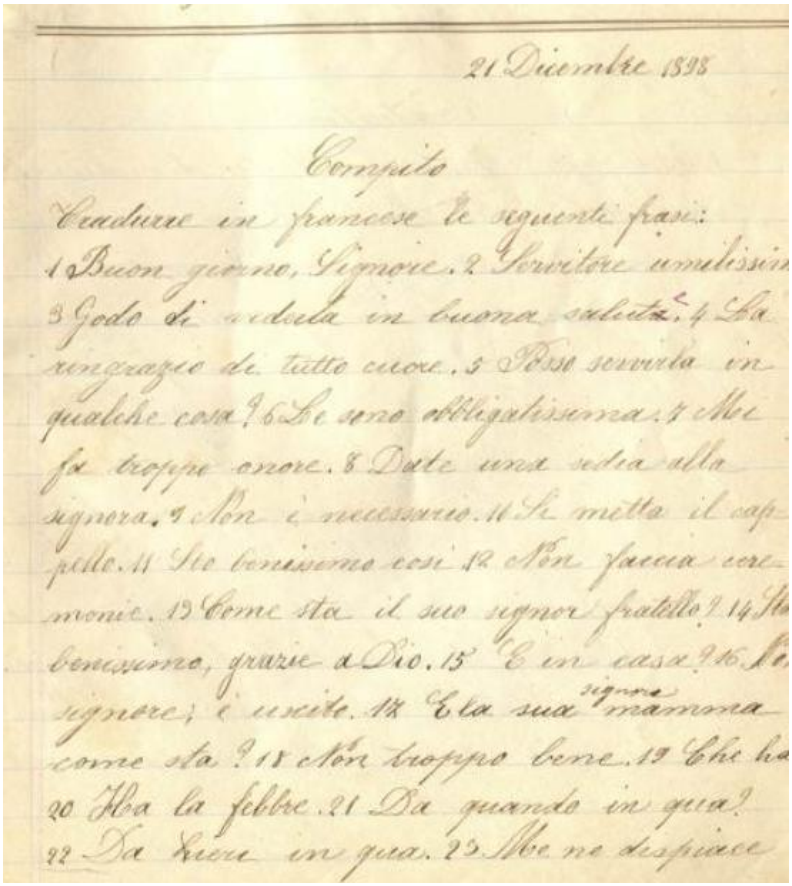
Les guides pour l'enseignant publiés à partir de 2019 approfondissent cette orientation et soulignent l'importance de l'automatisation du geste graphique. Le guide consacré à l'enseignement de la lecture et de l'écriture au CP rebondit sur la nécessité d'automatiser le geste d'écriture cursive pour « libérer les ressources de mémoire et d'attention nécessaires à la rédaction » (Ministère de l'Education Nationale, 2019). Pour le CE1, le guide publié la même année recommande des séances d'écriture centrées sur des tracés modèles proposés par l'enseignant, avec une attention particulière portée à la posture de l'élève, à la position du support d'écriture et à l'organisation de l'espace en classe (Ministère de l'Education nationale et de la Jeunesse, 2019).

Les programmes diffusés pour la rentrée 2025 clarifient encore davantage cette progression par niveaux. Un paragraphe est explicitement dédié à l'apprentissage de l'écriture cursive et à ses étapes et qui doit être débuté en moyenne section (Ministère de l'Education Nationale, 2025, p. 8) : au CP les lettres doivent être étudiées isolément puis en enchainement avant d'aborder l'apprentissage des majuscules au CE1, pour arriver à l'automatisation complète du geste graphique au CE2.

Alors que les programmes français contemporains affinent la progression de l'écriture cursive dès la maternelle, l'histoire de l'école italienne révèle une approche fondée sur d'autres enjeux et finalités. Dès l'introduction de l'obligation scolaire à la suite de l'Unité de l'Italie, les orientations institutionnelles préconisaient l'apprentissage et l'adoption d'une écriture lisible et rapide à partir des premières années de scolarisation. La priorité accordée par les autorités éducatives concernait davantage la lutte contre l'analphabétisme que l'enseignement de la calligraphie (Pani, 2012). Ainsi, dans les *Instructions et programmes pour l'enseignement de la langue italienne et de l'arithmétique dans les écoles primaires* (1867)[100], il est explicitement indiqué que les enseignements doivent viser l'apprentissage d'une écriture claire et utile dans la vie quotidienne, et que ce n'est pas l'objectif de l'école primaire de former des calligraphes (en italien, « *Non è ufficio delle scuole elementari il formare dei calligrafi*. »). L'écriture cursive était néanmoins adoptée de manière plus ou moins uniforme dans les écoles, comme nous pouvons encore l'observer dans des cahiers qui ont survécu à cette époque issus du corpus CoDiSSc (Revelli, 2024), et elle a continué à être enseignée tout au

[100] *Istruzioni e programmi per l'insegnamento della lingua italiana e dell'aritmetica nelle scuole elementari (R. D. del 10 ottobre 1867*) en italien. Le paragraphe sur les exercice d'écriture et de calligraphie en italien est le suivant : « Nelle due sezioni della prima classe i programmi propongono esercizi di scrittura, e non di calligrafia. La ragione di ciò non ha bisogno, pare, di spiegazione. Bensì non è forse inopportuno raccomandare che gli esercizi di scrittura nella prima classe sieno condotti in modo da preparare gli alunni agli esercizi di calligrafia prescritti per le classi seguenti; e gli esercizi di calligrafia sieno diretti in guisa da far acquistare agli alunni l'abitudine di una scrittura chiara e snella, che è di tanta utilità negli usi della vita, e non già da far disegnare laboriosamente lettere e parole con perdita di tempo prezioso. Non è ufficio delle scuole elementari il formare dei calligrafi. »

long de l'histoire scolaire italienne, attestant de la persistance de cette pratique au fil des décennies.


21 Dicembre 1898

Compito

Traduire en francese le seguenti frasi:
1 Buon giorno, Signore. 2 Servitore umilissim
3 Godo di vederla in buona salute. 4 La
ringrazio di tutto cuore. 5 Posso servirla in
qualche cosa? 6 Le sono obbligatissima. 7 Mi
fa troppo onore. 8 Date una sedia alla
signora. 9 Non è necessario. 10 Si metta il cap
pello. 11 Sto benissimo così 12 Non faccia cere
monie. 13 Come sta il suo signor fratello? 14 Si
benissimo, grazie a Dio. 15 È in casa? 16 No
signore; è uscito. 17 E la sua signora mamma
come sta? 18 Non troppo bene. 19 Che ha
20 Ha la febbre. 21 Da quando in qua?
22 Da ieri in qua. 23 Mi ne dispiace

Figure 95. Page d'un cahier contenu dans le corpus CoDiSSc daté entre 1879 et 1899. Cahier 0231 provenant de Vallée d'Aoste.

La présence de la cursive semble cependant rester implicite dans les programmes scolaires italiens analysés, publiés entre 2012 et 2025. Cette période constitue un repère pertinent pour l'analyse, les Indications nationales de 2012 (MIUR, 2012) étant les dernières directives détaillées disponibles avant la constitution de notre corpus. Les Indications de 2015, bien que publiées entre-temps, restent très succinctes et ne fournissent aucune précision concernant l'enseignement de l'écriture à l'école.

Si les Indications Nationales de 2015 ne mentionnent pas de manière spécifique l'enseignement de l'écriture à l'école, les *Nouvelles Indications Nationales* publiées en 2025 marquent toutefois un passage important dans l'imposition d'un style d'écriture au niveau national, qui peut être interprétée comme un retour à des orientations pédagogiques qui préconisent de manière explicite l'utilisation de la cursive. Ces indications semblent redonner à la calligraphie une place plus centrale dans les premières étapes d'apprentissage, notamment en tant que « facilitation du développement œil-main » et comme moyen d'éloigner les élèves des écrans (MIM, 2025).[101] Cette nouvelle orientation contraste avec les Indications de 2012, où chaque enseignant était libre

[101] « Nelle scuole del primo ciclo di istruzione la scrittura è fondamentale e va curata con particolare attenzione, a partire dall'apprendimento del corsivo e della calligrafia, perché agevola lo sviluppo della coordinazione oculomanuale, allontana i bambini dagli schermi e permette di tutelare gli spazi vitali dell'esperienza concreta, ingrediente necessario, specie nella scuola primaria, per affinare pensiero e ragionamento. » (Ministero dell'Istruzione e del Merito, 2025, p. 13)

d'introduire les pratiques d'écriture de manière graduelle et selon la méthode considérée la plus appropriée pour le groupe d'élèves, en veillant à développer la compétence d'écriture en même temps que celle de lecture (MIUR, 2012). En général, sur la base des observations empiriques conduites sur notre corpus et sur des corpus scolaires issus d'Italie (comme le corpus CoDiSSc), il apparait que l'utilisation de la cursive n'est pas imposée, et qu'au moins jusqu'en 2012, le choix de la méthode d'enseignement de l'écriture relevait largement de la liberté pédagogique de l'enseignant, à condition d'aborder en même temps les processus d'écriture et de lecture (Teruggi, 2019). Pourtant, la question reste ouverte sur la réelle nécessité des élèves de s'approprier si tôt de l'écriture cursive, compte tenu des évidentes difficultés que son acquisition comporte pour les jeunes enfants (Bara et al., 2011).

Comme déjà remarqué par la recherche *Lire et écrire au CP* (Goigoux, 2016), le temps dédié par semaine à la production écrite en CP (38 à 42 minutes) était toujours inférieur au temps consacré à la copie (46 à 50 minutes). Les enseignants centrent leur activité à ce niveau sur l'écriture du mot, en travaillant en particulier le codage orthographique. C'est seulement dans un deuxième temps que le travail sur la phrase, puis sur la textualité sont introduits (Doquet & Pilorgé, 2020). Seulement en fin d'année au CP les élèves devraient être capables de produire des courts textes (d'une à cinq lignes) toujours en lien avec l'apprentissage de la lecture. Au même temps, la norme assume une place centrale dans ce travail sur la langue : selon le Programme de français du cycle 2, les enseignants doivent indiquer « systématiquement la forme normée à chaque erreur de l'élève » dans la phrase d'écriture, ils doivent être « attentifs à l'orthographe dès les premiers écrits autonomes », donner des retours immédiats et « corriger ce qui n'est pas totalement automatisé » (Ministère de l'Education Nationale, 2025, p. 9). Cela semble confirmer cette attention continue placée sur la norme plutôt que sur le processus d'apprentissage en lui-même, qui devrait laisser de l'espace à l'erreur comme outil d'acquisition.

La dernière hypothèse concerne les différences dans l'orthographe des deux langues et son acquisition, et sur les contraintes qu'elles peuvent imposer aux apprenants. L'italien est une langue dont la restitution à l'écrit est largement transparente, caractérisée par une correspondance quasiment univoque entre graphème et phonème (Cossu et al., 1995, p. 11). Le français, à l'inverse, présente un système orthographique peu transparent, caractérisé par une correspondance entre graphèmes et phonèmes beaucoup plus riche et complexe. L'orthographe du français se caractérise aussi par « la forte homophonie des finales verbales et le caractère non audible des marques d'accord, contrairement aux autres langues romaines » (Garcia-Debanc et al., 2022, p. 7). La différence dans le degré de consistance (ou entropie)[102] entre les deux langues montre bien cette différence : 0,20 pour l'italien, 0,46 pour le français (Ziegler, 2018). Autrement dit, les élèves n'ont parfois aucun appui dans l'oralisation lorsqu'ils doivent effectuer des accords morphosyntaxiques ou choisir les marques verbales de personne adaptées au contexte. Cette complexité orthographique est susceptible de constituer un obstacle à la réalisation de textes plus longs et

[102] Nous définissons ici comme entropie la mesure de l'incertitude associée aux relations entre formes phonologiques et formes graphiques : plus ces relations sont univoques, plus l'entropie est faible ; à l'inverse, plus elles sont variables et ambiguës, plus l'entropie est élevée.

plus complexes, en comparaison avec des apprenants qui écrivent avec une écriture plus « simple ».

En conclusion, les différences de longueur observées entre les productions écrites des élèves français et italiens ne peuvent s'expliquer par un seul facteur. Elles semblent résulter de l'articulation entre les caractéristiques orthographiques propres aux deux langues et les pratiques d'enseignement de l'écriture. Les hypothèses avancées dans cette section, relatives aux styles graphiques, aux prescriptions institutionnelles et aux contraintes linguistiques, ne peuvent toutefois être ni confirmées ni infirmées à ce stade, notre analyse étant purement exploratoire et descriptive. Elles constituent néanmoins des pistes interprétatives que nous espérons voir explorées et approfondies par de futures recherches, notamment à partir du corpus *Scolinter*.

5. Conclusion

Dans ce chapitre, nous avons proposé une analyse qualitative des chaines de continuité référentielle observées dans le corpus annoté. Cette analyse avait pour objectif de compléter les résultats quantitatifs présentés dans le chapitre précédent en examinant plus précisément la manière dont les expressions référentielles sont mobilisées dans les textes narratifs produits par les élèves et comment elles sont organisées dans la linéarité du texte.

Comme nous l'avons observé à partir du nombre moyen de référents par texte (cf. Chapitre 7, §1.3), les textes du corpus s'organisent généralement autour d'un nombre limité de référents principaux, qui structurent la majorité des chaines de référence observées dans les récits. Cette organisation correspond à la structure narrative attendue dans les récits produits à partir de la consigne proposée aux élèves : un ou deux personnages principaux constituent le point d'ancrage du développement de l'histoire et se configurent comme les référents les plus saillants du texte. Cette observation permet également de confirmer la pertinence de la méthodologie d'annotation adoptée dans ce travail. En effet, le schéma d'annotation retient principalement les référents animés, qui correspondent précisément aux entités les plus saillantes dans ce type de narration. Ce choix permet de limiter l'annotation d'entités périphériques ou contextuelles et réduit ainsi la présence de singletons dans le corpus. Cette contrainte méthodologique contribue ainsi à limiter le bruit dans les données annotées, ce qui constitue un avantage dans la perspective d'une exploitation ultérieure du corpus, notamment pour des applications en Traitement Automatique des Langues, comme nous le discuterons dans le chapitre suivant.

L'analyse qualitative a également permis d'examiner plus précisément certaines configurations référentielles mises en évidence par les résultats quantitatifs. L'étude des mentions longues montre que ces expressions référentielles apparaissent rarement dans le corpus, mais qu'elles jouent un rôle particulier dans l'introduction et la caractérisation des personnages. Ces syntagmes nominaux plus développés apparaissent souvent dans les positions initiales des chaines et permettent d'associer au référent introduit certaines propriétés descriptives ou situationnelles, contribuant ainsi à la mise en place du cadre narratif. Ces observations présentent également un intérêt potentiel du point de vue didactique. Les mentions longues constituent en effet des configurations linguistiques dans lesquelles les élèves mobilisent différentes ressources descriptives (notamment

les expansions adjectivales) pour enrichir la présentation des personnages. L'identification de ces configurations pourrait ainsi servir de point d'appui pour des activités portant sur l'introduction des référents dans le récit et sur l'enrichissement des syntagmes nominaux dans l'écriture narrative.

À l'inverse, les positions suivantes dans la chaine sont plus fréquemment occupées par des expressions référentielles plus brèves, notamment des pronoms ou des syntagmes nominaux simples. L'alternance entre expressions référentielles plus développées et expressions plus économiques apparait ainsi comme un mécanisme central dans la gestion de la continuité référentielle dans les récits d'élèves.

L'analyse a également permis de mettre en évidence l'influence de certaines structures narratives sur l'organisation des chaines de référence. Dans de nombreux textes, les phases d'ouverture du récit présentent des configurations relativement stéréotypées, proches des structures narratives caractéristiques du conte, en particulier en présence des formules typiques d'ouverture des contes (*il était une fois, c'era una volta*). Dans ces contextes, les expressions référentielles sont fréquemment associées à des propositions relatives.

L'analyse des stratégies référentielles en fonction du nombre de personnages présents dans les textes met également en évidence l'influence de la configuration narrative sur l'organisation des chaines de référence. Dans les récits centrés sur un seul personnage, la gestion de la continuité référentielle apparait relativement simple : elle peut se développer de manière continue et les reprises sont souvent assurées par des formes peu informatives, notamment des pronoms en français ou des anaphores zéro en italien, sans que cela ne crée d'ambiguïté pour le lecteur car en absence de compétition entre référents. En revanche, lorsque deux personnages sont introduits dans un intervalle narratif restreint, la gestion des reprises devient plus complexe et les élèves mobilisent plus fréquemment des expressions nominales explicites afin de maintenir la distinction entre les référents. Cette tendance est particulièrement visible dans les configurations où les personnages partagent les mêmes propriétés morphosyntaxiques, par exemple lorsqu'ils sont du même genre grammatical, situation dans laquelle les pronoms (et en général, les mentions peu informatives et plutôt atténuées) peuvent devenir ambigus. Ces observations montrent que la gestion des chaines de référence dépend non seulement des propriétés linguistiques des expressions référentielles employées, mais également de la configuration narrative du texte et du nombre de référents qui doivent être maintenus simultanément dans le discours.

L'analyse qualitative a également permis de revenir sur la différence de longueur observée entre les textes des corpus français et italien, déjà mise en évidence dans les analyses quantitatives du chapitre précédent. En effet, les productions italiennes apparaissent en moyenne nettement plus longues que les textes français. L'examen du corpus suggère que cette différence ne peut être attribuée à un seul facteur, mais résulte probablement de l'articulation entre plusieurs dimensions. D'une part, les pratiques pédagogiques liées à l'enseignement de l'écriture diffèrent entre les deux systèmes scolaires : en France, l'apprentissage de l'écriture cursive et la maitrise du geste graphique occupent une place centrale dès les premières années de scolarisation, tandis qu'en Italie les pratiques observées semblent plus variées et moins strictement normées. D'autre part, les

caractéristiques linguistiques des deux langues peuvent également jouer un rôle. L'orthographe italienne, relativement transparente du point de vue des correspondances graphème-phonème, peut faciliter la production écrite et permettre aux élèves de produire des textes plus longs. À l'inverse, la complexité du système orthographique français, notamment en ce qui concerne les marques morphologiques non audibles et les accords grammaticaux, peut constituer une contrainte supplémentaire dans la production de textes écrits. Ces différentes hypothèses ne peuvent toutefois être confirmées à partir de nos seules données et doivent être considérées comme des pistes interprétatives qui mériteraient d'être explorées dans de futurs travaux.

Enfin, l'ensemble des analyses présentées dans ce chapitre permet de mettre en relation les tendances quantitatives observées dans le chapitre précédent avec les stratégies de gestion référentielle effectivement présentes dans les productions des élèves. L'articulation entre ces deux approches, quantitative et qualitative, permet ainsi de mieux comprendre les mécanismes linguistiques et discursifs qui sous-tendent la gestion de la continuité référentielle dans les récits d'élèves. À l'issue de ces résultats, il devient possible de dégager plusieurs implications relatives à la description linguistique des chaines de référence dans les écrits scolaires, ainsi qu'aux usages potentiels du corpus annoté. Ces résultats ouvrent également un ensemble de perspectives de recherche, tant du point de vue de l'exploitation du corpus en linguistique de corpus et en traitement automatique des langues que du point de vue de la didactique de l'écriture. Ces différentes implications et pistes de prolongement seront discutées plus en détail dans les conclusions générales de ce travail.

Chapitre 9. Conclusions

Ce travail de thèse avait pour objectif d'étudier la continuité référentielle dans un corpus de productions écrites d'élèves d'école primaire, à partir de l'annotation manuelle de ce phénomène et de son analyse quantitative et qualitative. Plus précisément, il visait à décrire la manière dont les élèves gèrent la présence et l'évolution des référents à travers l'analyse d'un corpus de textes narratifs, en examinant à travers quels types d'expressions référentielles ils introduisent, maintiennent ou réactivent les référents dans le discours, ainsi que la manière dont les chaines de référence s'organisent dans la linéarité du texte. Cette recherche s'inscrit ainsi à l'interface entre la linguistique, l'étude des corpus scolaires et le TAL. Elle propose à la fois un cadre théorique pour l'analyse de la continuité référentielle, une méthodologie d'annotation adaptée à des productions écrites d'élèves et une étude empirique et outillée fondée sur l'analyse d'un corpus annoté. L'objectif de ce travail est non seulement de contribuer à la description des formes et des mécanismes qui constituent la continuité référentielle dans les écrits scolaires, mais également de fournir des pistes de réflexion pour une didactique de la textualité et de l'écriture en langue première, qui prenne en compte l'enseignement explicite de la cohésion textuelle dès les premiers niveaux de scolarité.

L'ensemble du manuscrit s'organise autour de trois parties, correspondant aux différentes étapes de ce travail de recherche. La première partie est consacrée au cadre théorique et conceptuel nécessaire à l'étude de la référence dans les textes. Elle vise notamment à clarifier les notions centrales mobilisées dans ce travail, en particulier celles d'*anaphore*, *coréférence*, *chaine de référence* et *continuité référentielle*. La deuxième partie porte sur la constitution et l'annotation du corpus. Elle décrit la constitution et les caractéristiques du corpus d'écrits scolaires étudié, le corpus *Scolinter*, et présente la méthodologie adoptée pour l'annotation de la continuité référentielle, en proposant un protocole adapté aux spécificités des productions écrites d'élèves. Enfin, la troisième partie est consacrée à l'analyse empirique du corpus annoté. Elle propose une étude quantitative et qualitative des chaines de continuité référentielle produites par les élèves afin de mieux comprendre les mécanismes linguistiques mobilisés dans la gestion de la référence dans les années d'école primaire représentés dans notre corpus (niveaux 2 et 3, CE1 et CE2) dans des écoles françaises et italiennes.

Sur le plan théorique, l'adoption des modèles d'accessibilité et de continuité référentielle proposés par Ariel et Givón s'est révélée particulièrement pertinente pour la description des textes d'élèves du corpus étudié. En nous appuyant sur les travaux déjà menés sur le corpus *RésolCo*, nous avons pu confirmer l'applicabilité de ces approches à l'analyse de productions écrites d'apprenants. Toutefois, leur mise en œuvre dans le cadre de l'annotation d'un corpus scolaire fondé sur une consigne narrative spécifique, impliquant la présence potentielle de quatre personnages, à la différence de la consigne de *RésolCo* qui avait orienté certains choix d'annotation, a nécessité une formalisation méthodologique ultérieure de la notion de référent à annoter. Cette formalisation a également dû prendre en compte la dimension comparative du corpus, annoté dans deux langues aux propriétés morphosyntaxiques distinctes, le français et l'italien, tout en pensant à la

perspective de l'extension des annotations à une troisième langue. La conception et la stabilisation du guide d'annotation ont ainsi constitué une étape centrale de ce travail, permettant de traduire des catégories théoriques en procédures qui guident l'annotation, reproductibles et comparables entre deux langues différentes avec un potentiel d'application à une troisième langue.

Sur le plan méthodologique, ce travail a contribué à décrire les différentes étapes nécessaires à la constitution d'un corpus d'écrits scolaires annoté et comparable, dont le recueil s'est déroulé dans plusieurs pays. La réalisation d'un tel corpus implique en effet un ensemble d'opérations complexes, allant du recueil des productions manuscrites à leur numérisation, leur transcription et leur normalisation, jusqu'à leur annotation linguistique. Toutes ces étapes ont été menées selon des protocoles communs pour les trois langues du corpus, dont nous avons décrit les conventions, permettant ainsi de disposer de textes véritablement comparables. Dans ce contexte, le cœur de notre travail, l'annotation des chaines de référence, constitue une tâche particulièrement exigeante, qui nécessite de prendre en compte à la fois la diversité des formes linguistiques pouvant fonctionner comme expressions référentielles et les spécificités des productions d'élèves, souvent éloignées de la norme linguistique attendue. L'approche itérative adoptée pour l'élaboration du guide d'annotation et pour l'organisation des différentes étapes de rodage, de pré-campagne et de campagne d'annotation a permis de stabiliser progressivement les critères retenus et de garantir la cohérence des décisions prises par les annotateurs. Enfin, la phase d'adjudication a assuré la cohérence globale des annotations entre les deux langues ainsi qu'avec les règles d'annotation établies et stabilisées. Cette dernière étape a permis la constitution d'un *corpus gold* fiable.

Sur le plan empirique, l'analyse des chaines de référence dans le corpus annoté a permis de mettre en évidence plusieurs caractéristiques saillantes des stratégies référentielles mobilisées par les élèves dans leurs textes narratifs. La caractérisation du corpus, au-delà des annotations, a confirmé une différence déjà observée dans le corpus longitudinal concernant la longueur des textes entre le français et l'italien : les textes italiens apparaissent en moyenne plus longs que les textes français, avec toutefois une proximité entre les productions de CE2 en français et celles de CE1 en italien. Cette tendance se confirme globalement dans la distribution des mentions référentielles, plus nombreuses en italien, et dans l'indice de densité référentielle, légèrement plus élevé dans le cas de l'italien mais globalement similaire entre les deux langues.

Les résultats quantitatifs montrent tout d'abord que les récits produits par les élèves s'organisent autour d'un nombre relativement restreint de référents. Les textes du corpus comportent en moyenne entre trois et quatre référents, mais la narration se structure le plus souvent autour d'un ou deux personnages principaux. Ces référents constituent les entités les plus saillantes du récit et assurent la continuité narrative, tandis que d'autres entités apparaissent de manière plus ponctuelle et donnent lieu à des chaines plus courtes ou à des singletons. En général, les mentions courtes sont les plus fréquentes dans les textes, les mentions longues représentant une partie restreinte des annotations. L'analyse des expressions référentielles utilisées par les élèves met également en évidence certaines régularités dans l'organisation des chaines de référence. Les syntagmes nominaux, notamment indéfinis, jouent un rôle central dans l'introduction des

référents, tandis que les reprises ultérieures sont plus fréquemment assurées par des formes moins informatives, notamment des pronoms en français ou des anaphores zéro en italien. Cette distribution confirme en partie les prédictions des modèles théoriques de la continuité référentielle, selon lesquels le choix des expressions référentielles dépend du degré d'accessibilité du référent dans le discours, de sa distance ou de sa saillance dans le texte. Les analyses ont également montré que les mentions les plus longues apparaissent de manière relativement marginale dans le corpus, mais qu'elles remplissent souvent des fonctions spécifiques dans l'introduction et la caractérisation des référents. Les deux corpus présentent des principes organisationnels comparables dans la structuration des chaines de référence, notamment dans la similarité de patterns plus fréquents des trois premières mentions des chaines dans les deux langues.

Enfin, les analyses qualitatives ont permis d'identifier plusieurs difficultés spécifiques aux productions d'élèves, notamment certaines ambiguïtés dans l'interprétation des expressions référentielles ou des ruptures dans la continuité référentielle susceptibles de rendre le texte moins interprétable pour le lecteur. Ces phénomènes apparaissent en particulier dans les configurations où plusieurs personnages sont introduits dans un intervalle narratif restreint, ce qui peut entraîner une compétition référentielle entre plusieurs entités.

L'ensemble de ces analyses montre que les chaines de référence constituent un élément central dans la structuration des textes narratifs d'élèves. Elles ne se limitent pas à assurer la désignation des entités, mais participent pleinement à l'organisation du discours et à la cohérence textuelle. L'articulation entre les analyses quantitatives et qualitatives a permis de proposer une description à la fois globale et fine des mécanismes de la continuité référentielle, en reliant des tendances statistiques à des stratégies de construction de la référence et du discours narratif effectivement mises en œuvre dans les textes.

Notre travail met ainsi en évidence l'intérêt d'une approche fondée sur l'étude des chaines de référence pour analyser la cohérence et la cohésion textuelle dans les productions d'élèves. Nous avons ainsi pu fournir une description systématique des chaines annotées, tout en suivant les critères et en appliquant des méthodes d'analyse décrits dans la littérature, pour mettre en perspective notre description dans le panorama plus large des travaux existants en coréférence. L'exploitation d'outils issus du traitement automatique des langues a également permis d'automatiser ou de soutenir certaines analyses, notamment pour la catégorisation des types de mentions, et a contribué à orienter l'identification de phénomènes pertinents pour l'analyse qualitative.

L'annotation et l'analyse des chaines de référence constituent ainsi un observatoire privilégié pour étudier la manière dont les scripteurs en cours d'apprentissage organisent les entités dans leurs textes et développent progressivement des stratégies discursives plus complexes. Elles ouvrent enfin des perspectives à la fois pour la linguistique de corpus, pour le Traitement Automatique des Langues et pour la didactique de l'écriture, en permettant d'identifier des points d'appui pour l'enseignement explicite de la cohésion textuelle.

1. Quelles perspectives pour un travail multidisciplinaire ?

Le travail présenté dans cette thèse s'inscrit à l'intersection de trois champs scientifiques : la linguistique, la didactique de l'écriture en langue première et le Traitement Automatique des Langues. Cette étude exige en effet de mobiliser des cadres théoriques issus à la fois de la linguistique du texte, de la constitution et outillage de corpus et aux approches computationnelles de l'annotation et résolution de coréférence. Les résultats obtenus sont ensuite interprétés dans une perspective didactique qui vise au développement d'un enseignement de l'écriture intégrant de manière explicite la gestion référentielle.

La difficulté inhérente à un tel travail réside précisément dans sa dimension multidisciplinaire. L'analyse des productions écrites d'apprenants implique en effet de prendre en compte des phénomènes linguistiques et des contextes pédagogiques, ce qui relève de champs de recherche différents. La gestion de la référence dans les textes d'élèves peut ainsi être étudiée à partir de plusieurs perspectives complémentaires, notamment celles de la linguistique du texte, de la didactique de l'écriture, de la psycholinguistique ou encore de la recherche sur l'acquisition du langage. Dans le cadre de cette thèse, nous avons choisi d'aborder cette question principalement à partir d'une perspective de linguistique de corpus et d'annotation linguistique. Ce choix répond à un double objectif : d'une part, proposer une description systématique des phénomènes de continuité référentielle dans un corpus d'écrits scolaires ; d'autre part, construire une ressource annotée susceptible d'être mobilisée dans des recherches ultérieures. D'autres approches disciplinaires, notamment celles issues de l'acquisition du langage ou de la psycholinguistique, ont déjà abordé la question du développement de la cohérence textuelle et pourraient tirer parti de la ressource constituée dans ce travail pour approfondir ces questions sous d'autres angles d'interprétation.

Dans cette perspective, la ressource annotée produite dans le cadre de cette recherche représente une première étape vers la constitution d'un corpus multilingue permettant de suivre de manière longitudinale le développement des phénomènes de cohérence et de cohésion textuelle à l'école primaire et dans trois langues. À ce jour, cette étude exploite, à notre connaissance, la seule ressource présentant ces caractéristiques et permettant d'étudier le développement de l'écriture à l'école primaire dans trois langues romanes.

Cette ressource ouvre ainsi plusieurs perspectives de recherche qui s'inscrivent dans les trois domaines scientifiques mobilisés dans ce travail. Dans la suite de cette conclusion, nous présenterons d'abord les perspectives liées à la linguistique de corpus et au traitement automatique des langues, avant d'aborder les prolongements possibles de cette recherche du point de vue de la didactique de l'écriture.

2. Perspectives en linguistique de corpus et en TAL

Comme nous l'avons déjà souligné, l'annotation manuelle d'un corpus constitue une opération particulièrement coûteuse en temps et en ressources humaines, surtout lorsque l'objectif est la constitution des ressources linguistiques fiables et de grande taille. Dans le cadre de ce travail doctoral, dont la durée était limitée à trois années, nous avons fait le choix d'annoter un échantillon restreint des corpus français et italien. Bien que cet échantillon soit restreint en taille, il nous semblât suffisant pour tester nos hypothèses. Cette limitation constitue également l'un des points de départ des perspectives de recherche qui peuvent prolonger ce travail.

Une première perspective concerne l'extension de l'annotation à la troisième langue présente dans le corpus *Scolinter*, à savoir l'espagnol. Au moment de la rédaction de ce manuscrit, une nouvelle campagne d'annotation est en cours sur un échantillon de textes espagnols. Cette campagne s'appuiera sur l'expérience acquise lors des campagnes précédentes, en particulier sur le guide d'annotation désormais stabilisé. Une attention particulière sera portée au suivi des difficultés rencontrées par les annotateurs, notamment à travers la tenue d'un journal de bord partagé et la mise en place de dispositifs de communication permettant de documenter les hésitations d'annotation. L'objectif est ainsi de constituer un sous-corpus annoté en continuité référentielle, permettant d'élargir la comparaison à l'ensemble des langues romanes représentées dans le corpus longitudinal.

Une seconde perspective concerne l'augmentation de l'échantillon annoté. L'annotation d'un plus grand nombre de textes dans les trois langues permettrait de renforcer la robustesse des analyses quantitatives et d'explorer plus finement les variations observées entre niveaux scolaires. Dans cette perspective, il serait particulièrement intéressant d'étendre l'annotation manuelle et la réalisation d'un corpus *gold* à des niveaux scolaires plus avancés, notamment CM1 et CM2, afin d'examiner l'évolution des indices de continuité référentielle lors d'une étape plus avancée du développement de l'écriture.

La ressource annotée produite dans ce travail ouvre également des perspectives importantes dans le domaine du traitement automatique des langues. Les annotations manuelles réalisées constituent en effet un ensemble de données de référence susceptible d'être utilisé pour l'entraînement et l'évaluation de modèles de résolution de la coréférence. Cependant, il n'existe à notre connaissance aucun modèle de résolution de la coréférence adaptés aux écrits scolaires, qui présentent souvent des caractéristiques linguistiques spécifiques et des formes non normées. Cependant, les versions normalisées des textes pourraient être exploitées pour cette étape, dans le prolongement de la méthodologie d'annotation adoptée ici. L'hypothèse est que les modèles de résolution de la coréférence devraient être capables d'annoter nos textes normalisés, ayant été exposés au préalable à un échantillon annoté manuellement lors d'une phase de *fine tuning*. À terme, l'entraînement de modèles de résolution de la coréférence pourrait permettre d'annoter automatiquement une partie plus large du corpus *Scolinter*, facilitant ainsi l'analyse de volumes de données plus importants.

Dans une perspective comparative entre langues, plusieurs stratégies peuvent être envisagées pour le développement de ces outils. Une première consisterait à affiner des modèles existants pour la résolution de la coréférence, un modèle pour chacune des langues du corpus, afin de prendre en compte leurs spécificités linguistiques. Une seconde possibilité serait d'exploiter un modèle multilingue existant capable de traiter simultanément les différentes langues du corpus. Dans les deux cas, l'enjeu principal consistera à garantir la comparabilité des annotations produites automatiquement, afin que les différences observées entre corpus puissent être interprétées comme des phénomènes linguistiques plutôt que comme des biais liés au fonctionnement des modèles d'annotation. Le corpus *gold* pourra servir comme outil d'évaluation des annotations automatique et garantir le bon fonctionnement des modèles appliqués pour les objectifs de description linguistique poursuivis.

Enfin, les analyses qualitatives menées dans ce travail ouvrent également des perspectives pour de nouvelles recherches portant sur la relation entre stratégies de gestion référentielle et cohérence textuelle. Dans le cas des textes narratifs étudiés ici, souvent proches du genre du conte, la structuration des chaines de référence semble étroitement liée à l'organisation narrative des récits. Une perspective particulièrement prometteuse consisterait à annoter également certains éléments de la structure narrative dans les textes, afin d'examiner plus précisément les relations entre construction narrative et organisation des expressions référentielles.

Plus largement, la formalisation du guide d'annotation et l'explicitation des procédures méthodologiques développées dans ce travail offrent des garanties de reproductibilité et de transférabilité vers d'autres corpus d'écrits d'élèves narratifs. La ressource constituée peut ainsi servir de point d'appui à de futures recherches, tant dans le domaine de la linguistique que dans celui du TAL ou de la didactique de l'écriture, et nourrir des collaborations interdisciplinaires autour de l'étude empirique des écrits scolaires.

3. Perspectives en didactique

Comme nous l'avons vu précédemment, certains corpus scolaires annotés en continuité référentielle existaient déjà, mais les analyses réalisées à partir de ces ressources restent encore relativement limitées. Le travail mené dans le cadre du corpus *RésolCo* s'inscrit dans une perspective plus large d'étude des écrits scolaires ; toutefois, les résultats issus des annotations réalisées sur ce corpus ont donné lieu à un nombre relativement restreint de publications. Par ailleurs, le travail proposé par Perugini et Zucchini (2025) se situe encore à l'étape d'un test empirique visant à expérimenter un schéma d'annotation potentiel.

Si le corpus annoté dans le cadre de cette recherche ne permet pas encore une observation à très large échelle, la quantité de données actuellement disponibles ouvre néanmoins des perspectives intéressantes pour la poursuite de recherches dans une perspective didactique. Les analyses menées dans la troisième partie de ce travail ont en effet permis de mettre en évidence plusieurs régularités dans la manière dont les élèves gèrent les expressions référentielles dans leurs textes. Ces observations peuvent constituer un point d'appui pour réfléchir à des dispositifs pédagogiques visant à développer la cohésion et la cohérence des productions écrites, soit à travers l'observation

des annotations effectuées dans notre travail, soit à travers des activités de production et observation active des chaines de référence dans des textes narratifs produits par les élèves.

Une première piste de travail concerne ici la gestion simultanée de plusieurs référents dans un texte. Les analyses qualitatives montrent en effet que certaines difficultés rencontrées par les élèves apparaissent lorsque plusieurs personnages doivent être maintenus actifs dans le discours. Dans cette perspective, des activités de lecture ou d'observation de textes où deux référents du même genre grammatical contre l'observation de la gestion dans des cas des chaines de genre grammatical différent pourraient amener les élèves à identifier les différentes expressions référentielles associées aux personnages et à observer comment celles-ci se distribuent dans le texte pour maintenir la continuité référentielle selon les différentes configurations, et quelles stratégies mettre en place en particulier dans des cas où des zones d'ambiguïté pourraient surgir.

Une deuxième piste consiste à exploiter le rôle des différentes catégories d'expressions référentielles dans la structuration du texte. L'introduction d'un référent peut par exemple être réalisée à l'aide d'un syntagme nominal plus ou moins enrichi, parfois accompagné d'adjectifs ou d'autres expansions permettant de caractériser l'entité introduite. L'observation de ces configurations pourrait servir de point d'appui pour des activités visant à enrichir l'introduction des personnages dans les récits et à développer la capacité des élèves à préciser ou à caractériser les référents.

Une piste ultérieure concerne l'approfondissement de l'utilisation des expansions adjectivales au sein des chaines. Comme nous avons observé, celles-ci ne sont pas fréquentes aux niveaux scolaires analysés, tout en contribuant à la richesse lexicale et à la représentation du référent. Un travail didactique sur l'expansion adjectivale lors de l'introduction des référents pourrait véritablement supporter le développement d'un répertoire adjectival chez l'élève, fonctionnel à la narration et potentiellement à la différentiation entre référents lors des reprises ultérieures dans le texte.

Dans cette perspective, il pourrait également être intéressant de proposer aux élèves des activités de repérage des expressions référentielles dans des textes, sans mobiliser d'emblée une terminologie grammaticale complexe. L'objectif serait d'amener les élèves à observer la variation des expressions utilisées pour désigner une même entité au fil du texte et à comprendre comment ces variations contribuent à la cohésion et à la cohérence du discours. Ces observations pourraient ensuite servir de point d'appui pour introduire progressivement certaines notions grammaticales, comme la proposition relative, qui permet d'enrichir un syntagme nominal lors de l'introduction ou de la réactivation d'un référent, et qui semble déjà mobilisée aux niveaux scolaires analysés.

Plus largement, les résultats de cette recherche invitent à envisager la cohésion textuelle non seulement comme un objet d'analyse linguistique, mais également comme un objet d'enseignement pouvant être abordé à partir de l'observation de textes. Dans cette perspective, l'étude des expressions référentielles et des chaines de référence peut constituer un outil intéressant pour développer chez les élèves une réflexion sur l'organisation des textes et sur les moyens linguistiques permettant d'assurer leur interprétabilité.

Enfin, si cette recherche s'est concentrée principalement sur des textes narratifs, en raison de la nature des consignes utilisées dans le recueil du corpus, une autre perspective didactique consisterait à étendre ce type d'annotation et d'analyse à d'autres genres textuels. Le récit constitue certes un terrain privilégié pour observer la gestion des référents, mais le développement d'une compétence complète en littéracie implique également la capacité à lire et produire d'autres types de textes, notamment des textes explicatifs, descriptifs ou fonctionnels. L'étude de la gestion des expressions référentielles dans ces différents genres pourrait ainsi contribuer à mieux comprendre les stratégies discursives mobilisées par les élèves dans des situations d'écriture variées.

Dans cette perspective, les corpus scolaires annotés peuvent également constituer une ressource précieuse pour la formation des enseignants. L'observation de productions authentiques d'élèves permet en effet de mieux comprendre les stratégies linguistiques mobilisées par les scripteurs en cours d'apprentissage et d'identifier certaines difficultés récurrentes dans la gestion de la cohésion textuelle. Ces observations peuvent ainsi nourrir la réflexion didactique sur l'enseignement de l'écriture et sur les dispositifs pédagogiques susceptibles de soutenir le développement des compétences scripturales des élèves.

Références bibliographiques

Abeillé, A., Abeillé Anne linguiste, Godard Danièle, Delaveau Annie 19-, Gautier, A., & Gautier Antoine linguiste. (2021). *La grande grammaire du français*. Actes Sud Imprimerie nationale Éditions.

Achard-Bayle, G. (2016). Les référents évolutifs, objets et objets du discours. In L. Sarda, D. Vigier, & B. Combettes (Éds.), *Connexion et indexation : Ces liens qui tissent le texte*. ENS Éditions. https://doi.org/10.4000/books.enseditions.6875

Adam, J.-M. (1992). *Les textes : Types et prototypes: Récit, description, argumentation, explication et dialogue*. Armand Colin.

Adam, J.-M., & Petitjean, A. (1982). *Introduction au type descriptif*. https://doi.org/10.3406/prati.1982.1237

Alhamad, H. A., Shehab, M., Shambour, M. K. Y., Abu-Hashem, M. A., Abuthawabeh, A., Al-Aqrabi, H., Daoud, M. S., & Shannaq, F. B. (2024). Handwritten Recognition Techniques : A Comprehensive Review. *Symmetry*, *16*(6), Article 6. https://doi.org/10.3390/sym16060681

Alighieri, D. (1867). *La Divine Comédie. L'Enfer, traduction en vers français par M. François Villain Lami*. https://gallica.bnf.fr/ark:/12148/bpt6k6470857n

Antoine, J.-Y., Schang, E., Muzerelle, J., Lefeuvre, A., Pelletier, A., Désoyer, A., Landragin, F., Tellier, I., Villaneau, J., Eshkol, I., & Maurel, D. (2013). *Corpus ANCOR_Centre Présentation générale*. https://www.info.univ-tours.fr/~antoine/parole_publique/ANCOR_Centre/Pres_ANCOR_Centre.pdf

Apothéloz, D., & Reichler-Béguelin, M.-J. (1995). Construction de la référence et stratégies de désignation. *TRANEL. Travaux Neuchâtelois de Linguistique*, (23), 227-271.

Ariel, M. (1990). *Accessing noun-phrase antecedents* (first edition). Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9781315857473/accessing-noun-phrase-antecedents-rle-linguistics-grammar-mira-ariel

Ariel, M. (1991). The function of accessibility in a theory of grammar. *Journal of Pragmatics*, *16*(5), 443-463. https://doi.org/10.1016/0378-2166(91)90136-L

Asymétrie (statistiques). (2026). In *Wikipédia*. https://fr.wikipedia.org/w/index.php?title=Asym%C3%A9trie_(statistiques)&oldid=232443583

Bakhtine, M. (1984). *Esthétique de la création verbale*. Gallimard.

Bara, F., Morin, M.-F., Montésinos-Gelet, I., & Lavoie, N. (2011). Conceptions et pratiques en graphomotricité chez des enseignants de primaire en France et au Québec. *Revue française de pédagogie*, *176*(3), 41-56. https://doi.org/10.4000/rfp.3154

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., & Venturi, G. (2015). CItA : Un Corpus di Produzioni Scritte di Apprendenti l'Italiano L1 Annotato con Errori. *Proceedings of the 2nd Italian Conference on Computational Linguistics (CLiC-it)*, 31-35. https://iris.unito.it/retrieve/handle/2318/1532924/75495/#page=33

Barbagli, A., Lucisano, P., Dell'Orletta, F., Montemagni, S., & Venturi, G. (2016). CItA : An L1 Italian Learners Corpus to Study the Development of Writing Competence. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 88-95. https://aclanthology.org/L16-1014

Barletta, M. (2024). Annotation de la continuité référentielle dans un corpus scolaire – premiers résultats. In M. Balaguer, N. Bendahman, L.-M. Ho-Dac, J. Mauclair, J. G. Moreno, & J. Pinquier (Éds.), *35èmes Journées d'Études sur la Parole (JEP) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 26ème Rencontre des Étudiants*

Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL) (p. 28-41). ATALA & AFPC. https://inria.hal.science/hal-04622985

Barletta, M. (2025). *Guide pour l'annotation en continuité référentielle du corpus Scolinter*. https://hal.science/hal-05082398

Barletta, M., & Ponton, C. (2023, avril). *La question de la normalisation des écrits scolaires pour leur traitement automatique. Le cas de l'omission de mots.* [Présentation]. Bruit de fond ou valeur ajoutée? Gérer le bruit lors des traitements informatiques des corpus linguistiques, Grenoble, France. https://hal.science/hal-04090669

Barletta, M., & Ponton, C. (2025a). Analyse de la continuité référentielle dans le corpus d'écrits scolaires français et italien Scolinter. *Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN)*, *1 : articles scientifiques originaux*, 134-153.

Barletta, M., & Ponton, C. (2025b). La question de la normalisation des écrits scolaires pour leur traitement automatique. Le cas de l'omission de mots. *Corpus*, (26). https://doi.org/10.4000/1364v

Betthäuser, B. A., Bach-Mortensen, A. M., & Engzell, P. (2023). A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic. *Nature Human Behaviour*, *7*(3), 375-385. https://doi.org/10.1038/s41562-022-01506-4

Boiteau, A., Mathet, Y., & Widlöcher, A. (2025). Intégration des relations inter-référents dans l'annotation de la coréférence : Modèle et application. *Actes de CORIA-TALN-RJCRI-RECITAL 2025. Actes des 32ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : articles scientifiques originaux*, 437-450. https://talnarchives.atala.org/TALN/TALN-2025/101.pdf

Bonnemaison, K. (2018). *Anaphore et référence en production écrite : Étude de textes narratifs d'élèves de 9 à 11 ans, du CE2 au CM2.* [Thèse de doctorat, Université Toulouse le Mirail - Toulouse II]. https://hal.science/tel-01924645/file/Th%C3%A8se%20Karine%20Bonnemaison%20vol.1.pdf

Bonomi, I., Masini, A., Morgana, S., & Piotti, M. (2010). *Elementi di linguistica italiana*. Carocci.

Boré, C., & Elalouf, M.-L. (2017). Deux étapes dans la construction de corpus scolaires : Problèmes récurrents et perspectives nouvelles. *Corpus*, (16). https://doi.org/10.4000/corpus.2731

Boré, C., Roubaud, M.-N., & Elalouf, M.-L. (2018). *Corpus ÉMA, écrits scolaires*. https://hdl.handle.net/11403/ema-ecrits-scolaires-1/v5

Boudreau, S., & Kittredge, R. (2005). Résolution des anaphores et détermination des chaînes de coréférences : Différences entre variétés de textes. *Traitement Automatique des Langues*, *46*(1), 41-70.

Brissaud, C., Pasa, L., Ragano, S., & Totereau, C. (2016). Effets des pratiques d'enseignement de l'écriture en cours préparatoire. *Revue française de pédagogie*, *196*(3), 85-100. https://doi.org/10.4000/rfp.5079

Brissaud, C., & Ponton, C. (2026). *Scoledit*. https://doi.org/10.82270/scoledit/v1

Bronckart, J.-P. (1997). *Activité langagière, textes et discours. Pour un interactionnisme socio-discursif*. https://archive-ouverte.unige.ch/unige:37758

Brown, G., & Yule, G. (1983). *Discourse Analysis*. Cambridge University Press.

Brunato, D., & Dell'Orletta, F. (2015). ISACCO : A corpus for investigating spoken and written language development in Italian school–age children. In C. Bosco, S. Tonelli, & F. M. Zanzotto (Éds.), *Proceedings of the Second Italian Conference on Computational Linguistics CLiC-it 2015 : 3-4 December 2015, Trento* (p. 62-66). Accademia University Press. https://doi.org/10.4000/books.aaccademia.1324

Charolles, M. (1978). Introduction aux problèmes de la cohérence des textes : Approche théorique et étude des pratiques pédagogiques. *Langue française*, *38*(1), 7-41. https://doi.org/10.3406/lfr.1978.6117

Charolles, M. (1987). Contraintes pesant sur la constitution des chaînes de références comportant un nom propre. *Travaux du Centre de Recherches Sémiologiques, Université de Neuchâtel, Centre de Recherches Sémiologiques, 1987/53//29-55*. https://libra.unine.ch/handle/20.500.14713/57078

Charolles, M. (1988a). La gestion des risques de confusion entre personnages dans une tâche rédactionnelle. *Pratiques*, *60*(1), 75-97. https://doi.org/10.3406/prati.1988.1498

Charolles, M. (1988b). Les études sur la cohérence, la cohésion et la connexité textuelles depuis la fin des années 1960. *Modèles linguistiques*, *Tome X, Fasc. 2*, 45-66.

Charolles, M. (1988c). Les plans d'organisation textuelle : Périodes, chaînes, portées et séquences. *Pratiques*, *57*(1), 3-13. https://doi.org/10.3406/prati.1988.1468

Charolles, M. (1995). Cohésion, cohérence et pertinence du discours. *Travaux De Linguistique*, *29*, 121-151.

Charolles, M. (2001). Référents évolutifs et évolution de la référence. In *Les référents évolutifs entre linguistique et philosophie* (p. 39-97). Klincksieck. https://hal.archives-ouvertes.fr/hal-01404846

Charolles, M. (2002). *La référence et les expressions référentielles en français* (Ophrys).

Charolles, M. (2011). Cohérence et cohésion du discours. In K. H. ; C.Marello (Éd.), *Dimensionen der Analyse Texten und Diskursivent—Dimensioni dell'analisi di testi e discoursi* (p. 153-173). Lit Verlag. https://hal.archives-ouvertes.fr/hal-00665838

Charolles, M., & François, J. (1998). Les prédicats transformateurs et leurs patient : Fondements d'une ontologie naturelle. *Cahiers de Recherche Linguistique*, (11). https://hal.science/hal-01404729

Charolles, M., & Schnedecker, C. (1993). Coréférence et identité : Le problème des référents évolutifs. *Langages*, *112*, 106-126.

Chastain, C. (1975). Reference and Context. *Language, Mind, and Knowledge, University of Minnesota Press, Minneapolis.*, *7*, 194-269.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37-46. https://doi.org/10.1177/001316446002000104

Conte, M.-E. (1999). *Condizioni di coerenza : Ricerche di linguistica testuale* (Nuova ed. con l'aggiunta di due saggi / a cura di Bice Mortara Garavelli.). Edizioni dell'Orso.

Corblin, F. (1985). Les chaînes de référence : Analyse linguistique et traitement automatique. *Intellectica*, *1*(1), 123-143. https://doi.org/10.3406/intel.1985.851

Corblin, F. (1994). La condition de nouveauté comme défaut. *Faits de langues*, *2*(4), 147-153. https://doi.org/10.3406/flang.1994.951

Corblin, F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes. https://jeannicod.ccsd.cnrs.fr/ijn_00550962

Cornish, F. (1999). *Anaphora, discourse, and understanding : Evidence from English and French*. Oxford ; New York : Clarendon Press. http://archive.org/details/anaphoradiscours0000corn

Cossu, G., Gugliotta, M., & Marshall, J. C. (1995). Acquisition of reading and written spelling in a transparent orthography : Two non parallel processes? *Reading and Writing*, *7*(1), 9-22. https://doi.org/10.1007/BF01026945

Da Milano, F. (2025). *Per una descrizione della deissi testuale nel corpus Univers-ITA*. Franco Angeli. https://boa.unimib.it/handle/10281/549761

De Cesare, A.-M. (2010). Deittici. In *Enciclopedia dell'Italiano Treccani*. https://www.treccani.it/enciclopedia/deittici_(Enciclopedia-dell'Italiano)/

Delaborde, M. (2020). *Analyse en corpus de chaînes de coréférence : La coréférence non-stricte à l'épreuve de la linguistique outillée* [Thèse de doctorat, Université de la Sorbonne Nouvelle - Paris III]. https://theses.hal.science/tel-03425446

Dell'Orletta, F., Montemagni, S., & Venturi, G. (2011). READ–IT : Assessing Readability of Italian Texts with a View to Text Simplification. In N. Alm (Éd.), *Proceedings of the Second Workshop on Speech and Language Processing for Assistive Technologies* (p. 73-83). Association for Computational Linguistics. https://aclanthology.org/W11-2308/

Dinarelli, M., Niaouri, D., Lopez, F., Gonzalez-Saez, G., Nacklé, M., Esperança-Rodier, E., Rossi, C., Schwab, D., & Ballier, N. (2024). Context-Aware Neural Machine Translation Models Analysis And Evaluation Through Attention. *Revue TAL : traitement automatique des langues*, *64*(3). https://hal.science/hal-04581509

Doquet, C. (2020). Analyser linguistiquement l'écriture à l'école : EcriScol, un corpus génétique. In *CLUB Working Papers in Linguistics Volume 4* (Vol. 4, p. 127-140). https://hal.science/hal-02883152

Doquet, C., Enoiu, V., Fleury, S., & Mazziotti, S. (2017). Problèmes posés par la transcription et l'annotation d'écrits d'élèves. *Corpus*, (16), Article 16. https://doi.org/10.4000/corpus.2776

Doquet, C., Ho-Dac, L.-M., & Ponton, C. (2023, juillet 3). Un corpus de référence pour l'écriture de l'école à l'université : La ressource É-Calm. *Jounées Linguistique de Corpus*. https://shs.hal.science/halshs-04212830

Doquet, C., & Pilorgé, J.-L. (2020). La correction de copies au collège entre langue et discours : Une catégorisation syntactico-énonciative. *Repères. Recherches en didactique du français langue maternelle*, (62), Article 62. https://doi.org/10.4000/reperes.3274

Doquet, C., & Ponton, C. (2021). Écrire de l'école à l'université : Corpus, traitements, analyses outillées. Présentation. *Langue francaise*, *211*(3), 11-20. https://doi.org/https://doi.org/10.3917/lf.211.0011

Doquet, C., Revelli, L., & Moysan, A. (2021). Écriture et forme scolaire : Spécificités de transcription et de traitement. *Langue francaise*, *211*(3), 21-36.

Dubois, J. (avec Internet Archive). (1965). *Grammaire structurale du francais*. Paris : Librairie Larousse. http://archive.org/details/grammairestructu0000dubo

Ducrot, O. (1930-) A. du texte, & Todorov, T. (1939-2017) A. du texte. (1979). *Dictionnaire encyclopédique des sciences du langage / Oswald Ducrot, Tzvetan Todorov*. https://gallica.bnf.fr/ark:/12148/bpt6k4813280n

Dupret, C. (2022). *Le développement de la référence en récit à l'âge préscolaire et scolaire : Une analyse multifactorielle* [Thesis, Paris 3]. https://theses.fr/2022PA030027

Ediger, M. (2002). Assessing Handwriting Achievement. *Reading Improvement*, *39*(3), 103-110.

Elalouf, M.-L. (avec Bertucci, M.-M.). (2005). *Écrire entre 10 et 14 ans : Un corpus, des analyses, des repères pour la formation* (Vol. 1-1). SCÉRÉN-CRDP Académie de Versailles CDDP Essonne.

Elalouf, M.-L., Besnard, M., Boré, C., Castagnet-Caignec, S., Gerlaud, B., Koulibali, Y., Neyman, C., Roubaud, M.-N., Similowski, K., & Valma, E. (2025). Corpus EMA-écrits scolaires. *Scolagram - Revue de didactique de la grammaire*, (12, Exploitation d'un grand corpus d'écrits scolaires pour la recherche et la formation). https://scolagram.u-cergy.fr/index.php/content_page/item/1359-corpus-ema-ecrits-scolaires

Elalouf, M.-L., & Boré, C. (2007). Construction et exploitation de corpus d'écrits scolaires. *Revue française de linguistique appliquée*, (1), 53-70. https://doi.org/10.3917/rfla.121.0053

Engel, P., & Nef, F. (1988). Identité, Vague Et Essences. *Les Études philosophiques*, (4), 475-494.

Fairon, C., Klein, J., & Paumier, S. (2006). *Le langage SMS. Étude d'un corpus informatisé à partir de l'enquête « Faites don de vos SMS à la science »*.

Familier (esprit). (2024). In *Wikipédia*. https://fr.wikipedia.org/w/index.php?title=Familier_(esprit)&oldid=212246594

Farina, E., & Teruggi, L. (2021). Un corpus di testi trilingue per promuovere la riflessione sulla pratica didattica. *Ricerca e didattica per promuovere intelligenza comprensione e partecipazione*, *I tomo panel 1-2-3*, 263-279.

Federzoni, S. (2025). *Vers une typologie des chaînes de référence à la lumière de corpus annotés diversifiés* [Thèse de doctorat]. Université Toulouse Jean Jaurès.

Federzoni, S., Ho-Dac, L.-M., & Fabre, C. (2021). Coreference Chains Categorization by Sequence Clustering. In C. Braud, C. Hardmeier, J. J. Li, A. Louis, M. Strube, & A. Zeldes (Éds.), *Proceedings of the 2nd Workshop on Computational Approaches to Discourse* (p. 52-57). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.codi-main.5

Federzoni, S., Ho-Dac, L.-M., & Fabre, C. (2023). A linear approach of chain composition. In L. Gardelle, L. Vincent-Durroux, & H. Vinckel-Roisin (Éds.), *Reference. From conventions to pragmatics* (Vol. 228, p. 107-126). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.228.06fed

Federzoni, S., Ho-Dac, L.-M., & Rebeyrolle, J. (2020). Les chaînes topicales dans la ressource ANNODIS. *SHS Web of Conferences*, *78*, 11005. https://doi.org/10.1051/shsconf/20207811005

Federzoni, S., Rebeyrolle, J., Ho-Dac, L.-M., Bard, Y., & Garcia-Debanc, C. (2021). *Guide d'annotation RésolCo*. https://hal.science/hal-03297228

Ferrari, A. (2014). *Linguistica del testo. Principi, fenomeni, strutture*. Carocci.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378-382. https://doi.org/10.1037/h0031619

Fort, K. (2012). *Les ressources annotées, un enjeu pour l'analyse de contenu : Vers une méthodologie de l'annotation manuelle de corpus* [Thèse de doctorat, Université Paris-Nord - Paris 13]. https://theses.hal.science/tel-00797760

Garcia-Debanc, C., & Bonnemaison, K. (2014). La gestion de la cohésion textuelle par des élèves de 11-12 ans : Réussites et difficultés. *SHS Web of Conferences*, *8*, 961-976. https://doi.org/10.1051/shsconf/20140801349

Garcia-Debanc, C., & Bras, M. (2016). *Vers une cartographie des compétences de cohérence et de cohésion textuelle dans une tâche-problème de production écrite réalisée par des élèves de 9 -12 ans : Indicateurs de maitrise et progressivité*. *Recherches textuelles* (13). https://hal.science/hal-01987031

Garcia-Debanc, C., Ho-Dac, L.-M., Bras, M., & Rebeyrolle, J. (2017). Vers l'annotation discursive de textes d'élèves. *Corpus*, (16), Article 16. https://doi.org/10.4000/corpus.2783

Garcia-Debanc, C., Ho-Dac, L.-M., Federzoni, S., Bras, M., & Rebeyrolle, J. (2019). *« RésolCo » un corpus de manuscrits d'élèves et d'étudiants pour l'étude de la cohérence*. 10èmes Journées Internationale de la Linguistique de Corpus. https://hal.science/hal-02877122

Garcia-Debanc, C., Rebeyrolle, J., & Ho-Dac, L.-M. (2021). La continuité référentielle dans le corpus RÉSOLCO : Méthode d'annotation et premières analyses. *Langue francaise*, *211*(3), 99-114. https://doi.org/10.3917/lf.211.0099

Garcia-Debanc, C., Roubaud, M.-N., & Béchour, M. (2022). *La grammaire pour écrire—CE2 et cycle 3*. https://hal.science/hal-03972823

Gardelle, L., Rossi, C., & Vincent-Durroux, L. (2019). La gestion de l'anaphore en discours : Complexités et enjeux. *Cahiers de praxématique*, (72), Article 72. https://doi.org/10.4000/praxematique.5368

Gardent, C., & Manuélian, H. (2005). Création d'un corpus annoté pour le traitement des descriptions définies. *Revue TAL*, *46*(1), 115.

Giora, R. (2003). On Our Mind : Salience, Context, and Figurative Language. In *On Our Mind : Salience, Context, and Figurative Language*. https://doi.org/10.1093/acprof:oso/9780195136166.001.0001

Givón, T. (1983). *Topic continuity in discourse : A quantitative cross-language study*. Amsterdam ; Philadelphia : J. Benjamins Pub. Co. http://archive.org/details/topiccontinuityi0000unse

Goigoux, R. (2016). *Étude de l'influence des pratiques d'enseignement de la lecture et de l'écriture sur la qualité des premiers apprentissages.* https://ife.ens-lyon.fr/sites/default/files/2022-05/rapport-lire-et-ecrire.pdf

Grandi, N. (2025). Introduzione : Una lingua in declino? In N. Grandi, *FrancoAngeli Series – Open Access*. https://series.francoangeli.it/index.php/oa/catalog/book/1329

Grandi, N., Ballarè, S., Chiusaroli, F., Gallina, F., Pascoli, M., & Pistolesi, E. (2023). *Corpus Univers-ITA* [Jeu de données]. https://corpora.ficlit.unibo.it/CUSP/crystal/index.html#dashboard?corpname=UniverS-Ita. https://doi.org/https://doi.org/10.60760/unibo/univers-ita

Grice, H. P. (1975). Logic and conversation. *Syntax and Semantics*, *3*, 41-58. https://doi.org/10.1163/9789004368811_003

Grobol, L. (2020a). *Coreference resolution for spoken French* [Phdthesis, Université Sorbonne Nouvelle - Paris 3]. https://hal.archives-ouvertes.fr/tel-02928209

Grobol, L. (2020b). *Coreference resolution for spoken French* [Thèse de doctorat, Université Sorbonne Nouvelle - Paris 3]. https://hal.archives-ouvertes.fr/tel-02928209

Grobol, L., Tellier, I., de la Clergerie, É., Dinarelli, M., & Landragin, F. (2018, mai). ANCOR-AS : Enriching the ANCOR Corpus with Syntactic Annotations. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018. https://aclanthology.org/L18-1064

Guillaume, B., & Fort, K. (2013). Expériences de formalisation d'un guide d'annotation : Vers l'annotation agile assistée. *20e conférence sur le Traitement Automatique des Langues Naturelles*, 628-635. https://hal.science/hal-00840895

Guo, C. (2023). *Etude contrastive des chaînes de référence en français et en chinois : Étude de cas à partir d'un corpus parallèle composé de Jean-Christophe de Romain Rolland et de sa traduction par Lei Fu : volume 1* (Numéro 2023STRAC015) [Theses, Université de Strasbourg]. https://theses.hal.science/tel-04453175

Gutiérrez-Caceres, R., Farina, E., Teruggi, L. A., Brissaud, C., & Ponton, C. (2023). *Description and analysis of the use of evaluation devices in first grade primary narrative texts in three linguistic contexts*. Writing Research Across Borders, Trondheim, Norway.

Heiden, S. (2010). The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, 389-398. https://aclanthology.org/Y10-1044

Hirschman, L. (1998). MUC-7 Coreference Task Definition. In N. Chinchor (Éd.), *Proceedings of the 7th Message Understanding Conference*. https://www-nlpir.nist.gov/related_projects/muc/proceedings/co_task.html

Hirst, G. (1981). *Anaphora in Natural Language Understanding : A Survey* (Vol. 119). Springer. https://doi.org/10.1007/3-540-10858-0

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). *spaCy : Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303

Hoste, V. (2016). The Mention-Pair Model. In M. Poesio, R. Stuckardt, & Y. Versley (Éds.), *Anaphora Resolution : Algorithms, Resources, and Applications* (p. 269-282). Springer. https://doi.org/10.1007/978-3-662-47909-4_9

Hou, J. (2021). *Étude de la saillance référentielle dans un corpus de textes narratifs : Analyse contrastive des chaînes de référence en français et en chinois* [Thesis, Paris 3]. https://theses.fr/2021PA030060

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes : The 90% Solution. *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, 57-60. https://aclanthology.org/N06-2015

Istruzioni e programmi per l'insegnamento della lingua italiana e dell'aritmetica nelle scuole elementari (R. D. 10 ottobre 1867) (1867). https://www.museodellascuola.it/wp-content/uploads/2018/03/02-programmi-1867.pdf

Jacques, M.-P., & Rinck, F. (2017). Un « corpus de littéracie avancée : Résultat et point de départ. *Corpus*, (16), Article 16. https://doi.org/10.4000/corpus.2806

Jang, Y., Hong, S., Son, J., Park, S., Park, C., & Lim, H. (2025). *From Ambiguity to Accuracy : The Transformative Effect of Coreference Resolution on Retrieval-Augmented Generation systems* (arXiv:2507.07847). arXiv. https://doi.org/10.48550/arXiv.2507.07847

Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd éd.). https://web.stanford.edu/~jurafsky/slp3/

Karttunen, L. (1976). *Discourse Referents*. https://doi.org/10.1163/9789004368859_021

Kim, S. Y. (2023). *Les expressions référentielles et les chaînes de référence en français et en coréen* [Thèse de doctorat, Université de Strasbourg]. https://publication-theses.unistra.fr/public/theses_doctorat/2023/Kim_Seoyeoung_2023_ED520.pdf

Kleiber, G. (1990). Marqueurs référentiels et processus interprétatifs : Pour une approche « plus sémantique ». *Cahiers de linguistique française*, (11), 241-258.

Kleiber, G. (1997). Sens, référence et existence : Que faire de l'extra-linguistique ? *Langages*, (127), 9-37. https://doi.org/10.3406/lgge.1997.2123

Kleiber, G. (1999). *Anaphore associative et relation partie-tout : Condition d'aliénation et principe de congruence ontologique*. https://doi.org/10.3406/lfr.1999.6288

Kleiber, G. (2011). *Sémantique et pragmatique du déterminant possessif*. https://doi.org/10.3406/igram.2011.4142

Kleiber, G., Patry, R., & Ménard, N. (1993). Anaphore associative : Dans quel sens « roule »-t-elle ? *Revue québécoise de linguistique*, *22*. https://doi.org/10.7202/602773ar

Kleiber, G., & Vassiliadou, H. (2007). Sur les approches intuitives de la relation d'élaboration. *Scolia : Sciences Cognitives, Linguistiques et Intelligence Artificielle*, *22*(1), 147-161. https://doi.org/10.3406/scoli.2007.1109

Klie, J.-C., Bugert, M., Boullosa, B., Castilho, R. E. de, & Gurevych, I. (2018). The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5-9. http://tubiblio.ulb.tu-darmstadt.de/106270/

Kodandaram, S. R., Honnappa, K., & Soni, K. (2021). Masking private user information using Natural Language Processing. *International Journal of Advance Research in Computer Science and Management*, *7*, 1753-1763.

Koehn, P. (2005). Europarl : A Parallel Corpus for Statistical Machine Translation. *Proceedings of Machine Translation Summit X: Papers*, 79-86. https://aclanthology.org/2005.mtsummit-papers.11/

Krippendorff, K. (1980). *Content Analysis : An Introduction to its Methodology*. Thousand Oaks, Calif. : Sage. http://archive.org/details/contentanalysisi0000krip_y7k5

Kurtosis. (2025). In *Wikipédia*. https://fr.wikipedia.org/w/index.php?title=Kurtosis&oldid=229629779

Landragin, F. (2016a). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, (92), 11.

Landragin, F. (2016b). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, (92), 11.

Landragin, F. (2019). *Manuel d'annotation du projet DEMOCRAT (version 2.4.3), Projet ANR DEMOCRAT.* https://www.lattice.cnrs.fr/democrat/files/ANR-15-CE38-0008-DEMOCRAT_livrable_methodo.pdf

Landragin, F. (2020a). La saillance : Origines perceptives, applications linguistiques, enjeux interdisciplinaires. *Semen. Revue de sémio-linguistique des textes et discours*, (49), Article 49. https://doi.org/10.4000/semen.14430

Landragin, F. (2020b). *Rapport final du projet ANR Democrat, "Description et modélisation des chaînes de référence : Outils pour l'annotation de corpus et le traitement automatique"* [Research Report]. ANR (Agence Nationale de la Recherche - France). https://hal.archives-ouvertes.fr/hal-02533314

Landragin, F. (2021). Le corpus DEMOCRAT et son exploitation. Présentation. *Langages*, *224*(4), 11-24.

Landragin, F. (2022). Expressions référentielles et chaînes de référence en français : Le projet Democrat et son exploration des rapports entre linguistique textuelle et linguistique de corpus. *Écho des études romanes*, *18*(1), 49-65. https://doi.org/10.32725/eer.2022.004

Landragin, F., Glikman, J., Schnedecker, C., & Todirascu, A. (2024). Chaînes de référence dans le corpus Democrat : Une analyse en diachronie longue. *Corpus*, (25), Article 25. https://doi.org/10.4000/corpus.8581

Landragin, F., Poibeau, T., & Victorri, B. (2012). *ANALEC : A New Tool for the Dynamic Annotation of Textual Data*. 357. https://shs.hal.science/halshs-00698971

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., & Schwab, D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Éds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (p. 2479-2490). European Language Resources Association. https://aclanthology.org/2020.lrec-1.302

Lebarbé, T. (2009). Du corpus littéraire au corpus linguistique : Dématérialisation, restructuration, lectures rhizomatiques et analyses linguistiques des manuscrits. *Corpus*, (8), Article 8. https://doi.org/10.4000/corpus.1694

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics*, *39*(4), 885-916. https://doi.org/10.1162/COLI_a_00152

LIDILEM. (2025). *Scolinter*. https://hdl.handle.net/11403/scolinter/v1

Longo, L. (2013). *Vers des moteurs de recherche « intelligents » : Un outil de détection automatique de thèmes : méthode basée sur l'identification automatique des chaînes de référence* [These de doctorat, Strasbourg]. http://www.theses.fr/2013STRAC041

Longo, L., & Todirascu, A. (2009). Une étude de corpus pour la détection automatique de thèmes. *Texte et Corpus.*, (4), 143.

Longo, L., & Todirascu, A. (2014). Vers une typologie des chaînes de référence dans des textes administratifs et juridiques. *Langages*, *195*(3), 79-98. https://doi.org/10.3917/lang.195.0079

Longobardi, E., Spataro, P., Renna, M., & Rossi-Arnaud, C. (2014). Comparing Fictional, Personal, and Hypothetical Narratives in Primary School : Story Grammar and Mental State Language.

European Journal of Psychology of Education, *29*(2), 257-275. https://doi.org/10.1007/s10212-013-0197-y

Lopez, F. (2023). Approches neuronales pour la détection des chaînes de coréférences : Un état de l'art. In M. Candito, T. Gerald, & J. G. Moreno (Éds.), *Actes de CORIA-TALN 2023. Actes des 16e Rencontres Jeunes Chercheurs en RI (RJCRI) et 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL)* (p. 101-113). ATALA. https://hal.science/hal-04130236

Losh, M., & Capps, L. (2003). Narrative ability in high-functioning children with autism or Asperger's syndrome. *Journal of Autism and Developmental Disorders*, *33*(3), 239-251. https://doi.org/10.1023/a:1024446215446

Maillard, M. (1974). *Essai de typologie des substituts diaphoriques*. https://doi.org/10.3406/lfr.1974.5665

Manuélian, H., & Schnedecker, C. (2021). Introduction des référents et premier maillon de suites coréférentielles : La réalité des textes. *Langages*, *224*(4), 67-86. https://doi.org/10.3917/lang.224.0067

Martinelli, G., Barba, E., & Navigli, R. (2024). *Maverick : Efficient and Accurate Coreference Resolution Defying Recent Trends* (arXiv:2407.21489). arXiv. http://arxiv.org/abs/2407.21489

Mathet, Y., Widlöcher, A., & Métivier, J.-P. (2015). The Unified and Holistic Method Gamma (γ) for Inter-Annotator Agreement Measure and Alignment. *Computational Linguistics*, *41*(3), 437-479. https://doi.org/10.1162/COLI_a_00227

Mazziotti, S. (2021). *L'incidence du système linguistique : Étude des postures de correction des enseignants et des modalités de réécriture à l'école primaire en France et en Italie* [Phdthesis, Université de la Sorbonne nouvelle - Paris III ; Università degli studi (Bologne, Italie)]. https://theses.hal.science/tel-03737461

Mazziotti, S., Pallanti, L., Moysan, A., & Reversé, C. (2024). E-CALM. Analyser de grands corpus scolaires et universitaires : Des questions pour la recherche et pour la formation. *SHS Web of Conferences*, *186*, 01001. https://doi.org/10.1051/shsconf/202418601001

Mélanie-Becquet, F., Barré, J., Seminck, O., Plancq, C., Naguib, M., Pastor, M., & Poibeau, T. (2024). *BookNLP-fr, the French Versant of BookNLP. A Tailored Pipeline for 19th and 20th Century French Literature* [Application/pdf,application/xml]. *3*(1). https://doi.org/10.48694/JCLS.3924

Merlin, S. (2016). *The origins of 'anaphora' and 'deixis'. Ancient theories on articles and pronouns.* https://iris.univr.it/handle/11562/939737

Milner, J.-C. (1982). *Ordres et raisons de langue*. Éditions du Seuil.

MIM. (2025). *Nuove indicazioni 2025 Scuola dell'infanzia e Primo ciclo di istruzione Materiali per il dibattito pubblico*. https://www.mim.gov.it/documents/20182/0/Nuove+indicazioni+2025.pdf/cebce5de-1e1d-12de-8252-79758c00a50b?version=1.1&t=1741806754378

Ministère de l'Education Nationale. (2019). *Pour enseigner la lecture et l'écriture au CP*. https://eduscol.education.fr/document/1508/download?attachment

Ministère de l'Education Nationale. (2025). *Programme de français du cycle 2*. https://www.education.gouv.fr/sites/default/files/ensel135_annexe3.pdf

Ministère de l'Education Nationale de la Jeunesse et des Sports. (2025). *Programmes et horaires à l'école élémentaire*. Ministère de l'Éducation nationale. https://www.education.gouv.fr/programmes-et-horaires-l-ecole-elementaire-9011

Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche. (2015). *Bullettin officiel spécial n° 11 du 26 novembre 2015*. https://ww2.ac-poitiers.fr/eps/sites/eps/IMG/pdf/bo_spe_11_26-11-2015.pdf

Ministère de l'Education Nationale, & Direction générale de l'enseignement scolaire. (2015). *Programme d'enseignement du cycle des apprentissages fondamentaux (cycle 2)*. https://eduscol.education.fr/84/j-enseigne-au-cycle-2

Ministère de l'Education nationale et de la Jeunesse. (2019). *Pour enseigner la lecture et l'écriture au CE1—Un guide fondé sur l'état de la recherche*. https://eduscol.education.fr/document/1509/download

Miquet-Savin, H. (2024). *Vous écrivez ? Anaphorez ! Ou les anaphoriques dans les productions écrites de lycéens-nes* [Thèse de doctorat, Université Grenoble - Alpes]. https://hal.science/tel-04888254

MIUR. (2012). *Indicazioni nazionali per il curricolo della scuola dell'infanzia e del primo ciclo d'istruzione*. https://mim.gov.it/documents/20182/51310/DM+254_2012.pdf/1f967360-0ca6-48fb-95e9-c15d49f18831?version=1.0

Moysan, A. (2024). *Les élèves écrivent, les professeurs corrigent. Les pratiques langagières d'enseignants de français sur les copies d'élèves comme traces de leurs conceptions et de l'enseignement de l'écriture.* [Thèse de doctorat, Université de la Sorbonne nouvelle - Paris III]. https://theses.hal.science/tel-05037712

MUC Consortium. (1995a). Appendix D : Coreference Task Definition (v2.3). *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. MUC 1995. https://aclanthology.org/M95-1025

MUC Consortium. (1995b). Sixth Message Understanding Conference (MUC-6) : Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995. *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. MUC 1995. https://aclanthology.org/M95-1000

Muller, C. (2019). *Indéfinis et partitifs en français*. Presses universitaires de Bordeaux.

Muzerelle, J., Lefeuvre, A., Antoine, J.-Y., Schang, E., Maurel, D., Villaneau, J., & Eshkol, I. (2013). ANCOR, premier corpus de français parlé d'envergure annoté en coréférence et distribué librement. In ATALA (Éd.), *TALN'2013, 20e conférence sur le Traitement Automatique des Langues Naturelles* (p. 555-563). https://hal.archives-ouvertes.fr/hal-01016562

Ng, V. (2016). Advanced Machine Learning Models for Coreference Resolution. In M. Poesio, R. Stuckardt, & Y. Versley (Éds.), *Anaphora Resolution : Algorithms, Resources, and Applications* (p. 283-313). Springer. https://doi.org/10.1007/978-3-662-47909-4_10

Oberlé, B. (2016). *Étude des chaînes de référence dans les articles de recherche de format IMRaD : problèmes d'annotation, analyse quantitative et qualitative* [Mémoire de master 1].

Oberlé, B. (2018, mai). SACR : A Drag-and-Drop Based Tool for Coreference Annotation. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. LREC 2018. https://aclanthology.org/L18-1059

Oberlé, B. (2019). Types de chaînes de référence dans les articles de recherche de format IMRAD. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (25), https://doi.org/10.4000/discours.10487

Oberlé, B., Schnedecker, C., Baumer, E., Capin, D., Glikman, J., Guo, C., Revol, T., Todirascu, A., & Tushkova, J. (2018). Les chaînes de référence dans les textes encyclopédiques du 12e au 21e siècle : Étude longitudinale. *Travaux de linguistique*, *77*(2), 67-141. https://doi.org/10.3917/tl.077.0067

Obry, V., Glikman, J., Guillot-Barbance, C., & Pincemin, B. (2017). Les chaînes de référence dans les récits brefs en français : Étude diachronique (XIIIe – XVIe s.). *Langue française, Les chaînes de référence en corpus*, *2017/3*(195), 91-110. https://doi.org/10.3917/lf.195.0091

Pallotti, G., Borghetti, C., & Rosi, F. (2021). *Insegnare a scrivere nella scuola primaria : Il progetto « Osservare l'interlingua »*. https://cris.unibo.it/handle/11585/805993

Pani, L. (2012). Insegnare e imparare il corsivo in Italia oggi : Riflessioni di una paleografa. *Gazette du livre médiéval*, *58*, 17-36.

Pecorari, F. (2025). Punteggiatura e testualità nel corpus Univers-ITA. Fenomeni e problemi. In N. Grandi, *Franco Angeli Series – Open Access* (p. 244-258). https://series.francoangeli.it/index.php/oa/catalog/book/1329

Perdicoyianni-Paleologou, H. (2001). Le concept d'anaphore, de cataphore et de déixis en linguistique française. *Revue québécoise de linguistique*, *29*(2), 55-77. https://doi.org/10.7202/039441ar

Perrault, C. (1826). *Œuvres choisies de Charles Perrault / Contes de ma mère l'oye* (C. de Plancy, Éd.; 1826e éd., p. 21-24). Peytieux.

Perret, M. (2000). Quelques remarques sur l'anaphore nominale aux XIVe et XVe siècles. *L'information grammaticale*, *87*(1), 17-23. https://doi.org/10.3406/igram.2000.2740

Perugini, N., & Zucchini, E. (2025). Catene referenziali nel testo narrativo : Un'indagine esplorativa sulla scrittura di alunni monolingui e plurilingui. *STUDI ITALIANI DI LINGUISTICA TEORICA E APPLICATA*. https://www.studitlinguisticateoricappl.it/article/view/1536

Péry-Woodley, M.-P., Afantenos, S. D., Ho-Dac, L.-M., & Asher, N. (2011). La ressource ANNODIS, un corpus enrichi d'annotations discursives. *Revue TAL : traitement automatique des langues*, *52*(3), 71-101.

Poesio, M. (2016). Linguistic and Cognitive Evidence About Anaphora. In M. Poesio, R. Stuckardt, & Y. Versley (Éds.), *Anaphora Resolution : Algorithms, Resources, and Applications* (p. 23-54). Springer. https://doi.org/10.1007/978-3-662-47909-4_2

Poesio, M., & Artstein, R. (2008). Anaphoric Annotation in the ARRAU Corpus. In D. T. Nicoletta Calzolari (Conference Chair) Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (Éd.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

Poesio, M., Artstein, R., Uryupina, O., Rodriguez, K., Delogu, F., Bristot, A., & Hitzeman, J. (2013). *The ARRAU Corpus of Anaphoric Information* (p. 184570 KB) [Jeu de données]. Linguistic Data Consortium. https://doi.org/10.35111/Y3MR-HE10

Poesio, M., Grishina, Y., Kolhatkar, V., Moosavi, N., Roesiger, I., Roussel, A., Simonjetz, F., Uma, A., Uryupina, O., Yu, J., & Zinsmeister, H. (2018). Anaphora Resolution with the ARRAU Corpus. In M. Poesio, V. Ng, & M. Ogrodniczuk (Éds.), *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference* (p. 11-22). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0702

Ponton, C. (2018, juin). Constituer et analyser un corpus scolaire. L'approche Scoledit. *3ème école de Printemps Petale*. https://hal.univ-grenoble-alpes.fr/hal-01910629

Ponton, C., Brissaud, C., & Wolfarth, C. (2022). Évolution de la gestion de l'orthographe en production de textes entre 6 et 15 ans : Apports du TAL. *SHS Web of Conferences*, *143*, 01005. https://doi.org/10.1051/shsconf/202214301005

Ponton, C., Doquet, C., Fleury, S., Ho-Dac, L.-M., & Jacques, M.-P. (2022). *E-CALM* [Jeu de données]. https://hdl.handle.net/11403/e-calm/v2.1

Ponton, C., Gutiérrez-Caceres, R., Teruggi, L., Farina, E., Brissaud, C., & Wolfarth, C. (2021). Scolinter : Un corpus trilingue. L'exemple de la segmentation en mots. *Langue française*, *211*(3), 37-50. https://doi.org/10.3917/lf.211.0037

Ponton, C., Wolfarth, C., & Brissaud, C. (2019, novembre). Premières explorations textométriques d'un corpus scolaire longitudinal (CP-CM1). *Journées Linguistique de Corpus (JLC2019)*. https://shs.hal.science/halshs-03205233

Prince, E. F. (1981). Toward a taxonomy of given-new information. *Radical Pragmatics*. https://cir.nii.ac.jp/crid/1573950400281637248?lang=en

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza : A Python Natural Language Processing Toolkit for Many Human Languages. *arXiv:2003.07082 [cs]*. http://arxiv.org/abs/2003.07082

Quignard, M., Le Mené Guigourès, M., & Landragin, F. (2021). Élaboration du corpus DEMOCRAT : Procédures d'annotation et d'évaluation. *Langages*, *224*(4), 25-46. https://doi.org/10.3917/lang.224.0025

Recasens, M., Hovy, E., & Martí, M. A. (2010). A Typology of Near-Identity Relations for Coreference (NIDENT). In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Éds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA). https://aclanthology.org/L10-1103/

Recasens, M., Hovy, E., & Martí, M. A. (2011). Identity, non-identity, and near-identity : Addressing the complexity of coreference. *Lingua*, *121*(6), 1138-1152. https://doi.org/10.1016/j.lingua.2011.02.004

Revelli, L. (2011). *Scritture scolastiche dall'unità d'Italia ai giorni nostri : Studi e testimonianze*. Aracne.

Revelli, L. (2024). *Tra norma, usi e sentire comune : Pseudoregole dell'italiano attraverso il CoDiSSc.* (p. 47-61). https://doi.org/10.6092/unibo/amsacta/8065

Riegel, M., Pellat, J.-C., & Rioul, R. (2004). *Grammaire methodique du francais*. Presses Universitaries de france. http://archive.org/details/grammairemethodi0000mart

Riegel, M., Pellat, J.-C., & Rioul, R. (2018). *Grammaire methodique du francais*. Presses Universitaries de France. http://archive.org/details/grammairemethodi0000mart

Roubaud, M.-N., & Garcia-Debanc, C. (2014). *L'approche d' « anomalies » dans des textes narratifs d'élèves de fin d'école primaire (10-11 ans) : Quelques pistes pour la lecture des textes par les enseignants*. Peter Lang. https://doi.org/10.3726/978-3-0352-6464-7

Rousier-Vercruyssen, L., & Landragin, F. (2019). Interdistance et instabilité au sein des chaînes de référence : Indices textuels ? *Discours. Revue de linguistique, psycholinguistique et informatique*, (25). https://doi.org/10.4000/discours.10522

Salles, M. (2015). Chaînes de référence : La deuxième mention. L'exemple des entités inanimées dans les narrations littéraires. *Travaux de linguistique*, *71*(2), 111-133. https://doi.org/10.3917/tl.071.0111

Schang, E., Boyer-Pelletier, A., Muzerelle, J., Antoine, J.-Y., Eshkol, I., & Maurel, D. (2011). Coreference and anaphoric annotations for spontaneous speech corpos in French. *DAARC'2011, 8th Discourse Anaphora and Anaphor Resolu1on Colloquium*, 9 pp. https://hal.science/hal-00831414

Schmid, H. (1999). Improvements in Part-of-Speech Tagging with an Application to German. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, & D. Yarowsky (Éds.), *Natural Language Processing Using Very Large Corpora* (p. 13-25). Springer Netherlands. https://doi.org/10.1007/978-94-017-2390-9_2

Schnedecker, C. (1997). *Nom propre et chaînes de référence* (Klincksieck, Éd.; Vol. 21). Librairie Klincksieck. https://hal.archives-ouvertes.fr/hal-00808797

Schnedecker, C. (2005). Les chaînes de référence dans les portraits journalistiques : Éléments de description. *Travaux de linguistique*, *51*(2), 85-133. https://doi.org/10.3917/tl.051.0085

Schnedecker, C. (2006). *De l'un à l'autre et réciproquement.... Aspects sémantiques, discursifs et cognitifs des pronoms anaphoriques corrélés l'un/l'autre et le premier/le second*. De Boeck Supérieur. https://doi.org/10.3917/dbu.schne.2006.02

Schnedecker, C. (2014). Chaînes de référence et variations selon le genre. *Langages*, *195*(3), 23-42. https://doi.org/10.3917/lang.195.0023

Schnedecker, C. (2017). Les chaînes de référence : Une configuration d'indices pour distinguer et identifier les genres textuels. *Langue française*, *195*(3), 53-72. https://doi.org/10.3917/lf.195.0053

Schnedecker, C. (2019). De l'intérêt de la notion de chaîne de référence par rapport à celles d'anaphore et de coréférence. *Cahiers de praxématique*, (72), Article 72. https://doi.org/10.4000/praxematique.5339

Schnedecker, C. (2021). *Les chaînes de référence en français*. OPHRYS.

Schnedecker, C., & Landragin, F. (2014). Les chaînes de référence : Présentation. *Langages*, *195*(3), 3-22. https://doi.org/10.3917/lang.195.0003

Schnedecker, C., & Longo, L. (2012). Impact des genres sur la composition des chaînes de référence : Le cas des faits divers. In *SHS Web of Conferences* (Vol. 1). https://doi.org/10.1051/shsconf/20120100061

Schneuwly, B. (1995). Apprendre à écrire. Une approche socio-historique. In *La production de textes : Vers un modèle d'enseignement de l'écriture* (p. 73). Les Ed. Logiques. https://archive-ouverte.unige.ch/unige:34349

Schneuwly, B., & Dolz, J. (1997). *Les genres scolaires. Des pratiques langagières aux objets d'enseignement*. https://doi.org/10.3406/reper.1997.2209

Schulz, C. M. (2010). *Snoopy & les Peanuts 1969-1970* (Éd. intégrale, Vol. 1-1). Dargaud. 125401

Smith, R. (2007). An Overview of the Tesseract OCR Engine. *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*, 629-633. https://doi.org/10.1109/ICDAR.2007.4376991

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, *27*(4), 521-544. https://doi.org/10.1162/089120101753342653

Sperber, D., & Wilson, D. (1996). *Relevance : Communication and Cognition* (2nd edition). Wiley-Blackwell.

Stein, N. L., & Glenn, C. G. (1975). *An Analysis of Story Comprehension in Elementary School Children : A Test of a Schema*. https://eric.ed.gov/?id=ED121474

Stuckardt, R. (2016). Introduction. In M. Poesio, R. Stuckardt, & Y. Versley (Éds.), *Anaphora Resolution : Algorithms, Resources, and Applications* (p. 1-19). Springer. https://doi.org/10.1007/978-3-662-47909-4_1

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2020). Anaphora and Coreference Resolution : A Review. *Information Fusion*, *59*. https://doi.org/10.1016/j.inffus.2020.01.010

Teruggi, L. A. (2019). *Leggere e scrivere a scuola : Dalla ricerca alla didattica*. Carocci.

Testenoire, P.-Y. (2017). Transcrire des écrits scolaires : Entre philologie et génétique textuelle. *Corpus*, (16), Article 16. https://doi.org/10.4000/corpus.2762

Tutin, A. (2002). A corpus-based study of pronominal anaphoric expressions in French. In A. Branco, T. McEnery, & R. Mitkov (Éds.), *Proceedings of the 4th Discourse Anaphora and Anaphora Resolution Colloquium (DAARC 2002)*. Edições Colibri.

Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, É., Zaenen, A., Rayot, S., & Antoniadis, G. (2000). Annotating a large corpus with anaphoric links. *Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC2000)*, 2. https://hal.archives-ouvertes.fr/hal-00373327

UNICEF. (2021). *COVID-19 and School Closures : One year of education disruption* (p. 18). UNICEF. https://data.unicef.org/resources/one-year-of-covid-19-and-school-closures/

Voormann, H., & Gut, U. (2008). *Agile corpus creation*. *4*(2), 235-251. https://doi.org/10.1515/CLLT.2008.010

Widlöcher, A., & Mathet, Y. (2012). The Glozz Platform : A Corpus Annotation and Mining Tool. *Proceedings of the 2012 ACM Symposium on Document Engineering*, 171-180.

Wolfarth, C. (2019). *Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal* [Thèse de doctorat, Université Grenoble Alpes]. https://tel.archives-ouvertes.fr/tel-02517320

Wolfarth, C., Brissaud, C., & Ponton, C. (2018a). Transcrire et normer un corpus scolaire : Pour quelles analyses? In C. Brissaud, M. Dreyfus, & B. Kervyn (Éds.), *Repenser l'écriture et son évaluation au primaire et au secondaire* (p. 121-145). Presses universitaires de Namur. https://books.openedition.org/pun/5283?lang=en

Wolfarth, C., Brissaud, C., & Ponton, C. (2018b). Transcrire et normer un corpus scolaire : Pour quelles analyses? In C. Brissaud, M. Dreyfus, & B. Kervyn (Éds.), *Repenser l'écriture et son évaluation au primaire et au secondaire* (p. 121-145). Presses universitaires de Namur. https://books.openedition.org/pun/5283?lang=en

Wolfarth, C., Ponton, C., & Brissaud, C. (2018). Gestion de la morphographie verbale en production d'écrits : Que peut nous apprendre un corpus longitudinal ? *Repères. Recherches en didactique du français langue maternelle*, (57), https://doi.org/10.4000/reperes.1576

Wolfarth, C., Ponton, C., & Totereau, C. (2017). Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique. *Corpus*, *Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement* (16), https://doi.org/10.4000/corpus.2796

Zhou, E., & Choi, J. D. (2018). They Exist ! Introducing Plural Mentions to Coreference Resolution and Entity Linking. In E. M. Bender, L. Derczynski, & P. Isabelle (Éds.), *Proceedings of the 27th International Conference on Computational Linguistics* (p. 24-34). Association for Computational Linguistics. https://aclanthology.org/C18-1003/

Ziegler, J. C. (2018). Différences inter-linguistiques dans l'apprentissage de la lecture: *Langue Française*, *N° 199*(3), 35-49. https://doi.org/10.3917/lf.199.0035

Žitkus, V., Butkienė, R., & Butleris, R. (2024). Linguistically aware evaluation of coreference resolution from the perspective of higher-level applications. *Natural Language Engineering*, *30*(4), 821-850. https://doi.org/10.1017/S1351324923000293

Zucchini, E. (2023). *L'italiano neostandard nella lingua a scuola : Il caso dell'alternanza fra indicativo e congiuntivo.* [Thèse de doctorat, Alma Mater Studiorum - Università di Bologna]. https://doi.org/10.48676/unibo/amsdottorato/10505

Liste des tableaux

Tableau 1. Synthèse des taux de densité référentielle observés dans divers corpus. 32
Tableau 2. Distribution (%) des types d'expressions anaphoriques selon le genre textuel dans le corpus étudié par Tutin (2002). 36
Tableau 3. Distribution des mentions par type dans différents genres textuels. 38
Tableau 4. Composition du corpus RésolCo. Données tirées de la page internet http://redac.univ-tlse2.fr/corpus/resolco/telechargement.html, consultée le 01/02/2026. 66
Tableau 5. Tableau récapitulatif des corpus d'écrits scolaires italiens analysés dans ce chapitre. 82
Tableau 6. Tableau récapitulatif des corpus scolaires italiens d'écrits scolaires. 82
Tableau 7. Tableau de synthèse des caractéristiques des corpus français annotés en coréférence analysés. 98
Tableau 8. Distribution des élèves participants au recueil en France. 115
Tableau 9. Textes recueillis pour la partie française du corpus - corpus Scoledit 140
Tableau 10. Textes recueillis pour la partie italienne du corpus. Les deux recueils de 2018 et de 2019 ont été réalisés en parallèle. 140
Tableau 11. Distribution des textes recueillis pour le corpus espagnol. 141
Tableau 12. Distribution des écoles et des années scolaires - corpus espagnol. 141
Tableau 13. Textes disponibles sur la plateforme Scolinter. 141
Tableau 14. Distribution des élèves par école d'appartenance dans le corpus longitudinal. 142
Tableau 15. Caractérisation des textes du corpus français longitudinal. 143
Tableau 16. Caractérisation des textes du corpus italien longitudinal. 143
Tableau 17. Correspondances entre balises XML présentes dans les normalisations (côté transcripteur), leur équivalent dans le fichier exporté et l'élément utilisé pour remplacer chaque balise dans notre chaine de prétraitement. 147
Tableau 18. Comparaison entre distribution sur le corpus longitudinal et le sous-corpus *gold* de test sélectionné pour l'expérimentation. 149
Tableau 19. Tokens choisis pour remplacer chaque catégorie morphosyntaxique proposée lors de la normalisation. 152
Tableau 20. Extrait du texte CE1, 2449 dans sa *version* gold et dans les résultats obtenus à travers les trois méthodes testées. 152
Tableau 21. Versions *gold* et issue de la méthode FlauBERT du même texte (CM1, 3012, corpus français). 154
Tableau 22. Versions *gold* et issue de la méthode FlauBERT du même texte (corpus français, CE1, 2984). 154
Tableau 23 Versions de référence et méthode FlauBERT du même texte (corpus français, Texte CE2, 2983). 155
Tableau 24. Distribution des textes par nombre de tokens sur le corpus longitudinal français. 172
Tableau 25. Distribution des textes par nombre de tokens sur le corpus annoté français - première campagne d'annotation. 172
Tableau 26. Caractérisation de la longueur des textes du corpus annoté en nombre de tokens. 173
Tableau 27. Indices de description des chaines de référence dans le corpus annoté par niveau scolaire... 174
Tableau 28. Comparaison entre la distribution par longueur des textes dans le corpus longitudinal et dans le corpus à annoter du français. 181
Tableau 29. Caractérisation de la longueur des textes du corpus annoté en nombre de tokens. 192
Tableau 30. Nombre d'expressions référentielles annotées et distribution des chaines (≥3 mentions, paires, singletons) par langue et par niveau scolaire. 193
Tableau 31. Distribution des mentions annotées par texte (singletons, anaphores et chaines confondues). 194
Tableau 32. Longueur des mentions annotées (nombre de tokens) par langue et niveau scolaire. 195
Tableau 33. Nombre de référents annotés par texte : total, moyenne et étendue selon la langue et le niveau scolaire. 196
Tableau 34. Distribution des textes par nombre de référents - comparaison entre corpus. 197

Tableau 35. Indices de descriptions - longueur des chaines en termes de mentions. 199
Tableau 36. Calcul de la distance inter-maillonnaire sur les chaines du texte de l'exemple (68). 201
Tableau 37. Caractérisation des valeurs absolues d'interdistance- comparaison entre français et italien.. 202
Tableau 38. Caractérisation des valeurs du ratio d'interdistance- comparaison entre le corpus français et le corpus italien. 203
Tableau 39. Calcul de la densité référentielle : ratio entre le nombre de maillons et le nombre total de tokens du texte. 205
Tableau 40. Indices de description de la distribution des densités référentielles – comparaison entre langues. 206
Tableau 41. Résultats obtenus à partir du t test de Welch en comparant les indices de densité référentielle. 207
Tableau 42. Calcul de la densité référentielle : ratio entre le nombre de tokens inclus dans des maillons et le nombre total de tokens du texte. 211
Tableau 43. Indices de description de la distribution des densités référentielles (prise en compte de la longueur des maillons) – comparaison entre langues. 211
Tableau 44. Résultats obtenus à partir du t test de Welch en comparant les indices de densité référentielle (prise en compte de la longueur des maillons). 213
Tableau 45. Éléments pour le calcul des indices de stabilité, instabilité et CNSL pour les chaines du texte du corpus français, CE2, 3032. 218
Tableau 46. Calcul des indices de stabilité, instabilité référentielle et coefficient normalisé de stabilité lexicale. 219
Tableau 47. Description des valeurs du CNSL - comparaison entre les deux corpus. 220
Tableau 48. Distributions des mentions par type grammatical (toutes mentions confondues). 225
Tableau 49. Détail de la distribution des types de mentions dans les chaines - comparaison entre corpus et niveaux scolaires. 227
Tableau 50. Distribution des types de mentions en position 1 et 2 des chaines. 228
Tableau 51. Distribution des titres et des SN définis en position 1 et 2 des chaines. 228
Tableau 52. Distribution des mentions des paires anaphoriques/coréférentielles par type grammatical. . 232
Tableau 53. Distribution des singletons par type grammatical. 233
Tableau 54. Distribution des titres et des SN définis dans les chaines annotées. 253
Tableau 55. Caractérisation du corpus annoté - reprise d'une partie du Tableau 29. 290

Liste des figures

Figure 1. Illustration des liens anaphoriques dans l'exemple de texte (3). 14
Figure 2. Échelle de l'accessibilité de Ariel (1991). 44
Figure 3. Échelle de topicalité de Givón (1983). Traduction de Schnedecker (1997). 47
Figure 4. Texte du corpus RésolCo. Transcription (à gauche) accompagnée du scan (à gauche). 67
Figure 5. Texte du corpus RésolCo. Version normalisée (à gauche) accompagné du scan (à droite). 68
Figure 6. Version normalisée et annotée d'un texte aves les filtres possibles sélectionnés. 68
Figure 7. Exemple d'annotation des erreurs, tiré de Barbagli et al., 2015 (p. 32). 74
Figure 8. Représentation des types d'écrits présents dans le CoDiSSc issus du site web du projet (consulté le 20/08/2025). 75
Figure 9. Page 6/14 du cahier 0486 du corpus CoDiSsC, produit en 1938. 77
Figure 10. Exemples de modélisations relationnelles (1 à 4) et ensemblistes (5 à 7). Figure tirée de Boiteau et al., 2025, p. 438. 90
Figure 11. Tableau de synthèse des tâches de lecture-écriture tiré de Goigoux, 2016, p. 71 112
Figure 12. Consigne visuelle utilisée au CP pour le recueil des données du le projet « Lire et écrire au CP ». 113
Figure 13. Consigne visuelle utilisée du CE1 au CM2. 114
Figure 14. Exemple de support d'écriture utilisé en France au CP - lignes Seyes. 118
Figure 15. Exemple de support d'écriture utilisé en Italie - lignes larges. 118
Figure 16. Exemple de support d'écriture utilisé en Espagne. 118
Figure 17. Texte visualisé sur la plateforme Scoledit *(corpus français, CE1, 51)*. 126
Figure 18. Visualisation côté transcripteur du scan et de la transcription du texte sur la plateforme Scolinter *(corpus français, CE2, 2983).* 127
Figure 19. Extrait du scan et de la transcription d'un texte du corpus Scolinter *(corpus français, CE1, 52).* 127
Figure 20. Extrait de scan et transcription du même texte contenant un titre et son indication métatextuelle (*corpus italien, CE1, 735).* 128
Figure 21. Scan et transcription du même texte *(corpus italien, CE2, 826)*. Le titre se trouve en fin de texte, précédé d'une indication métatextuelle. 128
Figure 22. Extrait d'un scan et de la transcription du même texte (*corpus Scoledit, CM2, 103).* Exemple tiré de Wolfarth, 2019. 129
Figure 23. Extrait d'un scan et de la transcription du même texte *(corpus italien, CE2, 769).* 129
Figure 24. Extrait du scan et de la transcription du même texte *(corpus italien, CE2, 769).* 130
Figure 25. Extrait du scan et de la transcription du texte du corpus espagnol *(CE1, 988).* 130
Figure 26. Extrait du texte du corpus italien contenant des révisions et un ajout (*corpus italien, CE2, 826).* 131
Figure 27. Extrait du scan et de la transcription contenant un fragment illisible (*corpus français, CE1, 1926).* 132
Figure 28. Page d'édition d'un texte du corpus sur la plateforme Scolinter. 136
Figure 29. Test d'OCR effectué sur un extrait du texte du corpus français (CE1, 828), image et transcription attendue. 138
Figure 30. Extrait du test d'OCR sur le texte du corpus italien (CE2, 747), image et transcription attendue. 138
Figure 31. Nombre moyen de tokens par texte – comparaison entre corpus longitudinaux français et italien. 144
Figure 32. Longueur des textes dans le corpus longitudinal en français et en italien. Comparaison effectuée entre CE1 et CE2. 144
Figure 33. Scan d'un extrait de texte contenant une omission (corpus français, CE1, 1350). 148
Figure 34. Distribution des omissions par étiquette syntaxique -comparaison entre corpus longitudinal et corpus gold (ou de référence). 149

Figure 35. Analyse en dépendance d'une phrase tirée du texte CE1, S1350, générée par Spacy (v.3) en utilisant le modèle fr_core_news_lg. En haut, la phrase sans le mot omis, en bas la phrase avec le token « qui » réintroduit (entouré dans la représentation). 151
Figure 36. Métriques UAS, LAS, LS et précision sur le POS calculé en comparaison avec les 54 textes du *gold*. 153
Figure 37. Analyse en dépendance des deux versions effectué automatiquement à l'aide de SpaCy v. 3 et du modèle fr_core_news_lg 154
Figure 38. Phases de la création de corpus traditionnelle ou "à cascade". Image inspirée de Voorman et Gut (2008). 156
Figure 39. Phases de la création d'un corpus annoté selon la méthode agile. Figure inspirée de Voorman et Gut (2008). 157
Figure 40. Texte contenant un référent évolutif (corpus français, CE2, 1851). 162
Figure 41. Schéma illustrant les étapes du processus d'annotation de notre corpus sur le plan méthodologique. 166
Figure 42. Frise chronologique détaillante l'organisation du processus d'annotation entre octobre 2023 et janvier 2025. 166
Figure 43. Ecran de la fonctionnalité d'adjudication sur la plateforme INCEpTION. 178
Figure 44. Annotation sur INCEpTION (corpus français, CE2, 1323) contenante un clitique, « vous » à la phrase 3. 179
Figure 45. Annotation sur INCEpTION (corpus italien, CE2, 1427). 179
Figure 46. Distribution des textes en nombre de tokens - comparaison entre niveaux CE1 et CE2 du corpus longitudinal français et niveaux CE1 et CE2 du corpus sélectionné pour l'annotation de la deuxième campagne 182
Figure 47. Distribution des mentions par texte dans le corpus annoté en français et en italien. 194
Figure 48. Distribution des mentions annotés par longueur (nombre de tokens). Comparaison entre niveaux par langue. 196
Figure 49. Distribution des référents dans le corpus annoté. 197
Figure 50. Distribution des binômes de personnages les plus fréquents dans les textes. 198
Figure 51. Distribution de la longueur des chaines dans le corpus français. 200
Figure 52. Distribution de la longueur des chaines dans le corpus italien. 200
Figure 53. Distribution des chaines par taux de ratio d'interdistance dans le corpus français – tous niveaux confondus. 203
Figure 54. Distribution des chaines par taux de ratio d'interdistance dans le corpus italien – tous niveaux confondus. 204
Figure 55. Comparaison des densités référentielle et p-values obtenues selon le *test t de Welch*. 208
Figure 56. Distribution des textes en fonction de la densité référentielle - représentation de la distribution du corpus français. 209
Figure 57. Distribution des textes en fonction de la densité référentielle - représentation de la distribution du corpus italien. 209
Figure 58. Distribution de la densité référentielle dans les textes prenant en compte la taille des maillons - comparaison entre français et italien. 212
Figure 59. Calcul de la densité référentielle prenant en compte la longueur des mentions. Comparaison entre langues et entre niveaux scolaires pour chaque langue. 214
Figure 60. Distribution du CNSL sur les 184 chaines du corpus français. 220
Figure 61. Distribution du CNSL sur les 236 chaines du corpus italien. 221
Figure 62. Distribution des types de mentions sur les 5 premières maillons des chaines - comparaison entre langues. 229
Figure 63. Distribution des types de mentions sur les cinq premiers maillons des chaines dans le corpus français - comparaison entre CE1 et CE2 230
Figure 64. Distribution des mentions par type grammatical sur les cinq premiers maillons des chaines dans le corpus italien - comparaison entre CE1 et CE2 231

Figure 65. Distribution des *patterns* les plus fréquents des trois premières mentions dans 200 chaines de référence du corpus français. 234
Figure 66. Distribution des patterns les plus fréquents des trois premières mentions dans 198 chaines de référence du corpus italien. 236
Figure 67. Distribution des types de mentions sur les 5 premiers maillons des chaines - comparaison entre langues (reprise de la Figure 62). 252
Figure 68. Distribution des textes par nombre de référents dans le corpus annoté. Reprise de la Figure 49. 262
Figure 69. Schéma de la chaine dans le texte du corpus français, CE1, 3045. 264
Figure 70. Schéma de la chaine dans le texte du corpus français, CE1, 3046. 264
Figure 71. Schéma de la chaine du texte du corpus français, CE1, 2901. 265
Figure 72. Schéma de la chaine du texte du corpus français, CE2, 3055. 266
Figure 73. Schéma de la chaine du texte du corpus italien, CE1, 599. 267
Figure 74. Schéma de la chaine du texte du corpus italien, CE1, 627. 268
Figure 75. Schéma illustrant les chaines du texte du corpus français, CE2, 207. 271
Figure 76. Schéma illustrant les chaines du texte du corpus français, CE2, 1352. 272
Figure 77. Schéma illustrant les chaines du texte du corpus français, CE2, 3012. 273
Figure 78. Schéma illustrant les chaines du texte du corpus français, CE2, 1596 274
Figure 79. Schéma illustrant les chaines du texte du corpus français, CE2, 2893. 275
Figure 80. Schéma illustrant les chaines du texte du corpus français, CE2, 1866 277
Figure 81. Schéma illustrant les chaines du texte du corpus français, CE1, 534. 278
Figure 82. Schéma illustrant les chaines du texte du corpus français, CE2, 847. 279
Figure 83. Schéma illustrant les chaines du texte du corpus français, CE2, 536. 280
Figure 84. Schéma illustrant les chaines du texte du corpus français, CE1, 2983. 282
Figure 85. Schéma illustrant les chaines du texte du corpus français, CE2, 2983. 283
Figure 86. Schéma illustrant les chaines du texte du corpus français, CE1, 950. 284
Figure 87. Schéma illustrant les chaines du texte du corpus français,CE1, 100. 285
Figure 88. Schéma illustrant les chaines du texte du corpus italien, CE1, 325. 286
Figure 89. Schéma illustrant les chaines du texte du corpus italien, E1, 1380 287
Figure 90. Schéma illustrant les chaines du texte du corpus italien, CE1, 28. 288
Figure 91. Texte du corpus italien (CE1, 821). Exemple d'écriture en script majuscule. 291
Figure 92. Extrait du texte du corpus italien (CE2, 761). Exemple d'écriture en script. 291
Figure 93. Extrait du texte du corpus italien (CE2, 1423). Exemple d'écriture cursive. 292
Figure 94. Page d'un cahier contenu dans le corpus CoDiSSc daté entre 1879 et 1899. Cahier 0231 provenant de Vallée d'Aoste. 294

Annexes

Annexe A. Consigne corpus Univers-ITA

« Devi scrivere un testo di media lunghezza: tra le 250 e le 500 parole. Dovrai cercare di usare uno stile formale: quindi, scrivi nel modo più corretto possibile, come se scrivessi per un tuo professore. Proprio perché la rilevazione è totalmente anonima, sarà impossibile associare il testo alla tua persona e quindi, partecipando alla rilevazione, rinunci alla proprietà intellettuale su di esso. Il testo non sarà mai pubblicato integralmente e sarà utilizzato solo per scopi di ricerca. » (Grandi, 2025, p. 78)

Vous devez rédiger un texte de longueur moyenne : entre 250 et 500 mots. Vous devrez vous efforcer d'adopter un style formel : écrivez donc de la manière la plus correcte possible, comme si vous vous adressiez à votre professeur. Précisément parce que cette enquête est totalement anonyme, il sera impossible de relier le texte à votre personne et, par conséquent, en participant à cette enquête, vous renoncez à vos droits de propriété intellectuelle sur celui-ci. Le texte ne sera jamais publié dans son intégralité et ne sera utilisé qu'à des fins de recherche.

« Traccia: Immagina che il tuo corso di laurea abbia aperto un sondaggio rivolto a tutti gli studenti, con l'obiettivo di raccogliere opinioni sul funzionamento della didattica a distanza nei mesi di emergenza sanitaria. Scrivi un testo in cui esponi, in modo non schematico, i vantaggi e gli svantaggi della didattica a distanza, secondo il tuo punto di vista. . » (Grandi, 2025, p. 79)

Sujet : Imaginez que votre cursus universitaire ait lancé un sondage destiné à tous les étudiants, dans le but de recueillir des avis sur le fonctionnement de l'enseignement à distance pendant les mois d'urgence sanitaire. Rédigez un texte dans lequel vous exposez, de manière non schématique, les avantages et les inconvénients de l'enseignement à distance, selon votre point de vue.

(notre traduction en collaboration avec Deepl)

Annexe B. Lettre aux directeurs d'écoles italiens

UNIVERSITA' DEGLI STUDI DI MILANO - BICOCCA

Piazza dell'Ateneo Nuovo, 1 - 20126 - Milano

Milano, 2 maggio 2018

Al Dirigente Scolastico

dell'Istituto Comprensivo

.....

.....

L' Università di Milano Bicocca in collaborazione con l'Università di Grenoble Alpes sta attuando una ricerca longitudinale sulle competenze linguistiche degli alunni della scuola primaria. Nello specifico, scopo della ricerca è quello di confrontare l'evoluzione delle competenze ortografiche e testuali nei bambini francesi e italiani nella fascia di età che va dai 6 agli 8 anni. L'analisi dei dati raccolti in entrambi i paesi potrà avere delle ricadute positive sulla formazione degli insegnanti nell'ambito della didattica della lingua. A questo proposito il gruppo di ricerca di Milano Bicocca si impegna a dare una restituzione dei risultati, entro l'anno scolastico successivo, ai singoli Istituti Comprensivi che avranno aderito all'iniziativa.

La ricerca è strutturata in diverse fasi, la prima delle quali da svolgersi nel mese di maggio (seconda quindicina) 2018 e coinvolgerà le classi prime e seconde, alle quali verranno somministrate le seguenti prove: nelle classi prime, la produzione scritta individuale di una storia sulla base di quattro immagini (Tempi: 15 minuti), nelle classi seconda, la produzione scritta individuale di una storia sulla base di una o due immagini scelte tra quattro messe a disposizione (Tempi: 30 minuti).

Ai docenti di classe NON viene richiesto alcun impegno nella somministrazione e/o nella valutazione degli elaborati in quanto queste azioni verranno svolte dal gruppo di ricerca Milano Bicocca.

Vi ringraziamo per la collaborazione che vorrete offrire e siamo disponibili ad incontrare i docenti di classe per fornire ulteriori dettagli.

Responsabile della ricerca in Italia

Prof.ssa Lilia Andrea Teruggi

Annexe C. Description du protocole de recueil adopté par l'équipe italienne

Traduction du document d'origine en italien.

Les responsables du recueil ont demandé aux enseignants de retirer des murs de la classe ou cacher toute affiche susceptible d'aider les élèves lors de la rédaction du texte (ex. des affiches montrant les mots relatifs aux quatre personnages de l'histoire, des listes de verbes, des affiches rappelant des règles d'écriture etc.).

Après avoir distribué les feuilles contenantes la consigne, celle-là a été présenté à voix haute par la chercheuse en utilisant cette formulation (avec des variations éventuelles).

"Oggi ciascuno di voi scriverà una storia con uno o due personaggi"

"Aujourd'hui, chacun d'entre vous écrira une histoire avec un ou deux personnages.

La chercheuse demandait à un élève de nommer à vois hautes les personnages représentés pour être sûr que tous les élèves aient en tête les mêmes personnages

"**Io vi leggerò la consegna scritta sotto le immagini, ascoltate bene la consegna. Ecco 4 personaggi. Scegliete uno o due personaggi e raccontate una storia, cerchia il o i personaggi che hai scelto"**

« Je vais vous lire le texte écrit sous les images, écoutez bien le texte. Voici quatre personnages. Choisissez un ou deux personnages et racontez une histoire, entourez-le ou les personnages que vous avez choisis.

"**Avete guardato bene le immagini? Avete scelto uno o due personaggi? Quando avete scelto, cerchiate il vostro o i vostri personaggi. Prima di cominciare a scrivere pensate bene alla storia che volete raccontare"**

« Avez-vous bien regardé les images ? Avez-vous choisi un ou deux personnages ? Quand vous avez choisi, entourez votre ou vos personnages. Avant de commencer à écrire, réfléchissez bien à l'histoire que vous voulez raconter ».

"Avete 35 minuti per scrivere questa storia. Dovete lavorare da soli, senza chiedere ai compagni o alle maestre"

"Vous avez 35 minutes pour écrire cette histoire. Vous devez travailler seul, sans demander l'avis de vos camarades ou de vos enseignantes ».

Lors du déroulement de l'épreuve, les enseignants devaient répondre de manière la plus neutrale possible aux élèves posant des questions et ne devaient pas fournir d'aides aux élèves. Par exemple, à des questions sur l'orthographe d'un mot, les enseignants devaient répondre de manière encourageante mais sans interférer avec l'écriture de l'enfant « Tu sais comment l'écrire, et même si tu n'es pas sûr, ce n'est pas grave, écris comme tu sais et comme tu peux. »

Si des élèves terminaient l'épreuve à l'avance, leur était demandé de se reposer et de ne pas interférer avec le travail des camarades. Ce n'était pas permis de colorer les images constituantes la consigne, car cela aurait pu inciter certains élèves à interrompre leur test d'écriture. Les élèves n'étaient pas autorisés à se lever pour aller lire un livre.

Annexe D. Guide de transcription – corpus Scoledit et Scolinter

Encodage et problèmes de numérisation

Productions écrits – CP

1) Marqueur de retour à la ligne :

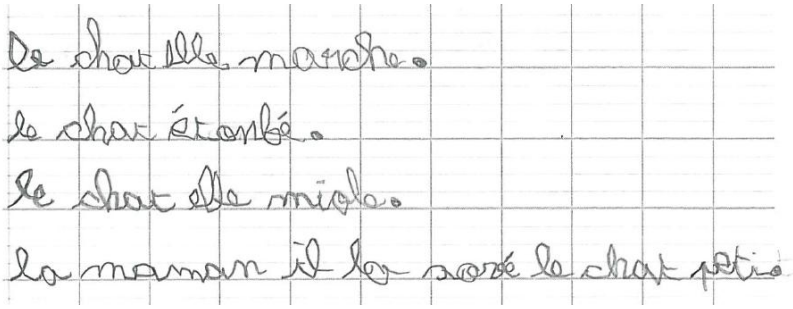
/ : Si le retour à la ligne est dû à la fin de ligne sur la page



(Production **1354**)

L'exemple ci-dessus sera transcrit :
Le petit chat est tonder il c'est / fait male au pate et a la téte il a / fait maou sa maman vien le / récupére le petit chat.

// : Si le retour à la ligne est jugé volontaire et non contraint par le support



(Production **1361**)

L'exemple ci-dessus sera transcrit :

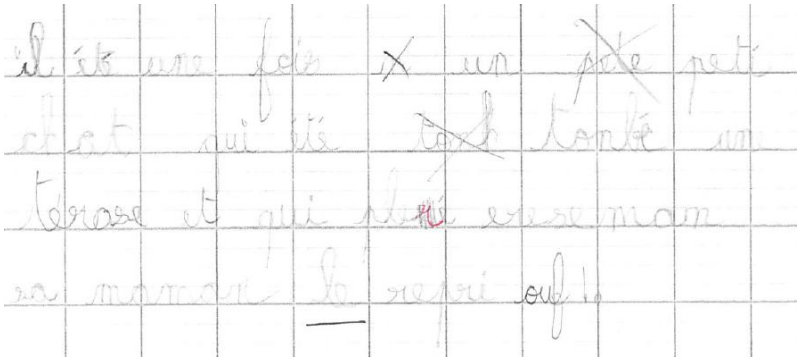
le chat elle marche. // le chat étonbé. // le chat elle miole. // la maman il la sové le chat peti.

/# : changement de page : sans changement de paragraphe

//# : changement de page : avec changement de paragraphe

2) Marques de révision :

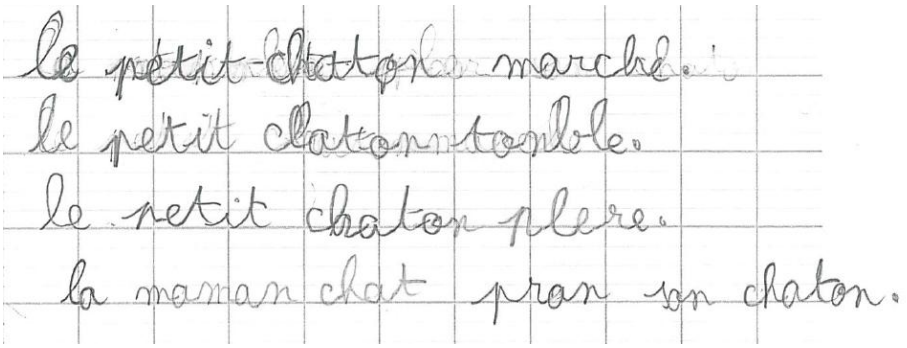
<revision/> : Pour tout élément comportant une rature, des traces de gomme ou des traces de réécriture.



(Production **1560**)

L'exemple ci-dessus sera transcrit :

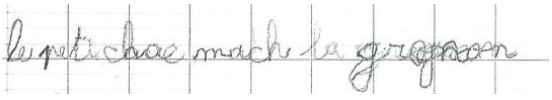
il été une fois <revision/> un <revision/> peti / chat qui été <revision/> tonbé une / terase et qui ple<revision/>ré ereseman / sa maman le repri ouf !.



(Production **1343**)

L'exemple ci-dessus sera transcrit :

<revision/> le petit chaton marche. // <revision/> le petit chaton tonble. // le petit chaton plere. // la maman chat pran son chaton.



(Production **2533**)

L'exemple ci-dessus sera transcrit :

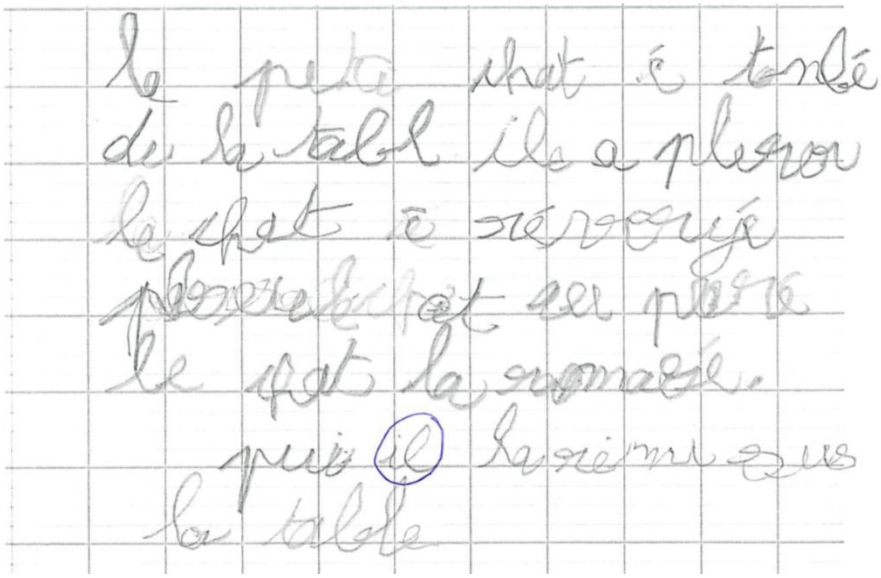
le peti chae mach la <revision/><illisible/>

3) Soulignement et les lettres entourées :

Le soulignement, tout comme les lettres ou mots entourés, étant parfois le fait du relecteur il n'est pas toujours possible de faire la différence entre un soulignement produit par l'élève de celui produit par l'adulte, nous n'en tiendrons pas compte.

4) **Segments illisibles : <illisible/ >**

Marquage des parties indéchiffrables de la production

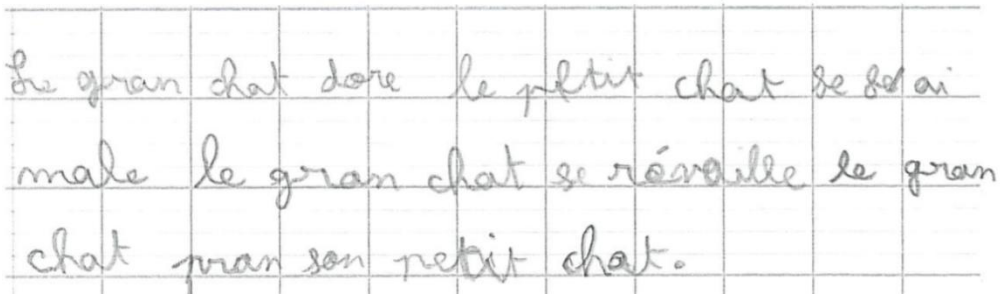


(Production **1341**)

Cette production sera transcrite :

le peti chat é tonbé / de la tabl ile a plerer / le chat é réveryé / <illisible/> <revision/> at <illisible/> pleré / le chat la ramasé. / pui il la remi sus / la table

5) Doute sur la transcription : <unsure>…</unsure>



(Production **1340**)

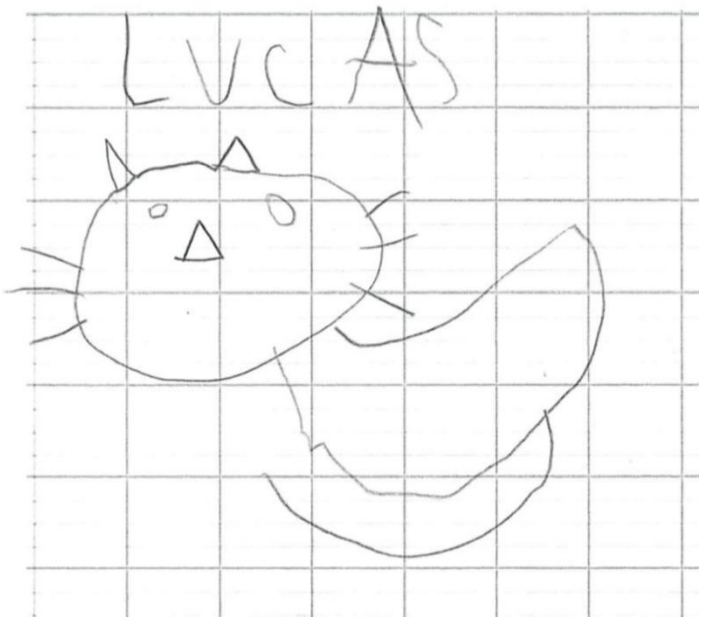
Dans cette production, le doute persiste pour « réveille », la transcription sera donc :

Le gran chat dore le petit chat se fe ai / male le gran chat se <unsure>révaille</unsure> le gran / chat pran son <revision/>petit chat.

6) Dessin et prénom

<dessin/> : présence de dessin dans la production

<prenom/> : écriture du prénom de l'enfant par l'enfant lui-même en place de la production



(Production **1288**)

Pour cette production, il ne sera fait mention que des balises <prenom/> et <dessin/>.

7) Ajout dans la production

<ajout> … </ajout> : trace de révision par ajout dans la production

Annexe E. Guide de normalisation – corpus Scoledit et Scolinter

Protocole pour l'élaboration d'une version normée de productions de texte, à partir de leur transcription

De la transcription à sa normalisation

1. Éléments retenus

Ne sont retenus dans la normalisation que les éléments textuels produits par l'enfant : les mots, ou groupes de lettres, et la ponctuation.

Dans le cas où le texte est écrit en majuscules, la normalisation se fait en minuscules (la majuscule est un marqueur de phrase).

2. Éléments non retenus

Sont absents de la normalisation les balises et autres marques spécifiques à la transcription, tels que <revision/>, /, //#. La mention de dessin ou tout autre élément extratextuel n'est pas conservé dans la normalisation.

Principes de normalisation

1. Segmentation en mots, orthographe et accords

L'orthographe est normée (*etait* → *était* et *foie* → *fois* dans l'exemple (1)).

(1) Il etait une foie (568, CP)
[normalisation] Il était une fois

(2) IL GATINO CAMINA (01, 1P)
[normalisation] Il gattino cammina

En cas de mot concerné par les rectifications orthographiques de 1990, le choix de l'apprenant est respecté.

(3) Ses maitres etait trés gentille. (1557, CE1)
[normalisation] Ses maitres étaient très gentils.

(4) c'est son pére quil fait le maître. (2937, CE1)
[normalisation] c'est son père qui fait le maître.

L'orthographe des interjections les plus courantes est normée (exemple (5)).

(5) comme par surprise BOOM le loup le mangea !!! (825, CM1)
[normalisation] comme par surprise BOUM le loup le mangea !!!

Pour *BOUM* et *Miaou*, on laissera les graphies françaises, qui figuraient sur les vignettes.

La segmentation en mots est rétablie (*lepren → le prend* dans l'exemple (5)).

(6) il lepren (563, CP)
[normalisation] il le prend
(7) SIÉ FATTO MALE (03-1P)
[normalisation] si è fatto male

(8) E·VAGIU·LA·MAMMA·LOVA·APREDERE (04-1P)
[normalisation] e va giù <segmentation/> la mamma lo va a prendere.

Les accords en genre, en nombre (*les chaton → les chatons* dans l'exemple (9)) et en personne (*tu est → tu es* dans l'exemple (11)) sont rétablis.

(9) et les chaton aussi (1143, CP)
[normalisation] et les chatons aussi

(10) pourquoi tu est venu (1866, CE2)
[normalisation] pourquoi tu es venu

3. Verbes et conjugaison

Les temps des verbes sont conservés, le critère oral primant sur le critère écrit : par exemple, dans l'exemple (12) on rétablit l'imparfait que l'élève a orthographié.

(11) il tombe et il pleurer. (586, CP)
[normalisation] Il tombe et il pleurait.

(12) MENTRE·LA·MAMMA·DEL·GATINO·DORMIVA·ILGATINO·CADE·E·SI·FECIE·MOLTO MALE (06-1P)
[normalisation] mentre la mamma del gattino dormiva il gattino cadde e si fece molto male

(en cas de doute, le passé est rétabli dans un contexte où tous les verbes sont au passé)

La cohérence des temps n'est pas rétablie (exemple (13) et exemple (14)).

(13) qan stest ler dalé fair la siest le pe ti chaton ne voulest pas fair la siet (1706, CP)
[normalisation] quand c'est l'heure d'aller faire la sieste <segmentation/> le petit chaton ne voulait pas faire la sieste

(14) C'era una volta un gattino e un gatto grande <segmentation/> il gatto cammina <unsure> si è perso </unsure> e il gatto grande dormiva e poi il gatto cade e svegliò il gatto grande <segmentation/> fine (09-1P)

La morphologie normée est rétablie (exemple (15) à (18)).

(15) il entendu (94, CE2)

[normalisation] il entendit

(16) SI METE A PIANGRE SVEIO UNA MAMMA DI GATO LO VISE E LO ADOTO (11-1P)
[normalisation] si mette a piangere e svegliò una mamma di gatto <omission type=”pronom”/> lo vide e lo adottò.

(17) elle a ouvris (70, CE2)
[normalisation] elle a ouvert

(18) L A MAMMA E I SUOI PICOLINI DORMETAVANO (14-1P)
[normalisation] la mamma e i suoi piccolini dormivano

2. Sur le plan lexical

Lorsqu’un mot est omis :

a. S’il est identifiable de manière non ambigue, il est rétabli. C’est le cas, par exemple, du *ne* de négation (exemple (19) ou du déterminant la dans l’exemple (20)).

(19) mai il a pas vus le trautoir (562, CP)
[normalisation] Mais il n’a pas vu le trottoir.

(20) la maman le pran par bouche (200, CP)
[normalisation] la maman le prend par la bouche

(21) LA MAMMA GATO·PRENBI E RITRNAO A LETO. (15-1P)
[normalisation] la mamma gatto lo prende e ritornano a letto.

b. S’il est non identifiable ou que les possibilités sont multiples, mais que la catégorie du mot absent est identifiable sans ambiguïté, il est marqué par une balise <omission type="TYPE"/>, où TYPE est remplacé par la catégorie du mot absent (exemples (22) à (24)).

(22) il été t'une foi un petit chat se promené acoté de sa litière (1055, CP)
[normalisation] Il était une fois un petit chat <omission type="pronom"/> se promenait à côté de sa litière

(23) [normalisation] si mette a piangere e svegliò una mamma di gatto <omission type=”pronom”/> lo vide e lo adottò. (11-1P)

(24) POI PERO IL GATTO E CADUTO E PIANGEVA E FA SBEGLIARE E LA MAMMA SBEGLIA I PICOLI GATTI (44-1P)
[normalisation] però il gatto è caduto e piangeva e fa svegliare <omission type= “nom”/> e la mamma sveglia i piccoli gatti e lo va a prendere e poi dormono tutti.

Dans cet exemple, une reprise pronominale est nécessaire pour lier la première proposition « Il était une fois un petit chat » et la deuxième proposition « se promenait à côté de sa litière ». Plusieurs possibilités peuvent être envisagées :

- Il était une fois un petit chat qui se promenait à côté de sa litière
- Il était une fois un petit chat, il se promenait à côté de sa litière

Pour ne pas avoir à choisir entre ces différentes possibilités, on indique la nécessité de la reprise pronominale par la balise <omission type="pronom"/>. Les catégories retenues pour cette balise sont les suivantes : verbe, nom, pronom, adjectif, adverbe, préposition, déterminant.

Lorsqu'un choix lexical est jugé incorrect de manière non ambigüe, le mot concerné est remplacé (*les/la* dans l'exemple (25), *voie/voix* dans l'exemple (26)).

(25) Il était une fois dans une village perdue une sorcière qui travailler dur, elle se lever à 3h00 du matin et finisser à 15h00 de l'apré-midi. Un jour elle eu une idée pour *les* renplasser de son travaille (1111, CM1)
[normalisation] Il était une fois dans un village perdu une sorcière qui travaillait dur, elle se levait à 3h00 du matin et finissait à 15h00 de l'après-midi. Un jour elle eut une idée pour *la* remplacer de son travail

(26) il manquer juste de l'abiller comme elle est lui mettre sa voie... (1111, CM1)
[normalisation] il manquait juste de l'habiller comme elle et lui mettre sa voix...

Lorsqu'un mot est identifié comme surnuméraire de manière non ambigüe et de manière non stylistique, il est supprimé (la deuxième répétition de « le chat » dans exemple (27) ou « fatto » dans l'exemple (28)).

(27) le chat pannicai le chat *le chat* dit « c'est pour quoi » (63, CE2)
[normalisation] le chat paniquait <segmentation/> le chat dit « c'est pourquoi »

(28) MIAI *FATTO* INSEGNARE UNA LE ZIONE CHE DEVO STARE ATTENTO (03-1P)
[normalisation] <dialogue> mi hai insegnato una lezione <segmentation> che devo stare attento </dialogue>.

Lorsque la répétition de la forme « très » est vue comme un procédé stylistique d'accentuation (exemple (16)), elle est conservée. Elle trouve sa correspondance dans « molto » ou « tanto » en italien.

(29) le loup avait faim très très faim (1155, CM1)
[normalisation] le loup avait faim très très faim

Les constructions issues de l'oral, telles que :

- GN{1} + Pronom{2} + Verbe + GN{2} (exemple (12)) où {2} renvoie à une entité unique,
- GN {1} + Pronom{1} + Verbe (exemple (13)) où {1} renvoie à une entité unique

sont conservées.

(30) Il le mange le chat (exemple non attesté)
[normalisation] Il le mange le chat

(31) Le cha il mache (61, CP)
[normalisation] Le chat il marche

(32) LA MAMMA LO TROVA CHE PIANGE E LO PRENDE
[normalisation] la mama lo trova che piange e lo prende (02-1P)

4. Organisation des mots dans la proposition

En cas d'inversion de mots, l'ordre est rétabli.

(33) Et le pour remercier le chat (67, CE2)
[normalisation] Et pour le remercier le chat

(34) (le 30 juin) 2100 elle pue l'emené au travaille (1111, CM1)
[normalisation] (le 30 juin 2100) elle put l'emmener au travail

5. Ponctuation et segmentation en propositions

La majuscule en début de production et le point en fin de production sont rétablis.

Dans le cas où un texte s'ouvre sur un titre, la majuscule est rétablie pour le titre. Elle est rétablie également pour le premier mot qui suit le titre.

(35) IL GATTINO QUANDO LA MAMMA GATTA CON I SUOI GATTINI DORMIRONO UN MICETTO SI SVELIO ANDO IN GIRO PER LA CASA (28-1P)
[normalisation] <titre> Il gattino </titre> Quando la mamma gatta con i suoi gattini dormivano un micetto si svegliò e andò in giro per la casa

Lorsqu'un point de fin de phrase est présent sans majuscule, la majuscule est rétablie.

(36) le cha et tonbe sur le tapi. le cha pler (48, CP)
[normalisation] le chat est tombé sur le tapis. Le chat pleure

(37) VA SUBITO A PRENDRLO E LO PORTA DAGLIALIRI GATTINI. E VISERO TUTTI FELICI CONTNTI. (31-1P)
[normalisation] va subito a prenderlo e lo porta dagli altri gattini. E[CB8.1] vissero tutti felici contenti.

Lorsqu'une ponctuation est manquante et que le mot suivant contient une majuscule, le point est rétabli.

(38) Le petit chat sest fé male Le peti chat a male (1531, CP)
[normalisation] Le petit chat s'est fait mal. Le petit chat a mal

En cas d'interrogation directe, le point d'interrogation est rétabli (exemple (25)).

(39) - bon jour la sorsiere commens allé vous. (1927, CE2)
[normalisation] <dialogue> - Bonjour la sorcière comment allez-vous ? </dialogue>

(40) GATTO COMEAFARE ACADERE DA LARBRO (19-1P)
[normalisation] Gatto come ha fatto a cadere dall'albero?

En cas d'absence de segmentation en propositions, celle-ci est marquée par la balise <segmentation/> (exemple (26)). Cette balise peut remplacer un signe de ponctuation ou un connecteur.

(41) [normalisation] C'era una volta un gattino che andava a fare una passeggiata sull'erba <segmentation> c'era anche la sua mamma poi la mamma aveva salvato il gattino. (29-1P)

Quand une proposition est la conséquence d'une proposition précédente ou que la coordination (ou la virgule si 3 éléments) est la plus vraisemblable, on rajoute et (français) ou e (italiano) dans le cas où deux propositions sont liées et on n'utilise pas la balise <segmentation/>

(42) PERO E CADYTO SIEFATMALE LA MAMMA METRE STAFA DORMEDO NON PSAFA CHE IL SUOFILIO CHE E CADUTO E SIEFATMALE MA SIERARCOTA CHE E CADUTO (19-1P)
[normalisation] però è caduto e si è fatto male <segmentation> la mamma mentre stava dormendo non pensava che il suo figlio che è caduto e si è fatto male ma si era accorta che è caduto.

(43) il cur il tonbe partere il pleur il saute sur / un notre chat (1173, CP)
[normalisation] Il court, il tombe par terre et il pleure <segmentation/> il saute sur un autre chat.

La balise <segmentation> est donc réservée aux cas où une pause forte s'impose.

(44) [normalisation] C'era una volta un gattino che un giorno si allontanò dalla sua mamma <segmentation> il gattino camminò e cadde da un gradino poi incominciò a piangere <segmentation> la mamma si svegliò e lo recuperò poi lo portò dai suoi fratelli. (20-1P)

Dans cet exemple, plusieurs versions normées pourraient être envisagées, en voici quelques exemples :

- Il court. Il tombe par terre. Il pleure. Il saute sur un autre chat.
- Il court, il tombe par terre, il pleure et il saute sur un autre chat.
- Il court et il tombe par terre. Il pleure et il saute sur un autre chat.

Pour marquer ce manque de segmentation en propositions sans choisir un marqueur parmi les multiples possibilités, on utilise la balise <segmentation/>.

Les connecteurs valent pour marque de segmentation, la balise <segmentation/> ne sera donc pas ajoutée.

(45) Le chat sailoigne de sa maman est des bébé pan dans que sa maman dor ansuit il tonbe est le bébé pleur est sa maman lanten pui sa maman ramas le chton. (1665, CP)
[normalisation] Le chat s'éloigne de sa maman et des bébés pendant que sa maman dort ensuite il tombe et le bébé pleure et sa maman l'entend puis sa maman ramasse le chaton.

(46) IL GATINO CAMINA E CADE POI PIANGE POI LA MAMMA GATA LO PRENDE (01-1P)
[normalisation] Il gattino cammina e cade poi piange poi la mamma gatta lo prende.

6. Le cas de la virgule

En cas d'énumération, on rétablit la ponctuation usuelle, à savoir : tous les items sont séparés par des virgules, sauf l'avant-dernier et le dernier qui sont séparés par le connecteur « et ».

(47) alima gata et paulète alèrent se coucher (2978, CE1)
[normalisation] Alima, Gata et Paulète allèrent se couche
(48) mes mon chat qui été dans la marmite a eu telleman pere que il est sorti de la marmote a monter lechelle et courer dans touts sanse. (2978, CE1)
[normalisation] mais mon chat qui était dans la marmite a eu tellement peur qu'il est sorti de la marmite, a monté l'échelle et courait dans tous les sens.

(49) Il étai une foix dans la fôré un loup qui mangé tout le monde qui avai mange / troi cochon (1189, CE1)
[normalisation] Il était une fois dans la forêt un loup qui mangeait tout le monde et qui avait mangé trois cochons

L'énumération est le seul cas où une virgule peut être ajoutée.

En cas de virgule surnuméraire (exemple (26)) ou non attendue (exemple (27)), elle est supprimée.

(50) - C'est vrais aucun mal ?, vous ne manger plus les animaux ? (2978, CE1)
[normalisation] - C'est vrai aucun mal ? Vous ne mangez plus les animaux ?

(51) le chat noir c'est, enfuit de sa cage. (70, CE2)
[normalisation] le chat noir s'est enfui de sa cage.

(52) E SIE FATTO MALLE E, LA MAMMA SIE SVEGLIATA (50-1P)
[normalisation] e si è fatto male e la mamma si è svegliata

En cas de virgule marquée par un signe de ponctuation non valide (exemple (33)), elle est rétablie.

(53) Et un beau jour comme dans les dessins animés. Les robots se sont rebellé. (1129, CE2)

[normalisation] Et un beau jour comme dans les dessins animés, les robots se sont rebellés.

Dans tous les autres cas, on n'ajoute aucun signe de ponctuation.

7. Dialogues

Les propos des personnages sont encadrés par les balises <dialogue> (début) </dialogue> (fin). Chaque nouveau tour de parole est encadré par ces balises.

(54) […] le chat dit : Bonjour méchante et orible sorciere j'ai apris que tu allais dominer le monde. Oui c'est sa et comment t'appelle tu je m'appelle jean - pomme de terre. Sa te dit con fase équipe ensemble […] (1336, CM1)

[normalisation] […] le chat dit : <dialogue> Bonjour méchante et horrible sorcière j'ai appris que tu allais dominer le monde. </dialogue> <dialogue> Oui c'est ça et comment t'appelles-tu ? </dialogue> <dialogue> Je m'appelle Jean-Pomme de terre. </dialogue> <dialogue> Ça te dit qu'on fasse équipe ensemble ? </dialogue> […]

(55) LA·MAMMA SIÉ SVEIATA E A DETTO STAI ATTENTO TI SEI FATTO MALE SI UN' PO T EL'O DETTO CHE DEVI STARE· ATTENTO MIAI FATTO INSEGNARE UNA LE ZIONE CHE DEVO STARE ATTENTO. (02-1P)

la mamma si è svegliata e ha detto <dialogue> stai attento ti sei fatto male </dialogue> <dialogue> sì un po' </dialogue> <dialogue> te l'ho detto che devi stare attento </dialogue> <dialogue> mi hai insegnato una lezione <segmentation> che devo stare attento </dialogue>.

Les ponctuations de dialogue existantes sont conservées, mais les ponctuations manquantes, comme la virgule, le point, le double-point, les guillemets ou encore les tirets, ne sont pas rétablies, sauf lorsqu'une marque double (ex. guillemets) est incomplète ou mal positionnée (exemple (33)), la deuxième marque est rétablie.

(56) il a dit « je vais te manger. Il prend (exemple non attesté)

[normalisation] il a dit <dialogue> « je vais te manger » </dialogue>. Il prend

(57) elle a dit «» je vait le raiqupérait. est anfin elle a réqupérait con chat. (937, CE1)

[normalisation] elle a dit <dialogue> « je vais le récupérer » </dialogue>. Et enfin elle a récupéré son chat.

En cas de proposition incise à l'intérieur d'un dialogue, la balise <dialogue> est fermée puis rouverte.

(58) - Mais quesque tu fais la dedans dit le loup sort vites ! (exemple non attesté)

[normalisation] <dialogue> - Mais qu'est-ce que tu fais là-dedans </dialogue> dit le loup <dialogue> sors vite ! </dialogue>

Les pensées ou idées des personnages sont également encadrées de balises.

(59) - elle eu une idée et si je me transphorma en chat !!! (1111, CM1)
[normalisation] Elle eut une idée <dialogue> et si je me transformais en chat !!! </dialogue>

La balise <dialogue> implique une segmentation, la balise <segmentation> est donc inutile dans le voisinage de la balise <dialogue>.

(60) et lui dit tu ma manqué (1337, CM1)
[normalisation] et lui dit <dialogue> tu m'as manqué </dialogue>

La majuscule est rétablie après un tiret marquant un nouveau tour de parole dans un dialogue.

(61) - ben je vais voir si le crocette son meilleur chez les voisin ! (1337, CM1)
[normalisation] <dialogue>- Ben je vais voir si les croquettes sont meilleures chez les voisins ! </dialogue>

8. Notations spécifiques

Lorsqu'un mot ou une phrase est non achevé en fin de production, on le ou la marque à l'aide de la balise <nonfini/>.

(62) et di maou maou qui r (1228, CP)
[normalisation] et dit « miaou miaou » qui r<nonfini/>

Les segments dont le sens ne peut être compris sont marqués à l'aide de la balise <incomprehensible/>.

(63) l ti chat pas c ese bébe cha (1363, CP)
[normalisation] le petit chat <incomprehensible/> bébé chat

Les titres en début de production sont encadrés par la balise <titre>.

(64) Le chat Il était une fois un chat (1845, CE2)
[normalisation] <titre>Le chat</titre> Il était une fois un chat

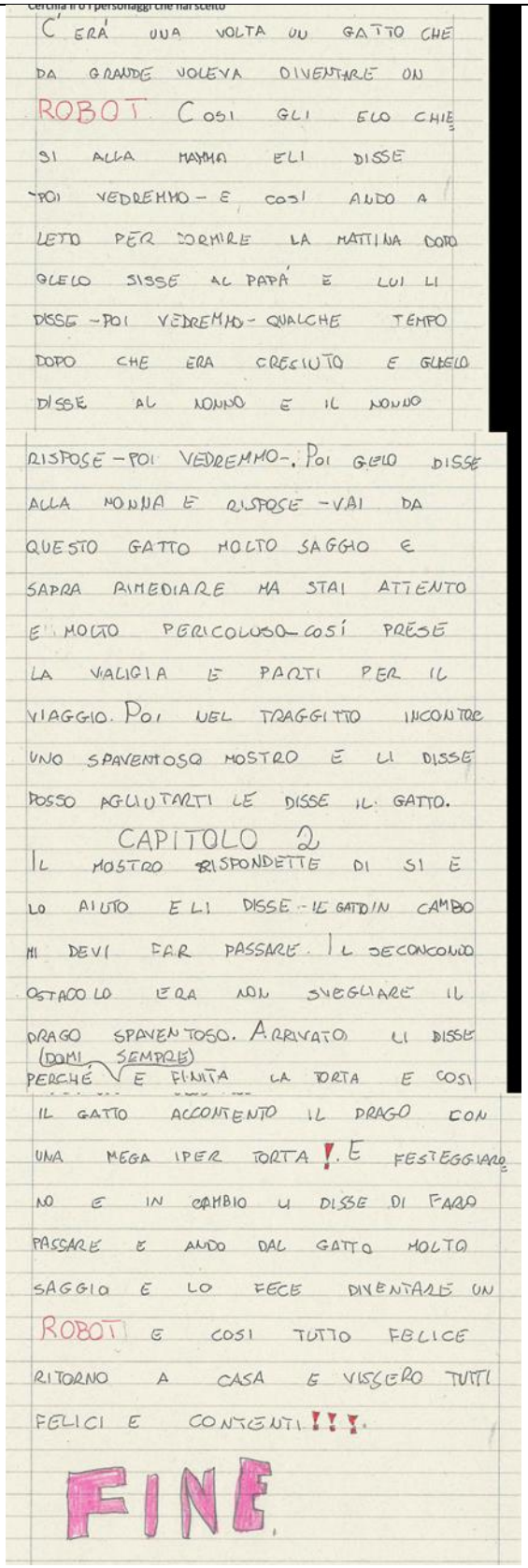
9. En cas de doute

En cas de doute du normalisateur, le passage concerné peut être encadré de la balise <unsure> … </unsure>.

(65) il marche sur au bor de la plake. (1352, CP)
[normalisation] Il marche <unsure> au bord </unsure> de la plaque.

Annexe F. Test de transcription automatique par OCR

Test effectué sur un texte du corpus italien (CE1, 747)



C'ERÁ UNA VOLTA UN GATTO CHE DA GRANDE VOLEVA DIVENTARE UN ROBOT. COSI GLI ELO CHIE SI ALLA MAMMA ELI DISSE "POI VEDREMMO - E COSI ANDO A LETTO PER DORMIRE LA MATTINA DOPO GLELO SISSE AL PAPÁ E LUI LI DISSE - POI VEDREMMO - QUALCHE TEMPO DOPO CHE ERA CRESIUTO E GLELO DISSE AL NONNO E IL NONNO RISPOSE - POI VEDREMMO -. POI GELO DISSE ALLA NONNA E RISPOSE - VAI DA QUESTO GATTO MOLTO SAGGIO E SAPRA RIMEDIARE MA STAI ATTENTO E MOLTO PERICOLOSO - COSÍ PRESE LA VIALIGIA E PARTI PER IL VIAGGIO. POI UEL TRAGGITTO INCONTRE UNO SPAVENTOSO MOSTRO E LI DISSE POSSO AGIUTARTI LE DISSE IL GATTO.

CAPITOLO 2

IL MOSTRO RISPONDETTE DI SI E LO AIUTO E LI DISSE - IE GATTO IN CAMBO MI DEVI FAR PASSARE IL DECONDO OSTACOLO ERA NON SVEGLIARE IL DRAGO SPAVENTOSO. ARRIVATO LI DISSE (DORMI SEMPRE) PERCHÉ È FINITA LA TORTA E COSI IL GATTO ACCONTENTO IL DRAGO CON UNA MEGA IPER TORTA !. E FESTEGGIARE NO E IN CAMBIO U DISSE DI FARO PASSARE E ANDO DAL GATTO MOLTO SAGGIO E LO FECE DIVENTARE UN ROBOT E COSI TUTTO FELICE RITORNO A CASA E VISSERO TUTTI FELICI E CONTENTI !!!.

FINE

LUI UNITO. UT PETSUTIabpI LUTTO TIdI SUCITO

Coi oi
DALL EPM VOR DIVEVTRE VU

ROBON Cosi GELO SAIL

Si ALLA: pun ELI pisshi t1
SO) VEDREMMO — E così ALDI A

Lema Pao Sole LA HAT So

Serio ASI Lol
DISSE —Pol VERREMM- QUALCHE TEMPO |

DOPo — CHE ERA CRECIUIO E GUEIO

Disk Rui Au paoli.

IMISPOGE — Tor vengeMMO-. Pai CA iii

ALLA MObUA ose Ve Da
QUESTO GATTO MOLTO SAGGIO €

SAPLA BIMEDIALE MASIAI ATTENTO.
El iiono ego oaò osi posi.
si

(i SALCIA: i PAS LER-iL

VIAGGIO. PES (DEL _TRAGELTIO \NCOL TRE
|

n

Na SpaEntoso MOSTO È Lo Nssbl.

Mosso AGUUTALTI LE DISSE iL. GATTO.

	CARCFOLOCL2, lo... Lo Aiuto E Li DISSE IE campIN CAMBO Ni bevi Eae pASsSmME |lL nENOoLD ARCO ELA KML TN ACION VBA | PRAGO SPAVER TOSO. A aRiratO — Li DISSE SEMPAE) | E PENNA CA TORTA E COS = —- wie SI. to er — i GATTO ACCONTENTO \LPRAGO cou | UNA een \Led ol Y. È: PESTEGGIAM no E IN coMBIO bi DPISCNIIITAA9 | FAURE E AUDI VAL GATTO MOLTO _--. (C Lo CEcE. PINENTAZE UN | O RoBdLi i cos) TUÙTIO Felice | RITORNO. A cASR — ese i iii.

Annexe G. Guide d'annotation en continuité référentielle du corpus Scolinter

Inspiré des guides pour l'annotation des corpus *RésolCo* et *DEMOCRAT*

1. Introduction

Ce guide définit le processus d'annotation en continuité référentielle des textes du corpus Scolinter : ceci est un corpus longitudinal comparable de productions écrites d'élèves du CP au CM2 en français, italien et espagnol (Ponton et al., 2021). Les annotations sont effectuées sur les versions normalisées des textes du corpus. Par normalisation, nous entendons ici une réécriture du texte proche de la norme de la langue cible , tout en restant au plus près du texte originel.

Ce guide est inspiré des méthodologies d'annotation déjà utilisées sur les corpus *RésolCo* (Garcia-Debanc et al., 2021)[1] et *DEMOCRAT* (Landragin, 2016b)[2].

2. Définition de la continuité référentielle et des chaines de référence

On définit comme chaine de référence **un ensemble d'au moins trois expressions linguistiques qui font référence à la même entité dans l'univers d'un texte donné** (Charolles, 1988c; Corblin, 1995; Poesio, 2016; Schnedecker, 2021; Schnedecker & Landragin, 2014). Pour notre étude, nous avons entrepris l'annotation de toutes les expressions faisant référence à une entité au sein du texte à analyser. Ceci inclut les mentions uniques (également appelées singletons), les paires de mentions anaphoriques ou non, ainsi que les chaines comportant au moins trois expressions. La notion de continuité référentielle est en lien avec celle de *chaine de référence*, dans le sens où les chaines de référence peuvent faire partie de ces éléments, de ces « (…) ***formes linguistiques utilisées par les élèves pour construire la cohésion référentielle*** » (Garcia-Debanc *et al.*, 2021, p. 5). Ces formes linguistiques indiquent habituellement un personnage ou un élément précis de l'histoire racontée.

3. Objectif de l'annotation

L'objectif est d'annoter toutes les mentions qui se réfèrent à un personnage ou une entité animée spécifique dans l'univers du texte analysé. Cela inclut toutes les expressions linguistiques qui les désignent, définies selon les critères énoncés ci-après. Par exemple, dans le texte suivant, il faudra annoter toutes les expressions linguistiques qui font référence à la sorcière et au loup dans le texte : noms, noms propres, pronoms, etc.

Les sections suivantes du guide préciseront davantage le type d'expressions à annoter.

[1] Le guide d'annotation du corpus *RésolCo* est disponible au lien suivant https://hodaclm.github.io/resolco/

[2] Les livrables du projet *DEMOCRAT*, dont le guide d'annotation, sont disponibles au lien suivant https://www.lattice.cnrs.fr/democrat/livrables.html

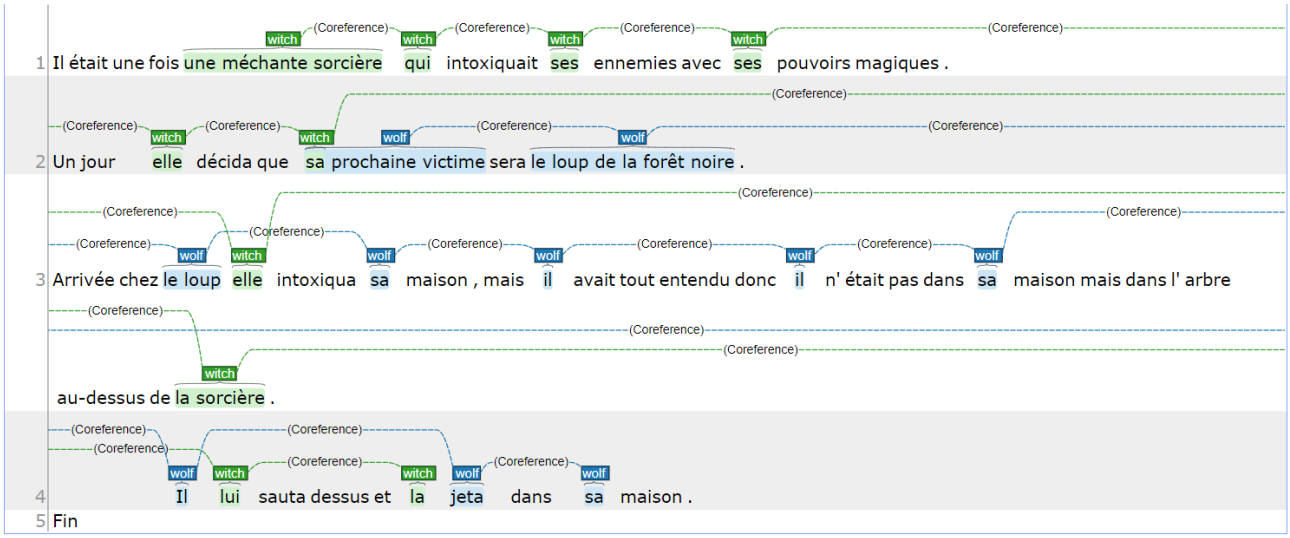


Figure 1. Texte CM2 S1143 fr (normalisation annotée sur INCEpTION).

4. Types de mentions annotées et règles pour l'annotation

4.1. - Définition des référents à annoter

L'annotation de la continuité référentielle dans ce projet cible tous les maillons relatifs aux **quatre personnages évoqués par la consigne** (Figure 2), ainsi que les maillons faisant référence à d'autres entités animées présentes dans les textes analysés (personnes, animaux, objets animés). Leur présence est annotée dès leur première occurrence, y compris dans le titre des textes.

Les quatre personnages mentionnés dans la consigne sont identifiables comme *chat*, *robot*, *sorcière* et *loup*. Ces personnages peuvent être désignés par leur nom commun correspondant (chat ou chatte, robot, sorcière, loup) lors de leur première mention, ou par un hyperonyme, un hyponyme, un synonyme (minou, magicienne), ou encore par un nom propre, avec ou sans le nom commun associé.

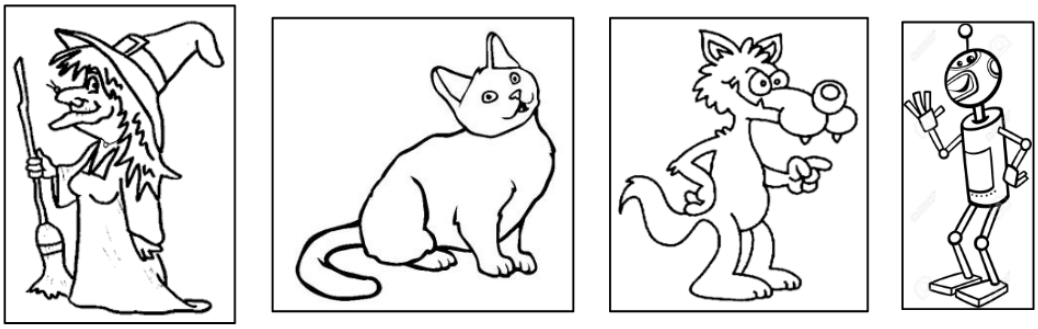


Figure 2. Consigne des productions écrites du corpus Scolinter du CE1 au CM2.

La présence des personnages sollicités par la consigne ainsi que des autres possibles entités animées que l'on retrouve dans les textes est annotée *dès sa première mention*, y compris dans le titre du texte.

Cette première mention doit permettre **d'identifier clairement le personnage indiqué et doit être reliée de manière non ambiguë à la chaine du personnage correspondant**. Cela inclut également les cas de cataphore pronominale si le rattachement à la chaine est jugé clair (voir exemple de la chaine de ext1 dans la Figure 3). De même vaut pour les personnages « autres », dont la présence peut être annoté dès que le rattachement est assez explicite, de manière anaphorique ou cataphorique.

En cas de rattachements ambigus, voir section 4.7 du guide.

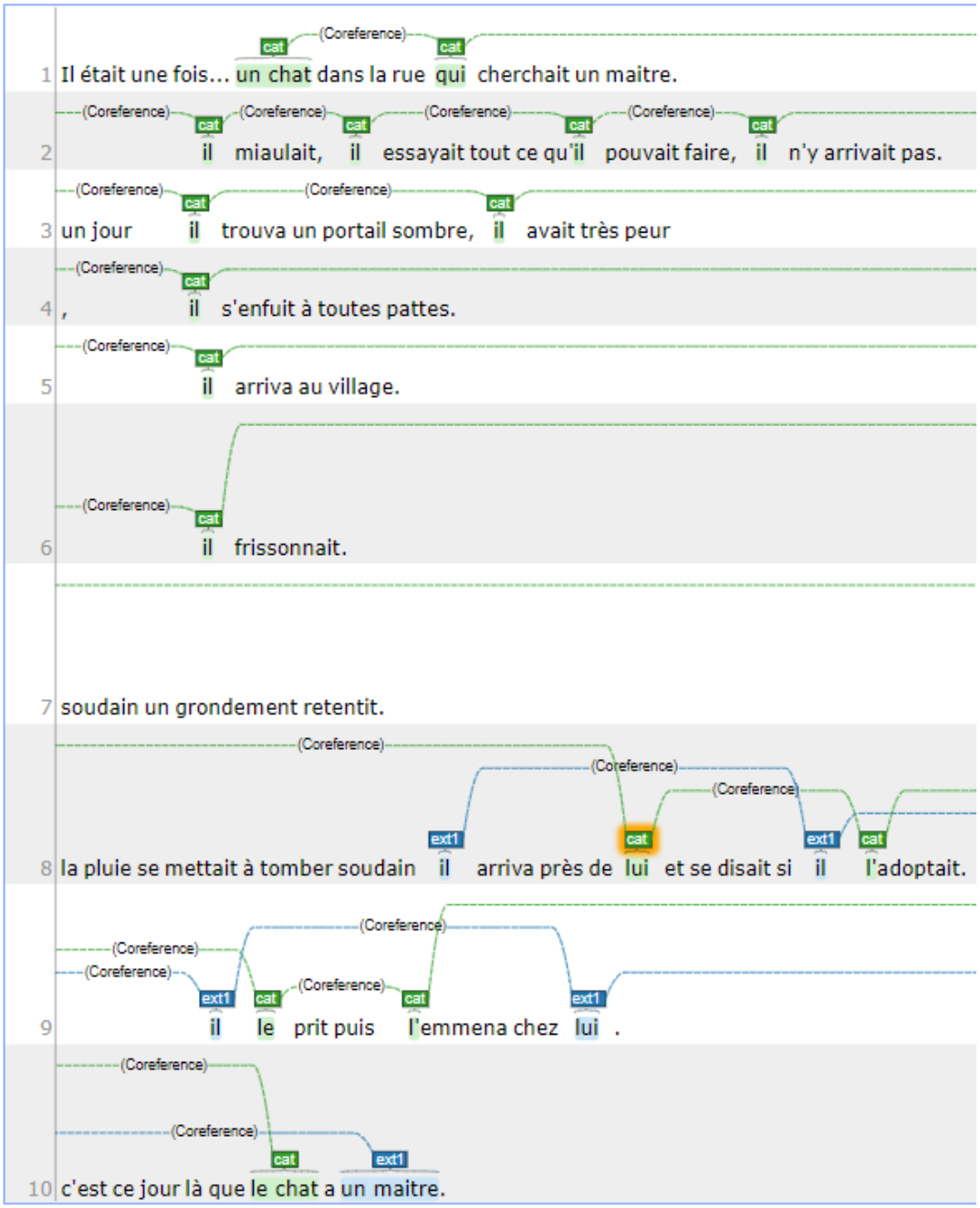


Figure 3. Texte CE1 1865 fr. (normalisation annotée sur INCEpTION).

Dans l'exemple de la figure 2, tiré du corpus français, on annote les maillons de ***chat*** ainsi que les maillons de ***maître (ext1)***. Attention : le premier maillon de maître (ligne 1) *n'est pas annoté* puisqu'il référence à un maître générique, et non au personnage mentionné par la suite dans le texte. Dans ces cas, on n'annote pas le maillon qui fait référence à l'entité en tant que classe.

On annote aussi les **singletons** (mentions qui indiquent seulement une fois un référent donné) et les **anaphores** (deux mentions qui indiquent un référent) afin d'effectuer par la suite des statistiques sur la présence de ces éléments, si ces mentions font partie du périmètre des entités que l'on annote selon les principes expliqués ci-dessus.

Pour rappel, on annote **toutes** les expressions qui font référence à un personnage donné dans un texte donné, selon les critères qui seront explicité ci-dessous.

4.2. - Etiquettes

Le jeu d'étiquettes proposé sert à ***annoter des entités ou personnages animés présents dans les textes***. On fait référence ici à des **entités animées, y compris des objets personnifiés si présents dans les textes**. Des tags ou étiquettes relatives aux quatre entités représentées dans la consigne seront disponibles dans le schéma d'annotation proposé sur INCEpTION : **cat**, **wolf**, **witch** et **robot** pour les personnages de la consigne[3] ; des autres tags seront également disponibles : **ext1**, **ext2** (puis **ext3**, **ext4**… selon les exigences du texte annoté) pour annoter ces entités animées qui ne sont pas indiquées par la consigne. Les annotateurs pourront rajouter si besoin des étiquettes.[4]

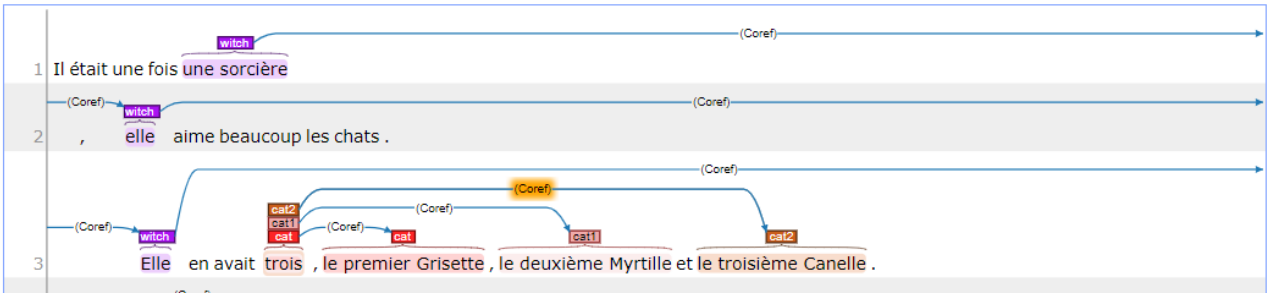


Figure 4. Texte CE2 1851 fr.

Dans le cas où plusieurs entités mentionnées par la consigne sont présentes dans le texte, on fera référence à ces entités comme **cat**, **cat1**, **cat2**… De même pour les autres entités dotées d'étiquette (ex. **witch** puis **witch1** etc.).

4.3. - Types de mentions

Dans notre corpus, nous avons pu identifier différents types de mentions à annoter. Les mentions sont des expressions langagières susceptibles de représenter le référent dans le texte.

[3] Vu que le projet inclut trois langues différentes, les étiquettes que l'on va utiliser seront en anglais pour ne pas choisir entre les trois langues du projet celle à adopter pour les étiquettes.

[4] Si cela était vrai lors de la première campagne, le tagset a été fermé lors de la deuxième campagne.

Les syntagmes nominaux

Les syntagmes nominaux (SN) sont toujours annotés. En font partie : les expressions nominales définies (le chat), indéfinies (un chat), démonstratives (ce chat) et possessives (son chat). Le maillon inclut le groupe nominal et l'ensemble des modifieurs antéposés ou postposés au nom ou au groupe nominal, comme :

1. les **articles** : dans le cas des expressions nominales définies, *l'article* qui précède le syntagme nominal fait toujours partie du maillon. Comme en français les articles contractés sont à inclure dans les maillons, en italien aussi les articles contractés font partie du syntagme (*del* gatto, *della* strega, *nel* robot) ;
2. les **adjectifs** postposés ou antéposés qu'on considère font partie du syntagme nominal (*une **petite** fille, un chat **mignon**, la sorcière **méchante**, les **deux** compagnons*) ;
3. les **adjectifs issus de participes** avec leurs éventuels compléments ;
4. les **syntagmes prépositionnels**, quelle que soit la préposition utilisée (une femme *de 80 ans*) ;
5. des appositions (*une sorcière **nommée Violaine**, una bambina **di nome Marcella***), dans ce cas-là, on annote le syntagme entier comme maillon unique ;
6. des noms propres quand ils suivent un nom commun de près. Ex. ***le chat Victor*** c'est un maillon unique ;
7. les **possessifs** (cf. Figure 5, par exemple [**[*son*]** *chat*] est un maillon unique, où **son** pourrait être aussi annoté de manière indépendante s'il représente un autre référent qu'il faut annoter dans le texte cf. *Les déterminants possessifs*).

Dans ce périmètre, ne rentrent pas « les propositions relatives, qui constituent un [type de] maillon distinct. » (guide RésolCo, p. 4)

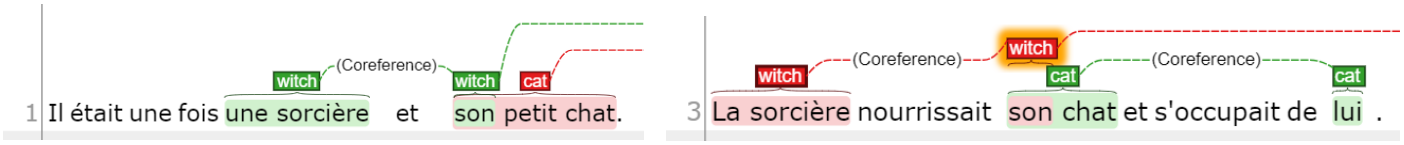


Figure 5. Texte CE2 107. Exemple d'annotation des maillons contenants un possessif.

Ne rentrent pas dans la limite du maillon de type syntagme nominal :

- **Les prépositions antéposées au syntagme** (**sauf les article contractés**), comme dans l'exemple dans la Figure 6, où la préposition « **de** » ne fait pas partie du maillon

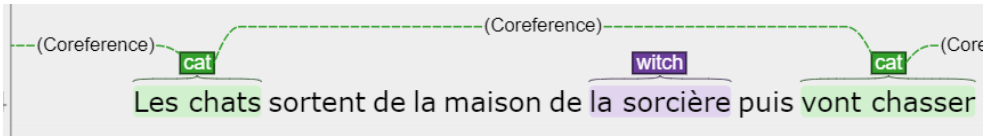


Figure 6. Texte CE2 1851, phrase 4.

Les noms propres

Les noms propres (NP) sont toujours annotés. Par exemple, nous considérons [X2] comme un maillon de type nom propre dans le texte illustré dans la Figure 7.



Figure 7. Texte CE2 202.

Un maillon NP peut être composé de plusieurs éléments. Si précédé d'un titre (Monsieur, Madame …), ce titre rentre dans les limites du maillon. Si le NP est introduit par un pronom relatif, le NP est annoté en plus du pronom relatif comme maillon à part.

Si le NP suit de près un syntagme nominal (SN), on l'annote en tant que partie du syntagme nominal.

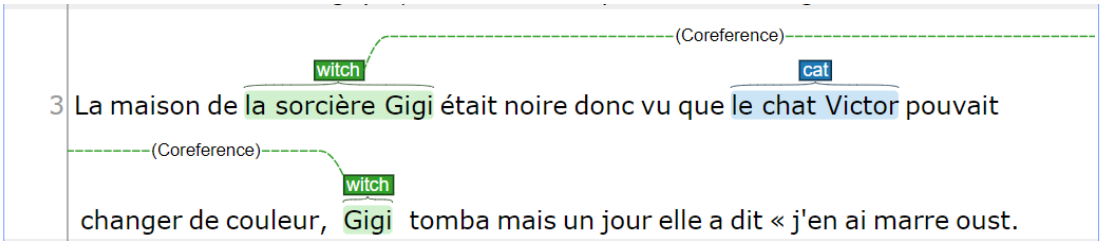


Figure 8. Texte CE2 212.

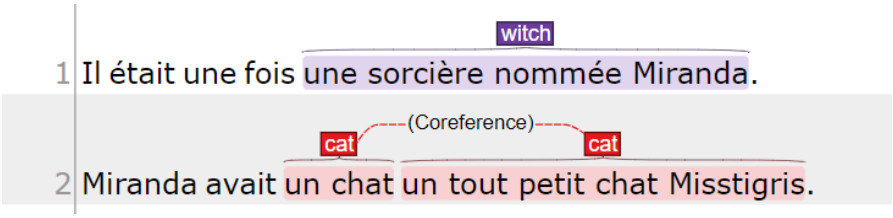


Figure 9. Texte CE2 2536.

Dans le cas où on retrouve deux SN adjacents, dont le deuxième commence par article, préposition ou déterminant, on considère les deux maillons comme séparés.

Les pronoms personnels, sujet ou objet et les pronoms relatifs

Les pronoms sont annotés quand ils font référence à un des référents qu'on annote. On annote :

1. les **pronoms personnels** sujet et objet (il, elle, lui …) ;
2. les **pronoms démonstratifs** (*ce*, *celui*, *celui-ci, celle-la*… en fr. ; *questo, quello, ciò, stesso*… en it.),
3. les **pronoms indéfinis** (*chacun* en fr., *ciascuno*, *chiunque* in it.),
4. les **pronoms corrélés** (*c'est à lui, l'un… l'autre* en fr. ; *l'uno… l'altro* en it.).
5. les **pronoms relatifs** qui modifient un syntagme nominal (ex. un chat ***qui*** s'appelle Victor) – on annote le pronom relatif seulement « lorsque le groupe nominal composant un maillon

est modifié par une proposition relative. Ce nouveau maillon est composé par le seul pronom relatif » (*guide Résolco*).

Les pronoms relatifs : en français on annote des maillons composés par un pronom relatif tels que **qui**, **dont**, **que**, **lequel, laquelle** lorsqu'ils modifient un maillon de type groupe nominal.

On n'annote pas :

1. les pronoms réfléchis clitiques (ex. Il **[se]** lave.) ;
2. les pronoms inclus dans des expressions figées ;
3. les pronoms impersonnels (ex. **[Il]** pleut.) ;
4. les pronoms négatifs cad les pronoms qui e désignent aucun élément (ex. **[Aucun]** de nous a pensé à prendre les clés.).

En revanche, les pronoms clitiques **objets** qui ne sont pas réfléchis et les réfléchis non clitiques sont annotés ex. Il **[le]** lave. Il se lave **[lui-même]**. sont annotés

Les déterminants possessifs

Dans le guide RésolCo, des « les expressions possessives référent au possesseur (…) seul le déterminant doit être annoté » (guide RésolCo) dans la chaine du possesseur mais vu qu'il modifie, attribue un trait en plus au référent qui est possédé, on le considère au même temps comme faisant partie du maillon nominal en tant que modifieur, comme dans l'exemple suivant.

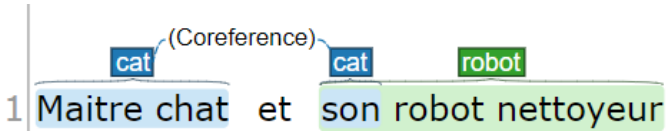


Figure 10. Texte CE2 214.

Si on retrouve dans le texte un déterminant possessif où le possesseur fait partie des référents à annoter, on relie le déterminant à la chaine du possesseur. Si le déterminant fait partie du syntagme nominal d'un référent qu'on annote, on le comprend aussi dans le syntagme nominal annoté, comme dans les exemples dans les figures 4, 5 et 10. Dans la figure 10 le token « ***son »*** est annoté et relié à la chaine de **cat** et fait aussi partie du maillon ***« son robot nettoyeur*** » annoté comme **robot**).

Anaphore zéro

Il faut annoter les verbes qui marquent la présence du sujet dans les cas d'anaphore zéro, c'est-à-dire les verbes dans des **phrases coordonnées** ou les **verbes à l'impératif** en français, dans des **phrases principales où le sujet n'est pas explicité** ou dans des **coordonnées** et dans des **verbes à l'impératif** en italien.

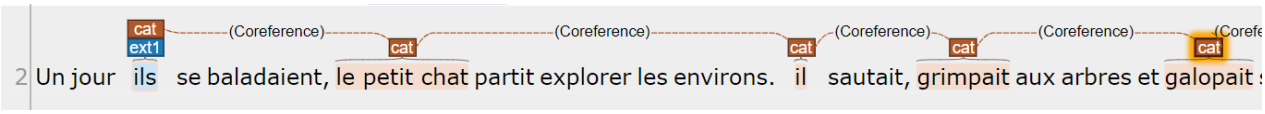


Figure 11. Texte CE2 841.

On est amené à annoter tout le groupe verbal, ce qui peut inclure les marques de négation, les auxiliaires éventuels, les pronoms réfléchis et certains pronoms clitiques dans les limites du maillon annoté.

4.4. Les référents multiples

Un référent multiple est une expression qui désigne un groupe d'individus à travers habituellement des expressions de forme plurielle (pronoms personnels au pluriel, SN pluriels ou SN collectifs). On annote les référents multiples comme maillons des différentes chaines desquelles ils font partie. Sur INCEpTION il est possible d'annoter plusieurs fois les mêmes tokens et d'y attribuer une étiquette différente pour chaque annotation.

Si l'expression inclut plusieurs référents annotés dans le texte, on peut empiler les annotations nécessaires à exprimer tous les référents inclus comme dans l'exemple dans la Figure 11.

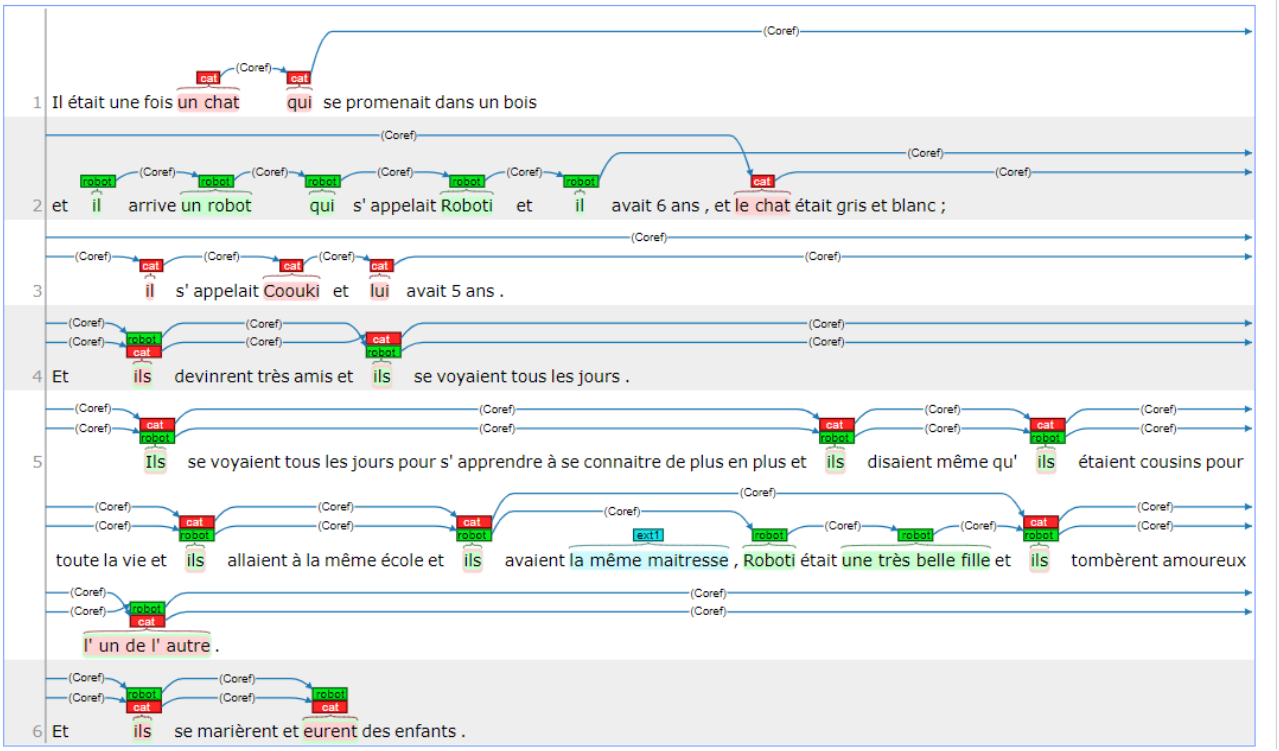


Figure 12. Texte de CE1 de l'élève 824.

4.5. Les référents évolutifs

Comme dans le cas du traitement fait dans le corpus Democrat, on annote un référent évolutif en incluant les maillons qui ont « évolué » dans la chaine du référent originaire, c'est-à-dire dans la chaine du premier référent avant avoir subi un changement. Ex. si un chat est transformé en loup et ensuite en robot, les maillons « loup » et « robot » vont faire partie de la chaine de référence du chat, et on va attribuer aux maillons l'étiquette **cat**. Dans l'exemple suivant, la sorcière est transformée en cuisse de poulet, donc on annote la deuxième expression en utilisant la même

étiquette utilisée pour le maillon de l'entité transformée.

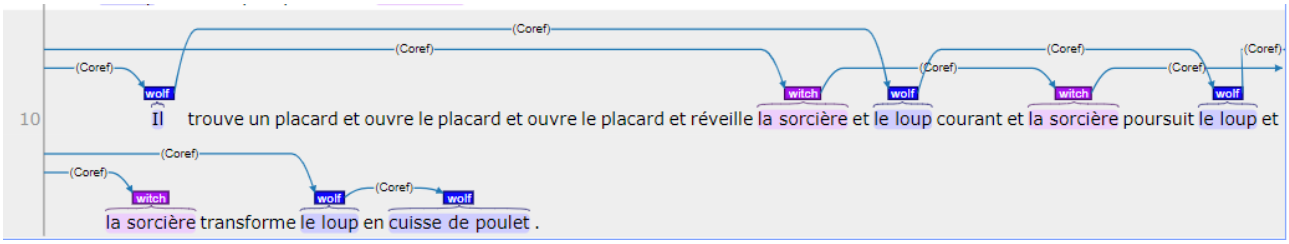


Figure 13. Texte CE2 1098.

4.6. Discours direct

Nous annotons les mentions selon les critères définis ci-dessus aussi dans les occurrences de discours direct. Les instances de discours direct sont habituellement signalées par le biais de ponctuation normée comme les guillemets, issue parfois de la phase de normalisation des textes. Dans le discours direct, on applique le principe d'identité et on étiquette selon l'entité à laquelle la mention fait référence. Dans l'exemple suivant, nous avons annoté seulement les maillons référents au personnage du loup, y compris à l'intérieur du discours direct.

4.7. Doutes dans le rattachement à une chaîne

Dans le cas de doutes dans le rattachement d'un maillon à une chaine définie, on vérifie son

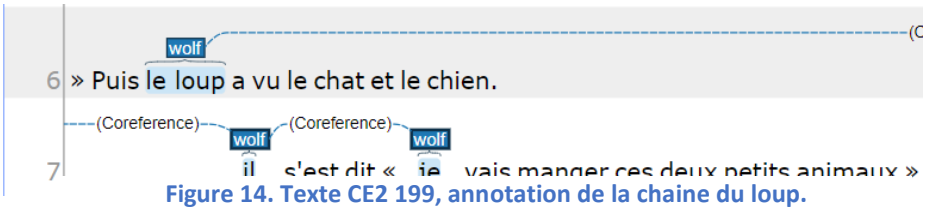


Figure 14. Texte CE2 199, annotation de la chaine du loup.

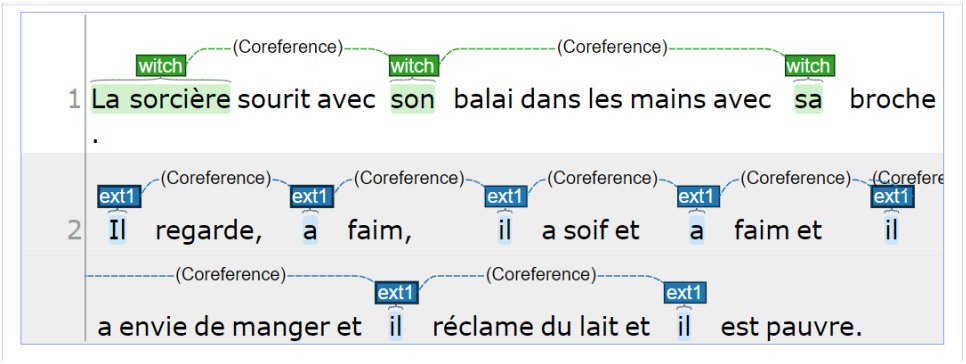


Figure 14. Texte CE1 211.

rattachement au dernier référent disponible, selon aussi des critères tels que la correspondance de genre et de numéro grammatical ou l'accord des traits. Si ce rattachement est incompatible, on constitue une nouvelle chaine qui puisse comprendre le maillon douteux et éventuellement les maillons suivants.

4.8. Référence générique, spécifique, particulière, anaphore associative et autres liens sémantiques (inspiré du manuel d'annotation de DEMOCRAT)

Dans des énoncés comme celui présent dans l'exemple dans la figure 16, on peut se retrouver face à un syntagme nominal qui renvoie à une même référence générique. Ces termes sont toujours considérés comme coréférents, même dans le cas de passage du singulier au pluriel ou changement de déterminant. Cependant, un référent générique ne peut pas être considéré comme coréférent avec un référent spécifique ou particulier.

Par exemple, dans ce cas, on n'annote pas « les chats noirs » car cette expression ne fait pas référence à des chats spécifiques mentionnés dans le texte, mais aux chats noirs de manière générale.

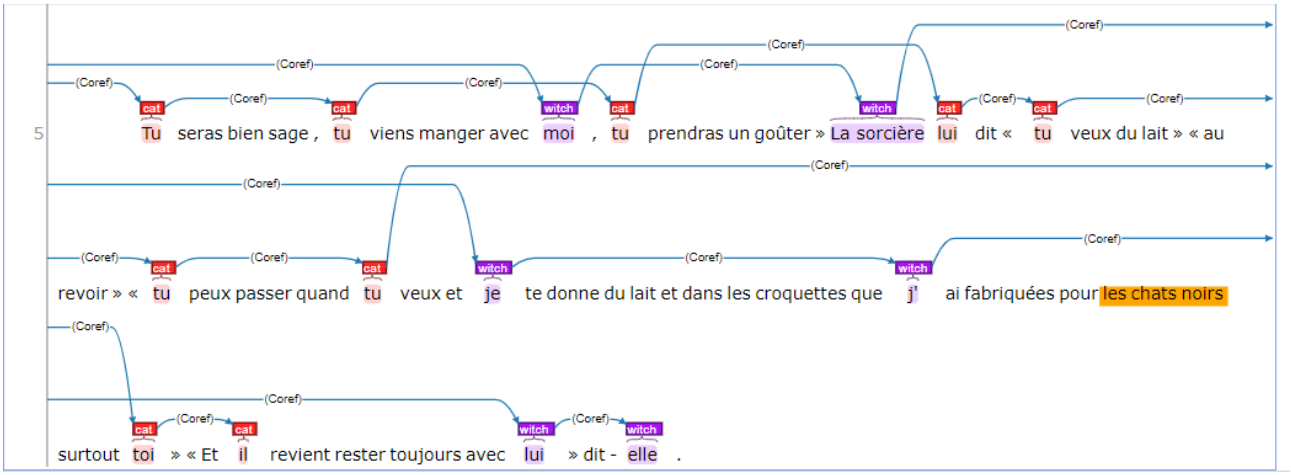


Figure 15. Texte CE1 S2986.

Bibliographie du guide

Charolles, M. (1988). Les plans d'organisation textuelle : Périodes, chaînes, portées et séquences. *Pratiques*, *57*(1), 3-13. https://doi.org/10.3406/prati.1988.1468

Corblin, F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Presses Universitaires de Rennes. https://jeannicod.ccsd.cnrs.fr/ijn_00550962

Garcia-Debanc, C., Rebeyrolle, J., & Ho-Dac, L.-M. (2021). La continuité référentielle dans le corpus RÉSOLCO : Méthode d'annotation et premières analyses. *Langue francaise*, *211*(3), 99-114. https://doi.org/10.3917/lf.211.0099

Klie, J.-C., Bugert, M., Boullosa, B., Castilho, R. E. de, & Gurevych, I. (2018). The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation. *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, 5-9. http://tubiblio.ulb.tu-darmstadt.de/106270/

Landragin, F. (2016). Description, modélisation et détection automatique des chaînes de référence (DEMOCRAT). *Bulletin de l'Association Française pour l'Intelligence Artificielle*, *92*, 11.

Poesio, M. (2016). Linguistic and Cognitive Evidence About Anaphora. In M. Poesio, R. Stuckardt, & Y. Versley (Éds.), *Anaphora Resolution : Algorithms, Resources, and Applications* (p. 23-54). Springer. https://doi.org/10.1007/978-3-662-47909-4_2

Ponton, C., Gutiérrez-Caceres, R., Teruggi, L., Farina, E., Brissaud, C., & Wolfarth, C. (2021). Scolinter : Un corpus trilingue. L'exemple de la segmentation en mots. *Langue française*, *211*(3), 37-50. https://doi.org/10.3917/lf.211.0037

Schnedecker, C. (2021). *Les chaînes de référence en français*. OPHRYS.

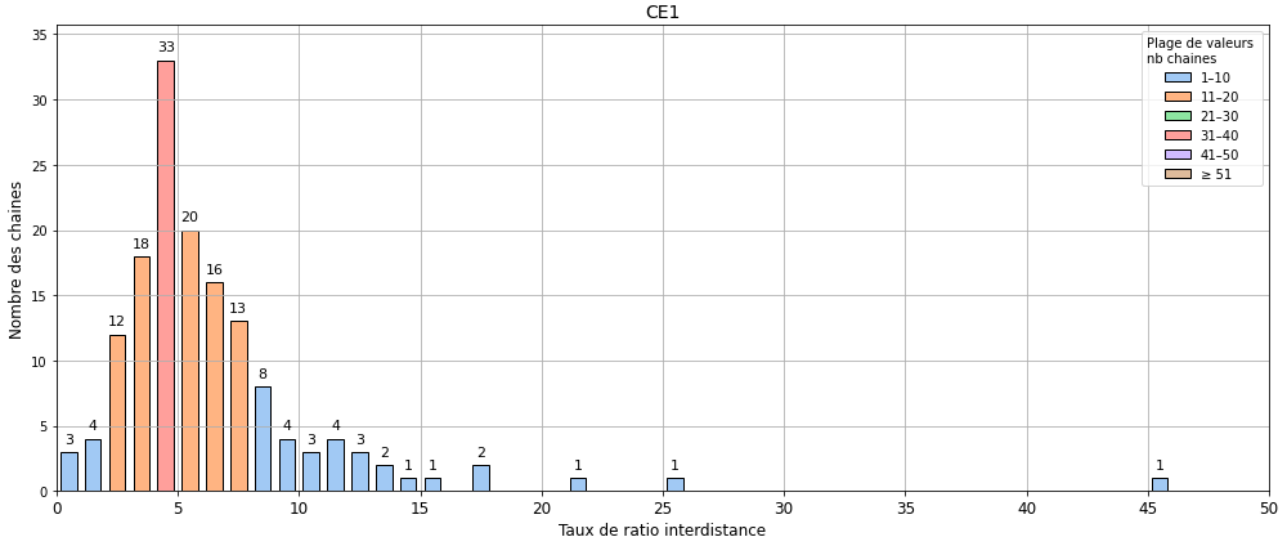
Schnedecker, C., & Landragin, F. (2014). Les chaînes de référence : Présentation. *Langages*, *195*(3), 3-22. https://doi.org/10.3917/lang.195.0003

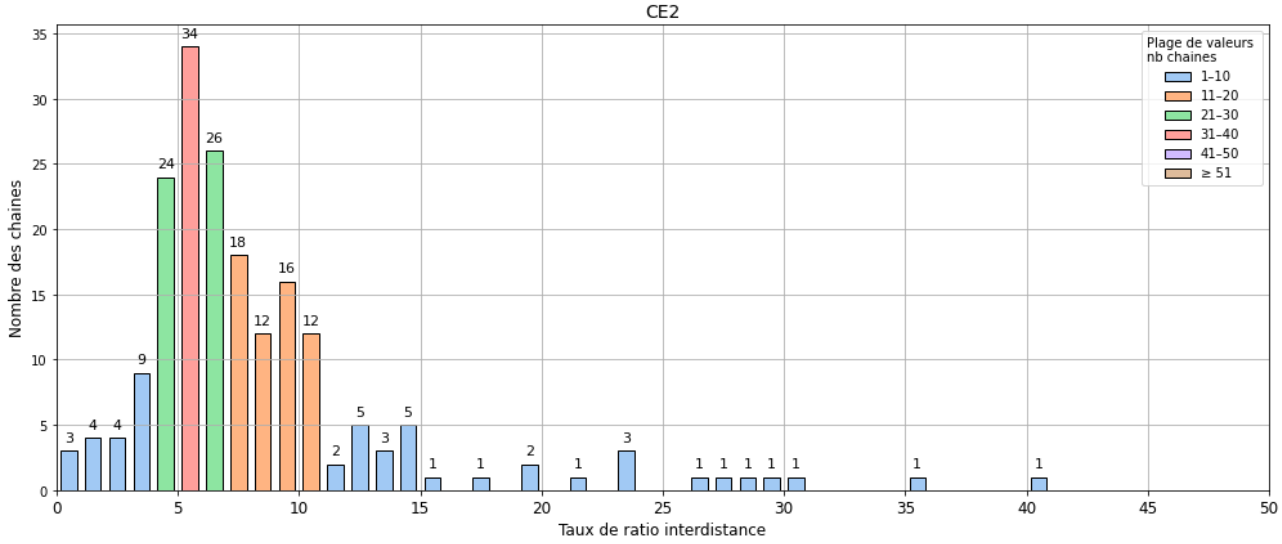
Widlöcher, A., & Mathet, Y. (2012). The Glozz Platform : A Corpus Annotation and Mining Tool. *Proceedings of the 2012 ACM Symposium on Document Engineering*, 171-180.

Annexe H. Statistiques

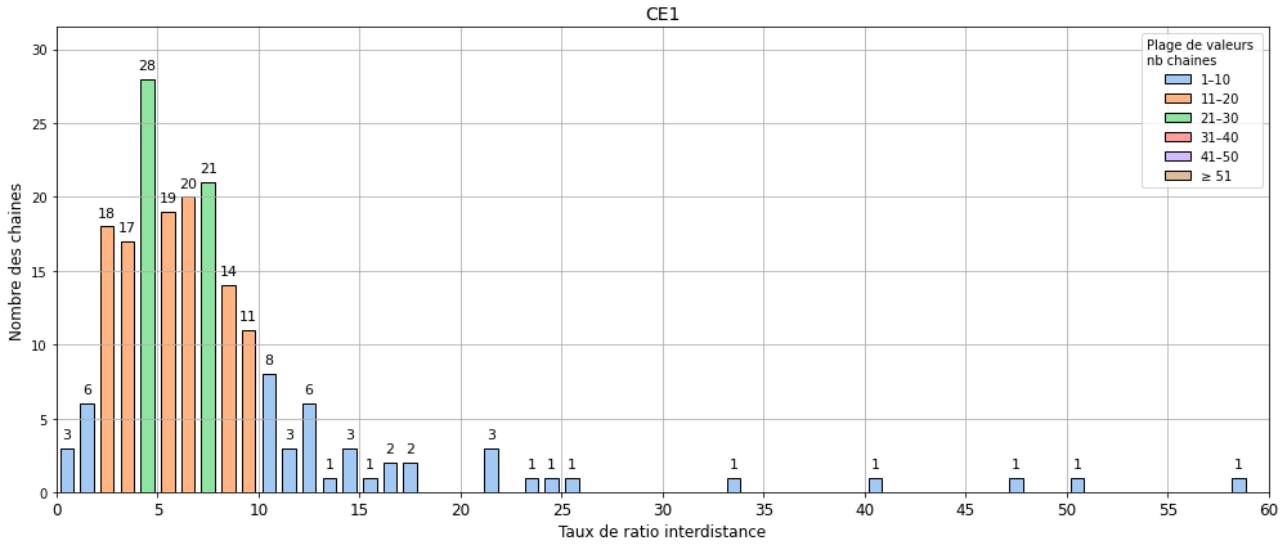
Distance intermaillonnaire – comparaison entre niveaux scolaires

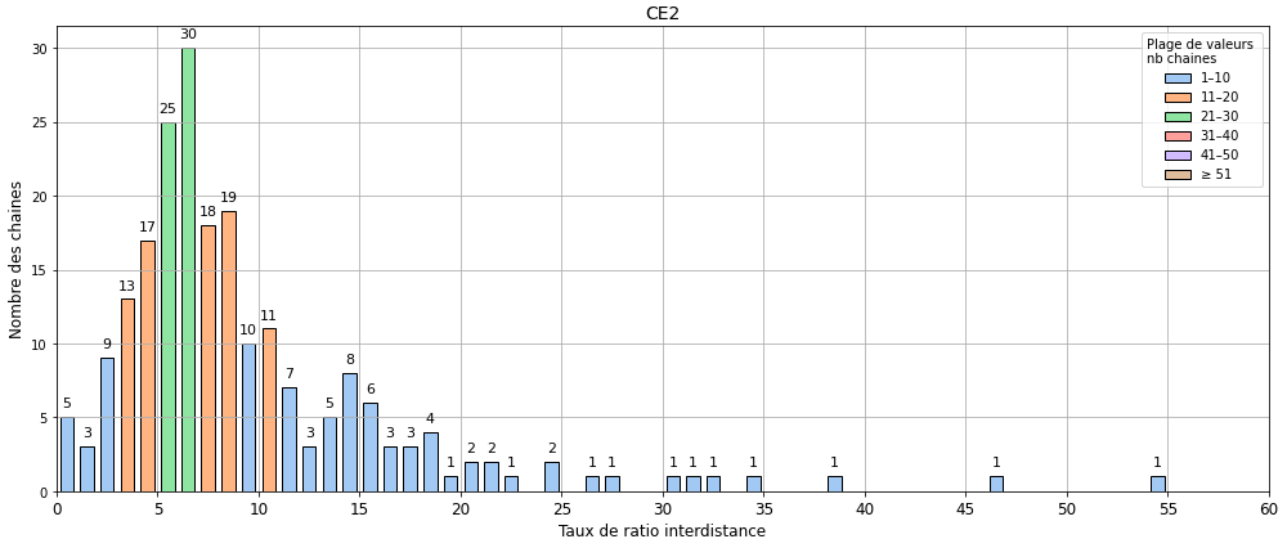
Corpus français





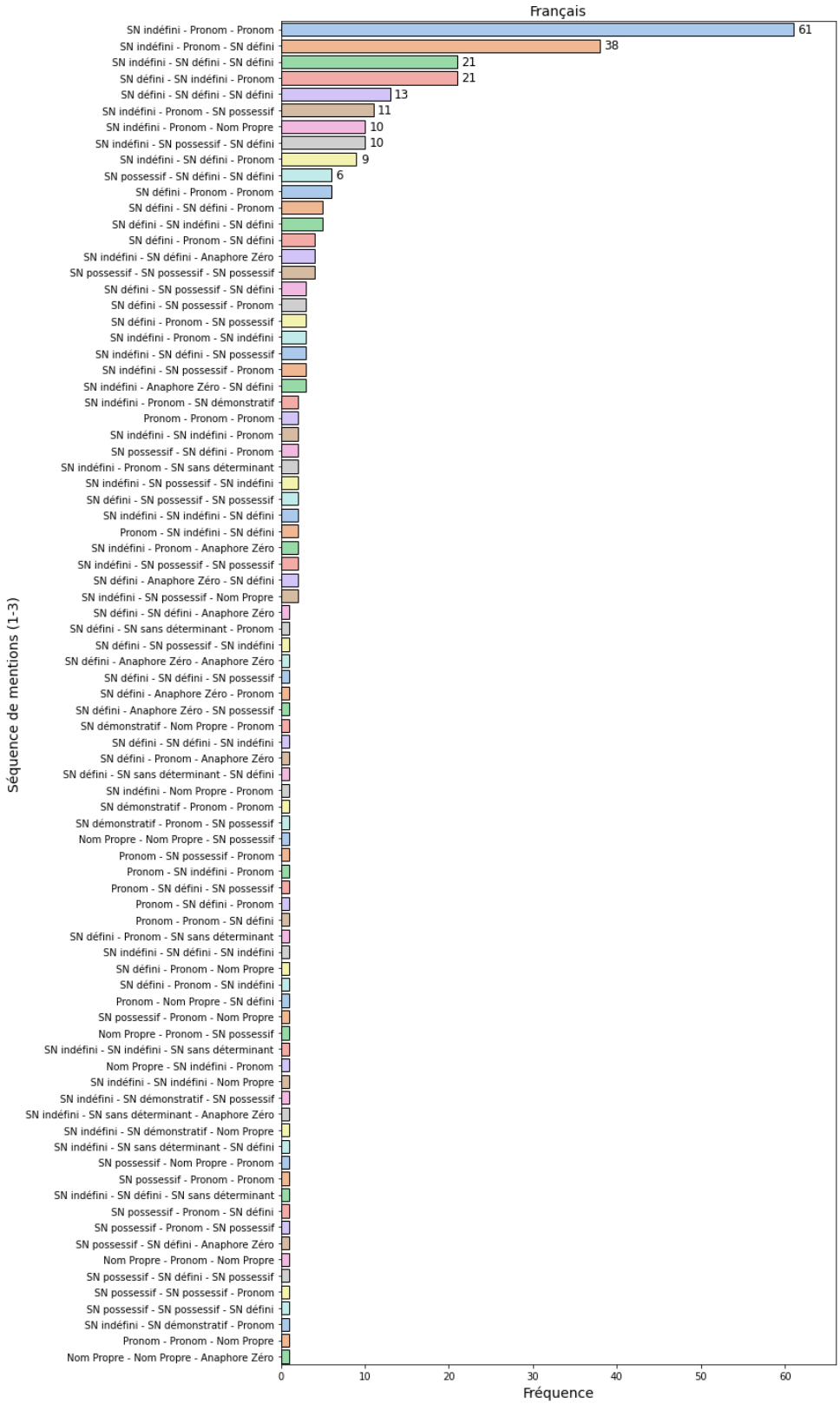
Corpus italien



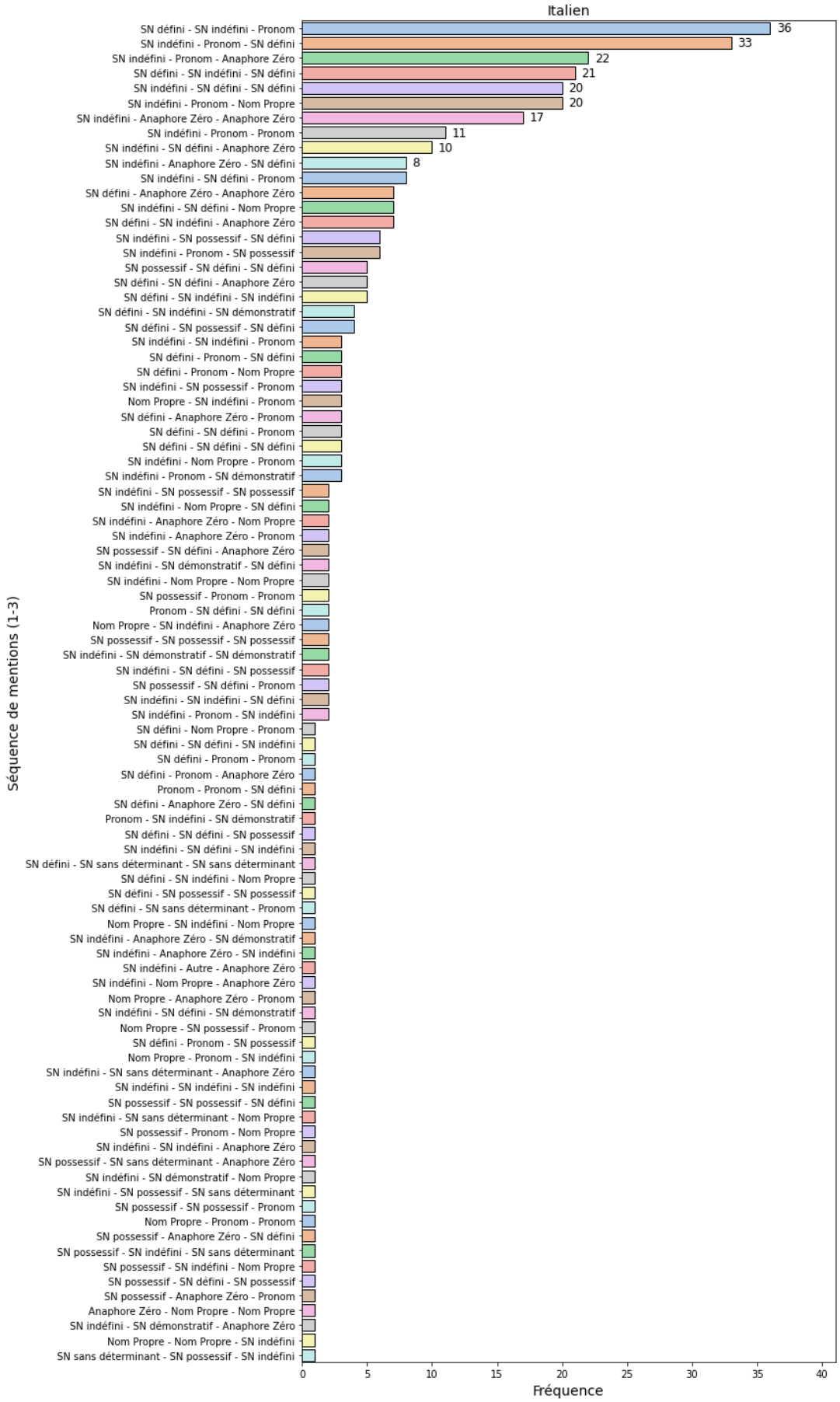


Annexe I. Séquences des trois premiers maillons

Corpus français

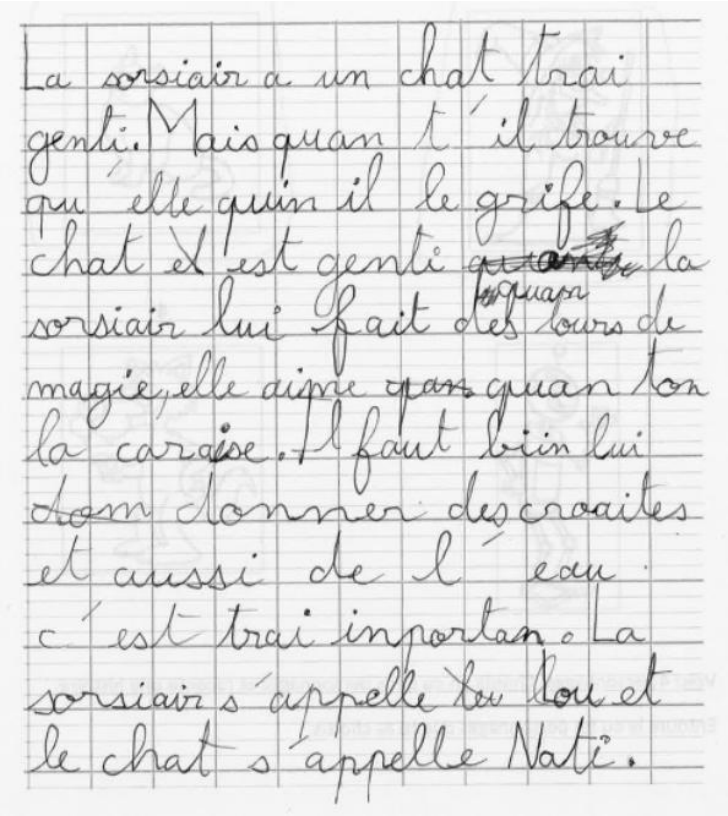


Corpus italien



Annexe J. Scan issu du corpus Scolinter

Texte en cursive mais lettres majuscules en script *(corpus français, CE1, 1518)*



Annexe K. Corpus des mentions longues

Français

NORM-EC-CE1-2015-14-D1-S1325-V1	un chat blanc avec des taches noires
NORM-EC-CE2-2016-14-D1-S1323-V1	un chat tout blanc aux beaux yeux bleus
NORM-EC-CE2-2016-37-D1-S1544-V1	un petit chat au pelage très doux
NORM-EC-CE1-2015-45-D1-S493-V1	la sorcière de chocolat , de vanille et de chantilly
NORM-EC-CE1-2015-67-D1-S1330-V1	la méchante sorcière nommée la méchante sorcière
NORM-EC-CE2-2016-14-D1-S1323-V1	au magnifique robot jaune , vert et bleu
NORM-EC-CE2-2016-20-D1-S113-V1	son armée de robots chevaliers en vie
NORM-EC-CE2-2016-28-D1-S3060-V1	un loup seul et sans famille
NORM-EC-CE2-2016-28-D1-S3060-V1	le seul loup de la forêt
NORM-EC-CE2-2016-37-D1-S1627-V1	quelque chose de noir caché derrière un arbre
NORM-EC-CE2-2016-57-D1-S1986-V1	un petit robot beau et intelligent
NORM-EC-CE2-2016-91-D1-S1160-V1	La maman du petit chaperon rouge
NORM-EC-CE2-2016-5-D1-S3026-V1	un gros chat tendre et dodu pour un loup aussi affamé
NORM-EC-CE2-2016-6-D1-S1991-V1	la sorcière la plus âgée de toutes les sorcières du monde

Italien

NORM_CE1_698	il gatto con il pelo nero
NORM_CE2_1438	una strega come le altre cattiva
NORM_CE1_200	un simpatico robottino, azzurro, giallo, rosso e blu
NORM_CE2_1665	un robot molto ma molto intelligente , giocherellone , divertente e studioso
NORM_CE2_1098	il custode di una foresta incantata
NORM_CE1_432	un'ombra buia , con occhi gialli , muso a punta e denti che spuntano dalla bocca
NORM_CE1_11	una vecchia strega di nome Riunione
NORM_CE1_432	un robot di nome Antenna di fuoco
NORM_CE2_844	un piccolo gattino di nome Bibi
NORM_CE2_1440	il suo amico gatto di nome Caddy
NORM_CE2_1104	un bambino di nome Giacomo e di 2 anni
NORM_CE1_1495	i due gattini Gimi e Sregino
NORM_CE2_1108	i due figli di nome Strega Dina e Lupo Nello
NORM_CE1_200	*la donna che mi ha costruito*
NORM_CE2_1676	*Gianni il bambino che è allergico ai gatti*
NORM_CE2_1101	*l' uomo che porta i bei sogni ai bambini*
NORM_CE2_1655	una strega molto cattiva e molto potente
NORM_CE2_1655	una squadra di supereroi molto forti
NORM_CE1_629	il lupo più cattivo della città
NORM_CE2_1438	una strega molto ma molto strana
NORM_CE2_1438	il lupo in forma di piccione
NORM_CE2_1516	un gatto nero come il carbone
NORM_CE2_835	dal suo amico Lucario il lupo

Annexe L. Corpus français annoté

CE1, 212

Il était une fois **[un petit chat blanc]**^{cat} **[qui]**^{cat} vivait dans la forêt. Un jour que **[le petit chat]**^{cat} se promenait et **[il]**^{cat} vit **[un loup]**^{wolf} **[qui]**^{wolf} était noir. **[Le loup]**^{wolf} était gentil, **[il]**^{wolf} voulait un ami et **[le petit chat blanc]**^{cat} voulait un ami. **[Le petit chat]**^{cat} et **[le loup]**^{wolf} sont amis.

CE1, 217

C'est **[un robot]**^{robot} **[qui]**^{robot} avait envie d'aller dans la forêt. Mais **[il]**^{robot} y voit **[un loup]**^{wolf} **[qui]**^{wolf} rêvait de manger **[le robot]**^{robot}. **[Le robot]**^{robot} se dit « **[je]**^{robot} vais essayer d'aller dans la forêt. » Le lendemain matin **[il]**^{robot} croise le regard et **[ils]**^{robot,wolf} se disent bonjour et **[ils]**^{robot,wolf} sont devenus amis pour toute la vie. **[Ils]**^{robot,wolf} sont les meilleurs amis inséparables.

CE1, 824

Il était une fois **[un chat]**^{cat} **[qui]**^{cat} se promenait dans un bois et **[il]**^{robot} arrive **[un robot]**^{robot} **[qui]**^{robot} s'appelait **[Roboti]**^{robot} et **[il]**^{robot} avait 6 ans, et **[le chat]**^{cat} était gris et blanc ; **[il]**^{cat} s'appelait **[Coouki]**^{cat} et **[lui]**^{cat} avait 5 ans. Et **[ils]**^{cat,robot} devinrent très amis et **[ils]**^{cat,robot} se voyaient tous les jours. **[Ils]**^{cat,robot} se voyaient tous les jours pour s'apprendre à se connaitre de plus en plus et **[ils]**^{cat,robot} disaient même qu'**[ils]**^{cat,robot} étaient cousins pour toute la vie et **[ils]**^{cat,robot} allaient à la même école et **[ils]**^{cat,robot} avaient **[la même maitresse]**^{ext1}, **[Roboti]**^{robot} était **[une très belle fille]**^{robot} et **[ils]**^{cat,robot} tombèrent amoureux **[l'un de l'autre]**^{cat,robot}. Et **[ils]**^{cat,robot} se marièrent et **[eurent]**^{cat,robot} des enfants.

CE1, 841

Il était une fois dans une maisonnette, vivaient **[un chat]**^{cat} et **[[sa]**$^{\mathbf{cat}}$ **maîtresse]**^{ext1}. Mais un jour **[une vilaine sorcière]**^{witch} voulait un chat, et **[elle]**^{witch} savait où en trouver un, **[elle]**^{witch} alla au magasin des animaux, mais il n'y avait plus de chat noir. **[Elle]**^{witch} devint en colère et **[fouilla]**^{witch} toutes les maisons, et **[trouva]**^{witch} la maison d'**[un chat noir]**^{cat}, **[elle]**^{witch} fouilla et **[elle]**^{witch} trouva **[le chat]**^{cat} qu'il **[lui]**^{witch} fallait, **[elle]**^{witch} **[le]**^{cat} vola et **[l']**^{cat} **[emprisonna]**^{witch} jusqu'à ce que **[la dame]**^{ext1} se rendit compte qu'**[elle]**^{ext1} avait perdu **[[son]**$^{\mathbf{ext1}}$ **chat]**^{cat}. **[Elle]**^{ext1} alla chez **[[sa]**^{ext1} **cousine la sorcière]**^{witch} et **[se rendit compte]**^{ext1} que **[[son]**$^{\mathbf{ext1}}$ **chat]**^{cat} était là. **[Elle]**^{ext1} devint furieuse et **[elle]**^{ext1} dit « Alors c'est **[toi]**^{witch} qui as volé **[[mon]**$^{\mathbf{ext1}}$ **chat]**^{cat} ! » **[[Sa]**$^{\mathbf{ext1}}$ **cousine]**^{witch} **[lui]**^{ext1} répondit « C'est que **[j']**^{witch} ai besoin d'un chat comme les autres sorcières et il n'y en avait plus dans le magasin ». « D'accord, **[j']**^{ext1} ai compris, **[tu]**^{witch} veux vivre avec **[moi]**^{ext1} ». « Oui ». « **[On]**^{ext1,witch} peut le faire ». « D'accord ». Fin.

CE1, 2977

C'est l'histoire d'**[un petit lapin]**[ext1] et **[[son]**[ext1] **copain le chat]**[cat]. **[Le lapin]**[ext1] et **[le chat]**[cat] étaient perdus. **[Ils]**[cat,ext1] marchèrent des secondes et des secondes, des minutes et des minutes, des heures et des heures, des jours et des jours. Et enfin **[ils]**[cat,ext1] voient **[quelqu'un]**[robot], **[ils]**[cat,ext1] s'approchent tout doucement et **[voient]**[cat,ext1] **[un robot]**[robot] mais **[le robot]**[robot] **[les]**[cat,ext1] a vus. Alors **[il]**[robot] attrape **[le chat]**[cat] et **[lui le lapin]**[ext1] court mais **[le robot]**[robot] **[l']**[ext1] attrapa.

CE1, 2983

Il était une fois **[un loup-garou]**[wolf] **[qui]**[wolf] voulait attaquer **[un chat]**[cat]. **[Le chat]**[cat] vivait dans une maison. **[Le chat]**[cat] était tout seul. Un jour **[le loup - garou]**[wolf] essayait de casser la maison **[du chat]**[cat]. Il faisait nuit et **[le chat]**[cat] **[il]**[cat] se réveilla. **[Il]**[cat] sortit pour voir ce qui se passait. **[Il]**[cat] ne voit rien, **[il]**[cat] va se recoucher. Le lendemain matin **[il]**[cat] posa un piège. Et un soir **[le loup-garou]**[wolf] tombe dans le piège. Le lendemain matin **[le chat]**[cat] dit « enfin » et **[il]**[cat] va dans **[sa]**[cat] maison.

CE1, 2984

[Un chat <u>tout</u> seul][cat] mais un jour **[il]**[cat] rencontra **[un loup]**[wolf] alors **[il]**[cat] courut puis **[se reperdit]**[cat]. **[Il]**[cat] trouva **[une girafe]**[ext1]. **[Il]**[cat] était plus soulagé. Puis **[les deux amis]**[cat,ext1] repartirent puis **[ils]**[cat,ext1] trouvent **[un lapin]**[ext2]. Puis **[ils]**[cat,ext1,ext2] repartirent, **[ils]**[cat,ext1,ext2] trouvèrent **[un lion]**[ext3]. **[Tous les <u>animaux</u>]**[cat,ext1,ext2] disent « au secours, **[le tigre]**[ext4] va **[nous]**[cat,ext1,ext2] manger » puis **[ils]**[cat,ext1,ext2] ont réussi à s'échapper. **[Ils]**[cat,ext1,ext2] retrouvèrent **[le loup]**[wolf] au même endroit puis la faim arrive.

CE1, 2986

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] vivait dans une maison hantée. Et dehors **[elle]**[witch] trouve **[un petit chat noir]**[cat] avec **[une souris]**[ext1] dans la bouche et **[elle]**[witch] **[lui]**[cat] dit « **[tu]**[cat] vas cracher **[cette souris]**[ext1]. Pauvre **[tu]**[cat] veux des croquettes et du lait ? » « **[j']**[cat] aime le lait » « viens dans **[ma]**[witch] maison hantée. **[Tu]**[cat] seras bien sage, **[tu]**[cat] viens manger avec **[moi]**[witch], **[tu]**[cat] prendras un goûter » **[La sorcière]**[witch] **[lui]**[cat] dit « **[tu]**[cat] veux du lait » « au revoir » « **[tu]**[cat] peux passer quand **[tu]**[cat] veux et **[je]**[witch] te donne du lait et dans les croquettes que **[j']**[witch] ai fabriquées pour les chats noirs surtout **[toi]**[cat] » « Et **[il]**[cat] revient rester toujours avec **[lui]**[witch] » dit**[-elle]**[witch].

CE1, 1337

Il était une fois **[un chat]**[cat] **[<u>qui</u>]**[cat] vivait dans un village mais **[un autre chat]**[cat1] arrivait dans le village mais **[le premier chat]**[cat] n'était pas content. **[L'autre chat]**[cat1] se faisait chouchouter aussi te chouchoute.

CE1, 1339

[Le petit chaton][cat] arriva dans la forêt et **[trouva]**[cat] **[une souris]**[ext1] et après **[il]**[cat] mangea **[la souris]**[ext1]. Le matin **[le chaton]**[cat] trouva **[un oiseau]**[ext2]. **[Il]**[cat] avait fini, du coup **[il]**[cat] mangea **[l'oiseau]**[ext2].

CE1, 1355

C'était une fois **[un chat abandonné]**[cat] **[qui]**[cat] se fit attraper par **[une fille]**[ext1] **[qui]**[ext1] avait **[des oiseaux]**[ext2]. **[Le chat]**[cat] un matin mangea **[les oiseaux]**[ext2].

CE1, 3012

{[La sorcière][witch] et **[le chat]**[cat]}

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] se promenait dans une forêt. **[Elle]**[witch] trouve **[un chat]**[cat], **[elle]**[witch] **[lui]**[cat] dit « **[je]**[witch] vais **[te]**[cat] prendre pour **[mon]**[witch] petit déjeuner » **[elle]**[witch] met **[le chat]**[cat] dans la casserole et **[la sorcière]**[witch] dit **[au chat]**[cat] « bien fait pour **[toi]**[cat] ! Fallait pas aller dans la forêt, fallait rester avec **[[ton]**[cat] **maitre]**[ext1] ou **[[ta]**[cat] **maitresse]**[ext1] » dit **[la sorcière]**[witch] à voix haute !

CE1, 2446

[Le chat][cat] est malade et **[va]**[cat] s'asseoir sur le bord de la fenêtre. **[Le robot]**[robot] fait de zumba. **[Le loup]**[wolf] veut manger **[les enfants]**[ext1]. **[La sorcière]**[witch] veut faire des potions.

CE1, 2449

[Le chat][cat] attend **[la souris]**[ext1] dans le petit trou. **[Il]**[cat] veut faire le piège pour **[la souris]**[ext1]. **[La souris]**[ext1] voit un fromage dans le piège et **[la souris]**[ext1] un frigo, **[le chat]**[cat] **[la souris]**[ext1]. **[Le chat]**[cat] court vite pour **[l']**[ext1] attraper. **[Le chien]**[ext2] se réveille. **[Le chat]**[cat] tourne la tête, **[le chat]**[cat] et **[le chien]**[ext2] aussi. Et **[la souris]**[ext1] fait attention aux pièges, **[est sauvée]**[ext1] et **[mange]**[ext1] tout le frigo.

Fin

CE1, 1829

C'est l'histoire d'**[un chat]**[cat] **[qui]**[cat] avait une belle maison. Mais un jour **[le chat]**[cat] alla se balader. Sur **[son]**[cat] chemin **[il]**[cat] rencontra **[un loup]**[wolf]. **[Le chat]**[cat] voulait devenir ami avec **[le loup]**[wolf]. Mais **[le loup]**[wolf] ne voulait pas devenir ami. **[Il]**[wolf] ne voulait pas devenir ami. **[Il]**[wolf] voulait manger **[le chat]**[cat]. **[Le loup]**[wolf] faisait comme s'**[il]**[wolf] voulait devenir ami. Quelques jours plus

tard **[le loup]**[wolf] alla chez **[le chat]**[cat] toute la journée. La nuit **[le loup]**[wolf] alla chez **[le chat]**[cat]. **[Le chat]**[cat] dormait, **[le loup]**[wolf] est rentré et **[a mangé]**[wolf] **[le chat]**[cat]. **[Le chat]**[cat] se réveille et **[donna]**[cat] un gros coup de griffe et **[ouvre]**[cat] le ventre **[du loup]**[wolf] et **[alla]**[cat] prendre un couteau et **[tua]**[cat] **[le loup]**[wolf] et **[le chat]**[cat] comprit que **[le loup]**[wolf] voulait **[le]**[cat] manger, **[il]**[cat] ne veut plus jamais avoir d'ami. Fin

CE1, 1830

Il était une fois **[un chat]**[cat] et **[une sorcière]**[witch]. **[La sorcière]**[witch] voulait transformer **[le chat]**[cat] en cheval mais un jour **[le chat]**[cat] se promenait dans la forêt, **[le chat]**[cat] entendait des bruits derrière **[lui]**[cat]. **[Le chat]**[cat] se retourna et **[dit]**[cat] à **[la sorcière]**[witch] « que fais-**[tu]**[witch] derrière **[moi]**[cat] ? » « Rien » dit **[la sorcière]**[witch]. Vu que **[le chat]**[cat] n'avait pas trop d'amis **[il]**[cat] dit à **[la sorcière]**[witch] « **[tu]**[witch] veux être **[mon]**[cat] amie ? » **[la sorcière]**[witch] dit « oui » et maintenant **[le chat]**[cat] avait beaucoup d'amis.

CE1, 1850

Il était une fois **[un chat]**[cat] et **[une sorcière]**[witch] **[qui]**[cat,witch] voulaient tellement avoir un balai, mais **[ils]**[cat,witch] n'avaient pas assez d'argent. Alors un jour **[la sorcière]**[witch] soupira et **[dit]**[witch] « Si seulement **[j']**[witch]avais un travail. »

« **[Je]**[cat] sais, pourquoi ne pas te fabriquer une baguette magique ? »

« Bonne idée **[[mon]witch chat]**[cat], **[tu]**[cat] sais que **[tu]**[cat] as des bonnes idées, **[je]**[witch] vais me fabriquer une baguette magique, mais d'ailleurs pourquoi ? » Et **[le chat]**[cat] répondit « bah parce que si **[tu]**[witch] veux un balai avec **[tu]**[witch] pourras faire ce que **[tu]**[witch] veux. » Et c'est ainsi que **[la sorcière]**[witch] s'en fabriqua une.

Fin.

CE1, 1125

Il était une fois **[un chat]**[cat] dans la rue. Il était **[le seul chat]**[cat], **[tous les autres chats]**[ext1] étaient bien au chaud dans **[leur]**[ext1] maison. **[Un loup]**[wolf] arriva dans la rue. **[Le loup]**[wolf] voit **[le chat]**[cat] dans la rue tout seul et gelé et à mourir de faim. **[Le loup]**[wolf] allait chercher à manger dans les poubelles. **[Il]**[wolf] trouva un poisson. **[Il]**[wolf] amena le poisson vers **[le chat]**[cat]. **[Le chat]**[cat] mangea le poisson. **[Le chat]**[cat] chercha un endroit pour s'abriter de la pluie et l'orage. **[Il]**[cat] trouva une chatière, **[il]**[cat] entra par la chatière et **[voit]**[cat] **[un chat comme [lui]cat]**[cat2]. **[Il]**[cat2] avait l'air gentil pour **[le chat]**[cat]. **[Il]**[cat2] avait un lit. **[Le chat]**[cat] alla dans le lit **[du chat comme [lui]cat]**[cat2]. **[Le chat comme [lui]cat]**[cat2] dormit dans le même lit que **[le chat]**[cat]. **[Le chat]**[cat] se leva à 7 heures du matin. **[Le loup]**[wolf] avait dormi dans une niche. Comme il avait eu **[un chien]**[ext1] **[le loup]**[wolf] en avait profité de dormir dans la niche. **[Le loup]**[wolf] se leva à sept heures du matin comme l'heure **[du**

chat][cat]. **[Il]**[cat] part de la maison. **[Il]**[cat] trouve à manger donc **[il]**[cat] mange du poisson. **[Le loup]**[wolf] mange un os. **[Il]**[cat] trouva un abri pour rester pour toujours. **[Il]**[cat] était bien.

CE1, 2530

Un jour, il y avait **[une sorcière]**[witch] **[qui]**[witch] vivait avec **[un chat]**[cat]. Mais le seul problème **[du chat]**[cat] c'est que **[la sorcière]**[witch] oublie de **[lui]**[cat] donner à manger. **[La sorcière]**[witch] est habillée avec un chapeau noir, un balai et des yeux bleus. **[Le chat]**[cat] est tout marron. **[Il]**[cat] s'embêtait beaucoup. Mais une semaine plus tard, **[il]**[cat] se fâche de moins en moins, c'était bizarre. **[Elle]**[witch] **[lui]**[cat] donnait **[son]**[cat] lait. **[Le chat]**[cat] était content. Et maintenant **[ils]**[cat,witch] s'aiment beaucoup. Alors **[ils]**[cat,witch] s'offraient beaucoup de cadeaux. Mais au réveillon de Noël, **[ils]**[cat,witch] n'avaient plus à manger alors **[ils]**[cat,witch] décidèrent d'aller tuer **[un enfant]**[ext1]. Et **[ils]**[cat,witch] mangèrent **[l'enfant]**[ext1].

CE1, 1317

Il y a **[une sorcière]**[witch] **[qui]**[witch] habitait dans une vieille maison. **[Elle]**[witch] dormait. Il y avait **[un loup]**[wolf] **[qui]**[wolf] s'appelait **[Ismaël]**[wolf]. **[Il]**[wolf] volait des bijoux et des bonbons. **[La sorcière]**[witch] **[qui]**[witch] s'appelait **[Eva]**[witch] se réveillait et **[dit]**[witch] « Où ils sont **[mes]**[witch] bijoux et **[mes]**[witch] bonbons ? » **[Eva]**[witch] entendait une voix et **[appela]**[witch] **[la police]**[ext1]. **[Eva]**[witch] dit à **[la police]**[ext1] « il y a **[un voleur]**[wolf] **[qui]**[wolf] habite juste à côté et **[qui]**[wolf] **[m']**[witch] a volé **[mes]**[witch] bijoux et **[mes]**[witch] bonbons. » « OK, répondit **[la police]**[ext1]. **[J']**[ext1] arrive ! Où **[il]**[wolf] est ? **[je]**[ext1] vais **[le]**[wolf] tuer. » « **[Il]**[wolf] est parti par là. » « OK ! **[Je]**[ext1] vais **[le]**[wolf] tuer » répondit **[la police]**[ext1]. Pan ! voilà, **[Eva]**[witch] va être tranquille. **[Eva]**[witch] dormait tranquille et **[Eva]**[witch] passa une bonne nuit sans **[le loup méchant]**[wolf]. **[Eva]**[witch] dit que ça fait du bien. **[Elle]**[witch] prend **[son]**[witch] déjeuner tranquille. **[Eva]**[witch] dit « c'est très bien »

fin

CE1, 1324

[La sorcière][witch] transforme **[le loup]**[wolf] et **[part]**[witch] avec **[son]**[witch] balai magique. Le lendemain matin **[la sorcière]**[witch] revient et **[voit]**[witch] **[le loup]**[wolf] qu'**[elle]**[witch] avait transformé. Et **[elle]**[witch] **[le]**[wolf] retransforme et **[elle]**[witch] part. Avec **[son]**[witch] balai magique.

CE1, 1325

Il était une fois **[une sorcière]**[witch] ***[qui]***[witch] cherchait un chat. **[Elle]**[witch] cherchait dans tous les magasins. Un jour **[elle]**[witch] trouva **[un chat blanc avec des taches noires]**[cat]. **[Elle]**[witch] jouait tout le temps avec **[lui]**[cat]. Et **[ils]**[cat,witch] devinrent meilleurs amis pour toujours. Et **[le chat]**[cat] eut **[des chatons]**[cat1] avec **[une chatte blanche]**[cat2].

CE1, 1518

[La sorcière]witch a un chat très gentil. Mais quand **[il]**cat trouve quelqu'un **[il]**cat le griffe. **[Le chat]**cat est gentil quand **[la sorcière]**witch **[lui]**cat fait des tours de magie, **[elle]**cat aime quand on **[la]**cat caresse. Il faut bien **[lui]**cat donner des croquettes et aussi de l'eau c'est très important. **[La sorcière]**witch s'appelle **[Lou]**witch et **[le chat]**cat s'appelle **[Nati]**cat.

CE1, 569

Il était une fois **[un chat]**cat **[qui]**cat voulait prendre des vacances sur une île. Car **[il]**cat était **[garagiste]**cat. Alors pour aller sur cette île **[il]**cat prit une barque et **[rama]**cat jusqu'à l'île. […]

CE1, 573

[Le robot]robot marche bizarrement et **[va]**robot chercher un ballon de foot et **[le robot]**robot s'appelle **[Zlata]**robot et **[joue]**robot au foot.

CE1, 584

C'est l'histoire d'**[un chat]**cat **[qui]**cat se promenait dans la forêt. **[Il]**cat rencontra **[un loup]**wolf. **[Le loup]**wolf dit « bonjour **[le chat]**cat, que viens - **[tu]**cat faire dans la forêt ? » « **[moi]**cat **[je]**cat me promenais dans la forêt » **[le chat]**cat était bizarre, **[il]**cat n'avait pas peur **[du loup]**wolf. « **[Tu]**cat n'as pas peur de **[moi]**wolf ? » dit **[le loup]**wolf. « Ben non **[tu]**wolf vois, **[je]**cat n'ai pas peur de **[toi]**wolf. » « Mais tout le monde a peur de **[moi]**wolf. » « Mon oeil personne n'a peur de **[toi]**wolf. »

Fin.

CE1, 100

Il était une fois **[un chat persan]**cat. **[Il]**cat se promenait le long du trottoir. Un jour **[il]**cat trouve **[un maitre]**robot. C'était **[un robot]**robot. **[Le robot]**robot était gentil, très gentil. **[Il]**cat s'est trouvé bien chanceux. Mais le lendemain matin **[le robot]**robot abandonna **[le chat]**cat.

CE1, 102

Il était une fois **[une très méchante sorcière]**witch. **[Elle]**witch avait fait **[un robot]**robot. **[La sorcière]**witch en fit **[[son]witch esclave]**robot. Mais **[le robot]**robot ne voulait pas alors **[il]**robot s'enfuit. Mais **[la sorcière]**witch **[le]**robot suivait. Mais **[le robot]**robot courait trop vite. **[La sorcière]**witch tomba boum. Et **[la sorcière]**witch s'en alla chez **[elle]**witch. **[Le robot]**robot était sauvé par **[un autre robot]**robot1.

CE1, 114

Il était une fois **[une chatte]**cat **[qui]**cat habitait dans une belle maison, et un jour **[quelqu'un]**witch vint, c'est **[la sorcière]**witch **[qui]**witch était très gentille, **[elle]**witch avait **[deux belles filles]**ext1. **[Elles]**ext1 étaient gentilles comme **[[leur]ext1 maman]**witch. Et **[[leur]ext1 papa]**ext2 ouvrit la porte tout

doucement. **[Les filles]**[ext1] tenaient **[la chatte]**[cat] **[qui]**[cat] tremblait de peur mais **[fut]**[cat] rassurée parce que c'est **[[leurs]**[cat] **maîtresses]**[ext1] et **[le maître]**[ext2] et il y a **[la maîtresse]**[witch] **[qui]**[witch] était fatiguée. Et **[la chatte]**[cat] tourne la tête et **[ne voulut plus]**[cat] dormir. **[Elle]**[witch] dort profondément. **[Elles]**[ext1] **[les filles]**[ext1] dorment **[toutes les deux]**[ext1] dans la même chambre. **[Le papa]**[ext2] ne dort pas encore, **[il]**[ext2] regarde la télé et **[se couche]**[ext2] à minuit et avant de se coucher **[il]**[ext2] vit **[un papillon]**[ext3] **[qui]**[ext3] fait de la lumière.

CE1, 560

Il était une fois **[trois cochons]**[ext1,ext2,ext3] **[qui]**[ext1,ext2,ext3] construisaient **[leur]**[ext1,ext2,ext3] maison le premier une maison en paille, le deuxième une maison en bois et le troisième une maison en brique. Et **[le loup]**[wolf] arriva devant la maison en paille et **[le loup]**[wolf] souffla fort et la maison s'envola et **[le cochon]**[ext1] part chez **[[son]**[ext1] **frère]**[ext2] et **[le loup]**[wolf] arriva devant la maison en bois et **[le loup]**[wolf] souffla fort et la maison s'envole et **[les deux cochons]**[ext1,ext2] vont chez **[[leur]**[ext2,ext1] **frère]**[ext3] et **[le loup]**[wolf] arriva devant la maison en […].

CE1, 93

Il était une fois **[un petit chat]**[cat] **[qui]**[cat] s'appelait **[Mimi]**[cat]. **[Il]**[cat] vivait dans une maison avec **[[son]**[cat] **frère]**[cat1]. Un jour **[il]**[cat] fut vendu au magasin puis **[il]**[cat] ne savait pas comment faire sans **[[son]**[cat] **frère]**[cat1] mais **[il]**[cat] ne se fit pas adopter alors **[il]**[cat] préférait rester chez **[lui]**[cat] en toute sécurité dehors dans le jardin mais **[le chien]**[ext1] *d'à côté]* n'arrêtait pas de **[l']**[cat]embêter. « Mais alors comment faire ? **[le]**[ext1] tuer, **[le]**[ext1] faire vendre au marché, ou demander à **[l']**[ext1]adopter ? » « **[Je]**[cat] le sais comment faire pour sortir de cette maison ! » dit **[le chat]**[cat] à **[[son]**[cat] **frère]**[cat1]. « Pourquoi ne pas **[le]**[ext1] faire fuir ? » « Comment ? » **[lui]**[cat] demande **[[son]**[cat] **frère]**[cat1]. « Et bien en **[lui]**[ext1] faisant peur » dit **[Mimi]**[cat] « **[tu]**[cat] n'as qu'à demander **[au grand maître des chats]**[cat2], qu'en dis - **[tu]**[cat] ? » « Non ! **[Je]**[cat] ne veux pas ! **[Je]**[cat] ne veux pas ! **[J']**[cat]ai trop peur » « Ah ! Ah ! Ah. **[T']**[cat] as peur. Pfuh **[t']**[cat]es bête, c'est **[un chat]**[cat2] **[qui]**[cat2] **[nous]**[cat,cat1] a fait vivre alors **[tu]**[cat] disais que **[tu]**[cat] avais peur. » « Bien sûr que non » dit **[le chat]**[cat] « bon est-ce que **[tu]**[cat] comprends l'histoire maintenant ? » « oui » « alors **[tu]**[cat] vois. » Fin.

CE1, 1188

Il était une fois **[un monsieur]**[ext1] et **[un chat]**[cat] **[qui]**[cat,ext1] se promenaient dans une forêt. **[Ils]**[cat,ext1] trouvent des lapins, des chiens, des renards et des écureuils et quand **[ils]**[cat,ext1] étaient sortis il faisait nuit donc **[ils]**[cat,ext1] restèrent ici. Le lendemain il faisait gris donc **[ils]**[cat,ext1] allèrent se fabriquer une maison.

CE1, 1190

Il était une fois **[une sorcière très gentille]**witch **[qui]**witch avait **[une chatte]**cat qui avait fait **[3 chatons]**cat1,cat2,cat3. **[Les chatons]**cat1,cat2,cat3 et **[cette chatte]**cat, **[la plus petite]**cat1 s'appelle **[charlotte]**cat1 et **[la moyenne]**cat2 s'appelle **[Ela]**cat2 et **[[leur]cat1,cat2 grande soeur]**cat3 s'appelle **[Lili]**cat3 et **[[leur]cat1,cat2,cat3 maman]**cat s'appelle **[Emma]**cat et enfin **[la sorcière]**witch s'appelle **[Luna]**witch.

Fin

CE1, 1197

Il était une fois **[un loup]**wolf **[qui]**wolf avait rencontré **[une sorcière]**witch **[qui]**witch était gentille. **[La sorcière]**witch a voulu faire un marché. Le marché c'était de **[lui]**witch donner de l'argent et **[la sorcière]**witch devait **[lui]**wolf donner à manger. **[La sorcière]**witch a voulu voir **[[sa]witch maman]**ext1 et **[le loup]**wolf avait faim.

CE1, 1531

C'est l'histoire d'**[un chat]**cat **[qui]**cat vivait dans une maison et **[il]**cat adorait aller dans **[sa]**cat panière mais un jour **[sa]**cat panière n'était plus là alors **[le chat chat]**cat s'est dit : " Pourquoi **[ma]**cat panière n'est plus là ? elle est bien quelque part " dit **[le chat]**cat.

CE1, 1554

Il était une fois **[un chat]**cat **[qui]**cat vivait dans une maison toujours propre avec **[une vieille dame]**ext1. **[Le chat]**cat **[lui]**cat aussi était toujours propre et bien élevé. **[Il]**cat adorait se balader, **[il]**cat avait **[trois petits]**cat1,cat2,cat3 **[une]**cat1 **[qui]**cat1 s'appelait **[Ésia]**cat1, **[l'autre]**cat2 **[qui]**cat2 s'appelait **[Jan]**cat2 et **[le troisième]**cat3 s'appelait **[Ulise]**cat3. **[Tous les trois]**cat1,cat2,cat3 étaient sages, **[ils]**cat1,cat2,cat3 jouaient beaucoup à la balle. **[Tous les voisins]**ext2 adoraient **[ces chats]**cat,cat1,cat2,cat3.

CE1, 1561

Il était une fois **[une sorcière]**witch **[qui]**witch avait **[un chat noir]**cat. **[Elle]**witch avait une maison avec un gros chaudron et plusieurs potions. Un jour, **[elle]**witch vit **[un enfant]**ext1 **[qui]**ext1 était tout seul, **[elle]**witch **[lui]**ext1 dit « Bonjour **[petit enfant]**ext1. Pourquoi es - **[tu]**ext1 tout seul ? » et **[l'enfant]**ext1 dit « : « **[Mes parents]**ext2 **[m']**ext1 ont abandonné ». Dès qu'**[il]**ext1 vit que **[la sorcière]**witch était méchante **[il]**ext1 **[lui]**witch donna un grand coup de pied.

CE1, 493

Il était une fois **[une vieille sorcière méchante]**witch et plus loin habitait **[une chatte blanche]**cat. Et **[cette chatte]**cat alla voir **[la sorcière]**witch et puis **[la sorcière]**witch **[la]**cat transforma en **[crapaud]**cat...

En **[souris]**[cat], en **[poisson baveux]**[cat] mais sans faire exprès **[la sorcière]**[witch] donna les pouvoirs à **[la chatte]**[cat] et **[la chatte]**[cat] fit ABRACADABRI ! en rigolant... Et **[elle]**[cat] ensorcela **[la sorcière de chocolat, de vanille et de chantilly]**[witch]... **[Et de noix de coco]**[witch] bien sûr et **[se fit]**[cat] un gâteau au chocolat, à la vanille et à la chantilly ! Et voilà comment se finit l'histoire... **[La chatte]**[cat] rentra chez **[elle]**[cat] et **[elle]**[cat] se régala sans parler de **[la sorcière]**[witch] **[qui]**[witch] s'est fait... Manger !

CE1, 2938

[Le chat][cat] part en promenade dans la forêt. **[Il]**[cat] rencontre **[le loup]**[wolf] et **[le loup]**[wolf] veut **[le]**[cat] manger. **[Le chat]**[cat] court à toute vitesse dans le château. **[Le chat]**[cat] part dans **[son]**[cat] lit puis **[le chat]**[cat] est sorti dans la fenêtre, **[il]**[cat] crie « si **[tu]**[wolf] ne **[me]**[cat] manges pas **[je]**[cat] **[t']**[wolf]ouvrirai la porte » **[le loup]**[wolf] dit « **[je]**[wolf] ne **[te]**[cat] mange pas, **[tu]**[cat] peux descendre » et **[le chat]**[cat] descend et **[le loup]**[wolf] **[le]**[cat] mange.

CE1, 3026

Il était une fois **[un loup tout gris, affamé]**[wolf]. **[Il]**[wolf] avait très faim, très très faim. Un jour, **[il]**[wolf] croisa **[un chat]**[cat] **[qui]**[cat] avait l'air bien appétissant. **[Ils]**[cat,wolf] devinrent vite de vrais amis. Mais un jour, **[le loup]**[wolf] avait beaucoup trop faim, donc **[il]**[wolf] décida de manger **[le chat]**[cat]. Puis, un jour, **[le chat]**[cat] était au bord de l'eau. Pendant ce temps, **[le loup]**[wolf] caché derrière un buisson, **[l']**[cat] observait. Soudain, **[le loup]**[wolf] marcha sur la pointe des pieds, où était **[le chat]**[cat], et d'un coup, **[il]**[wolf] **[le]**[cat] mangea.

Fin

CE1, 3045

Il était une fois **[un robot]**[robot]. **[Il]**[robot] vivait dans des endroits sous terre. **[Il]**[robot] était très grand. **[Il]**[robot] ne lâchait jamais **[sa]**[robot] pelle. **[Il]**[robot] voulait aller rejoindre le noyau de la terre pour voir ce qu'il y avait à l'intérieur. Mais **[il]**[robot] n'y arrivait pas alors **[il]**[robot] était obligé de remonter mais cela prenait des années et des années mais **[il]**[robot] n'y arrivait pas.

CE1, 3046

[Un chat][cat] **[qui]**[cat] a une queue noire et **[qui]**[cat] chasse tout le temps les souris. **[Il]**[cat] joue beaucoup à chasser les souris. Mais **[il]**[cat] n'aime pas les chiens et **[il]**[cat] n'aime pas se battre. Le seul truc qu'**[il]**[cat] aime c'est manger.

CE1, 3049

Il était une fois **[un robot]**[robot] **[qui]**[robot] voulait aller à l'école. Quand **[le robot]**[robot] alla à l'école **[il]**[robot] se fit plein d'amis. Et **[il]**[robot] était très fort dans tous les domaines. Et **[la maitresse]**[ext1] était très gentille.

CE1, 1166

Il y a **[une sorcière]**[witch] **[qui]**[witch] voulait manger **[le loup]**[wolf] et **[qui]**[witch] voulait **[le]**[wolf] mettre dans le feu et après **[il]**[wolf] **[lui]**[witch] demande : " **[tu]**[witch] ne peux pas **[me]**[wolf] mettre dans le feu, **[je]**[wolf] suis **[ton]**[witch] meilleur ami ".

CE1, 1985

[Le robot][robot] et **[le chat]**[cat] vont jouer aux cartes. **[Lilou le chat]**[cat] a mis l'as de pique, **[Robi]**[robot] a mis le roi de coeur.

CE1, 1996

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] avait fait une potion magique. **[Elle]**[witch] a voulu mettre **[un lapin]**[ext1] pour **[le]**[ext1] transformer en **[chat]**[ext1] mais **[elle]**[witch] est tombée avec **[le chat]**[ext1]. **[Elle]**[witch] est transformée en **[chat]**[witch].

CE1, 2003

{[La sorcière][witch] **[qui]**[witch] voulait un prince**}**

[Une sorcière][witch] vole sur **[son]**[witch] balai. À **[sa]**[witch] maison **[elle]**[witch] s'endort et **[rêve]**[witch] d'avoir un prince. Un jour **[elle]**[witch] a une idée, **[elle]**[witch] va transformer **[[son]**[witch] **chat]**[cat] en **[prince]**[cat]. The end !

CE1, 2004

Il était une fois **[un très gentil loup]**[wolf] **[qui]**[wolf] vivait dans la forêt. Et un beau jour **[un petit garçon]**[ext1] **[qui]**[ext1] se promenait dans la forêt avec **[[son]**[ext1] **chat]**[cat]. Mais **[le chat]**[cat] avait très peur et puis **[il]**[cat] s'enfuit. **[Le loup]**[wolf] vit **[le chat]**[cat] et **[dit]**[wolf] « **[n'aie pas]**[cat] peur, **[je]**[wolf] **[te]**[cat] protégerai ». « Mais à une condition, que **[[ton]**[cat] **maitre]**[ext1] **[me]**[wolf] donne de la viande » « C'est d'accord mais comment vais **[-je]**[cat] prendre la viande ? » « Eh bien **[tu]**[cat] dis à **[[ton]**[cat] **maitre]**[ext1] que **[tu]**[cat] as faim et **[il]**[ext1] **[t']**[cat] en donnera. » Et tous les jours **[le chat]**[cat] alla chercher de la viande chez **[[son]**[cat] **maitre]**[ext1] et **[le loup]**[wolf] était bien content. Fin

CE1, 2034

{[La chatte noire][cat]}

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] vivait avec **[une chatte noire, Lola]**[cat]. **[Elle]**[cat] était toute noire. Un jour **[la chatte]**[cat] avait disparu, mais au bout de quelques jours **[la chatte]**[cat] était toujours là. « Hum... étrange ? » Mais ça **[lui]**[witch] remonta vite le moral, car **[Lola]**[cat] avait envie de jouer avec **[elle]**[witch].

CE1, 2035

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] était dans **[sa]**[witch] maison. Une fois **[elle]**[witch] a décidé d'aller mettre un sort au village donc **[elle]**[witch] a pris **[son]**[witch] balai magique et **[alla]**[witch] mettre un sort au village et même sur le château de **[la princesse]**[ext1]. Après **[elle]**[witch] est rentrée chez **[elle]**[witch] pour préparer de la potion magique pour en donner au village et au château pour qu'**[ils]**[ext1,ext2] meurent **[tous]**[ext1,ext2].

CE1, 221

Il était une fois dans une maison vivait **[une sorcière]**[witch] **[qui]**[witch] était très méchante et **[elle]**[witch] faisait des potions magiques. **[La sorcière]**[witch] était si contente qu'**[elle]**[witch] en faisait tous les jours et un jour pendant qu'**[elle]**[witch] faisait une potion **[un chat]**[cat] rentra par **[sa]**[witch] fenêtre et **[la sorcière]**[witch] était [non fini].

CE1, 236

C'est **[un chat]**[cat] **[qui]**[cat] regarde et **[qui]**[cat] est assis et **[il]**[cat] tourne la tête. C'est **[un loup]**[wolf] **[qui]**[wolf] regarde et **[qui]**[wolf] a des dents très pointues et **[qui]**[wolf] a un gros nez et des et **[qui]**[wolf] a de grosses oreilles. Et des griffes.

CE1, 1302

{[La sorcière][witch] et **[son]**[witch] balai.}

Depuis l'anniversaire de **[la sorcière]**[witch], **[elle]**[witch] sort toujours **[son]**[witch] balai et **[sa]**[witch] baguette magique. « you hou hou hou ! " » s'écrie **[la sorcière]**[witch].

{**[Le petit chat blanc]**[cat].}

Dans une ferme, **[le fermier]**[ext1] chassait de **[[ses]**[ext1] **vaches]**[ext2] **[un petit chat blanc]**[cat]. **[Il]**[ext1] a **[des poules]**[ext3], **[des coqs]**[ext4] et **[des poussins]**[ext5]. **[Il]**[ext1] a aussi **[un âne]**[ext6] et **[[son]**[ext1] **âne]**[ext6] s'appelle **[Pascalu]**[ext6]. **[Le petit chat]**[cat] buvait le lait de **[la vache]**[ext2] et **[le fermier]**[ext1] **[le]**[cat] chassait encore une fois alors **[le petit chat]**[cat] allait vers **[les poules]**[ext3] et **[le fermier]**[ext1] **[le]**[cat] prit et **[le]**[cat] **[mit]**[ext1] dans une niche fermée à clé.

CE1, 1330

{[La méchante sorcière]witch est devenue gentille.}

Il était une fois **[une méchante sorcière]**witch **[qui]**witch avait **[un chat]**cat. **[[Son]witch chat]**cat **[l']**witchavait griffée, dommage **[elle]**witch n'avait plus de pouvoir, **[elle]**witch voulait **[lui]**cat donner un sort. Mais **[la méchante sorcière nommée la méchante sorcière]**witch est gentille sans **[ses]**witch sorts.

CE1, 1332

{[La méchante sorcière]witch}

[La méchante sorcière]witch préparait un potage. **[Un petit garçon]**ext1 jouait dans **[son]**ext1 jardin. Soudain **[la méchante sorcière]**witch **[le]**ext1 prenait et **[le]**ext1 **[mettait]**witch dans **[son]**witch potage. **[Le méchant renard]**ext2 **[Un renard]**ext2 était en train de chasser **[un animal]**ext3, et soudain **[il]**ext2 **[l']**ext3 attrapait et **[le]**ext3 **[mangeait]**ext2.

CE1, 2886

Il était une fois : il y avait **[un robot]**robot à la décharge. Avant c'était **[un simple humain]**robot. **[Il]**robot a été transformé par **[une sorcière]**witch. Un jour **[Flique]**robot1 rentra à la décharge, BOUM !... **[Il]**robot1 fonça dans **[quelque chose solide et résonnant]**robot. C'était **[un autre robot]**robot. « Qui es - **[tu]**robot ? » dit **[Flique]**robot1 « Je suis **[Fleque]**robot **[un robot]**robot transformé par **[une sorcière]**witch ». « Quelle sorcière ? » demanda **[Flique]**robot1 « par **[Crapouillate]**witch ». « Oh C'est marrant **[moi]**robot1 aussi. » « **[J']**robot ai une idée, **[on]**robot,robot1 va construire des armes à la décharge. »

CE1, 2893

[Le petit chat]cat joue avec **[sa]**cat balle et **[[sa]cat maman]**cat1 **[le]**cat regarde et voici **[le garçon]**ext1 **[qui]**ext1 prend **[le petit chat]**cat et **[le]**cat **[ramène]**ext1 dans **[sa]**ext1 chambre et **[la chatte]**cat1 Mioul dans toute la maison **[court]**cat1 après **[le garçon]**ext1 et hop la porte se ferme et **[elle]**cat1 pleure et **[elle]**cat1 pleure mais **[elle]**cat1 voit un trois et **[elle]**cat1 court, **[elle]**cat1 prend **[le chat]**cat et **[elle]**cat1 court à **[sa]**cat1 maison.

CE1, 2901

[Une sorcière]witch se promène dans la rue et tout d'un coup **[elle]**witch s'envole vers **[sa]**witch maison imaginaire dans l'espace. **[Elle]**witch est arrivée à **[sa]**witch maison, **[elle]**witch rentre et **[elle]**witch se couche. Le matin **[elle]**witch va chercher le pain et des croissants pourris et **[elle]**witch prépare **[sa]**witch

potion magique. C'est l'hiver, il fait froid et il y a du vent. **[La sorcière]**witch ne sort pas. Enfin l'été il y a un gros soleil dehors et il fait très chaud dehors.

CE1, 2905

Il était une fois **[un chat perdu]**cat. **[Il]**cat dormait dans une poubelle. **[Il]**cat mangeait des peaux de bananes. **[Il]**cat buvait de la peinture.

CE1, 1969

{[Les deux amis]cat,robot**}**

Il était une fois **[une robot]**robot **[qui]**robot s'appelle **[Roti]**robot. **[Il]**robot était seul puis en fin d'après-midi **[il]**robot rencontre **[Kitty]**cat **[un chat]**cat. **[Roti]**robot **[lui]**cat demande si **[elle]**cat veut jouer avec **[lui]**robot. **[Kitty]**cat **[lui]**robot dit « oui bien sûr ». Puis demain **[ils]**cat,robot jouent ensemble à la corde à sauter et aux bulles puis **[Kitty]**cat va chez **[Roti]**robot et **[ils]**cat,robot mangent tous ensemble puis **[vont dormir]**cat,robot.

CE1, 1981

{[Le robot]robot et **[le chat]**cat.**}**

Il était une fois **[le chat]**cat et **[le robot]**robot **[qui]**cat,robot jouent au ballon en se faisant des passes et après **[le chat]**cat dit : « **[on]**cat,robot va manger. Il est 12h00 » après **[ils]**cat,robot vont gouter et ainsi est terminée mon histoire. Bisou à tout le monde nos amours.

CE1, 55

Il était une fois **[une sorcière]**witch **[qui]**witch avait faim. **[Elle]**witch sortit dehors et se promène. **[Elle]**witch voit **[un loup]**wolf et **[dit]**witch « **[je]**witch vais **[te]**wolf transformer en chat. **[Je]**witch vais me régaler miam miam » **[elle]**witch **[le]**wolf transforme en **[un autre loup]**wolf puis **[court]**wolf, **[court]**wolf aussi vite que possible. **[La sorcière]**witch le suit « **[je]**witch **[t']**wolf ai attrapé, **[je]**witch **[te]**wolf croque mais la peau revient mais **[je]**witch **[l']**wolf ai peut-être fait en **[sucette]**wolf **[qui]**wolf revient mais pourtant **[je]**witch **[l']**wolf ai transformé en **[je]**witch ne sais pas. ». L'histoire n'est pas finie.

CE1, 69

Il était une fois **[un loup]**wolf **[qui]**wolf s'ennuyait alors **[il]**wolf décida d'aller dans la forêt. Une fois dans la forêt, **[il]**wolf aperçut **[une sorcière]**witch. **[Il]**wolf alla voir **[la sorcière]**witch et **[il]**wolf **[lui]**witch dit « bonjour **[sorcière]**witch, **[viens]**witch chez **[moi]**wolf **[on]**witch,wolf va boire du thé ». « Bon d'accord » et **[le loup]**wolf une fois chez **[lui]**wolf **[il]**wolf mangea **[la sorcière]**witch. Fin.

CE1, 655

Il était une fois **[un loup]**^{wolf} et **[une sorcière]**^{witch}. **[La sorcière]**^{witch} était méchante. **[Le loup]**^{wolf} était très vieux mais **[il]**^{wolf} était gentil. Un beau jour **[la sorcière]**^{witch} **[lui]**^{wolf} jeta un sort. **[Le loup]**^{wolf} est devenu méchant mais **[il]**^{wolf} voulait se venger de **[la sorcière]**^{witch}. Un jour **[il]**^{wolf} **[la]**^{witch} tue et le sort se brisa et aussi ça **[lui]**^{wolf} a rendu **[sa]**^{wolf} jeunesse qu'**[elle]**^{witch} **[lui]**^{wolf} a prise avec **[sa]**^{wolf} gentillesse. Depuis **[il]**^{wolf} voulait habiter le village.

CE1, 771

[Le chat]^{cat} sort de la maison puis **[va]**^{cat} chez **[le voisin]**^{ext1}, **[attend]**^{cat} devant la porte et **[la chatte]**^{cat1} sortit. **[Ils]**^{cat,cat1} vont se promener puis **[rentrent]**^{cat,cat1} chez **[eux]**^{cat,cat1} puis **[ils]**^{cat,cat1} vont dormir. Le lendemain matin **[le chat]**^{cat} part avec **[[son]**$^{\mathbf{cat}}$ **maitre]**^{ext2} au marché puis **[ils]**^{cat,ext2} vont faire une promenade.

CE1, 950

Il était une fois **[un chat]**^{cat} **[qui]**^{cat} se promenait dans la forêt mais soudain **[un loup]**^{wolf} a mangé des cailloux. C'est pour reconnaitre **[son]**^{cat} chemin **[du chat]**^{cat}. **[Il]**^{cat} se retourne. **[Le chat]**^{cat} a griffé **[le loup]**^{wolf} et **[tombe]**^{wolf} par terre boum. Donc **[le chat]**^{cat} est perdu mais **[le loup]**^{wolf} a fait semblant de tomber alors **[il]**^{wolf} se déguise en **[chat]**^{wolf}. **[Le chat]**^{cat} se retourne et **[voit]**^{cat} **[un énorme chat]**^{wolf} donc **[le loup]**^{wolf} et **[le chat]**^{cat} sont devenus amis. Fin de l'histoire.

CE1, 982

[Le chat]^{cat} se baladait mais **[le loup]**^{wolf} venait en ville pour chercher à manger. **[Il]**^{wolf} voulait manger des bonbons et des carambars mais ils coûtaient trop cher et en plus personne ne voulait **[lui]**^{wolf} en donner alors **[ils]**^{ext1} **[lui]**^{wolf} disaient « **[va]**^{wolf} - t'en ! » Mais **[le loup]**^{wolf} répondit en râlant « bon si **[vous]**^{ext1} ne voulez pas **[m']**^{wolf}en donner **[je]**^{wolf} vais ailleurs ». Mais **[le loup]**^{wolf} au même moment que **[le chat]**^{cat} se regardent et **[se disent]**^{cat,wolf} bonjour. **[Tous les enfants]**^{ext2} se disent « que c'est bizarre de voir **[un chat]**^{cat} et **[un loup]**^{wolf} se faire un câlin ». Et **[les enfants]**^{ext2} **[eux]**^{ext2} aussi disent bonjour **[au loup]**^{wolf} et **[au chat]**^{cat}.

CE1, 1925

Il était une fois **[un loup affamé]**^{wolf}. **[Il]**^{wolf} rêvait de manger un petit chaton, ces pauvres petits chatons de rue. Un jour **[un chaton]**^{cat} s'en alla dans la forêt et **[le loup]**^{wolf} **[qui]**^{wolf} rêvait d'en manger un **[le]**^{cat} trouva et **[le]**^{cat} **[mangea]**^{wolf} !

CE1, 2820

Il était une fois **[un petit chat]**^{cat} **[qui]**^{cat} se promenait dans la forêt. Et un jour alors qu'**[il]**^{cat} se promenait **[il]**^{cat} rencontre **[un robot]**^{robot}. **[Le robot]**^{robot} dit « **[pauvre petit chat]**^{cat}, **[tu]**^{cat} es perdu ? » « Non **[j']**^{cat} ai juste envie de me promener. » « Ha bon » dit **[le robot]**^{robot} « pourquoi **[tu]**^{cat} as envie de te promener ? » « parce que **[j']**^{cat} ai envie. » « **[Viens]**^{cat} avec **[moi]**^{robot} dans **[mon]**^{robot} laboratoire, **[j']**^{robot} ai plein de potions mais **[j']**^{robot} en ai perdu une. » « Laquelle de potion ? » « La plus belle de toutes les potions, c'est celle qui rend amoureux. » « D'accord. » **[le chat]**^{cat} **[qui]**^{cat} se promène.

CE1, 1231

[Le loup]^{wolf} **[qui]**^{wolf} veut manger **[le robot]**^{robot}}

Il était une fois **[un robot]**^{robot} **[qui]**^{robot} courut parce que **[le loup]**^{wolf} veut **[le]**^{robot} manger. **[Le loup]**^{wolf} veut manger aussi **[François Hollande]**^{ext1}.

CE1, 1693

Il était une fois **[une sorcière très gentille]**^{witch} **[qui]**^{witch} aimait les enfants. **[Elle]**^{witch} donnait des petits cadeaux. **[Elle]**^{witch} était **[maîtresse]**^{witch} pour apprendre à être gentille. Et il y avait **[un chaton]**^{cat} dans **[sa]**^{witch} classe. Fin !

CE1, 816

C'est l'histoire d'**[une sorcière très méchante]**^{witch}, **[elle]**^{witch} écrabouillait tous les crapauds, **[elle]**^{witch} mangeait les enfants et **[décrochait]**^{witch} les dents des crocodiles. Un jour **[un loup]**^{wolf} passait par là. **[La sorcière]**^{witch} voulait **[lui]**^{wolf} prendre **[ses]**^{wolf} dents, mais **[le loup]**^{wolf} se jeta sur **[elle]**^{witch} ! Et on ne **[la]**^{witch} reverra plus. Fin.

CE1, 818

Il était une fois **[une sorcière]**^{witch} **[qui]**^{witch} avait un balai qui puait. Il était une fois **[un loup]**^{wolf} **[qui]**^{wolf} avait du fric.

CE1, 937

[Une sorcière]^{witch} a trouvé **[un chat]**^{cat}. **[La sorcière]**^{witch} a dit **[au chat]**^{cat} "**[ne va pas]**^{cat} trop loin, **[tu]**^{cat} peux aller où **[tu]**^{cat} veux que **[je]**^{witch} **[te]**^{cat} voie". **[Le chat]**^{cat} ne veut pas alors **[il]**^{cat} dit « oui » et **[il]**^{cat} va un peu plus loin et **[il]**^{cat} dit "au moins **[je]**^{cat} suis allé trop loin". **[La sorcière]**^{witch} ne voit plus **[le chat]**^{cat}. **[Elle]**^{witch} prend **[ses]**^{witch} jumelles et **[elle]**^{witch} voit qu'**[il]**^{cat} est sur la route et **[elle]**^{witch} se dit "**[je]**^{witch} vais **[le]**^{cat} récupérer". Et enfin **[elle]**^{witch} a récupéré **[[son]**^{witch} **chat]**^{cat}.

CE2, 202

{[Un chat][cat] et **[un robot]**[robot]}

Un jour par accident **[un chat]**[cat] monta sur un arbre. **[Pauvre chat]**[cat] **[il]**[cat] ne pouvait plus descendre. **[Un robot]**[robot] **[qui]**[robot] s'appelait **[X2]**[robot] passait par là. **[Il]**[robot] voulut rattraper **[le petit chat]**[cat] mais échec. **[Il]**[robot] était trop petit. Le lendemain **[X2]**[robot] rencontra **[un savant]**[ext1] **[qui]**[ext1] **[lui]**[robot] avait donné un mécanisme pour allonger **[ses]**[robot] bras. **[Il]**[robot] retourna voir **[le petit chat]**[cat] et **[il]**[robot] puis **[le]**[cat] sauva. **[Le chat]**[cat] revenait voir **[le robot]**[robot]. Malheureusement un jour **[X2]**[robot] tomba dans un centre de pièces détachées. **[Pauvre X2]**[robot], **[il]**[robot] fut complètement anéanti mais **[un garçon]**[ext2] **[qui]**[ext2] s'appelait **[Emilien]**[ext2] trouva **[X2]**[robot]. **[Il]**[ext2] **[l']**[robot] emmena chez **[lui]**[ext2] pour **[le]**[robot] réparer. **[Il]**[ext2] essaya des jours, des jours et des jours mais impossible de **[le]**[robot] réparer. **[Emilien]**[ext2] allait au magasin de bricolage. **[Il]**[ext2] trouva les bons outils et **[put]**[ext2] réparer **[X2]**[robot] et tout est bien qui finit bien. FIN

CE2, 203

Hier **[un chat]**[cat] est allé chercher **[la sorcière]**[witch] pour le volcan. **[Le chat]**[cat] dit « **[dépêchez]**[witch] il **[nous]**[cat,witch] reste 30 minutes » **[la sorcière]**[witch] « allons **[chat]**[cat] il faut se dépêcher **[allons]**[cat,witch] maintenant il **[nous]**[cat,witch] reste 20 minutes 10 minutes » fin

CE2, 204

{[La sorcière][witch] et **[le loup]**[wolf].}

Il était une fois **[une petite fille rouge]**[ext1] **[qui]**[ext1] entra dans les bois sombres. **[Elle]**[ext1] voit une maison en bonbon. **[La petite fille]**[ext1] toqua à la porte toc toc toc **[la petite fille]**[ext1] entra dans la maison. **[Elle]**[ext1] monta dans l'escalier et **[elle]**[ext1] voit une lumière et c'était une marmite cachée derrière la porte. Il y avait **[une sorcière]**[witch]. **[La petite fille]**[ext1] se retourne et voilà que **[la sorcière]**[witch] se fait dévorer par **[le loup]**[wolf]. **[La petit fille]**[ext1] sort une mitraillette de **[sa]**[ext1] poche et **[la petite fille]**[ext1] transperce **[le loup]**[wolf] de la jusqu'à la tête.

CE2, 207

[Une sorcière][witch] avait une maison noire et **[un chat noir]**[cat]. **[Il]**[cat] se cachait dans la maison pour ne pas qu'on **[le]**[cat] voie. **[Il]**[cat] ne sortait que la nuit. Un jour **[la sorcière]**[witch] en avait marre. Avec **[sa]**[witch] baguette magique **[elle]**[witch] **[l']**[cat]a transformé en **[chat vert]**[cat] comme ça **[elle]**[witch] **[le]**[cat] voyait tout le temps.

CE2, 211

Il était une fois **[un petit chat]**[cat] **[qui]**[cat] s'appelait **[Moustache]**[cat]. **[Il]**[cat] dormait tranquillement sur le canapé, **[il]**[cat] avait entendu du bruit sous le canapé, c'était **[une petite souris]**[ext1]. **[La souris]**[ext1] alla dehors et **[le chat]**[cat] rencontre **[le loup]**[wolf] en chemin. **[Le chat]**[cat] est parti en courant pour arriver à la maison. Pas loin il y avait **[une sorcière]**[witch] avec **[son]**[witch] balai magique. **[La souris]**[ext1] et **[le chat]**[cat] sont ressortis et **[la sorcière]**[witch] arrive avec **[son]**[witch] balai magique et une baguette magique à la main et **[tous les deux]**[cat,ext1] **[ils]**[cat,ext1] se sont reposés.

CE2, 855

Dans une forêt il y avait une maison où habitaient **[une sorcière]**[witch] et **[[son]witch chat]**[cat]. **[[Son]witch chat]**[cat] était noir avec des yeux perçants qui faisaient peur à un chien. Et **[la sorcière]**[witch] ce n'était pas mieux, **[elle]**[witch] attrapait **[un enfant]**[ext1] et "hop" **[elle]**[witch] **[le]**[ext1] mettait directement dans une cage avec des serpents, des asticots... Et **[elle]**[witch] fermait la cage à double tour. Comme ça **[l'enfant]**[ext1] ne pouvait pas s'échapper. Sinon **[elle]**[witch] **[le]**[ext1] mettait directement dans la marmite. Mais un jour, **[elle]**[witch] attrapa **[un enfant]**[ext2] et **[elle]**[witch] **[le]**[ext2] mit dans la cage. **[Elle]**[witch] voulait aller chercher **[un enfant]**[ext3] à la ville qui était très très très loin. **[Elle]**[witch] était bien pressée, du coup **[elle]**[witch] oublia de fermer la cage et du coup **[elle]**[witch] avait demandé à **[[son]witch chat]**[cat] de surveiller la cage mais **[le chat]**[cat] a vu **[une souris]**[ext4] passer, du coup **[le chat]**[cat] **[la]**[ext4] poursuit. **[Le chat]**[cat] ouvre la porte et **[poursuit]**[cat] **[la souris]**[ext4]. Du coup **[le petit enfant]**[ext2] sort de **[sa]**[ext2] cage et comme **[la sorcière]**[witch] avait laissé les clés **[le petit enfant]**[ext2] ferme la porte à double tour et **[la sorcière]**[witch] avait laissé **[sa]**[witch] baguette, **[elle]**[witch] jette un sort qui bloque la porte. Et ensuite **[la sorcière]**[witch] revient avec **[3 enfants]**[ext5]. Mais **[elle]**[witch] ne peut pas rentrer. Du coup **[elle]**[witch] pose le sac où il y a **[l'enfant]**[ext3] et **[les 3 enfants]**[ext5] s'échappent et **[le petit enfant]**[ext2] **[qui]**[ext2] est dans la maison appelle **[les 3 autres]**[ext5] pour venir. Puis **[la sorcière]**[witch] voit **[[son]witch chat]**[cat] revenir avec **[une souris]**[ext4] alors **[la sorcière]**[witch] dit à **[[son]witch chat]**[cat] « **[t']**[cat] aurais dû la surveiller, **[j']**[witch] avais laissé la cage ouverte ».

CE2, 2983

C'est l'histoire d'**[un chat]**[cat] **[qui]**[cat] se faisait toujours embêter par **[un loup]**[wolf]. **[Ce méchant loup]**[wolf] n'arrêtait pas d'aller chez **[le chat]**[cat] et de mettre le bazar. **[Ce pauvre animal]**[cat] dor mais un jour **[Rouqui le chat]**[cat] décida de mettre des pièges de partout. Pour que **[Maléficore]**[wolf] ne rentre plus, **[Rouqui]**[cat] met des tapettes à souris. **[Il]**[cat] mit des trous avec des feuilles dessus tout autour de la maison **[du chat]**[cat]. Un soir **[Maléficore]**[wolf] vient, **[il]**[wolf] voit les feuilles donc **[il]**[wolf] rentre vite dans la maison **[du chat]**[cat], **[il]**[wolf] croit que c'était un piège les feuilles. **[Il]**[wolf] avança et **[il]**[wolf] marcha sur une tapette à souris, **[il]**[wolf] cria « aie aie aie » **[il]**[wolf] sort vite de la maison mais **[il]**[wolf] tomba dans un trou et **[il]**[wolf] resta coincé. **[Le chat]**[cat] vient voir **[le loup]**[wolf], **[Maléficore]**[wolf] dit « **[sors]**[cat] **[-moi]**[wolf] d'ici s'il **[te]**[cat] plait ». **[Le chat]**[cat] dit « oui mais **[tu]**[wolf] arrêtes de **[m']**[cat]embêter » **[le loup]**[wolf] dit « oui ». **[Le chat]**[cat] **[le]**[wolf] sort et **[ils]**[cat,wolf] deviennent amis.

CE2, 1339

{[Le chat][cat] et **[la sorcière]**[witch]}

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] s'appelait **[Baba - yaca]**[witch]. Un jour **[[son]**[**witch**] **chat Loulanae]**[cat] avait disparu

« - ho ! non **[j']**[witch] ai perdu **[[mon]**[**witch**] **chaton de dix semaines]**[cat] ». **[La sorcière]**[witch] était très très triste, **[la mère de [Loulanae]**[**cat**]**]**[cat1] était triste **[elle]**[cat1] aussi. **[La mère]**[cat1] alla **[la]**[witch] consoler sur **[ses]**[witch] genoux même **[[son]**[**witch**] **balai]**[ext1] **[qui]**[ext1] s'appelle **[Frédie]**[ext1]. Mais le lendemain matin **[le chaton]**[cat] revenait.

[La sorcière][witch] répétait toujours « - quel miracle ! **[Loulanae]**[cat] est revenu. »

CE2, 1346

{[Le chat Fiti][cat]} était **[un chat gris]**[cat], **[il]**[cat] était roux et très câlin, **[il]**[cat] jouait toujours avec **[sa]**[cat] baballe. Un jour **[il]**[cat] vit **[[son]**[**cat**] **copain Toufi]**[cat1], **[il]**[cat1] était comme **[lui]**[cat], **[il]**[cat1] allait chez **[le vétérinaire]**[ext1] pour **[sa]**[cat1] visite des 5 mois mais **[Fiti]**[cat] attend longtemps longtemps longtemps puis **[il]**[cat] vit la voiture de **[Toufi]**[cat1]. **[Il]**[cat] court vers la voiture mais **[Toufi]**[cat1] n'était pas dans la voiture. **[La maîtresse de [Toufi]**[**cat1**]**]**[ext2] dit à **[Fiti]**[cat] « **[Toufi]**[cat1] est encore **[au vétérinaire]**[ext1]. **[Il]**[cat1] reviendra demain soir » alors **[Fiti]**[cat] attend demain soir « **[Toufi]**[cat1] **[tu]**[cat1] es enfin là, qu'est -ce que **[tu]**[cat1] as à la patte ? » demande **[Fiti]**[cat] » « un plâtre, **[je]**[cat1] me suis cassé une patte » « Hooo ! alors **[tu]**[cat1] peux courir ou pas ? » « Non **[je]**[cat1] ne peux pas tant que **[mon]**[cat1] plâtre n'est pas enlevé. **[J']**[cat1] irai l'enlever dans 2 mois » « quoi ? deux mois, **[tu]**[cat1] veux rire ? » « NON NON » alors **[Fiti]**[cat] attend que les deux mois soient passés.

Aujourd'hui **[Toufi]**[cat1] va enlever **[son]**[cat1] plâtre. **[Il]**[cat1] revint « **[Toufi]**[cat1] on va jouer ? »

« NON **[le vétérinaire]**[ext1] en enlevant **[mon]**[cat1] plâtre **[m']**[cat1] a ouvert la patte, **[il]**[ext1] **[me]**[cat1] l'a recousue mais ça saigne alors **[je]**[cat1] ne dois pas courir » alors **[Tifi]**[cat] s'ennuie pendant 5 années. Puis un jour **[Toufi]**[cat1] vint attaquer **[Tifi]**[cat] « qu'est -ce qui **[t']**[cat1] arrive ? » « **[viens]**[cat], **[on]**[cat,cat1] va jouer. »

FIN

CE2, 1352

{[Le chat][cat] et **[la sorcière]**[witch]}

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] s'appelait **[Craboudine]**[witch]. **[Elle]**[witch] avait **[un chat]**[cat], **[il]**[cat] s'appelait **[Grifdure]**[cat], **[il]**[cat] était très intelligent. **[Ils]**[cat,witch] restaient toujours ensemble, rien ne pouvait **[les]**[cat,witch] séparer. Un jour catastrophe **[le chat]**[cat] avait fait une grosse bêtise en jouant, **[il]**[cat] avait renversé la potion de **[[sa]**[**cat**] **maîtresse]**[witch], **[elle]**[witch] était si furieuse qu'**[elle]**[witch] abandonna **[[son]**[**witch**] **chat]**[cat].

CE2, 3004

{[Le loup de [Mina]^{ext1}]^{wolf}}

A l'école, **[Mina]**^{ext1} **[me]**^{ext2} dit que **[[son]^{ext1} chien]**^{wolf} a aboyé bizarrement. « Mais comment ? » demande **[Anaïs]**^{ext3}. **[Mina]**^{ext1} **[lui]**^{ext3} répond « comme un loup », « **[on]**^{ext1,ext2,ext3} pourrait aller **[le]**^{wolf} voir ce soir ? » « Oh » dit **[Mina]**^{ext1}. Le soir **[on]**^{ext1,ext2,ext3} y va. « Oh on dirait **[un loup]**^{wolf} ! » « D'accord c'est pour ça. »

CE2, 3012

Il était une fois **[un chat abandonné]**^{cat}. Un jour il y avait **[une sorcière très méchante]**^{witch} mais avec les chats **[elle]**^{witch} était trop gentille. Mais avec les humains non. **[Elle]**^{witch} rentra chez **[elle]**^{witch} avec **[le chat]**^{cat} et **[elle]**^{witch} demanda **[au chat]**^{cat} « veux-**[tu]**^{cat} **[me]**^{witch} donner un poil de **[toi]**^{cat} ? C'est pour les méchants humains » **[le chat]**^{cat} répond « bon ben allez, si **[tu]**^{witch} veux » et **[la sorcière]**^{witch} cracha **[son]**^{cat} poil pour le mettre dans la grande soupe pour tuer les humains. Et **[la sorcière]**^{witch} s'appelait **[Océane]**^{witch}, **[la sorcière]**^{witch} était très maléfique mais seulement qu'avec les humains. Et il y avait **[le chat]**^{cat} que **[la sorcière]**^{witch} avait adopté, **[le chat]**^{cat} car un jour **[il]**^{cat} était abandonné et maintenant **[le chat]**^{cat} est à **[la sorcière]**^{witch} et **[la sorcière]**^{witch} était méchante avec **[le chat]**^{cat}.

CE2, 2453

{[La Sorcière]^{witch} et **[[son]^{witch} chat]**^{cat}}

[Une sorcière]^{witch} veut que **[[son]^{witch} chat]**^{cat} parle. **[Elle]**^{witch} fait la potion mais il **[lui]**^{witch} manque 1 ingrédient. Donc **[elle]**^{witch} demande à **[[son]^{witch} chat]**^{cat} d'aller le chercher et **[elle]**^{witch} a dit : « si **[tu]**^{cat} ne le ramèneras pas avant 2 jours **[je]**^{witch} **[te]**^{cat} tuerai ! » Donc **[le chat]**^{cat} se mit tout de suite. **[Il]**^{cat} ne le ramène pas tout de suite et **[la sorcière]**^{witch} allait tuer **[[son]^{witch} chat]**^{cat}. Mais **[elle]**^{witch} se dit « Comment **[elle]**^{witch} va faire pour aller chercher les ingrédients ? »

CE2, 1845

{[Le chat]^{cat}}

Il était une fois **[un chat]**^{cat} **[qui]**^{cat} était perdu dans la forêt. **[Il]**^{cat} recherchait **[[sa]^{cat} patronne]**^{ext1} et un jour **[il]**^{cat} trouvait une maison. Et **[il]**^{cat} faisait mia mia et il y a **[une dame]**^{ext2} **[qui]**^{ext2} sortait dehors et **[elle]**^{ext2} s'appelait **[Carine]**^{ext2}. **[Elle]**^{ext2} avait 70 ans et **[le chat]**^{cat} **[il]**^{cat} disait mia mia au minou. **[Le pauvre]**^{cat} **[il]**^{cat} recherchait **[[sa]^{cat} patronne]**^{ext1} « **[entre]**^{cat} **[[mon]^{ext2} petit]**^{cat}. **[Tu]**^{cat} as soif ? » « mia » « **[mange]**^{cat} **[[mon]^{ext2} petit]**^{cat} » « **[je]**^{ext2} vais t'appeler **[Paul]**^{cat}. Et **[je]**^{ext2} vais **[t']**^{cat} emmener au cabinet pour animaux pour voir **[ton]**^{cat} âge » « bonjour **[il]**^{cat} a quel âge **[ce petit Paul]**^{cat} ? » « **[il]**^{cat} a 3 ans » « merci, au revoir **[madame]**^{ext2} » « ça va **[minou]**^{cat} ? » fin de l'histoire.

CE2, 1848

{[Le loup][wolf] et **[le chaperon rouge]**[ext1]}

« Bonjour **[loup]**[wolf], ça va ? » « oui ça va et **[toi]**[ext1] ? » « oui » « **[Tu]**[wolf] vas où ? chez **[[ma]**[ext1] **grand-mère]**[ext2] ? **[t']**[wolf] as quelques minutes pour passer chez **[moi]**[ext1] ? » « Oui mais pas longtemps d'accord » « comme ça **[tu]**[wolf] pourras boire un petit cocktail » « oui » « mais aussi ce n'est pas tout, aussi une tarte » « oui mais vite alors » « oui ben **[j']**[wolf] arrive mais c'est où au juste ? » « en face ».

CE2, 1851

Il était une fois **[une sorcière]**[witch], **[elle]**[witch] aime beaucoup les chats. **[Elle]**[witch] en avait **[trois]**[cat,cat1,cat2], **[le premier Grisette]**[cat], **[le deuxième Myrtille]**[cat1] et **[le troisième Canelle]**[cat2]. **[Les chats]**[cat,cat1,cat2] pouvaient parler car **[la sorcière]**[witch] **[leur]**[cat] avait donné le don. **[Les chats]**[cat,cat1,cat2] sortent de la maison de **[la sorcière]**[witch] puis **[vont chasser]**[cat,cat1,cat2] les souris. **[Les trois chats]**[cat,cat1,cat2] voient **[une souris]**[ext1], dès qu'**[ils]**[cat,cat1,cat2] s'approchent de **[la souris]**[ext1], « **[la souris]**[ext1] se transforme en **[monstre hideux]**[ext1] » ! disent **[-ils]**[cat,cat1,cat2]. **[Ils]**[cat,cat1,cat2] courent en voyant **[le monstre]**[ext1], **[ils]**[cat,cat1,cat2] se perdent dans la forêt « **[On]**[cat,cat1,cat2] s'est perdus dans la forêt ! »

[Les chats][cat,cat1,cat2] marchent en quête de maison. Tout à coup **[ils]**[cat,cat1,cat2] voient une maison au milieu de la forêt. **[Ils]**[cat,cat1,cat2] pensent qu'elle pourra **[les]**[cat,cat1,cat2] abriter pendant la nuit. Pendant ce temps **[la sorcière]**[witch] remarque que **[les chats]**[cat,cat1,cat2] ne sont plus là. **[Elle]**[witch] se met en route avec **[son]**[witch] balai volant. **[Elle]**[witch] va dans la forêt. Au matin **[les chats]**[cat,cat1,cat2] se réveillent et **[retournent]**[cat,cat1,cat2] dans la forêt et **[voient]**[cat,cat1,cat2] **[un balai volant]**[ext2]. « **[Balai]**[ext2] **[ramène]**[ext2] **[-nous]**[cat,cat1,cat2] dans la ville ». **[Les chats]**[cat,cat1,cat2] rentrent chez **[eux]**[cat,cat1,cat2], **[ils]**[cat,cat1,cat2] ne voient pas **[[leur]**[**cat2,cat,cat1**] **mère]**[witch] dans la maison. **[La mère [des chats]**[**cat1,cat,cat2**]**]**[witch] rentre désespérée, à la maison **[elle]**[witch] **[les]**[cat,cat1,cat2] voit très contents.

CE2, 1323

Il était une fois **[un robot]**[robot] **[qui]**[robot] se promenait à la montagne. **[Il]**[robot] rencontre **[un chat tout blanc aux beaux yeux bleus]**[cat] et **[le robot]**[robot] demanda **[au tout petit chat]**[cat] « Bonjour que faites **[-vous]**[cat] **[M. Le chat]**[cat] ? »

« **[Je]**[cat] joue à cache-cache avec **[[mes]**[**cat**] **copains]**[ext1], et **[vous]**[robot] que faites **[-vous]**[robot] **[M. Le robot]**[robot] ? »

« **[Je]**[robot] me balade » s'exclama **[le robot]**[robot].

« Voulez -vous bien être **[mon]**[cat] ami ? » demanda **[le chat]**[cat] **[au magnifique robot jaune, vert et bleu]**[robot]. « Oui **[je]**[robot] veux bien mais par contre **[on]**[cat,robot] se tutoie. »

« D'accord **[je]**[cat] veux bien **[te]**[robot] tutoyer mais à condition que ce soir **[on]**[cat,robot] va faire du patin à glace, à la patinoire du petit village d'en-dessous. »

« D'accord » dit **[le robot]**[robot]. »

Et c'est pour ça que les chats sont des animaux domestiques.

FIN

De l'histoire **[du chat]**[cat] et **[du robot]**[robot].

CE2, 1329

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] était plutôt gentille, et **[qui]**[witch] aimait beaucoup les chats, alors un jour **[elle]**[witch] fit apparaitre **[un chat]**[cat], mais **[le chat]**[cat] était très coquin. Un jour **[la sorcière]**[witch] fit une bouillie de crapauds et d'escargots et **[elle]**[witch] alla chercher les araignées. Pendant qu'**[elle]**[witch] alla les chercher **[le chat]**[cat] alla prendre une grosse poignée de sel et la mit dans le chaudron. Quelques secondes après **[la sorcière]**[witch] arriva, **[le chat]**[cat] retourna vite dans **[son]**[cat] panier, **[elle]**[witch] mit les araignées dans le chaudron et **[gouta]**[witch] la soupe. **[Elle]**[witch] la trouva beaucoup trop salée, **[elle]**[witch] s'énerva et puis **[se demanda]**[witch] qui avait pu bien mettre du sel dans le chaudron, puis **[elle]**[witch] chercha mais **[ne trouva pas]**[witch] le coupable.

Le lendemain pour le déjeuner **[elle]**[witch] fit une pâte et **[mit]**[witch] des vers de terre mais **[le chat]**[cat] était bien trop occupé pour faire une phrase, à la place **[il]**[cat] fit **[sa]**[cat] toilette.

FIN !

CE2, 1512

Il était une fois **[un chat très gentil]**[cat]. **[Il]**[cat] avait **[des amis]**[cat1], **[ils]**[cat,cat1] jouent à cache-cache. **[Les chats]**[cat1] **[ils]**[cat1] se cachent dans la forêt et **[l'autre chat]**[cat] comptait jusqu'à 30 et après **[il]**[cat] a fini de compter, **[il]**[cat] cherche **[les chats]**[cat1] quand tout à coup **[une sorcière]**[witch] **[elle]**[witch] dit « **[tu]**[cat] fais quoi ici ? » avec la voix bizarre, **[le chat]**[cat] dit « je cherche **[[mes]**[**cat**] **amis]**[cat1] », avec la voix un peu peur, **[elle]**[witch] dit « **[je]**[witch] vais **[te]**[cat] transformer en citrouille » mais **[le chat]**[cat] en même temps **[il]**[cat] court et **[il]**[cat] cherchait **[[ses]**[**cat**] **amis]**[cat1] mais **[il]**[cat] ne trouve pas alors **[il]**[cat] crie **[LES AMIS]**[cat1] AU SECOURS, et **[les amis]**[cat1] aident, **[ils]**[cat1] mettent des coups ici des coups par là et **[la sorcière]**[witch] avait mal de partout mais **[la sorcière]**[witch] avait pris **[sa]**[witch] baguette magique pour enlever tous les bobos et **[elle]**[witch] dit « **[vous]**[cat,cat1] ne pouvez pas **[me]**[witch] tuer » « eh **[les chats]**[cat1] il faut détruire **[sa]**[witch] baguette » mais **[ils]**[cat,cat1] n'arrivent pas mais tout à coup **[un énorme chat]**[cat2] et **[il]**[cat2] dit « **[vous]**[cat,cat1] avez besoin de **[moi]**[cat2] ? » **[les chats]**[cat,cat1] disent oui. **[Ils]**[cat,cat1,cat2] cassaient la baguette magique. **[Ils]**[cat,cat1,cat2] cassent la baguette magique et **[le chat]**[cat2] autre chose, **[les chats]**[cat,cat1] disent « oui, quoi, attaquer **[la sorcière]**[witch] ? et **[il]**[cat2] l'a fait » **[les chats]**[cat,cat1] disent oui oui, **[on]**[cat,cat1] a attaqué **[sorcière]**[witch]. Fin

CE2, 562

Il était une fois **[un loup]**[wolf] et **[[son]**[**wolf**] **fidèle serviteur le robot]**[robot]. **[Le robot]**[robot] faisait tout ce que **[le loup]**[wolf] demandait mais un jour **[le robot]**[robot] dit **[au loup]**[wolf] : « **[Je]**[robot] m'en vais. » **[Le loup]**[wolf] dit « : - Pourquoi pars - **[tu]**[robot] ? » **[Le robot]**[robot] dit : « **[Je]**[robot] prends grève, **[je]**[robot] suis fatigué » dit **[le robot]**[robot]. « Mais un robot est fait de fer, il n'est jamais fatigué, » dit **[le loup]**[wolf] « Mais **[je]**[robot] prends grève quand même. **[Je]**[robot] prends grève de 100 ans. » **[Le loup]**[wolf] reste bouche bée devant **[le robot]**[robot] **[qui]**[robot] s'en va. Fin.

CE2, 576

{[La sorcière][witch] et **[[sa]**[**witch**] **chatte]**[cat].}

[La sorcière][witch] et **[[sa]**[**witch**] **chatte]**[cat] font une potion magique avec des plantes et des serpents et des myrtilles tellement **[elles]**[cat,witch] étaient les plus gentilles et des fois **[étaient]**[cat,witch] des fois méchantes. **[La sorcière]**[witch] s'appelle **[Justine]**[witch] et **[[sa]**[**witch**] **chatte]**[cat] s'appelle **[Chillie]**[cat].

CE2, 1596

Il était une fois **[un chat]**[cat] et **[une sorcière]**[witch]. **[Le chat]**[cat] était sur **[son]**[cat] tapis pendant que **[la sorcière]**[witch] est en train de travailler, et d'un coup **[le chat]**[cat] miaule, **[la sorcière]**[witch] **[lui]**[cat] donne de la pâté de chat. **[Le chat]**[cat] miaule toujours donc **[la sorcière]**[witch] met **[le chat]**[cat] dehors. **[Le chat]**[cat] miaule, **[saute]**[cat] par la fenêtre et d'un coup **[la sorcière]**[witch] crie et **[le chat]**[cat] miaule encore plus fort donc **[la sorcière]**[witch] refait rerentrer **[le chat]**[cat] et **[le chat]**[cat] ne miaule plus et **[le chat]**[cat] s'endort et quand **[le chat]**[cat] se réveille **[la sorcière]**[witch] était avec **[lui]**[cat]. Pour le petit-déjeuner **[le chat]**[cat] mange des croquettes et **[la sorcière]**[witch] mange un bol de lait et deux tartines de beurre et confiture. **[La sorcière]**[witch] et **[le chat]**[cat] sortent dehors. **[La sorcière]**[witch] donne la balle **[au chat]**[cat] et **[le chat]**[cat] fait la passe à **[la sorcière]**[witch] et ainsi de suite, **[la sorcière]**[witch] dit **[au chat]**[cat] « c'est l'heure de manger **[[mon]**[**witch**] **petit chat]**[cat] » **[elle]**[witch] **[lui]**[cat] a préparé des croquettes.

CE2, 3055

Il était une fois **[un chat]**[cat] **[qui]**[cat] avait oublié **[sa]**[cat] nourriture dans un voyage. Alors **[il]**[cat] prenait la nourriture des gens. 2 semaines plus tard **[le chat]**[cat] n'avait plus faim. Alors **[le chat]**[cat] va se coucher. Et après **[il]**[cat] avait encore faim et **[il]**[cat] mange enfin **[sa]**[cat] nourriture.

CE2, 113

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] était seule, un jour **[elle]**[witch] créa **[un robot maléfique]**[robot] puis **[le robot]**[robot] en vie. **[Il]**[robot] combattait **[d'autres robots maléfiques]**[robot1]. **[Il]**[robot] **[les]**[robot1] détruisait avec un bouclier, une épée et une lance, puis **[la sorcière]**[witch] créa **[[son]**[**witch**] **armée de robots chevaliers en vie]**[robot2], puis **[elle]**[witch] créa **[son]**[witch] château créé par

[[ses]witch robots]robot2 puis **[ils]**robot2,witch font la guerre, contre **[une armée de robots romains]**robot3 puis **[il]**robot perd la bataille de l'Angleterre remportée par **[les robots chevaliers]**robot2 et **[[leurs]robot2 alliés]**ext1... !

FIN !

CE2, 97

Il était une fois **[une sorcière]**witch habitant dans un château fort. Il était immense. Il y avait de grandes tours. **[Elle]**witch s'appelait **[Camila]**witch. Un soir **[elle]**witch allait se promener, tout d'un coup **[elle]**witch croisa **[un chat]**cat, **[il]**cat était tout noir et **[il]**cat avait une tache blanche. **[Il]**cat miaula parce qu'**[il]**cat avait faim, alors **[la sorcière]**witch disait **[au balai]**ext1 « descends **[balai]**ext1 » et **[le balai]**ext1 descendait et **[la sorcière]**witch ramena **[le chat blanc et noir]**cat au château et **[elle]**witch **[lui]**cat donna du lait et un peu de croquettes **[au chat]**cat. **[Elle]**witch **[le]**cat faisait dormir avec **[elle]**witch dans **[son]**witch lit, **[il]**cat regardait la télévision avec **[la sorcière]**witch, **[il]**cat prenait la douche avec **[elle]**witch. Un jour **[la sorcière]**witch allait se promener et **[elle]**witch vit **[un loup]**wolf et **[lui]**wolf **[dit]**witch « bonjour **[loup]**wolf, **[tu]**wolf vas bien ? » « oui **[je]**wolf vais bien » et **[la sorcière]**witch reprit **[son]**witch chemin jusqu'au château.

FIN

FIN !!!

CE2, 1173

Il était une fois **[une sorcière]**witch, et **[[son]witch chat noir]**cat. Un bon matin commence. **[Elle]**witch prépara une potion magique. **[La sorcière]**witch mit une grenouille, des cailloux et une pièce de tout en 1. Puis **[le chat noir]**cat s'enfuit car **[la sorcière]**witch s'enfuit.

CE2, 3057

Il était une fois **[une sorcière]**witch qui s'appela **[Amanda]**witch, **[elle]**witch habitait dans une maison à l'orée du la forêt, avec **[[son]witch chat Mistigri]**cat. **[Ils]**cat,witch voulaient transformer **[les enfants]**ext1 en crapauds. Pendant ce temps **[la sorcière]**witch prépara le diner, et **[Mistigri]**cat **[lui]**cat jouait avec **[[ses]cat copines les souris]**ext1. Après le diner, **[il]**cat prépara une grande fête. Il y avait de la musique. Juste à côté de **[leur]**cat,witch maison habitait **[un loup]**wolf. Quand **[il]**wolf entend la musique **[il]**wolf a dit « c'est pas fini ce boucan ? », **[la sorcière]**witch dit « excuse, **[on]**cat,witch fait une fête ». « Une fête ? » dit **[le loup]**wolf « mais **[moi]**wolf **[j']**wolf adore les fêtes ! » « **[vous]**wolf voulez venir ? » **[le loup]**wolf répondit « oui ». Mais **[la sorcière]**witch se rend compte qu'**[elle]**witch n'a pas transformé **[les enfants]**ext1, **[le loup]**wolf dit « mais ce n'est pas grave, **[on]**cat,witch,wolf s'amuse ».

CE2, 3060

{[Le loup Cédric][wolf] veut se marier}

Il était une fois **[un loup seul et sans famille]**[wolf], **[il]**[wolf] voulait absolument une famille vu qu'**[il]**[wolf] était **[le seul loup de la forêt]**[wolf]. Une heure deux heures passées toujours aucun loup. **[Le loup Cédric]**[wolf] en a eu marre donc **[il]**[wolf] décidait de quitter la forêt. Et d'aller au village. **[Les habitants]**[ext1] étaient terrorisés. **[Le loup Cédric]**[wolf] ne savait pas que les humains avaient peur des loups. **[Il]**[wolf] a vu **[une femelle loup]**[wolf1] qu'on a capturée. **[Il]**[wolf] **[la]**[wolf1] sauva et **[ils]**[wolf,wolf1] tombèrent amoureux **[l'un de l'autre]**[wolf,wolf1] et **[ils]**[wolf,wolf1] se marièrent et **[ils]**[wolf,wolf1] ont eu **[2 enfants]**[ext2].

Fin.

CE2, 1188

Il était une fois **[un robot]**[robot] **[qui]**[robot] avait **[un chat]**[cat]. Mais dans la forêt **[le robot]**[robot] et **[le chat]**[cat] ne savaient pas ce qu'il y avait. Il y avait **[un loup un méchant loup]**[wolf]. Et **[lui]**[wolf] il savait ce qu'il y avait dans la cabane. **[Il]**[wolf] savait qu'il y avait **[un robot]**[robot] et **[un chat]**[cat]. Un jour **[le robot]**[robot] ne trouvait plus **[[son]**[**robot**] **chat]**[cat]. **[Le loup]**[wolf] **[le]**[cat] **[lui]**[robot] avait piqué. **[Il]**[wolf] **[l']**[cat] avait emporté chez **[lui]**[wolf]. Dans **[sa]**[wolf] maison **[il]**[wolf] ne **[lui]**[cat] donnait que du thon. **[Le chat]**[cat] aimait bien comme tous les chats. De son côté **[le robot]**[robot] pleurait car **[ce chat]**[cat] était magique pour **[lui]**[robot]. Donc un jour **[il]**[robot] allait chez **[le loup]**[wolf] à cinq heures du matin pour aller chercher **[[son]**[**robot**] **chat]**[cat]. **[Le loup]**[wolf] dormait avec **[le chat]**[cat] **[qui]**[cat] commençait à tomber dans la marmite **[du loup]**[wolf]. **[Le robot]**[robot] récupère **[[son]**[**robot**] **chat]**[cat] et quand **[le loup]**[wolf] chercha dans **[sa]**[wolf] marmite **[il]**[wolf] criait : « non ! »

CE2, 1197

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] vivait dans les montagnes et **[qui]**[witch] détestait les robots. Un jour **[un robot]**[robot] part dans les montagnes, **[[son]**[**witch**] **chat]**[cat] avait vu **[le robot]**[robot].

[Le chat][cat] est allé prévenir **[la sorcière]**[witch] et **[la sorcière]**[witch] était vraiment très furieuse. **[La sorcière]**[witch] a fait apparaitre **[le loup géant]**[wolf], **[le loup]**[wolf] a tué **[le robot]**[robot], puis **[le robot]**[robot] s'est reconstruit et **[il]**[robot] a tué **[la sorcière]**[witch], **[le chat]**[cat] puis **[le loup]**[wolf]. Puis **[il]**[robot] est reparti chez **[lui]**[robot].

CE2, 1212

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] avait **[un chat]**[cat]. **[Le chat]**[cat] était avec **[la sorcière]**[witch] dans une maison en pain d'épices. **[Quelqu'un]**[wolf] frappe à la porte. **[La sorcière]**[witch] ouvrit la porte et **[un loup]**[wolf] entra. **[Le loup]**[wolf] entre dans la maison et **[voit]**[wolf] **[le chat]**[cat]. **[Le chat]**[cat] voit **[le loup]**[wolf]. Et **[le chat]**[cat] court dans la maison. **[La sorcière]**[witch] dit « stop ! » **[Le robot]**[robot] est l'ami de **[la sorcière]**[witch]. **[Le robot]**[robot] voit **[le chat]**[cat] et **[le loup]**[wolf], **[il]**[robot] s'en va. **[La sorcière]**[witch]

dit **[au robot]**[robot] que c'est **[[mes]**[**witch**] **amis]**[cat,wolf]. Et **[le robot]**[robot] rentre dans, et **[ils]**[cat,robot,witch,wolf] font la fête.

CE2, 1531

[Le chat][cat] joue avec **[sa]**[cat] balle. **[Il]**[cat] la lance de tous les côtés. **[Le robot]**[robot] arrive et puis **[il]**[robot] joue avec **[le chat]**[cat]. Mais **[le chat]**[cat] n'est pas d'accord. **[Il]**[cat] **[le]**[robot] mord. Mais **[le robot]**[robot] **[le]**[cat] met dehors. **[Le chat]**[cat] décide d'aller jouer tout seul dehors. Mais **[le chat]**[cat] s'ennuie tout seul. Mais **[le robot]**[robot] s'ennuie aussi. **[Il]**[robot] n'a rien à faire alors **[il]**[robot] fait rentrer **[le chat]**[cat] et puis **[ils]**[cat,robot] jouent ensemble à la balle. Fin

CE2, 1538

Il était une fois **[un petit chat]**[cat] **[qui]**[cat] était abandonné dans la forêt. Trois jours après **[le petit chat]**[cat] ne trouvait toujours pas le chemin, donc **[il]**[cat] n'avait rien à manger ni à boire. Mais un jour **[un enfant]**[ext1] et **[[ses]**[**ext1**] **parents]**[ext2] trouvèrent **[le petit chat]**[cat]. **[Ils]**[ext1,ext2] **[le]**[cat] trouvèrent très joli, **[les parents]**[ext2] dirent de **[le]**[cat] laisser **[le petit chat]**[cat] mais **[l'enfant]**[ext1] **[le]**[cat] voulait tellement que **[les parents]**[ext2] dirent oui pour adopter **[le petit chat]**[cat] **[décidèrent]**[ext1,ext2] alors de **[l']**[cat] appeler **[Mimi]**[cat]. **[Mimi]**[cat] a rencontré **[une chatte]**[cat1] et puis **[la chatte]**[cat1] et **[Mimi]**[cat] sont devenus amoureux et puis **[ils]**[cat,cat1] ont fait **[7 bébés]**[cat2,cat3,cat4,cat5,cat6,cat7,cat8]. **[Ces bébés]**[cat2,cat3,cat4,cat5,cat6,cat7,cat8] s'appellent **[Mini]**[cat2], **[Pêche]**[cat3], **[Abricot]**[cat4], **[Pomme]**[cat5], **[Menthe]**[cat6], **[Bela]**[cat7] et **[Cerise]**[cat8]. Fin

CE2, 1544

Il était une fois **[un petit chat au pelage très doux]**[cat] **[qui]**[cat] vivait avec **[un robot grand]**[robot], **[il]**[robot] faisait 1 mètre 50. **[Le chat]**[cat] était content d'avoir **[un maitre aussi gentil]**[robot], en plus **[il]**[cat] avait beaucoup de croquettes. Mais bon **[le robot]**[robot] n'allait pas souvent aux courses. Je vais vous raconter un moment où **[le chat]**[cat] parle **[au robot]**[robot] en allant aux courses. « Est -ce que **[tu]**[robot] peux prendre beaucoup de croquettes ? Parce qu'**[on]**[cat,robot] ne va pas souvent aux courses ! »

« Non **[je]**[robot] prendrai comme d'habitude et puis ça **[te]**[cat] suffit bien non ? »

« Non ça ne **[me]**[cat] suffit pas, même qu'à chaque fois **[j']**[cat] en raffole. »

« Et bah tant pis pour **[toi]**[cat] mais **[moi]**[robot] **[je]**[robot] trouve que **[tu]**[cat] en as assez. »

« **[Tu]**[robot] es très méchant ! Pour la peine **[je]**[cat] **[te]**[robot] quitte. »

Voilà c'est la fin des courses.

[Le chat][cat] trouva une maitresse qui n'était pas un robot.

Et **[le robot]**[robot] **[une chatte]**[cat1] !

Finalement tout est bien qui finit bien !

FIN

CE2, 1554

Une nuit alors que **[la sorcière]**witch prépara une potion. **[Le chat abandonné]**cat regarda ce que **[la sorcière]**witch mit dans la potion. **[Il]**cat vit **[la sorcière]**witch mettre dans la potion du poison. **[La sorcière]**witch attrapa **[le chat]**cat. **[Il]**cat essaya de s'échapper mais **[il]**cat n'y arriva pas. **[Il]**cat devint **[le chat de [la sorcière]witch]**cat. Mais un jour **[un autre chat]**cat1 arriva et **[tua]**cat1 **[la sorcière]**witch. **[Le chat abandonné]**cat remercia **[l'autre chat]**cat1. **[Les deux chats]**cat,cat1 tombaient amoureux. **[Il]**cat partit de la maison de **[la sorcière]**witch. **[Le chat]**cat1 demanda à **[son maitre]**ext1 s'**[il]**ext1 pouvait adopter **[le chat abandonné]**cat. **[Le maitre]**ext1 dit « oui. Mais à condition que **[vous]**cat,cat1 soyez sages. » « OK. »

FIN

CE2, 1627

{[Le chat de [la sorcière]witch]cat}

Il était une fois **[une sorcière malheureuse]**witch car **[elle]**witch n'avait pas de chat. Un jour **[elle]**witch décida de partir à la recherche d'un chat noir et aux yeux rouges, un moment **[elle]**witch trouva **[quelque chose de noir caché derrière un arbre]**ext1. **[Elle]**witch tira dessus... Et **[la chose]**ext1 se retourna d'un coup, **[la sorcière]**witch prit peur, **[elle]**witch cria : « OHAA ! » de terreur. **[La sorcière]**witch ne voulut plus jamais avoir un chat noir aux yeux rouges.

CE2, 1669

[Le loup]wolf part chasser **[des cochons]**ext1 après **[il]**wolf **[l']**ext1 emporta à **[sa]**wolf maison et **[le]**ext1 fait cuire à la marmite. **[Le loup]**wolf n'est pas content parce qu'il pleut et **[il]**wolf rentre chez **[lui]**wolf.

[**Le chat]**cat rentre chez **[le loup]**wolf sans faire exprès et **[le loup]**wolf invita **[le chat]**cat à diner et à dormir. **[Le loup]**wolf attrape des chats. **[Le loup]**wolf mange **[le chat]**cat.

CE2, 500

Il était une fois **[une sorcière]**witch et **[un chat]**cat **[qui]**cat,witch s'aimaient beaucoup. Un jour **[ils]**cat,witch partirent à la forêt, **[le chat]**cat vit **[un papillon]**ext1, et comme **[le chat]**cat adorait les papillons, **[il]**cat **[le]**ext1 suivit. **[La sorcière]**witch continua **[son]**witch chemin et **[ne se rendit compte]**witch de rien. **[Le chat]**cat peu à peu rentrait dans le noir total, et **[ne vit plus]**cat **[le papillon]**ext1. **[Il]**cat suivit ses traces et **[arriva]**cat enfin sur le bon chemin de **[la sorcière]**witch, mais **[la sorcière]**witch était rentrée dans une maison pour mettre cette maison en dégât, **[elle]**witch ressortit de là, **[se rendit compte]**witch qu'**[elle]**witch n'était plus avec **[le chat]**cat dans la maison et puis **[elle]**witch vit **[le chat]**cat, comme **[le chat]**cat **[l']**witch avait vue et **[ils]**cat,witch vécurent heureux toute **[leur]**cat,witch vie ensemble.

CE2, 3026

{[Le loup affamé]wolf}

Il était une fois, dans la forêt, **[un loup affamé]**wolf **[Il]**wolf cherchait de quoi se nourrir. **[Il]**wolf avait faim, très faim. **[Il]**wolf cherchait **[sa]**wolf nourriture de partout. Dans les ruelles, les prairies, les forêts, etc. Mais **[il]**wolf ne trouvait rien. Pourtant un jour, **[il]**wolf vit dans les herbes hautes d'une prairie, **[un chat]**cat, **[un gros chat tendre et dodu pour [un loup aussi affamé]wolf]**cat. Vite ! **[Le chat]**cat **[le]**wolf regardait, **[le loup]**wolf se cachait dans les hautes herbes. **[Il]**wolf avançait jusqu'**[au chat]**cat lentement mais sûrement. Et puis, **[il]**wolf bondit sur **[le chat]**cat **[qui]**cat n'avait pas le temps de s'enfuir et **[le]**cat **[mangea]**wolf. Et, depuis ce jour, **[le loup]**wolf cherchait un coin pour digérer **[le chat]**cat car **[il]**wolf avait aussi avalé les poils **[du chat]**cat.

Fin

CE2, 3027

« Bonjour **[je]**robot m'appelle **[Robotec]**robot et **[toi]**cat ? » dit **[le robot]**robot.

« **[Moi]**cat **[je]**cat m'appelle **[Chichat]**cat, » dit **[le chat]**cat.

« **[Tu]**robot veux jouer avec **[moi]**cat ? » dit **[le chat]**cat.

« Oui mais par contre pas de pelote de laine, » dit **[le robot]**robot.

« Pourquoi ? » dit **[le chat]**cat.

« Parce que **[je]**robot suis allergique à la laine et après **[j']**robot ai plein de boutons, » dit **[le robot]**robot.

« D'accord ! » dit **[le chat]**cat.

« Mais alors **[tu]**robot veux qu'**[on]**cat,robot joue à quoi ? » dit **[le chat]**cat.

« **[Je]**robot sais, à chat perché ! » dit **[le robot]**robot.

« C'est **[toi]**cat **[le chat]**cat, » dit **[le robot]**robot.

CE2, 3032

{[Le chat]cat et **[la sorcière]**witch}

Il était une fois **[un petit chat]**cat et **[une sorcière]**witch, **[la sorcière]**witch est très méchante avec les chats. Un jour **[elle]**witch voit **[le chat]**cat **[dit]**witch « Oh **[un chat]**cat, **[il]**cat **[me]**witch tente bien pour faire une soupe ». **[Elle]**witch dit **[au chat]**cat « **[viens]**cat **[mon minou]**cat. **[Je]**witch ne vais pas **[te]**cat manger. **[Entre]**cat dans **[ma]**witch maison. **[Je]**witch vais **[te]**cat donner à manger » **[le chat]**cat entre dans la maison, **[il]**cat voit des tas de bonnes choses à manger comme des gâteaux pour chat, des souris mais **[il]**cat voit aussi un chaudron. **[Il]**cat s'approcha et **[la sorcière]**witch **[le]**cat poussa et **[il]**cat

tombe dans le chaudron. **[La sorcière]**witch alla chercher des crapauds, des yeux d'araignées mais **[elle]**witch ne se doutait pas que **[le chat]**cat était très très malin. **[Il]**cat réussit à sortir du chaudron à temps, **[se cacha]**cat et **[la sorcière]**witch arriva. **[Elle]**witch prend ce qu'**[elle]**witch est allée chercher, **[elle]**witch le jetait dans le chaudron, **[prend]**witch **[son]**witch livre de magie et **[lut]**witch une potion magique : « Bygoudi Bygouda **[je]**witch veux que **[ce chat]**cat se transforme en soupe. » Mais **[elle]**witch ne voit pas d'explosion, **[elle]**witch ne vit plus **[le chat]**cat, **[elle]**witch regarda dans toute **[sa]**witch maison et **[elle]**witch ne retrouva pas **[le chat]**cat. Mais **[le chat]**cat appela **[la police]**ext1 et **[leur]**ext1 **[dit]**cat : « **[Venez]**ext1 vite **[je]**cat suis dans une maison, c'est la maison de **[Mme je suis méchant]**witch en gros **[Madame sauvag]**witche ». **[Le policier]**ext1 arrive et **[dit]**ext1 : « ah oui c'est **[la sorcière]**witch. **[Ce chat]**cat avait raison de **[m']**ext1appeler, **[elle]**witch est très dangereuse. » **[Il]**ext1 alla à la porte et **[toqua]**ext1 : toc toc toc. **[La sorcière]**witch ouvrit et **[le policier]**ext1 **[lui]**witch dit « **[vous]**witch êtes en état d'arrestation pour violence sur chat sans défense. » **[Il]**ext1 sortit les menottes et **[lui]**witch **[attacha]**ext1 aux poignets. **[Il]**ext1 **[la]**witch met dans la voiture et **[le chat]**cat peut se balader maintenant dans les rues de New York.

Fin de l'histoire.

CE2, 3034

Il était une fois **[une sorcière très méchante]**witch. Un jour **[elle]**witch transforma **[un chien]**ext1 en **[robot]**ext1 mais **[le robot très serviable]**ext1 pouvait faire plein de choses. Je vais vous les nommer : **[il]**ext1 joue avec les petits-enfants, le ménage, aider les personnes en difficulté...

CE2, 1143

Il était une fois **[une sorcière très très laide]**witch. **[Elle]**witch avait des verrues partout, et **[sa]**witch maison était remplie de chauve - souris, il y a aussi des fantômes ! Mais **[la sorcière]**witch **[elle]**witch était très méchante mais aussi farceuse ; quand **[[son]witch chat]**cat arrivait dans la cuisine **[elle]**witch **[lui]**cat jetait une tarte aux araignées !

CE2, 1986

Il était une fois **[un petit robot beau et intelligent]**robot. **[Il]**robot s'appelait **[Sam]**robot. Mais il y avait aussi **[une méchante sorcière]**witch, **[elle]**witch s'appelait **[Babagara]**witch. **[Elle]**witch était moche, **[elle]**witch avait un bouton sur le nez et il y avait aussi **[un loup]**wolf. **[Lui]**wolf **[il]**wolf ne pensait qu'à manger. C'est là que l'histoire commença. **[Le robot]**robot habite une maison en fer. **[La sorcière]**witch habite dans un chaudron géant. **[Le pauvre loup]**wolf n'avait pas d'habitat, **[il]**wolf avait le malheur. **[Le robot]**robot rencontra **[le loup]**wolf, **[le loup]**wolf voulut **[la]**robot manger alors **[le robot]**robot décida de faire la bagarre. Alors **[le robot]**robot VS **[le loup]**wolf. **[La sorcière]**witch **[qui]**witch passait par là **[les]**robot,wolf **[entendit]**witch... À SUIVRE.

CE2, S1991

[Une vieille sorcière des bois]witch et **[[son]witch chat]**cat s'appelaient **[Marie-Memèr]**witch et **[[son]witch chat Catoumane]**cat. **[Il]**cat était très méchant avec les gens qui **[les]**cat,witch entouraient. Mais un jour **[la sorcière]**witch **[qui]**witch faisait le mal **[une autre sorcière]**witch1 arriva en ville et **[essaya]**witch1 de **[l']**witchattraper mais **[Marie-Memèr]**witch ne se laissa pas faire et **[elles]**witch,witch1 se lancèrent un combat de magie. Le lendemain le combat commença mais **[une très très veille sorcière]**witch2 arriva, **[elle]**witch2 avait 364 ans, c'était **[la sorcière la plus âgée de toutes les sorcières du monde]**witch2. Donc **[elle]**witch2 menotta **[les deux sorcières]**witch,witch1 et **[les]**witch,witch1 **[jeta]**witch2 en prison !

CE2, 2003

« **[Je]**robot dois faire rire les gens » répétait sans arrêt **[le robot de [Pr Chat]cat]**robot « **[j']**cat ai mis **[[mon]cat robot]**robot sur mode clown » alors **[il]**cat essayait de se rectifier mais **[le robot]**robot commença à dire « " 5 - 4 - 3 - 2 - 1 Boum ! » Alors **[il]**cat en reconstruit **[un autre]**robot1. **[Celui - ci]**robot1 dit « **[Moi]**robot1 devoir détruire **[[mon]robot1 créateur]**cat » Alors **[il]**cat essaya de sortir de **[son]**cat laboratoire mais la porte était bloquée. Alors **[il]**cat se dit « **[j']**cat ai travaillé pendant trois mois sur **[ce robot]**robot1 et **[je]**cat dois **[le]**robot1 détruire » alors **[il]**cat prit la caisse de grenades et la **[lui]**robot1 **[lança]**cat dessus. Alors **[il]**cat essaya de faire un avion mais l'avion se crasha et **[lui]**cat aussi. On **[le]**cat retrouva 100 ans plus tard mort noyé.

CE2, 1098

Il était une fois **[un loup très gourmand]**wolf. **[Il]**wolf mangeait du porc, du cochon, et les dindes et un jour de la nuit d'Halloween **[une terrible sorcière]**witch **[qui]**witch volait toute la nuit. Et un soir **[le loup]**wolf voyait quelque chose. **[Il]**wolf n'avait pas mangé depuis 5 jours donc **[il]**wolf croyait que c'était **[une cuisse de poulet]**witch. Donc **[le loup]**wolf décida de **[la]**witch suivre. **[Il]**wolf voyait **[la sorcière]**witch rentrer dans la maison maudite. **[Le loup]**wolf réfléchissait et **[il]**wolf rentre dans la maison maudite. **[Il]**wolf essaye de trouver **[la cuisse de poulet]**witch et **[la cuisse de poulet]**witch c'était **[la sorcière]**witch. Mais **[le loup]**wolf ne savait pas que c'était **[la sorcière]**witch. **[Il]**wolf trouve un placard et **[ouvre]**wolf le placard et **[ouvre]**wolf le placard et **[réveille]**wolf **[la sorcière]**witch et **[le loup]**wolf courant et **[la sorcière]**witch poursuit **[le loup]**wolf et **[la sorcière]**witch transforme **[le loup]**wolf en **[cuisse de poulet]**wolf.

CE2, 1099

Il était une fois **[un loup]**wolf **[qui]**wolf mangeait toutes les personnes qui passaient dans la forêt. Et comme ça tous les jours. Mais **[il]**wolf ne savait pas que **[la sorcière]**witch rôdait dans les parages. **[La sorcière au pouvoir maléfique]**witch. **[Elle]**witch peut voir tout ce qui passe dans la forêt. Un jour **[elle]**witch décide que ça ne peut pas continuer. « Après il n'y aura plus d'habitant. **[Je]**witch vais **[le]**wolf voir et **[vais [lui]wolf donner]**witch une bonne raclée. » **[La sorcière]**witch va voir **[le loup]**wolf et **[le]**wolf

[transforme]witch en **[tout mignon petit chat]**wolf. Et comme **[la sorcière]**witch a **[un animal de compagnie]**wolf. Et si un jour comme celui -là **[il]**wolf aura le droit au même sort. « Et **[je]**witch pourrai donner ce sort avec **[[mon]witch chat]**cat. Ça sera très marrant de lancer des sortilèges aux vilains méchants loups. Et les autres chats iront dans le village le plus proche. Et dans la forêt il n'y aura plus de vilain méchant loup. » Et **[la sorcière]**witch fait bien **[son]**witch boulot. Et maintenant **[les villageois]**ext1 **[l']**witch appellent **[la sorcière aux chatons]**witch.

CE2, 1100

Il était une fois **[une sorcière]**witch **[qui]**witch préparait **[sa]**witch potion, ensuite **[elle]**witch dit des mots bizarres. Et un jour, **[un chat]**cat s'était perdu et **[trouva]**cat la maison de **[la sorcière]**witch, la porte était un peu ouverte, **[le chat]**cat entra. **[La sorcière]**witch **[qui]**witch préparait une autre potion **[l']**cata vu, et **[dit]**witch « **[Un chat]**cat chez **[moi]**witch ? pourquoi ? » dit**[-elle]**witch. **[Elle]**witch **[lui]**cat jeta un sort pour que **[le chat]**cat puisse parler, et dire ce qu'il **[lui]**cat est arrivé. Du coup **[le chat]**cat dit « Miaou » et ensuite **[il]**cat a raconté à **[la sorcière]**witch ce qu'il **[lui]**cat est arrivé. Quand **[il]**cat finit **[la sorcière]**witch pleurait et **[dit]**witch « **[Tu]**cat peux rester avec **[moi]**witch » **[le chat]**cat répondit « Vraiment ? » - « Oui » **[elle]**witch sécha **[ses]**witch larmes et **[ils]**cat,witch restèrent ensemble toute **[leur]**cat,witch vie.

CE2, 1101

{[Le chat]cat et **[monsieur Robeau]**robot**}**

[Mr Robot]robot va dans la forêt cueillir des champignons et **[rencontre]**robot **[un chat]**cat. **[Mr Robot]**robot dit « **[Tu]**cat n'es pas à la maison ? » **[Le chat]**cat dit « Non, **[j']**cat ai été abandonné dans la rue. » **[Mr Robot]**robot dit « **[Viens]**cat chez **[nous]**robot si **[tu]**cat veux ». « Non ! pourquoi ? » « **[je]**cat n'ai pas le droit de faire ça, c'est la règle n° 1 » « ah bon ! pourquoi donc ? ! » « **[je]**cat ne sais pas, **[on]**cat,robot va chercher si **[tu]**cat veux. » « D'accord ! pourquoi pas ? » 1 an plus tard **[il]**cat doit trouver pourquoi. **[Il]**ext1 **[nous]**cat jette dans la rue. **[Le Robot]**robot dit « **[moi]**robot **[je]**robot **[te]**cat donnais tout ce que **[tu]**cat veux » « tout ce que **[je]**cat veux ? » « bien sur que oui » c'est mieux d'être en une maison que dans une forêt. **[Ils]**cat,robot rentrent dans la maison **[du robot]**robot.

FIN

CE2, 1333

{[La sorcière]witch**}**

Il était une fois **[une sorcière]**witch **[qui]**witch vivait dans les bois. **[Elle]**witch a une maison noire. Un jour **[elle]**witch rencontra **[un chat]**cat. Et **[elle]**witch se mit à **[lui]**cat faire peur. Et **[le chat]**cat décide de se venger. **[Il]**cat entre dans la maison de **[la sorcière]**witch. **[Il]**cat entre dans le placard, et **[mange]**cat la bonne nourriture. Et **[la sorcière]**witch ouvre le placard. Et **[la sorcière]**witch **[lui]**cat jette un sort et **[le chat]**cat devient noir. Et **[le chat]**cat devient le compagnon de **[la sorcière]**witch. Et **[ils]**cat,witch furent

de bons amis. Et un jour **[un loup]**[wolf] toqua à la porte. Et **[le loup]**[wolf] essaya de manger **[le chat]**[cat] et **[le chat]**[cat] sauta par la fenêtre. Et **[le loup]**[wolf] fut attrapé et **[la sorcière]**[witch] **[lui]**[wolf] jeta un sort et **[le loup]**[wolf] partit et **[la sorcière]**[witch] eut perdu **[[son]**[**witch**] **chat]**[cat]. Et **[la sorcière]**[witch] était toute seule et **[[son]**[**witch**] **chat]**[cat] était revenu.

CE2, 1349

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] se baladait dans la forêt. **[Elle]**[witch] se dit « est -ce que **[j']**[witch] achète des yeux de crapauds ou des ailes de chauve - souris ? » Et tout d'un coup **[un petit loup]**[wolf] apparut. **[La sorcière]**[witch] se dit « peut-être que **[je]**[witch] pourrai **[l']**[wolf] adopter » se dit **[-elle]**[witch] « **[viens]**[wolf] **[je]**[witch] vais **[t']**[wolf] emmener chez **[moi]**[witch]. » En arrivant chez **[elle]**[witch] **[elle]**[witch] **[lui]**[wolf] donne une cuisse de poulet. **[Le petit loup]**[wolf] casse tout dans la maison de **[la sorcière]**[witch] « **[tu]**[wolf] arrêtes sinon **[je]**[witch] **[te]**[wolf] mets dehors ». **[Le petit]**[wolf] se calme tout de suite, **[la sorcière]**[witch] prépare une potion magique. **[Le petit loup]**[wolf] pousse **[la sorcière]**[witch] remonte et **[lui]**[wolf] **[dit]**[witch] « maintenant **[tu]**[wolf] arrêtes » voici la fin de l'histoire **[du loup]**[wolf] et **[la sorcière]**[witch].

CE2, 2883

Il était une fois **[une très méchante sorcière]**[witch] avec **[[son]**[**witch**] **chat noir]**[cat] **[qui]**[cat,witch] habitaient dans une cabane près du village. Un jour **[la sorcière]**[witch] rencontra **[un robot]**[robot] **[qui]**[robot] **[lui]**[witch] dit « il y a **[un loup]**[wolf] dans les parages **[qui]**[wolf] veut **[me]**[robot] manger » **[la sorcière]**[witch] dit « ce n'est pas grave tant qu'**[il]**[wolf] ne **[me]**[witch] mange pas » **[le robot]**[robot] repartit triste. **[Le loup]**[wolf] **[qui]**[wolf] était caché dans un arbre essaya de sauter sur **[le robot]**[robot] pour **[le]**[robot] manger mais **[le robot]**[robot] courut et **[le loup]**[wolf] s'écrasa par terre, **[le robot]**[robot] **[qui]**[robot] courait tous les jours s'arrêta pour regarder si **[le loup]**[wolf] **[le]**[robot] suivait. **[Il]**[robot] arriva près très près de la cabane de **[la méchante sorcière]**[witch].

CE2, 2886

Il était une fois **[un tout petit chat]**[cat] **[qui]**[cat] s'était perdu dans les bois. **[Il]**[cat] demandait à **[tous ceux]**[ext1] qu'**[il]**[cat] voyait : « est -ce que **[vous]**[ext1] savez où **[nous]**[cat,ext1] sommes ? » Jusqu'à aller dans une très vieille maison. Et dans cette maison il y avait **[un loup déguisé en grand-mère]**[wolf] **[qui]**[wolf] disait en voyant **[le chat]**[cat] : « la chance commence à tourner, plus besoin d'attendre **[le chaperon rouge]**[ext2] pour manger ? » en se léchant les moustaches ! **[le chat affolé]**[cat] courut aussi vite qu'**[il]**[cat] put. Mais **[il]**[cat] s'arrêta brusquement devant **[une sorcière]**[witch] **[qui]**[witch] dit **[au grand méchant loup]**[wolf] d'un air menteur que **[ce chat]**[cat] était à **[elle]**[witch]. Mais **[le loup]**[wolf] disait « quand **[vous]**[witch] **[m']**[wolf] aviez invité **[j']**[wolf] ai fouillé de partout pour voir s'il y avait **[le chaperon]**[ext2] mais il n'y avait que dalle, même pas de chat alors... » « **[Vous]**[wolf] avez fait quoi ! » gronda **[la sorcière]**[witch] : « Abracadabra en camembert **[tu]**[wolf] te transformeras ! » et depuis que **[le loup]**[wolf] disparu... **[La sorcière]**[witch] et **[le chat]**[cat] vécurent heureux.

CE2, 2893

[Un petit chat][cat] se promène dans la forêt et tout à coup **[le chat]**[cat] trouve une maison au loin, et **[le chat]**[cat] s'approche doucement vers la maison. Quand **[il]**[cat] est en face **[il]**[cat] a peur et après un bon moment **[il]**[cat] est rentré. Une fois dedans **[il]**[cat] rencontre **[une sorcière]**[witch] **[qui]**[witch] est très fâchée, quand **[il]**[cat] est rentré **[la sorcière]**[witch] a fermé la porte, et **[la sorcière]**[witch] jetait des sorts mais **[le chat]**[cat] esquivait tous les sorts et un long moment **[le chat]**[cat] était énervé alors **[le chat]**[cat] **[lui]**[witch] saute à la gorge, **[il]**[cat] **[la]**[witch] mord et un long moment **[le chat]**[cat] continue à faire bobo à **[la sorcière]**[witch] et encore et un long moment **[la sorcière]**[witch] **[elle]**[witch] est morte.

CE2, 2905

Il était une fois **[une sorcière]**[witch] et **[[son]**[**witch**] **chat]**[cat] **[qui]**[cat,witch] préparaient une potion magique qui avait la faculté d'empoisonner **[des gens d'Angleterre]**[ext1] et **[la reine d'Angleterre]**[ext2] pour avoir la richesse, la couronne, et la loi. Un jour tellement **[elle]**[ext2] était vieille qu'**[ils]**[cat,witch] n'ont même pas eu besoin de la potion magique donc **[la sorcière]**[witch] est maintenant **[reine d'Angleterre]**[witch]. À **[elle]**[witch] les richesses. À **[elle]**[witch] la couronne et la loi. Même **[le chat]**[cat] a une petite couronne.

CE2, 2906

Il était une fois **[un chat]**[cat] **[qui]**[cat] vivait chez **[un roi]**[ext1] car c'était le cadeau de **[[sa]**[**ext1**] **fille]**[ext2]. Mais un jour **[le chat]**[cat] partit, **[il]**[cat] voulait voir la mer, **[le roi]**[ext1] **[lui]**[cat] dit : « C'est stupide et **[tu]**[cat] ne sais pas tous les dangers qu'il y a là - bas ». **[Le chat]**[cat], **[lui]**[cat] **[il]**[cat] s'en fichait et **[il]**[cat] partit. Avant de partir **[le chat]**[cat] dit à **[la princesse]**[ext2] : « **[Je]**[cat] m'en vais, **[j']**[cat] en ai marre de la vie où on se la coule douce. **[Je]**[cat] vais à la mer ». **[La princesse]**[ext2] se mit à pleurer, **[le chat]**[cat] était déjà parti. Une fois à la mer **[il]**[cat] voulait rentrer au château mais **[il]**[cat] ne savait pas où c'était. Après une longue marche **[il]**[cat] retrouva le château et **[la princesse]**[ext2] était très contente. Fin.

CE2, 2907

Il était une fois **[un petit chaton]**[cat] **[qui]**[cat] vivait dans la rue. Un jour **[une fille]**[ext1] passait par ici. **[Elle]**[ext1] voit une queue, **[elle]**[ext1] se rapproche et **[elle]**[ext1] a vu **[une grosse boule de poils]**[cat]. **[Elle]**[ext1] **[la]**[cat] prend et **[la]**[cat] **[ramène]**[ext1] chez **[elle]**[ext1]. **[Elle]**[ext1] prend bien soin **[du chaton]**[cat]. Quelques ans plus tard **[elle]**[ext1] est adulte et **[le chaton]**[cat] est maintenant **[un chat]**[cat]. **[Elle]**[ext1] est maintenant **[une sorcière]**[ext1] mais **[une gentille sorcière]**[ext1] et **[le chat]**[cat] de plus en plus gentil. Mais maintenant voilà **[une petite famille]**[cat,ext1]. **[Le chat]**[cat] et **[la sorcière]**[ext1]. Fin.

CE2, 1959

[Ce chat][cat] veut des caresses, **[il]**[cat] est mignon. On voit qu'**[il]**[cat] veut plein de libertés. **[Cette sorcière]**[witch] a l'air gentille plus ou moins. **[Elle]**[witch] peut voler sur **[son]**[witch] balai magique et **[elle]**[witch] peut faire des potions : sort de gel, potion de lenteur, de trou noir, de téléportation ou de double. **[Elle]**[witch] peut faire bonne cuisinière dans les restaurants.

CE2, 1962

Il était une fois **[un chat]**[cat] **[qui]**[cat] avait un bracelet qui faisait de drôles de formes comme des carrés, de petits ronds et des fleurs et **[il]**[cat] rencontre **[une femelle rat]**[ext1], **[elle]**[ext1] était très belle et **[ils]**[cat,ext1] se mariaient et **[font]**[cat,ext1] plein d'**[enfants]**[ext2]. Il était une fois **[un loup en colère]**[wolf], **[il]**[wolf] était très très très énervé parce qu'**[il]**[wolf] n'a pas mangé depuis 5 jours. **[Il]**[wolf] n'a pas réussi à en attraper un seul et **[le loup]**[wolf] a vu **[un enfant]**[ext3] passer par la montagne et **[il]**[wolf] **[le]**[ext3] mange et **[le loup]**[wolf] a dit « **[j']**[wolf] avais une faim de loup » fin.

CE2, 48

Il était une fois **[une méchante sorcière]**[witch] **[qui]**[witch] avait **[un chat]**[cat] et **[un loup]**[wolf] **[qui]**[wolf] n'aimait pas **[cette sorcière]**[witch] et **[[son]**[witch] **chat]**[cat]. Et **[le loup]**[wolf] essaie de se débarrasser d'**[eux]**[cat,witch]. Mais **[il]**[wolf] n'y arrive pas parce que **[la méchante sorcière]**[witch] avait une baguette magique. **[Elle]**[witch] **[le]**[wolf] transforme en **[hippopotame]**[wolf] et **[le chat]**[cat] rigole avec **[la sorcière]**[witch] après **[ils]**[cat,witch] sont endormis et **[le loup]**[wolf] marchait doucement dans **[leur]**[cat,witch] chambre et **[il]**[wolf] a réussi à prendre la baguette et **[le loup]**[wolf] voulait transformer **[les deux]**[cat,witch] en grenouille mais cette baguette ne marche que sur les sorciers et **[la sorcière]**[witch] se réveille et **[elle]**[witch] avait un pouvoir magique pour que **[ce loup]**[wolf] ne soit plus là et la nuit tombait et **[les copains [du loup]**[wolf]**]**[wolf1] frappaient sur la porte de **[la méchante sorcière]**[witch]. Et **[les loups]**[wolf1] cherchaient **[[leur]**[wolf1] **copain]**[wolf] et **[la sorcière]**[witch] **[leur]**[wolf1] dit qu'il n'y a pas de loup ici et **[les loups]**[wolf1] sont partis.

CE2, 52

Il était une fois **[un chat]**[cat] **[qui]**[cat] s'appelait **[Chatchat]**[cat]. **[Il]**[cat] habitait dans une toute petite maison. Un jour **[le chat]**[cat] a décidé d'aller se promener dans la forêt. **[Il]**[cat] a vu **[un loup]**[wolf] et **[il]**[cat] est parti en courant. **[Le loup]**[wolf] **[le]**[cat] poursuivit et **[l']**[cat] **[attrapa]**[wolf]. **[Il]**[wolf] **[l']**[cat] emmenait chez **[lui]**[wolf]. **[Le loup]**[wolf] était très fatigué après avoir couru et **[il]**[wolf] s'endort. **[Le chat]**[cat] était dans tous ses états. **[Le loup]**[wolf] avait oublié de fermer la porte. **[Le chat]**[cat] a vu la porte ouverte et **[il]**[cat] est parti en courant. **[Il]**[cat] est arrivé chez **[lui]**[cat] et **[a fait]**[cat] une sieste. **[Le loup]**[wolf] se leva et **[ne voit plus]**[wolf] **[le chat]**[cat] **[qui]**[cat] est parti. **[Le loup]**[wolf] dit « C'est pas grave, **[je]**[wolf] peux en trouver d'autres chats ».

CE2, 83

Il était une fois **[une petite fille]**[ext1] **[qui]**[ext1] s'appelait **[le petit chaperon]**[ext1]. Tous les matins **[elle]**[ext1] allait se promener dans les bois. Un jour, **[[sa]**[ext1] **mère]**[ext2] **[lui]**[ext1] dit « **[[Nos]**[ext1,ext2] **voisins]**[ext3] **[m']**[ext2] ont dit qu'en suivant le chemin pour aller chez **[[ta]**[ext1] **grand-mère]**[ext4] **[ils]**[ext3] avaient trouvé des tonnes de buissons plein de fraises. Si **[tu]**[ext1] n'as rien à faire, **[vas]**[ext1] - y avec **[ton]**[ext1] panier. » **[La petite fille]**[ext1] **[qui]**[ext1] n'avait rien à faire y alla. Sur le chemin **[elle]**[ext1] rencontra **[un loup]**[wolf]. **[Le loup]**[wolf] **[lui]**[ext1] dit « où vas - **[tu]**[ext1] comme ça ? » « **[Je]**[ext1] suis le chemin de chez **[[ma]**[ext1] **grand-mère]**[ext4] » « **[Va]**[ext1] par ce chemin il sera plus court. » **[Le loup]**[wolf] courut sur le chemin qui rejoignait le chemin que **[le chaperon]**[ext1] avait pris. Mais **[le chaperon]**[ext1] trouva une carte dans **[son]**[ext1] panier. **[Elle]**[ext1] se rendait compte que **[le loup]**[wolf] **[l']**[ext1] avait piégée. **[Elle]**[ext1] retourna chez **[elle]**[ext1] voir **[[sa]**[ext1] **mère]**[ext2]. **[Le loup]**[wolf] était déçu. En plus il n'y avait plus de fraise, **[le chaperon]**[ext1] avait eu de la chance.

FIN

CE2, 1152

J'avais **[un chat]**[cat] **[qui]**[cat] s'appelle **[Ravza]**[cat]. **[Je]**[ext1] **[le]**[cat] fais jouer avec, c'est **[[mon]**[ext1] **ami]**[cat]. Un jour **[je]**[ext1] suis sorti avec **[lui]**[cat] et **[on]**[cat,ext1] était allé au parc et **[elle]**[cat] faisait de la balançoire et puis c'est là où **[j']**[ext1] ai appris que **[[mon]**[ext1] **chat Ravza]**[cat] savait parler la langue humaine. **[Je]**[ext1] **[lui]**[cat] ai dit en chuchotant « **[tu]**[cat] sais parler la langue humaine ! » **[Elle]**[cat] **[me]**[ext1] répondit « Oui **[je]**[cat] sais parler comme vous mais **[tu]**[ext1] ne le dis à personne » « d'accord, promis. Promis, juré, craché ».

CE2, 1155

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] était très méchante. **[Elle]**[witch] détruisait tout. Un jour **[elle]**[witch] vit **[un enfant]**[ext1] **[qui]**[ext1] était très gentil. **[Il]**[ext1] s'appelait **[Enzo]**[ext1]. **[La sorcière]**[witch] n'aimait pas les enfants gentils alors **[elle]**[witch] voulut **[le]**[ext1] détruire. Mais **[l'enfant]**[ext1] était tellement gentil qu'**[elle]**[witch] n'y arrivait pas. **[L'enfant]**[ext1] **[lui]**[witch] dit « pourquoi veux - **[tu]**[witch] **[me]**[ext1] détruire ? » **[elle]**[witch] répondit « **[je]**[witch] ne sais pas » **[l'enfant]**[ext1] **[lui]**[witch] dit « pourquoi es - **[tu]**[witch] méchante ? » **[elle]**[witch] répondit « car **[je]**[witch] suis comme ça » **[Enzo]**[ext1] **[lui]**[witch] dit « as - **[tu]**[witch] un ami ? » **[La sorcière]**[witch] dit « non ». **[L'enfant]**[ext1] **[lui]**[witch] dit « **[moi]**[ext1] aussi » puis **[il]**[ext1] dit « veux - **[tu]**[witch] être **[mon]**[ext1] amie ? » **[La sorcière]**[witch] dit « oui ». A partir de ce jour **[la sorcière]**[witch] est devenue gentille et **[a arrêté]**[witch] de faire du mal.

CE2, 1160

Il était une fois **[un petit chaperon rouge]**[ext1] **[qui]**[ext1] avait **[une grand-mère]**[ext2]. **[[Sa]**[ext1] **grand-mère]**[ext2] était très malade. **[La maman [du petit chaperon rouge]**[ext1]**]**[ext3] **[lui]**[ext1] a dit : « **[Va]**[ext1] chez **[[ta]**[ext1] **grand-mère]**[ext2] **[qui]**[ext2] est malade et **[va]**[ext1] **[lui]**[ext2] apporter cette tarte et ce petit pot de beurre et **[fais attention]**[ext1] à **[toi]**[ext1] car il y a **[le loup]**[wolf] dans la forêt. » Et **[la petite]**[ext1]

y alla. **[Le loup]**[wolf] arriva avant **[elle]**[ext1] chez **[la grand-mère]**[ext2] et **[la]**[ext2] **[dévora]**[wolf] et **[s'est déguisé]**[wolf] en **[vieille dame]**[wolf]. **[La jeune fille]**[ext1] entra chez **[[sa]**[ext1] **grand-mère]**[ext2] et **[le loup]**[wolf] **[l']**[ext1] avala.

CE2, 1164

Il était une fois **[une sorcière]**[witch] **[qui]**[witch] vivait dans un château affreux. **[La sorcière]**[witch] avait un nez laid et **[[son]**[witch] **chat]**[cat] était tout gris. **[Elle]**[witch] va dans la forêt et **[rencontre]**[witch] **[une jeune fille]**[ext1] **[qui]**[ext1] cueille des champignons. **[La sorcière]**[witch] voulait manger **[la fille]**[ext1]. **[Elle]**[witch] s'est cachée derrière un arbre et quand **[la fille]**[ext1] se retourne **[la sorcière]**[witch] **[lui]**[ext1] saute dessus. **[Elle]**[witch] retourne dans **[son]**[witch] château et **[mange]**[witch] avec **[le chat]**[cat].

CE2, 688

{[Le chat][cat] et **[le loup]**[wolf]}

Il était une fois **[un chat]**[cat] **[qui]**[cat] entend **[le loup]**[wolf] crier et **[le]**[wolf] **[voit]**[cat] devant **[lui]**[cat] : « bonjour **[petit chat]**[cat], comment ça va ? » « HE HE bien. Pourquoi ? » « Parce que **[je]**[wolf] vais **[te]**[cat] manger AAA ! » « NON ! Mais est -ce que **[tu]**[wolf] aimes les griffes ? » « C'est quoi ? » « C'est ça » « HAAAAAAA ça fait mal » « Alors **[pars]**[wolf] si **[tu]**[wolf] ne veux pas que **[je]**[cat] recommence » « OK ! Au revoir »

Et **[le petit loup]**[wolf] grogna.

Fin !

CE2, 1866

{[Le chat][cat] et **[la sorcière]**[witch]}

Il était une fois **[un chat]**[cat] **[qui]**[cat] était dans la forêt, **[il]**[cat] cherchait à manger : des souris, des crapauds. **[Il]**[cat] trouva une balle devant une maison, **[il]**[cat] voulait voir ce qu'il y avait dans cette maison vieille. **[Le petit chat]**[cat] essayait de pousser la porte. Mais **[il]**[cat] était trop petit. **[Il]**[cat] essaya de trouver quelque chose pour aller dans la maison. **[Il]**[cat] trouva une balle de tennis et une corde épaisse et longue. **[Il]**[cat] prit la corde et la **[mit]**[cat] dans **[sa]**[cat] bouche, puis **[il]**[cat] l'accrocha à la poignée de la porte et **[tira]**[cat]. La porte s'ouvrit et **[il]**[cat] voit **[une sorcière]**[witch] avec un balai magique, un chapeau de sorcière, une robe noire. **[Il]**[cat] entra dans la maison et **[la sorcière]**[witch] **[le]**[cat] voit, **[elle]**[witch] **[lui]**[cat] a dit avec une voix pointue « pourquoi **[tu]**[cat] es venu dans **[ma]**[witch] maison **[petit rat]**[cat] ? » « Mais non **[je]**[cat] ne suis pas un rat, **[je]**[cat] suis **[un chat]**[cat]. » « **[je]**[witch] vais **[te]**[cat] transformer en crapaud. » **[Elle]**[witch] prononçait **[sa]**[witch] formule et **[le]**[cat] **[transforma]**[witch] en crapaud et **[les deux]**[cat,witch] sont amis.

CE2, 1926

Il était une fois un jour **[une sorcière]**witch **[qui]**witch trouva **[un chat]**cat devant **[sa]**witch porte et **[la sorcière]**witch a voulu garder **[le chat]**cat. Et **[la sorcière]**witch dit « maintenant **[tu]**cat vas t'appeler **[Bonbon]**cat » parce que **[la sorcière]**witch adore les bonbons, dans **[son]**witch jardin il y avait des bonbons partout. 2 semaines après **[une petite fille]**ext1 et **[[son]ext1 papa]**ext2 toquent « toc toc » **[le papa]**ext2 disait « avez **[-vous]**witch vu **[le chat]**cat ? » **[la sorcière]**witch dit « non » mais **[elle]**witch mentait. Puis **[le chat]**cat venait à la porte. **[La petite fille]**ext1 criait « **[Nougat]**cat » parce que **[le chat]**cat s'appelait comme ça. **[La sorcière]**witch disait « non **[vous]**ext1,ext2 vous trompez de **[chat]**cat », et **[la sorcière]**witch claquait la porte.

3 semaines après **[la petite fille]**ext1, **[[son]ext1 papa]**ext2 et **[les policiers]**ext3 toquent et **[les policiers]**ext3 disaient « **[ouvre]**witch ». **[La sorcière]**witch ouvrit. **[Le chat]**cat venait encore une fois. **[La sorcière]**witch disait « il y a un problème ». **[Le policier]**ext3 disait « **[vous]**witch êtes en état d'arrestation pour vol ». **[Le policier]**ext3 met **[la sorcière]**witch dans la voiture. **[La petite fille]**ext1 cria « **[Nougat]**cat » depuis ce jour -là, **[la petite fille]**ext1 rentre tout le temps avec **[Nougat]**cat et **[elle]**ext1 ne laissait plus **[Nougat]**cat seul.

CE2, 816

{[Le robot malin]robot}

Il était une fois **[un robot]**robot **[qui]**robot vivait dans **[sa]**robot maisonnette, au pays des robots. Un jour, **[il]**robot décida d'aller se promener dans la forêt. **[Il]**robot y croisa **[toutes sortes d'animaux]**ext1,ext2,ext3,ext4 (**[cerf]**ext1, **[biche]**ext2, **[coccinelle]**ext3, **[renard]**ext4). Au bout de quelques minutes de marche, **[il]**robot vit **[un loup féroce]**wolf **[qui]**wolf était prêt à **[le]**robot manger. Courageusement, **[il]**robot dit « - Ne **[me]**robot **[mangez]**wolf pas ! **[je]**robot ne suis qu'**[un tas de ferraille]**robot ! » dit **[-il]**robot « **[Vous]**wolf ne pouvez pas **[me]**robot manger ! » **[Le loup]**wolf dit « Ce n'est pas **[toi]**robot que **[je]**wolf veux manger ! Ce sont **[la biche]**ext2, **[le cerf]**ext1 et **[le renard]**ext4 que **[je]**wolf veux manger ! **[Dis]**robot **[-moi]**wolf où ils sont » **[Le robot]**robot entendit un grand BOUM ! « **[Les chasseurs]**ext5 ! » s'écria -t **[-il]**robot. Et **[il]**robot courut en direction **[des chasseurs]**ext5. **[Le loup]**wolf **[qui]**wolf attendait **[sa]**robot réponse finit par **[le]**robot pourchasser. Puis ils arrivèrent près des chasseurs. Les chasseurs virent **[le loup]**wolf et tirèrent. **[Le loup]**wolf fut touché et les chasseurs remercièrent **[le robot]**robot pour **[son]**robot courage.

Fin.

CE2, 834

Il était une fois **[une chatte abandonnée]**cat, **[elle]**cat était très belle pourtant, mais un jour **[ce robot nommé Hubert]**robot trouva **[cette chatte magnifique nommée Moumoune]**cat, puis après avoir réfléchi **[Hubert]**robot dit tout fort « oui, **[je]**robot **[te]**cat veux » peu après **[il]**robot **[la]**cat prit dans **[ses]**robot bras et **[l']**cat **[amena]**robot dans **[sa]**robot maison. **[Ils]**cat,robot s'aiment tellement fort qu'**[ils]**cat,robot ne se séparaient plus. Fin

CE2, 936

« Bonjour **[le chat]**^{cat}, **[je]**^{ext1} vais **[te]**^{cat} déchiqueter. En un morceau, **[je]**^{ext1} ne vais pas **[te]**^{cat} déchiqueter, c'est une blague » dit **[le chat]**^{cat} « au revoir » « attends **[mon vieil ami]**^{cat}, **[je]**^{ext1} ne suis pas vieux, **[tu]**^{cat} n'as pas compris le mot. »

CE2, 940

{[La chatte magique]^{cat}}

Un beau jour d'un bon matin, **[une chatte]**^{cat} se promenait dans les bois, **[un loup]**^{wolf} **[la]**^{cat} dévora.

Mais comme **[cette chatte]**^{cat} était magique **[elle]**^{cat} en sortit et **[vit]**^{wolf} qu'**[elle]**^{cat} était magique, **[elle]**^{cat} parlait !

Alors, **[elle]**^{cat} alla voir **[[son]**$^{\mathbf{cat}}$ **ami le chasseur]**^{ext1}, puis **[il]**^{ext1} tua **[le loup]**^{wolf} et depuis ce jour-là, **[ils]**^{cat,ext1} restèrent toujours ensemble. Fin

Annexe M. Corpus italien annoté

CE1, 11

{[La strega felice]witch}

Un giorno che non era ancora giorno **[una vecchia strega di nome Riunione]**witch si svegliò alle 7:00 per vestirsi, fare colazione e lavare i denti.

Ma **[voleva]**witch una cosa.

Cosa?

[Un amichetto]cat per passare il tempo!

Ma quale?

[Riunione]witch così prese la **[sua]**witch scopa e **[andò]**witch al supermercato delle streghe, **[vide]**witch solo gatti: gatti rossi, gatti a macchie, ma **[le]**witch interessò proprio **[un gatto]**cat, **[non era]**cat rosso o a macchie ma **[era]**cat **[un gatto nero]**cat. Ma solo un punto era diverso dietro l'orecchio era bianco ed era quello che **[lo]**cat rendeva speciale.

Passarono anni e anni, **[la strega]**witch era vecchia anzi vecchissima invece **[il gatto]**cat non invecchiava.

[La strega]witch se ne andò, **[il gatto]**cat restò solo ma **[era]**cat nel cuore, pur essendo triste **[era]**cat felice.

CE1, 1376

C'era una volta **[un robot]**robot **[che]**robot si chiamava **[Arbo]**robot. Poi c'era **[una strega]**witch **[che]**witch si chiamava **[Eva]**witch.

Un giorno **[Arbo]**robot nel **[suo]**robot laboratorio segreto fece un'invenzione che scambia i corpi o li moltiplica.

Allora **[Arbo]**robot provó a moltiplicare il **[suo]**robot corpo allora **[chiese]**robot a **[Eva]**witch di schiacciare il tasto di accensione e girare la manopola a due copie e **[Eva]**witch lo fece e la macchina fece quel che richiesto e l'esperimento è riuscito.

CE1, 1380

C’era una volta **[Valentina una gattina]**cat. E un giorno **[una strega di nome Emma]**witch **[gli]**cat voleva fare un incantesimo per far**[la]**cat diventare cattiva. **[La gattina poverina]**cat correva sempre perché **[aveva paura]**cat **[della strega Emma]**witch.

Dopo 1 anno **[la strega]**witch è riuscita a catturare **[la gattina]**cat proprio il 31 dicembre. Mentre i fuochi d'artificio suonavano **[la strega]**witch correva di qua e di là.

CE1, 1470

C'era una volta **[un lupo]**^{wolf} e **[una strega]**^{witch}, **[quella strega]**^{witch} voleva catturare **[un lupo cattivo]**^{wolf} ma **[il lupo]**^{wolf} non aveva paura di nessuno.

Un giorno **[il lupo]**^{wolf} è andato a caccia di cibo. **[La strega]**^{witch} ha preparato una trappola per far amicizia.

E **[gli]**^{wolf} **[fa]**^{witch} mangiare una zuppa magica che **[lo]**^{wolf} faceva trasformare in **[una strega]**^{wolf} ma **[il lupo]**^{wolf} ha pensato che **[devo]**^{wolf} far**[lo]**^{wolf} chiamare. Dopo tre minuti **[il lupo]**^{wolf} vede che **[la signora]**^{witch} non è una signora ma è **[una strega]**^{witch} e dopo **[il lupo]**^{wolf} chiamò **[i [suoi]^{wolf} amici]**^{ext1} e dopo **[i [suoi]^{wolf} amici]**^{ext1} **[lo]**^{witch} picchiava ancora. **[La strega]**^{witch} promise che **[non farà]**^{witch} nessuna trappola **[a tutti]**^{ext1,wolf}.

CE1, 1471

C'era una volta **[un lupo di color rosso]**^{wolf} **[che]**^{wolf} giocava sempre con **[gli altri [suoi]^{wolf} amici]**^{ext1} e **[giocavano]**^{ext1,wolf} a nascondino ma a **[lui]**^{wolf} non **[gli]**^{wolf} piaceva giocare a nascondino quindi **[giocarono]**^{ext1,wolf} a ce l' hai, **[tutti]**^{ext1,wolf} si rincorrono insieme.

CE1, 1482

{[Un gatto]^{cat} e **[un lupo]**^{wolf} }

Una notte erano a casa soli soletti e ad un tratto punto **[sentirono]**^{cat,wolf} **[qualcuno]**^{ext1} bussare alla porta e dopo **[il lupo]**^{wolf} andò ad aprire e dopo **[hanno visto]**^{cat,wolf} **[un altro lupo]**^{ext1}.

CE1, 1493

C'era una volta **[una strega]**^{witch} e **[un lupo]**^{wolf} **[che]**^{witch,wolf} si chiamavano **[Scuro]**^{wolf} e **[Maria]**^{witch}.

E un giorno **[il lupo]**^{wolf} e **[la strega]**^{witch} stavano andando al castello **[del re]**^{ext1} **[che]**^{ext1} si chiamava **[Enea]**^{ext1}.

[Entrarono]^{witch,wolf} nel castello **[del re Enea]**^{ext1} e **[il re Enea]**^{ext1} disse: "cosa **[ci fate]**^{witch,wolf} nel **[mio]**^{ext1} castello?" "**[Io]**^{wolf} e **[la strega]**^{witch} siamo venuti nel castello perché **[abbiamo saputo]**^{witch,wolf} che **[avete]**^{ext1} **[qualcosa di speciale]**^{ext2} **[che]**^{ext2} sarebbe **[la principessa d'oro]**^{ext2}". "Sì, ok **[avete indovinato]**^{witch,wolf} però **[non [la]^{ext2} darò]**^{ext1} a **[voi due]**^{witch,wolf}".

"Allora **[[ce ne andiamo]^{wolf}]**^{witch} via dal castello" però **[il lupo]**^{wolf} era furbo quindi **[loro due]**^{witch,wolf} andarono nella stanza dei gioielli. E poi **[il lupo]**^{wolf} e **[la strega]**^{witch} prendevano **[la principessa d' oro]**^{ext2} e **[scappavano]**^{witch,wolf} via da dentro il castello **[del re Enea]**^{ext1}.

E poi **[la strega]**witch e **[il lupo]**wolf tornarono nella **[loro]**witch,wolf tana e **[videro]**witch,wolf dentro al sacco e dentro non c'era **[la principessa d' oro]**ext2 ma dentro non c'era quello che **[avevano rubato]**witch,wolf.

Ma dentro c'era **[il re Enea]**ext1 **[che]**ext1 chiamò **[le [sue]ext1 guardie]**ext3.

Però **[il lupo]**wolf e **[la strega]**witch avevano un piano, **[loro]**witch,wolf si facevano arrestare poi quando erano in prigione **[il lupo]**wolf e **[la strega]**witch, **[il lupo]**wolf invece aveva preso dalla **[loro]**witch,wolf tana una motosega che ruppe le sbarre della prigione e poi **[loro]**witch,wolf tornarono nel **[loro]**witch,wolf rifugio e poi al castello **[del re Enea]**ext1**[che]**ext1 era arrabbiato perché **[le loro guardie]**ext3 stavano dormendo per terra.

E *poi* **[il lupo]**wolf e **[la strega]**witch erano felici e contenti e anche **[il re Enea]**ext1 e **[le loro guardie]**ext3.

CE1, 1495

C'era una volta **[una strega]**witch **[che]**witch si chiama **[Marina]**witch, **[era]**witch sempre da sola e **[andò]**witch a prendere **[un gatto]**cat che **[lo]**cat chiamò **[Sregino]**cat e **[lo]**cat **[portò]**witch a casa **[sua]**witch **[della strega Marina]**witch e **[gli]**cat **[dà]**witch da mangiare il pesce e un po' di acqua e poi **[andavano]**cat,witch a fare una passeggiata e **[videro]**cat,witch **[la [loro]witch,cat amica Nella]**ext1 con **[il [suo]ext1 amico gattino Gimi]**cat1 e **[i due gattini Gimi e Sregino]**cat,cat1 giocano a palla fino al tramonto e poi **[ritornano]**cat,cat1,ext1,witch tutti a casa.

CE1, 15

C'era una volta **[una strega]**witch **[che]**witch viveva in una casetta in montagna e **[viveva]**witch da sola. Un giorno mentre **[volava]**witch sulla **[sua]**witch scopa **[vide]**witch **[un gattino]**cat. **[Lo]**cat **[guardò]**witch e **[vide]**witch che **[il micetto]**cat non aveva il collare quindi **[decise]**witch di portar**[lo]**cat con **[sé]**witch. Mentre passavano gli anni **[il gattino]**cat crebbe e **[la strega]**witch notò che **[era]**cat una femmina quindi **[la]**cat **[chiamò]**witch **[Bianchina]**cat perché **[era]**cat tutta bianca. **[Il gattino]**cat un giorno decise di uscire di casa da solo. **[La strega]**witch uscì a cercar**[lo]**cat e **[lo]**cat **[trovò]**witch. Quando **[tornarono]**cat,witch a casa **[la strega]**witch diede del cibo **[al micetto]**cat e **[lo]**cat **[fece]**witch riposare. **[Il gattino]**cat si promise che **[non sarebbe mai più uscito]**cat da solo e nella notte **[sognò]**cat che **[la strega]**witch **[gli]**cat comprò tanti biscottini e **[si svegliò]**cat ed **[era incinta]**cat e il giorno dopo **[fece]**cat **[dieci cuccioli]**cat2 e **[la strega]**witch disse "che **[bei cuccioli]**cat2". **[Gli]**cat2 **[preparò]**witch le cucce e **[gli]**cat2 **[mise]**witch **[i cuccioli]**cat2 a dormire e **[andò]**witch a dormire anche **[lei]**witch" buona notte a tutti gattini e bambini. **[La strega]**witch e **[il gattino]**cat vi vogliono bene "

CE1, 155

C'era una volta **[una gattina]**cat **[che]**cat passava per strada.

A mezzogiorno passò **[un bambino]**ext1 e **[disse]**ext1 **[alla mamma]**ext2 "**[la]**cat prendiamo?" "Certo", rispose **[la mamma]**ext2. Quando **[arrivarono]**cat,ext1,ext2 a casa **[le]**cat **[diede]**ext1 del cibo, carote,

croccantini e latte. **[La gattina]**[cat] era così contenta che **[non aveva mai visto]**[cat] **[un bambino]**[ext1] **[che]**[ext1] **[la]**[cat] curava, **[la]**[cat] **[coccolava]**[ext1] e perfino **[la]**[cat] **[portava]**[ext1] a passeggio e **[la]**[cat] **[chiamò]**[ext1] **[Luna]**[cat].

Il giorno dopo **[non [la]**[cat] **trovò più]**[ext1] e **[chiamò]**[ext1] e **[chiamò]**[ext1] e **[chiamò]**[ext1] e alla fine **[la]**[cat] **[ritrovò]**[ext1], **[la]**[cat] **[riportò]**[ext1] a casa e **[la]**[cat] **[coprì]**[ext1] con una copertina rosa e gialla. **[Il bambino]**[ext1] non aveva mai visto **[una gattina]**[cat] così.

CE1, 162

Nel bosco lontano dalla città viveva **[un lupo]**[wolf]. **[Amava]**[wolf] correre, saltare, curiosare...

Un giorno **[vide]**[wolf] uno stagno e **[provò]**[wolf] a entrarci dentro. **[Si buttò]**[wolf], **[stava]**[wolf] per annegare quando all' improvviso **[un robot di metallo]**[robot] **[gli]**[wolf] salvò la vita. **[Il lupo]**[wolf] **[gli]**[robot] disse "grazie, grazie mille!" **[il robot]**[robot] "di niente, ma la prossima volta **[non provarci]**[wolf] più" e da quel momento **[rimasero]**[robot,wolf] amici per sempre.

CE1, 182

C'era una volta **[un gatto arrabbiato]**[cat] perché nessuno giocava con **[lui]**[cat]. Un giorno arrivò **[un robot]**[robot] da un altro pianeta. **[Il gatto]**[cat] si spaventò e **[andò]**[cat] **[dal [suo]**[cat] **amico cane]**[ext1] ma **[il cane]**[ext1] era dentro casa a dormire. **[Il robot]**[robot] prese **[il gatto]**[cat] e **[disse]**[robot] **"[vuoi]**[cat] diventare **[mio]**[robot] amico" **[il gattino]**[cat] rispose **[al robot]**[robot] sì, **[sarò]**[cat] il **[tuo]**[robot] amico ma però **[non [mi]**[cat] **farai arrosto]**[robot]" "**[un robot]**[robot] che fa **[un gatto]**[cat] arrosto mai. **[Io]**[robot] curo **[i gatti]**[cat], **[do]**[robot] **[loro]**[cat] da mangiare, **[faccio]**[robot] **[loro]**[cat] una casa per vivere con **[me]**[robot]" "Quindi **[mi]**[robot] **[accetterai]**[cat]"" "sì **[ti]**[robot] **[accetterò]**[cat]" disse **[il gatto]**[cat] "se **[mi]**[robot] **[accetterai]**[cat] **[avrai]**[cat] una famiglia gentile" "**[mi piace]**[cat] avere una famiglia gentile" e vissero **[tutti e due]**[cat,robot] felici e contenti. Ciao, ciao ci vediamo alla prossima storia.

CE1, 185

C'era una volta **[una strega]**[witch] e **[aveva]**[witch] **[un gatto nero]**[cat]. **[La strega]**[witch] era allegra e buona, **[le piaceva]**[witch] mangiare i dolci.

Un giorno **[il gatto furbo]**[cat] si nascose sotto il letto.

Un giorno dopo **[la strega]**[witch] si accorse che **[il gatto]**[cat] era sparito così **[lo]**[cat] **[cercò]**[witch] tutta la notte ma **[non [lo]**[cat] **trovò]**[witch]. **[Gli]**[cat] **[voleva]**[witch] così bene che **[cercò]**[witch] tutto il giorno finché **[si addormentò]**[witch] in un lampo.

Intanto che **[la strega]**[witch] dormiva **[il gatto]**[cat] si spostò nell' armadio delle pozioni. **[La strega]**[witch] **[lo]**[cat] cercò ancora un giorno e **[si addormentò]**[witch] di nuovo. **[Il gatto]**[cat] si spostò e **[si mise]**[cat] sulle gambe **[della strega]**[witch].

Quando **[si alzò]**[cat] **[corse]**[witch] di gioia **[la strega]**[witch] e **[festeggiarono]**[cat,witch] tutta la notte.

E la **[trasformano]**[cat,witch] in una tradizione speciale. Infine **[la strega]**[witch] disse **[al gatto]**[cat] « **[ti]**[cat] **[terrò]**[witch] al sicuro ». Ma era un sogno però **[lei]**[witch] ha imparato che si amano anche gli animali.

In compagnia con **[gli amici [del gatto]**[cat]**]**[ext1] festeggiano. **[La strega]**[witch] e **[il gatto]**[cat] vissero felici e contenti e uniti.

CE1, 186

C'era una volta **[un gatto birichino]**[cat] **[che]**[cat] andava sempre in giro senza il permesso **[dei [suoi]**[cat] **genitori]**[ext1]. Un giorno **[cadde]**[cat] in una buca buia che C'era dentro **[una strega]**[witch] molto ma molto cattiva **[che]**[witch] **[lo]**[cat] rapì. **[Il gatto]**[cat] tutto impaurito cercò di scappare ma **[la strega]**[witch] glie**[lo]**[cat] impedì, **[lo]**[cat] **[portò]**[witch] dove c'erano tutti gli animali. Dopo un po' di giorni **[diventò]**[cat] il **[suo]**[witch] schiavo e **[fa]**[cat] tutto quello che **[gli]**[cat] **[dice]**[witch]. Un giorno arrivò **[un ladro]**[ext2] e **[lo]**[cat] **[portò]**[ext2] via in una casa abbandonata!!!

[La strega][witch] diventò brava e **[il gatto]**[cat] si liberò dalla magia e tutto tornò normale tranne una cosa, che **[il gatto]**[cat] era totalmente scomparso, **[non si trovava]**[cat] più da nessuna parte ma era colpa dell'incantesimo ma un giorno **[lo]**[cat] trovarono **[i suoi genitori]**[ext1] ed era tornato tutto alla normalità e vissero tutti felici e contenti.

CE1, 200

C'era una volta **[un gatto]**[cat] solo nella neve e molto infreddolito. **[Quel gatto]**[cat] era tutto nero. Un giorno arrivò **[un simpatico robottino, azzurro, giallo, rosso e blu]**[robot]. **[Il gatto]**[cat] **[gli]**[robot] chiese « chi sei **[tu]**[robot]? » « **[un robottino simpatico e carino]**[robot] » **[rispose]**[robot]. « E **[tu]**[cat] che ci fai qui tutto solo nella neve? » **[chiese]**[robot]. « **[Mi]**[cat] hanno abbandonato » **[rispose]**[cat]. « **[Ti]**[cat] **[capisco]**[robot], ***[la donna che mi ha costruito]***[ext1] **[mi]**[robot] ha rimpiazzato per **[un robot dell'ultimo modello]**[ext2]. » « **[Mi]**[cat] è venuta un ' idea! » **[Disse]**[cat]. « Perché **[non partiamo]**[cat,robot] per andare a Roma? **[Ho sentito]**[cat] che si fanno delle prove per il duo più bravo! » « Sì » e **[[si misero]**[robot,cat] in cammino fino ad arrivarci. **[Vinsero]**[cat,robot] la gara e **[diventarono]**[cat,robot] ricchi, e vissero per sempre felici e contenti.

CE1, 210

C'era una volta **[un gattino molto affamato]**[cat] e un giorno **[il gatto]**[cat] vede **[una strega cattiva]**[witch] allora **[il gatto]**[cat] vede **[la strega]**[witch] **[che]**[witch] aveva in mano un pesce e **[gli]**[witch] **[chiese]**[cat] "**[posso]**[cat] mangiare il **[suo]**[witch] pesce?" e **[lei]**[witch] **[gli]**[cat] rispose "no" allora **[il gatto]**[cat] se ne andò via ancora affamato. Una sera **[il gatto]**[cat] vede un castello che c'era su un pesce allora **[il gatto]**[cat] entrò e c'era lì **[la strega]**[witch] allora **[lo]**[cat] **[cacciò]**[witch] via. Un giorno arrivò ancora **[la strega]**[witch] e **[lo]**[cat] **[vede]**[witch] morto e **[gli]**[cat] **[dice]**[witch] "non sarà che **[è morto]**[cat] perché non **[gli]**[cat] **[ho dato]**[witch] il pesce?" e **[lei]**[witch] apre la bocca **[del gatto]**[cat], **[gli]**[cat] **[dà]**[witch] il pesce e **[si è svegliato]**[cat] allora **[erano diventati]**[cat,witch] amici.

E **[facevano]**[cat,witch] molte cose insieme.

CE1, 212

C'era una volta **[un robot]**[robot] **[che]**[robot] lavorava in una ditta di camion, **[lo]**[robot] pagavano tanto e questo lavoro **[gli]**[robot] piaceva però un giorno **[investì]**[robot] **[un gattino]**[cat] e **[se ne accorse]**[robot], **[pianse]**[robot] e **[pianse]**[robot] per tanto tempo. A un certo punto però **[vide]**[robot] che stava finendo l'olio e **[incominciò]**[robot] a lavorare da cassiere. Quel lavoro **[gli]**[robot] piaceva ancora di più e **[aveva]**[robot] la stessa paga ma al lavoro **[andava]**[robot] sempre in taxi allora un giorno **[comprò]**[robot] una Fiat Panda 4 per 4 del 2008 motore multijet, 75 cavalli, diesel, velocità massima 180 chilometri orari e funzionava bene e l'**[aveva pagata]**[robot] solo 5000 dollari e a un certo punto **[volle avere]**[robot] una casa perché **[abitava]**[robot] in un hotel. La **[comprò]**[robot] e **[spese]**[robot] solo 130000 dollari, un vero affare ma poi **[si stufò]**[robot] di fare il cassiere e **[provò]**[robotbene] e **[gli]**[robot] piacque e **[lo]**[robot] pagavano molto di più e poi **[si stufò]**[robot] della solita Fiat Panda e **[comprò]**[robot] una Citroen C3, 100 cavalli, diesel, 5 marce nuova e la **[pagò]**[robot] 12700 dollari e piano piano **[continuò]**[robot] a fare evoluzioni.

CE1, 217

C'era una volta **[un gatto]**[cat] **[che]**[cat] girava per la foresta. Un giorno **[vide]**[cat] una casa fatta tutta di caramelle. Una notte **[un lupo]**[wolf] **[lo]**[cat] vide dormire e da quel momento **[il lupo]**[wolf] ha avuto in mente di mangiare **[il gatto]**[cat]. La mattina dopo **[lui]**[cat] è andato in un lago ma sott'acqua e prima che **[scappasse]**[cat] **[il lupo]**[wolf] **[l']**[cat]ha mangiato e **[un pescatore]**[ext1] vide tutto poi **[ha chiamato]**[ext1] **[un poliziotto]**[ext2], **[il poliziotto]**[ext2] ha chiamato **[l'esercito]**[ext3], **[l'esercito]**[ext3] chiama **[un cacciatore]**[ext4] e **[il cacciatore]**[ext4] liberò **[il gatto]**[cat].

CE1, 275

C'era **[un gatto]**[cat] **[che]**[cat] si era perduto in un bosco e vicino al bosco c'era una casa con scritto "ciao sono **[la strega]**[witch] ahahah!" E **[il gatto]**[cat] si avvicinò e la porta si aprì da sola e c'era **[la strega]**[witch] e anche **[un lupo]**[wolf] **[che]**[wolf] stava preparando una minestra per **[la strega]**[witch] però **[il gatto]**[cat] si spaventò e **[urlò]**[cat]. **[Scappò]**[cat] e **[il lupo]**[wolf] **[lo]**[cat] sentì e quindi **[il lupo]**[wolf] si girò e **[fece]**[wolf] così "hmmm **[che buon gatto]**[cat] da mangiare" allora **[il gatto]**[cat] correva come un ghepardo e **[il lupo]**[wolf] anche **[lui]**[wolf] correva come un ghepardo e **[il lupo]**[wolf] stava per toccar**[lo]**[cat] ma **[un robot]**[robot] **[l']**[wolf] ha afferrato **[il lupo]**[wolf] e **[gli]**[wolf] **[disse]**[robot] "addio" e **[il robot]**[robot] **[gli]**[wolf] mise l' elettricità e **[il lupo]**[wolf] morì e **[il gatto]**[cat] rimarrà sempre vivo.

CE1, 279

C'era una volta **[un gatto]**[cat] **[che]**[cat] si chiamava **[Viola]**[cat] **[che]**[cat] era seduto in terra e **[stava miagolando]**[cat] seduto in un bosco e **[un lupo]**[wolf] è venuto e **[ha detto]**[wolf] "ciao, perché **[sei]**[cat] in

giro in questi boschi?" e poi **[il gatto]**[cat] ha detto che questo posto è bello e ha detto **[il lupo]**[wolf]" in queste parti del bosco ci sono **[i lupi cattivi]**[wolf1] **[che]**[wolf1] potrebbero mangiar**[ti]**[cat] tipo **[io]**[wolf]" e **[il gatto]**[cat] è andato a correre **[dalla [sua]**[cat] **padrona]**[ext1] e **[il lupo]**[wolf] è andato nel bosco e ancora **[il gatto]**[cat] è tornato nel bosco e **[il lupo]**[wolf] è tornato e ha detto **[il lupo]**[wolf] "**[mi spiace]**[wolf] per tutto" e **[il gatto]**[cat] accetta le **[sue]**[wolf] scuse e tutti restano tutti felici e contenti.

CE1, 28

C'era una volta **[una gatta selvaggia]**[cat] e un giorno **[incontrò]**[cat] **[una strega]**[witch]. **[La gatta]**[cat] **[gli]**[witch] disse: « **[Mi]**[cat] **[fai]**[witch] diventare magica? » e **[la strega]**[witch] accettò e **[l']**[cat]**[ha fatta diventare]**[witch] magica e **[la gatta]**[cat] se ne andò.

[La gatta][cat] dopo non trovò più casa **[sua]**[cat], **[non si ricordava]**[cat] più che aveva la magia ma **[la strega]**[witch] **[la]**[cat] trovò e **[la]**[cat] **[portò]**[witch] a casa **[sua]**[witch] e **[la gatta]**[cat] disse: « Grazie! » e **[la strega]**[witch] disse: « Prego! » e **[la gatta]**[cat] diventò **[l'animale [della strega]**[witch]**]**[cat] e **[il gatto]**[cat] andava sempre in giro con **[la [sua]**[cat] **padrona]**[witch] **[che]**[witch] era **[la strega]**[witch] ma un giorno **[la strega]**[witch] preparò la pappa per **[la gatta]**[cat] che era il tonno e **[la gatta]**[cat] lo mangiò e vissero tutti felici e contenti...

CE1, 284

Tanto tempo fa c'era **[un robot]**[robot] **[che]**[robot] cammina e **[vede]**[robot] **[una befana]**[witch], meno male **[la befana]**[witch] non **[l']**[robot]aveva visto, **[lui]**[robot] non sapeva che esistono le befane, lo **[andò]**[robot] a dire **[ai [suoi]**[robot] **amici robot]**[robot1] "ehi **[robot]**[robot1], lo sapete che esistono le befane, **[volete venire]**[robot1] con **[me]**[robot]?" "sì" **[dicono]**[robot1]. Lo disse **[agli altri robot]**[robot1] e **[gli altri]**[robot1] hanno detto "sì". **[Si sono avventurati]**[robot,robot1] nel bosco per rivedere **[quella befana]**[witch] "e **[gli altri]**[robot1] dove sono?" disse **[il robot]**[robot] che ha visto **[la befana]**[witch] "prima **[erano]**[robot1] 140 adesso sono 30" **[un altro robot]**[robot2] ha detto "**[si saranno persi]**[robot1]" "no, perché **[eravamo agganciati]**[robot,robot1] a una catena" disse **[il robot]**[robot] **[che]**[robot] aveva visto **[la befana]**[witch] "oh! **[Eccola]**[witch] lì! **[La befana]**[witch] e proprio lì!" "È vero, **[eccola]**[witch] lì, **[abbiamo staccato]**[robot1] la catena, **[siamo corsi]**[robot1] **[dalla befana]**[witch] e **[abbiamo chiamato]**[robot1] **[i soccorsi]**[ext1]. **[I soccorsi]**[ext1] avevano un jet che sparava reti, pum!!! **[Hanno sparato]**[ext1] la rete e **[l']**[witch]**[abbiamo presa]**[ext1]".

CE1, 285

C'era una volta **[una strega]**[witch] con **[il [suo]**[witch] **assistente robot]**[robot] **[che]**[robot] **[la strega]**[witch] **[lo]**[robot] obbligava a fare tutto. Un giorno però **[le]**[witch] dice **[il robot]**[robot] " **[io]**[robot] non voglio più fare tutto " ma **[la strega]**[witch] fa una magia e quella magia era che **[lui]**[robot] non poteva più parlare. **[Il robot]**[robot] è davvero triste, **[la strega]**[witch] se la gode. La notte **[il robot]**[robot] senza fare rumore se ne andò ma **[la strega]**[witch] di mattino notò che **[lui]**[robot] non C'era . Prima **[faceva colazione]**[witch], **[la strega]**[witch] bloccò tutto il quartiere tranne la **[sua]**[witch] casa però **[lei]**[witch] non aveva notato che

[non aveva bloccato]witch **[il robot]**robot. **[La strega]**witch cade in un bel sonno poi **[il robot]**robot si era svegliato, **[si bagna]**robot, **[la strega]**witch baciò **[il robot]**robot e quando **[lo]**robot **[baciò]**witch **[il robot]**robot poi diventò **[un principe]**robot e **[la strega]**witch **[una principessa]**witch.

CE1, 295

C'era una volta **[una strega]**witch **[che]**witch voleva prendere il cuore di **[un gatto]**cat. **[Provò]**witch con un sacco di trappole ma niente, **[il gatto]**cat era troppo intelligente. Ma un giorno **[gli]**cat **[ha fatto]**witch una trappola e **[il gatto]**cat ci è cascato. **[La strega]**witch si era trasformata **[nella padrona [del gatto]cat]**witch e **[gli]**cat **[ha dato]**witch dei croccantini avvelenati poi **[lo]**cat **[portò]**witch nella foresta ma **[la strega]**witch era vecchia e **[si era dimenticata]**witch il coltello a casa. Intanto passò **[un robot]**robot, ha visto **[il gatto]**cat e **[lo]**cat **[portò]**robot in ospedale. **[La strega]**witch tornò e quando **[non ha trovato]**witch **[il gatto]**cat **[si era arresa]**witch. Dopo due settimane **[il gatto]**cat guarì e **[il robot]**robot **[lo]**cat accompagnò a casa.

CE1, 297

C'era una volta **[una strega cattiva]**witch **[che]**witch aveva **[un gatto]**cat ma un giorno **[la strega]**witch cacciò via **[il gatto]**cat e poi **[un lupo]**wolf **[lo]**cat trovò e **[lo]**cat **[stava]**wolf per mangiare ma "sapete che è successo?" Che **[il gatto]**cat graffiò **[il lupo]**wolf e **[il lupo]**wolf dovette coprirsi gli occhi e **[il gatto]**cat riuscì a scappare ma **[sbatté]**cat contro un albero ma **[si rialzò]**cat e **[tornò]**cat **[dalla strega]**witch ma **[la strega]**witch era così arrabbiata che **[lo]**cat **[calciò]**witch, bamm! Ma **[il gatto]**cat questa volta "sapete chi incontrò?" **[Uno scoiattolo]**ext1 **[che]**ext1 **[gli]**cat strappava tutti i **[suoi]**cat peli e **[il povero scoiattolo]**ext1 venne mangiato **[dal gatto]**cat ma **[al gatto]**cat **[gli]**cat venne un gran mal di pancia. E poi **[era tornato]**cat **[dalla strega]**witch e **[la]**witch **[trovò]**cat felice e vissero tutti felici e contenti tranne **[il lupo]**wolf perché **[non è riuscito]**wolf a mangiare **[il gatto]**cat ma tranquilli perché **[ha mangiato]**wolf **[un topo]**ext2. Era un po' brutto ma **[al lupo]**wolf **[gli]**wolf era piaciuto mangiare **[il topo]**ext2. E **[si era divertito]**wolf tantissimo a inseguire **[il topo]**ext2 perché **[il topo]**ext2 pensava che era una gara ma invece era un modo per mangiar**[lo]**ext2.

CE1, 307

C'era una volta **[una strega]**witch **[che]**witch abita in un castello. E un giorno **[incontra]**witch **[un gatto]**cat. E **[lo]**cat **[prende]**witch in mano **[il gatto]**cat e **[lo]**cat **[fa diventare]**witch tutto nero. Proprio come **[la strega]**witch era cattiva e **[iniziò]**witch a fare le pozioni. **[Il gatto]**cat aveva paura quindi **[scappa]**cat.

E **[graffia]**cat **[la strega]**witch e **[alla strega]**witch **[le]**witch è uscito il sangue. **[Si sente male]**witch, **[decide]**witch di iniziare a essere gentile e adesso **[incontra]**witch **[tutte le persone]**ext1 e **[alle altre persone]**ext1 **[dice]**witch "ciao".

E **[dice]**witch "grazie" e così finisce la storia.

CE1, 308

[Il lupo][wolf] e **[il rubbotik]**[robot] hanno una casa bruttissima, il tetto in dentro e la casa brutta doveva cadere a pezzi.

E **[la strega]**[witch] **[gli]**[robot] fece una pozione magica e si rompe e **[scappò]**[witch] via con la scopa e **[lo]**[robot] **[portò]**[witch] via e **[lo]**[robot] **[mise]**[witch] in una caverna buio pesto.

CE1, 31

C'era una volta **[un gattino di nome Desi]**[cat].

[Il piccolo][cat] era:

carino, giocherellone e simpatico. **[Desi]**[cat] ha **[una padroncina]**[ext1] **[che]**[ext1] chiama **[Sara]**[ext1].

Ma un brutto giorno **[Desi]**[cat] si era perso inseguendo **[una farfalla]**[ext2]. **[Il gattino]**[cat] tremava dalla paura ma a un tratto **[sentì]**[cat] una voce e **[capì]**[cat] subito che era **[Sara]**[ext1] **[che]**[ext1] **[lo]**[cat] stava cercando, **[il piccolo]**[cat] andò verso la voce e alla fine **[la]**[ext1] **[vide]**[cat].

Poi tornati a casa insieme **[andarono]**[cat,ext1] a letto e vissero tutti felici e contenti.

CE1, 319

C'era una volta **[un lupo]**[wolf] **[che]**[wolf] ogni giorno doveva alzarsi e correre per vedere se **[la strega Gegi]**[witch] aveva portato dei dolcetti ma ogni giorno **[continuava]**[wolf] a guardare se **[la strega Gegi]**[witch] aveva portato qualcosa ma niente, **[non trovò]**[wolf] neanche una mosca. Allora un giorno **[disse]**[wolf] piagnucolando "**[io]**[wolf] non ho fatto niente di male, **[io]**[wolf] sono bravo, guarda **[quei bambini]**[ext1] come festeggiano per i dolcetti" allora **[il lupo]**[wolf] iniziò a urlare, a strillare e a piangere dicendo "non è giusto, **[non ho fatto]**[wolf] niente". **[[Sua]**[**wolf**] **mamma]**[ext2] **[lo]**[wolf] sgridò e **[disse]**[ext2] "**[sei]**[wolf] troppo monello, come **[permetti]**[wolf] che **[la strega Gegi]**[witch] **[ti]**[wolf] porta qualcosa?" **[Il lupo]**[wolf] disse "non è vero che **[sono]**[wolf] monello, **[io]**[wolf] sono bravissimo a calcio, a scuola, a casa, in giro, insomma **[io]**[wolf] sono bravo da tutte le parti". Un giorno dopo **[la strega]**[witch] passo di lì e **[vide]**[witch] **[il lupo]**[wolf] piangere e **[disse]**[witch] "forse **[sono stata]**[witch] un po' crudele". **[Il lupo]**[wolf] **[la]**[witch] vide, **[l']**[witch]**[acchiappò]**[wolf] e **[disse]**[wolf] "perché **[non [mi]**[**wolf**] **porti]**[witch] i dolci eeee?" **[la strega]**[witch] cercava di liberarsi e **[il lupo]**[wolf] **[la]**[witch] legò a una bara di legno a forma di cilindro e **[disse]**[wolf] "se **[non [mi]**[**wolf**] **porti]**[witch] i dolci **[ti]**[witch] succede questo" **[la strega]**[witch] disse "**[posso]**[witch] rimediare" **[il lupo]**[wolf] disse "no **[non puoi]**[witch] rimediare". **[Accese]**[wolf] la pentola bollente mettendo dentro **[la strega]**[witch]. **[Il lupo]**[wolf] disse "**[vedi]**[witch] cosa succede se **[mi]**[wolf] **[prendi]**[witch] in giro". **[Prese]**[wolf] un ago e gli mise dentro un pezzo di carne, **[tirò]**[wolf] fuori **[la strega]**[witch]. **[La strega]**[witch] disse "grazie" **[il lupo]**[wolf] disse "**[mi]**[wolf] **[dai]**[witch] un altro ago". Insieme fecero il barbecue "però **[Gegi]**[witch] devi andare prima che la **[mia]**[wolf] mamma e il **[mio]**[wolf] papà **[ti]**[witch] facciano del male" **[la strega]**[witch] disse "no" **[il lupo]**[wolf] disse "**[ti]**[witch] **[prego]**[wolf] sennò succede che ti fai male". **[La strega]**[witch] non **[lo]**[wolf] ascoltò ma **[sentì]**[witch] dei passi. **[Gegi]**[witch] disse

"questi sono i **[tuoi]**[wolf] genitori" **[lupo]**[wolf] disse "sì, **[scappa]**[witch] ora". **[La strega]**[witch] in un batter d' occhio scappò. **[La mamma di lupo]**[ext2] disse "cosa succede qua?" **[il papà]**[ext3] disse "che fracasso!" "cosa succede?" rispose **[la mamma]**[ext2]. **[Il lupo]**[wolf] fece finta di niente., faceva finta di guardare la tv. L' accese e **[guardò]**[wolf] una partita, era Napoli e Milan. **[La mamma]**[ext2] disse "cosa ci fai ancora in piedi, è ora di dormire" **[il lupo]**[wolf] disse "va bene **[mamma]**[ext2]" **[andò]**[wolf] a letto quando tutto fu silenzioso **[lupo]**[wolf] aprì un occhio guardando se **[mamma]**[ext2] e **[papà]**[ext3] dormivano, **[guardò]**[wolf] in salotto e vide un mare di dolcetti, **[fece]**[wolf] una chiamata a **[Gegi]**[witch] e disse "pronto, sono **[lupo]**[wolf], grazie per i dolci" **[strega]**[witch] disse "non c'è di che" **[lupo]**[wolf] disse "potresti fare anche un incantesimo" **[strega]**[witch] disse "sì che **[posso]**[witch]" **[lupo]**[wolf] disse "perfetto, **[puoi]**[witch] darci una casa al mare poi sulla sabbia poi una finestra e poi" **[strega]**[witch] disse "ho capito" **[lupo]**[wolf] disse "devo recitare qualcosa" **[strega]**[witch] disse "no, **[io]**[witch] devo recitare palimbo palumbo una casa al mare vorrei buu". **[Lupo]**[wolf] non crede ai **[suoi]**[wolf] occhi, una casa stupenda al mare. "Ora però **[vai]**[witch] che chiamo i **[miei]**[wolf] genitori" "ok" "**[mamma]**[ext2], [p**apà]**[ext3] non ci crederete mai, guardate una casa al mare. **[La strega]**[witch] ha fatto un incantesimo". Tutti con la bocca aperta. **[Lupo]**[wolf] entrò subito nell'acqua, **[lupo]**[wolf] disse "è gelata". **[Mamma]**[ext2] sulla sdraio e **[papà]**[ext3] sulla sabbia (omission?). **[Lupo]**[wolf] chiamò **[Gegi]**[witch], diventarono tutti amici. Fine e vissero felici e contenti.

CE1, 325

C'era una volta **[un robot chiamato Mattia]**[robot] e **[il robot]**[robot] va nel bosco e sull' albero c'era **[un gatto]**[cat] e **[il Mattia]**[robot] disse "**[vieni]**[cat] giù" **[al gatto]**[cat] e **[il gatto]**[cat] veniva giù. **[Mattia]**[robot] disse "**[ti]**[cat] **[chiamerò]**[robot] **[Silvestro]**[cat]". **[Mattia]**[robot] e **[Silvestro]**[cat] vanno a casa. **[Mattia]**[robot] dava il latte a **[Silvestro]**[cat] e dopo **[giocavano]**[cat,robot].

CE1, 327

{[Il lupo][wolf] e **[il robot]**[robot].}

C'era una volta in un bosco **[un lupo]**[wolf] **[che]**[wolf] era molto cattivo perché **[prendeva]**[wolf] sempre in giro **[gli animali]**[ext1] e alcune volte **[li]**[ext1] **[mangiava]**[wolf]. Un giorno nel bosco capitò **[una strega]**[witch]. **[Il lupo]**[wolf] **[che]**[wolf] **[la]**[witch] spiava disse "gnam, con **[questa strega]**[witch] ci farò un arrosto". Poi **[il lupo]**[wolf] balzò **[sulla strega]**[witch], **[la]**[witch] **[prese]**[wolf] e **[la]**[witch] **[portò]**[wolf] nella **[sua]**[wolf] tana. Una volta rinchiusa **[il lupo]**[wolf] leccandosi i baffi disse "che bel pranzetto" poi **[disse]**[wolf] **[alla strega]**[witch] "**[aspetta]**[witch]**[mi]**[wolf] qui **[io]**[wolf] intanto vado a pensare al condimento". Una volta che **[il lupo]**[wolf] se ne andò **[la strega]**[witch] **[che]**[witch] era furba pronunciò una formula e **[uscì]**[witch] dalla gabbia però **[il lupo]**[wolf] se ne accorse e **[corse]**[wolf] in casa. **[Il lupo]**[wolf] disse "**[sei]**[witch] in trappola" però **[la strega]**[witch] disse "lo vedremo. Dan din don **[robot]**[robot]!". **[Il lupo]**[wolf] diventò **[un robot]**[wolf] e così **[non prendeva più in giro]**[wolf] **[gli animali]**[ext1] e visse felice e contento.

CE1, 34

C’era una volta **[una strega]**witch ma non era cattiva anzi **[era]**witch buonissima e **[aveva]**witch **[un bel gattino]**cat **[che]**cat si chiama **[Nemo]**cat, **[era]**cat bianco e con gli occhi castani e **[adorava]**cat le bambine.

Un giorno davanti alla casa **[della strega]**witch arrivò **[uno stregone buono]**ext1 con **[un gattino]**cat1 **[che]**cat1 si chiamava **[Salem]**cat1, **[era]**cat1 marrone e **[aveva]**cat1 gli occhi azzurri.

Quando **[lo]**ext1 vide **[la strega]**witch se ne innamorò subito, anche quando **[la]**witch vide **[lo stregone]**ext1 se ne innamorò subito e **[le]**witch **[chiese]**ext1 di sposar**[lo]**ext1. **[La strega]**witch accettò, le nozze ci furono e **[si sposarono]**ext1,witch e vissero felici e contenti.

CE1, 349

C’era una volta **[una strega]**witch **[che]**witch viveva in un castello magico. Oggi **[uscì]**witch dal castello e **[andò]**witch in giro e **[incontrò]**witch **[un lupo]**wolf e disse **[la strega]**witch "che ci fai **[tu]**wolf qui?" e **[il lupo]**wolf disse "che ci fai **[tu]**witch.” “**[Io]**witch stavo facendo una passeggiata adesso **[ti]**wolf **[lancio]**witch un incantesimo" **[disse]**witch e la magia non funzionò e **[la strega]**witch se ne va al castello. Quando **[entrò]**witch nel castello **[vide]**witch **[il lupo di prima]**wolf "ma che ci fai **[tu]**wolf?" "che ci fai **[tu]**witch?" e **[diventarono]**witch,wolf amici "**[accomodati]**wolf" "sì. Grazie". **[La strega]**witch diventò buona e **[invitò]**witch **[una mamma]**ext1 e **[fecero]**ext1,witch una storia "**[andiamo]**ext1,witch nella foresta" "sì. **[Andiamo]**ext1,witch". **[Lei]**witch andò e **[incontrò]**witch **[un altro lupo]**wolf1 e "sì, ciao **[vuoi venire]**wolf1?" "sì, va bene" e **[andò]**witch al castello "**[andiamo]**ext1,witch ancora nella foresta" e **[incontrò]**witch **[10 lupi]**wolf2 "ok, **[andiamo]**ext1,witch nella nella foresta" "ok, però è l' ultima volta che **[andiamo]**ext1,witch" e **[incontrò]**witch **[5 lupi]**wolf3 "adesso **[andiamo]**ext1,witch al castello" e vissero felici e contenti.

CE1, 357

C’era una volta **[un gatto]**cat **[che]**cat passeggiava nel bosco e **[era]**cat felice. A un certo punto **[vede]**cat **[un'ombra]**ext1 e **[si spaventa]**cat e **[va]**cat dietro un albero e **[vede]**cat **[un gufo]**ext1. E **[disse]**cat: - era solo **[un gufo]**ext1, **[io]**cat non ho paura dei gufi - disse **[il gatto]**cat. Dopo **[vede]**cat **[un'altra ombra]**ext2 e **[l'ombra]**ext2 urlò. Ma **[lui]**cat non si spaventa e alla fine era **[un leone]**ext2 e **[uscì]**cat dal bosco e **[vede]**cat un bel parco giochi e **[lui]**cat trovò **[un altro gatto]**cat1 e **[giocarono]**cat,cat1. Dopo **[mangiarono]**cat,cat1 e di notte **[dormirono]**cat,cat1 ma di notte **[un topo]**ext3 **[li]**cat,cat1 sveglia e **[i gatti]**cat,cat1 **[lo]**ext3 inseguono ma **[si trovarono]**cat,cat1 a casa **[dei ladri]**ext4. **[Entrano]**ext4 ma non erano ladri erano **[la famiglia del gattino]**ext4. E **[il gatto]**cat disse: -chi sono **[loro]**ext4?- " e **[il gattino]**cat1 rispose "è **[la [mia]cat1 famiglia]**ext4" e **[il gatto]**cat andò via e in giro nel bosco alle 9:00 di sera c’era **[un robot di nome Ficcagenio]**robot e **[il gatto]**cat fece amicizia e vissero tutti felici e contenti.

CE1, 360

C’era una volta **[una strega]**witch **[che]**witch aveva **[un gatto]**cat e una sera venne **[un robot]**robot e **[rubò]**robot **[il gatto]**cat e ne **[mise]**robot **[un altro con le telecamere]**cat1 e **[la strega]**witch si accorse subito che **[il gatto]**cat era scomparso perché **[non mangiava]**cat1 e **[non miagolava]**cat1 così **[prese]**witch la scopa e **[volò]**witch via nel bosco e **[diceva]**witch " **[Laila, Laila]**cat " che era il nome **[della gatta]**cat. A un certo punto **[sbatté]**witch contro una cosa invisibile e **[vede]**witch la porta aperta e così **[pensa]**witch di entrare e **[si guarda]**witch un po' in giro e a un verto punto **[trova]**witch **[il gatto]**cat legato a una sbarra di ferro con su scritto "esperimenti" e così **[formula]**witch una formula magica "e ora staccati corda e vieni da **[me]**witch" e così **[lo]**cat mette sulla corda e **[volano]**cat,witch via e con **[il gatto falso]**cat1 e **[lo]**cat **[rimette]**witch al posto **[dell' altro]**cat1 e vissero felici e contenti.

CE1, 386

C’era una volta **[un gatto di nome Malic]**cat. **[Viveva]**cat a Parigi, **[non aveva]**cat padrone. Mentre **[camminava]**cat **[vide]**cat un volantino che rappresentava un negozio di giocattoli. **[Malic]**cat si incuriosisce e **[va]**cat nel negozio misterioso, **[vide]**cat molte cose che **[non aveva mai visto]**cat ma c’era **[una cosa]**robot che **[lo]**cat **[attirava]**robot, era **[un robot]**robot. **[Decise]**cat di prender**[lo]**robot, **[uscì]**cat dal negozio e **[lo]**robot **[portò]**cat nel **[suo]**cat posticino preferito vicino a un cantiere. A un certo punto **[il robot]**robot si svegliò! E **[disse]**robot "ciao, **[tu]**cat chi sei?" **[il gatto]**cat rispose "chi, **[io]**cat?" "sì" disse **[il robot]**robot "io sono **[Malic]**cat" **[rispose]**cat "ciao **[Malic]**cat. **[Io]**robot mi chiamo **[Sfasac]**robot". **[I due]**cat,robot decisero di fare una passeggiata, purtroppo **[Sfasac]**robot non conosceva molto le strada di Parigi. **[Malic]**cat invece vagava per tutto il mondo. **[Lo]**robot **[porta]**cat nei posti più famosi. **[Si divertivano]**cat,robot molto insieme. **[Passarono]**cat,robot un po' di mesi ma un giorno **[Malic]**cat riconobbe **[un uomo]**ext1, era **[il [suo]cat padrone]**ext1! **[Malic]**cat corse ad abbracciar**[lo]**ext1. **[L'uomo]**ext1 **[lo]**ext1 riconobbe **[Malic]**cat. **[Sfasac]**robot ci mise un po' per prender**[ci]**ext1 confidenza ma alla fine non era poi così male. **[L'uomo]**ext1 **[li]**cat,robot portò in casa **[sua]**ext1 e vissero tutti felici e contenti.

CE1, 394

C’era una volta **[un lupo]**wolf. Tutti avevano paura di **[lui]**wolf. **[Il povero lupo]**wolf era sempre solo e **[non sapeva]**wolf che cos'era il divertimento.

Un bel giorno passò di lì **[un gatto]**cat. **[Il gatto]**cat chiese **[al lupo]**wolf "perché **[sei]**wolf triste?" **[Il lupo]**wolf rispose "perché nessuno **[mi]**wolf vuole" allora **[il gatto]**cat disse "non **[ti]**wolf preoccupare, **[puoi]**wolf essere **[mio]**cat amico" disse **[il gatto]**cat.

Sulla faccia **[del lupo]**wolf venne un sorriso, **[il gatto]**cat portò **[il lupo]**wolf al ruscello, **[fecero]**cat,wolf un bagno e **[il lupo]**wolf si divertì tanto anzi tantissimo. **[Gli animali]**ext1 avevano visto **[il lupo]**wolf e **[il gatto]**cat giocare nell'acqua e **[decisero]**ext1 di aggiungersi anche **[loro]**ext1.

Alla fine **[diventarono]**cat,ext1,wolf tutti amici grazie **[al gatto]**cat ma nessuno sapeva che **[il gatto]**cat era magico.

Morale non fatevi ingannare dalle apparenze prima dovete conoscere le persone.

CE1, 396

C'era una volta **[un gatto di campagna]**^{cat} **[che]**^{cat} viveva dentro a un cespuglio con **[i [suoi]**^{cat} **cuccioli]**^{cat1,cat2,cat3} e un giorno **[la mamma]**^{cat} decise di andare a fare una passeggiata nel bosco mentre **[i 3 cuccioli]**^{cat1,cat2,cat3} stavano dormendo.

E **[la mamma]**^{cat} cominciò a marciare.

E dopo aver fatto il giro **[tornò]**^{cat} a casa ma **[i cuccioli]**^{cat1,cat2,cat3} non c'erano.

Allora **[la mamma preoccupata]**^{cat} decise di cercar**[li]**^{cat1,cat2,cat3}. **[Cercò]**^{cat} e **[cercò]**^{cat} ma **[non [li]**^{cat3,cat2,cat1} **trovò]**^{cat}.

Ad un tratto **[vide]**^{cat} una cosa, e gli **[chiese]**^{cat} "ehi **[tu]**^{cat1} chi sei?!" **[Lo straniero]**^{cat1} **[lo]**^{cat} guardò ed era **[il cucciolo]**^{cat1}! **[La mamma]**^{cat} corse ad abbracciar**[lo]**^{cat1}.

E insieme **[si misero a cercare]**^{cat,cat1} **[l'altro cucciolo]**^{cat2}. E videro **[un branco di lupi]**^{wolf} che rincorrevano **[un gatto]**^{cat2}. Era **[il cucciolo]**^{cat2}! Poi **[la mamma]**^{cat} pure **[lei]**^{cat} si mise a rincorrere **[il cucciolo]**^{cat2} e **[riuscì]**^{cat} a prender**[lo]**^{cat2}. Poi pochi metri di distanza c'era un burrone e **[tutti i lupi]**^{wolf} caddero tranne **[uno]**^{wolf} **[che]**^{wolf} sembrava buono. E infatti lo **[era]**^{wolf}. E **[li]**^{cat} **[aiutò]**^{wolf} a cercare **[l'altro cucciolo]**^{cat3}.

Videro **[una cosa bianca]**^{cat3} su un albero.

E **[si misero]**^{cat,cat1,cat2,wolf} uno sopra l'altro ed quello era **[il cucciolo]**^{cat3}!

[Il cucciolo]^{cat3} scese e tutti corsero a casa ed anche **[il lupo]**^{wolf}. E a casa **[festeggiarono]**^{cat,cat1,cat2,cat3,wolf}! E vissero tutti felici e contenti! Fine

CE1, 398

C'era una volta **[un gatto tutto solo]**^{cat}, un giorno **[il gatto]**^{cat} incontrò **[un lupo]**^{wolf} e **[si spaventò]**^{cat}, poi **[il lupo]**^{wolf} **[gli]**^{cat} disse "**[io]**^{wolf} sono buono mica come gli altri lupi che sono cattivi, **[vuoi]**^{cat} giocare con **[me]**^{wolf}?", "sì va bene a cosa **[giochiamo]**^{cat,wolf}?" "A nascondino" "ok però non si può andare più lontano da quegli alberi" "va bene" "allora **[iniziamo]**^{cat,wolf}, si conta fino a 20, 1, 2, 3 **[arrivo]**^{cat}!" "**[Trovato]**^{cat}, tocca a **[te]**^{cat} contare". "**[Mi]**^{wolf} è venuta fame" "anche a **[me]**^{cat} " "**[andiamo]**^{cat,wolf} al ristorante" dice **[il lupo]**^{wolf}, **[il gatto]**^{cat} dice "**[io]**^{cat} ordino il pesce", "invece **[io]**^{wolf} la carne, gnam che buono!", "ok, **[possiamo andare]**^{cat,wolf} a casa a dormire" "**[ci vediamo]**^{cat,wolf} domani alle 10.00" buona notte! E vissero tutti felici e contenti.

CE1, 405

C'era una volta **[un gatto]**^{cat} **[che]**^{cat} passeggiava tranquillamente in un piccolo sentiero e **[arrivò]**^{cat} in una casa che **[gli]**^{cat} sembrava un'antenna. **[Il gatto]**^{cat} si stupì e **[volle entrarci]**^{cat} dentro la

casetta. **[Trovò]**[cat] **[un lupo]**[wolf] **[che]**[wolf] faceva finta di essere malato e **[il gatto]**[cat] cascò nel tranello che **[gli]**[cat] aveva fatto **[il lupo]**[wolf] perché **[voleva mangiar[lo]**[cat]**]**[wolf].

Dopo un po' **[il lupo]**[wolf] si scoraggiò di mangiare **[il gatto]**[cat] allora **[gli]**[cat] **[disse]**[wolf] tutta la verità.

[Il gatto][cat] **[lo]**[wolf] sgridò molto ma poi **[divennero]**[cat,wolf] migliori amici e **[hanno sconfitto]**[cat,wolf] molti predatori ma un giorno **[il gatto]**[cat] fece uno scherzo **[al lupo]**[wolf], **[si perse]**[cat] volontariamente e **[il lupo]**[wolf] si spaventò tantissimo!

Il giorno dopo **[il gatto]**[cat] ritornò, e **[il lupo]**[wolf] **[gli]**[cat] chiese "dove **[sei stato]**[cat]?"

"A combattere" "oh **[che bravo micio]**[cat], e ora **[guardiamo]**[cat,wolf] un po' di TV" "ok" disse **[il gatto]**[cat]. La storia si conclude così, **[il gatto]**[cat] e **[il lupo]**[wolf] andarono a letto e **[fecero]**[cat,wolf] un sogno uguale.

E vissero tutti felici e contenti!

CE1, 414

C'era una volta **[un robot]**[robot] **[che]**[robot] cercava un amico solo che **[gli altri]**[ext1] non volevano essere amici **[del robot]**[robot]. **[Girava]**[robot] nei paesi, **[cercava]**[robot] da tutte le parti e alla fine **[aveva trovato]**[robot] **[un gatto]**[cat] **[che]**[cat] si chiamava **[Willy]**[cat] e dopo **[Willy]**[cat] e **[il robot]**[robot] stavano sempre attaccati e vissero felici e contenti.

CE1, 424

Un giorno **[un gatto]**[cat] stava camminando su un marciapiede. Dopo un po' di tempo **[si accorse]**[cat] che erano passate 3 ore da quando **[aveva iniziato]**[cat] a camminare, ma **[il gatto]**[cat] non era ancora stanco e quindi ancora **[continuò]**[cat] a camminare. Poi dopo un po' di tempo **[incontrò]**[cat] **[un altro gatto]**[cat1], e **[iniziarono]**[cat,cat1] a giocare insieme. Poi **[il gatto]**[cat] si accorse che era passata un'ora. E poi **[ritornò]**[cat] nella sua gabbietta però la **[chiude]**[cat] a chiave, per **[non avere]**[cat] pericolo. Dopo 50 ore arrivò **[la strega]**[witch], anche se la gabbietta era chiusa a chiave con un incantesimo **[aprì]**[witch] la gabbietta, però **[il gatto]**[cat] stava dormendo con l'incantesimo che **[ha fatto]**[witch] alla gabbietta funzionò anche per **[il gatto]**[cat], e **[si svegliò]**[cat]. Poi **[il gatto]**[cat] per rimettersi a letto **[alla strega]**[witch] **[le]**[witch] diede una graffiata in faccia e **[la strega]**[witch] morì e **[il gatto]**[cat] rimase sano e salvo e **[si mise]**[cat] di nuovo a letto ma **[non si addormentò]**[cat] e quindi dopo 10 ore **[gli]**[cat] era venuto sonno e **[dormì]**[cat] per 30 ore e dopo **[si svegliò]**[cat] ancora sano e salvo.

CE1, 432

Una sera buia nel castello di **[un robot di nome Antenna di fuoco]**[robot], **[qualcuno]**[wolf] bussò alla porta, **[Antenna di fuoco]**[robot] terrorizzato aprì la porta, si vedeva **[un' ombra buia, con occhi gialli, muso a punta e denti che spuntano dalla bocca]**[wolf]. **[Quell' ombra sconosciuta]**[wolf] entrò nel castello, era **[un lupo]**[wolf]! **[Antenna di fuoco]**[robot] scappò con tutte le forze e **[si nasconde]**[robot]

dietro il divano, **[il robot]**[robot] lanciò: bicchieri, scatole e matite, il rumore era forte PINC ZAC BOOM ma non c'era niente da fare, allora **[Antenna di fuoco]**[robot] lanciò un coltello e **[il lupo]**[wolf] morì e da quel giorno i lupi non esistettero più.

CE1, 451

C'era una volta una casa. Dentro ci vivevano **[un gatto nero]**[cat] e **[una strega di nome Vanvel]**[witch] **[che]**[witch] ogni giorno prendeva **[una lucertola]**[ext1] e **[la]**[ext1] **[trasformava]**[witch] in **[un corvo]**[ext1] per spiare **[il grande re Carlo]**[ext2], però un brutto giorno **[la strega]**[witch] vide che **[il gatto]**[cat] stava male e **[si accorse]**[witch] che **[aveva bisogno]**[cat] del peperoncino piccante che è a circa 100 chilometri da lì. **[Il gatto nero]**[cat] stava sempre più male e quindi **[la strega]**[witch] va nel bosco, **[montò]**[witch] sulla **[sua]**[witch] scopa e **[partì]**[witch] lasciando il camino acceso per **[il gatto]**[cat]. Quando **[la strega Vanvel]**[witch] si accorse che non c'era e **[capì]**[witch] che c'era un passaggio segreto. E prima **[si sedette]**[witch] sopra un fungo a macchie marroni e poi uno a macchie oro ma niente però **[vide]**[witch] scintillare un fungo che era argento con le macchie rosse, **[si sedette]**[witch] sopra e **[scivolò]**[witch] in un albero e lo **[vide]**[witch], però davanti a **[lei]**[witch] arrivò **[un tizio strano]**[ext3] **[che]**[ext3] non **[le]**[witch] permetteva di arrivare al peperoncino che prima **[lo]**[ext3] **[doveva sconfiggere]**[witch] e **[usò]**[witch] la magia. **[Non ce la fece]**[witch] finché non **[usò]**[witch] l'incantesimo più potente che **[conosceva]**[witch] e poi **[prese]**[witch] il peperoncino e lo **[portò]**[witch] **[dal gatto nero]**[cat]. Quando **[arrivò]**[witch] a casa **[corse]**[witch] subito a fare la pappa **[al gatto]**[cat] con il peperoncino e subito **[guarì]**[cat] e vissero per sempre felici e contenti.

CE1, 525

C'era una volta **[un robot]**[robot] **[che]**[robot] non aveva amici perché **[era]**[robot] l'unico al mondo. Quindi **[era]**[robot] triste e un giorno **[un gatto magico]**[cat] **[gli]**[robot] ha detto "**[stai]**[robot] tranquillo **[avrai]**[robot] degli amici" rispose **[il gatto magico]**[cat] e **[gli]**[robot] **[dice]**[cat] ancora "2 o 3 giorni e **[avrai]**[robot] degli amici" e dopo 2 giorni **[il robot]**[robot] si ritrova con **[tantissimi altri robot]**[robot1] **[che]**[robot1] **[gli]**[robot] dissero "**[vuoi]**[robot] essere **[il [mio]**[robot1] **nuovo amico]**[robot]?" e **[lui]**[robot] dice "sì, sì, sì, sì" e **[il robot]**[robot] torna **[dal gatto]**[cat] e **[il robot]**[robot] dice "grazie! grazie mille **[gatto magico]**[cat] **[non so]**[robot] come ringraziar**[ti]**[cat]" e **[il gatto magico]**[cat] rispose "prego, non c' è di che" e vissero tutti felici e contenti soprattutto **[il robot]**[robot] con **[i [suoi]**[robot] **amici]**[robot1].

CE1, 533

C'era una volta **[una strega]**[witch] **[che]**[witch] si chiamava **[Pazzerella]**[witch]. Un giorno **[disse]**[witch] "è da tanto che **[non faccio]**[witch] magie", e **[si incamminò]**[witch] in un parco ; c'era **[una famiglia]**[ext1] con **[un gatto]**[cat] e **[disse]**[witch] fra sè e sè "**[che bel gatto]**[cat], **[mi]**[witch] **[potrebbe servire]**[cat] per fare delle magie". La notte **[andò]**[witch] a casa **[loro]**[ext1], **[rapì]**[witch] **[il gatto]**[cat] e **[se ne andò]**[witch], appena **[arrivati]**[cat,witch] a casa **[sua]**[witch] **[si sdraiò]**[witch] sul letto e **[si addormentò]**[witch]. La mattina **[si svegliò]**[witch] alle 5:00 per fare magie, **[fece]**[witch] diventare **[il gatto]**[cat] viola, giallo, verde, e perfino

azzurro. A un certo punto **[si fermò]**witch e **[disse]**witch "ma che strano **[questo gatto]**cat non graffia". **[Andò]**witch al supermercato a comprare una gabbia per mettere dentro **[il gatto]**cat poi **[andò]**witch dal veterinario e **[gli]**ext2 disse **[Pazzerella]**witch **[al veterinario]**ext2 "**[questo gatto]**cat non graffia". **[Il veterinario]**ext2 disse "per forza, è **[un gatto da stregone]**cat e **[lei]**witch e **[una strega]**witch". **[Tornarono]**cat,witch a casa e **[continuò]**witch a fare magie per sempre e vissero felici e contenti.

CE1, 534

C'era una volta **[una strega]**witch **[che]**witch si chiamava **[Elena]**witch, **[aveva]**witch i capelli marroni, gli occhi azzurri. Un giorno a **[questa strega Elena]**witch **[le]**witch viene una idea di trasformare **[il gatto Thomas]**cat in un umano allora **[preparò]**witch una pozione magica che era fatta di rospi, uccelli magici e galline. Il giorno dopo **[il gatto Thomas]**cat diventò **[un umano]**cat e la pozione **[lo]**cat aveva fatto anche parlare, ma il **[gatto Thomas]**cat era molto ma molto arrabbiato perché **[non voleva essere]**cat un uomo e allora **[la strega Elena]**witch il giorno dopo preparò un'altra pozione magica che **[Thomas il gatto]**cat era tornato alla normalità ma **[la strega]**witch era molto felice perché **[aveva fatto diventare]**witch **[un gatto]**cat in **[un umano]**cat e vissero felici e contenti.

CE1, 536

{[La strega]witch e **[il gatto]**cat}

C'erano una volta **[una strega]**witch e **[un gatto]**cat **[che]**cat,witch non andavano molto d'accordo insieme. Un giorno **[la strega]**witch fece una pozione da usare contro **[il gatto]**cat, ma **[il gatto]**cat era molto furbo e **[si nascose]**cat sotto il letto **[della strega]**witch, **[la strega]**witch cerca cerca e cerca ma **[non trovò]**witch niente. **[La strega]**witch disse "forse **[so]**witch dove si è nascosto **[il [mio]witch gatto]**cat, sotto il **[mio]**witch letto." **[Il gatto]**cat sentì le parole **[della strega]**witch e **[disse]**cat tra sé e sé "aiuto, aiuto dove **[mi posso nascondere]**cat? **[la strega]**witch sta arrivando!" **[il gatto]**cat andò nella **[sua]**cat cuccia a far finta di dormire. Però **[il gatto]**cat era molto ingenuo. E così **[restò]**cat lì, **[la strega]**witch nel passare vide **[il gatto]**cat e **[gli]**cat **[versò]**witch la pozione. Ma però **[la strega]**witch aveva sbagliato e **[lo]**cat **[fece diventare]**witch **[una persona]**cat. **[La persona]**cat si svegliò e **[disse]**cat "grazie **[strega]**witch per **[aver[mi]cat liberato]**witch dall'incantesimo" **[la strega]**witch disse "**[ho sbagliato]**witch a fare la pozione perché ci **[ho messo]**witch delle trecce di persone", **[la persona]**cat disse "fa niente, anzi **[mi]**cat **[hai fatto]**witch una sorpresa". **[La strega]**witch disse "**[facciamo pace]**cat,witch" e **[la persona]**cat disse "certo che **[facciamo pace]**cat,witch!" e vissero tutti felici e contenti.

CE1, 540

C'era una volta **[un gatto]**cat e **[una strega]**witch **[che]**cat,witch un bellissimo giorno di sole con tantissimi uccellini che fischiavano sugli alberi decisero di andare a fare una passeggiata e all'improvviso **[la strega]**witch sentì un rumore dietro a un cespuglio e **[si erano spaventati]**cat,witch moltissimo, **[il gatto]**cat è saltato in aria. **[Sono andati]**cat,witch vicino a quel cespuglio e **[hanno visto]**cat,witch **[il [loro]cat,witch amico robottino]**robot **[che]**robot stava dormendo, **[l']**robot**[avevano**

svegliato][cat,witch] e **[gli]**[robot] ha detto **[la strega]**[witch] "perché **[stavi dormendo]**[robot] in un cespuglio?" **[Il robottino]**[robot] rispose "perché **[un lupo cattivo]**[wolf] **[mi]**[robot] ha buttato giù la casa e **[ho trovato]**[robot] un materasso vecchio vecchio e **[mi sono addormentato]**[robot] lì" e **[il gatto]**[cat] **[gli]**[robot] disse "**[avevi]**[robot] tantissimo sonno" **[il robottino]**[robot] rispose "sì, perché **[sono andato]**[robot] a mangiare da **[un [mio]**[**robot**] **amico]**[ext1] in città". Ma 5 minuti dopo. […]

CE1, 544

{[La strega][witch] e **[il robot]**[robot]}

C'era una volta **[una strega]**[witch] **[che]**[witch] aveva inventato **[un robot]**[robot] per aiutar**[la]**[witch] a preparare una pozione molto difficile. **[La strega]**[witch] mandò **[il robot]**[robot] nella foresta grigia. **[La strega]**[witch] **[gli]**[robot] aveva detto di prendere le zampe di ragno, la bava di lucertola e forfora di lucertola allora **[la strega]**[witch] mescolò gli ingredienti e **[diventò]**[witch] come tutte le altre persone.

CE1, 546

[Il robot][robot] sta giocando con **[i [suoi]**[**robot**] **amici]**[ext1] però un giorno arrivò **[uno stregone cattivo]**[ext2] e **[ha preso]**[ext2] **[tutti i [suoi]**[**robot**] **amici]**[ext1] e poi **[li]**[ext1] **[ha portati]**[ext2] nel **[suo]**[ext2] castello. Ora **[li]**[ext1] **[salverà]**[robot] tutti. "ora **[vado]**[robot] al castello". Ora **[è arrivato]**[robot] al castello, **[eccolo]**[robot], ora **[combattiamo]**[ext2,robot] e a quel punto **[ha vinto]**[robot] e **[ha salvato]**[robot] **[tutti i [suoi]**[**robot**] **amici]**[ext1]. Fine della storia.

CE1, 585

{[Il lupo cattivo][wolf].}

È un giorno di pioggia e **[il lupo Silvestro]**[wolf] e **[i [suoi]**[**wolf**] **amici]**[ext1] sono andati a cercare **[la famosa gatta Lucia]**[cat].

[Lucia][cat] si era nascosta in una casa spaventosissima che nessuno entrava.

[Silvestro][wolf] e **[i [suoi]**[**wolf**] **amici]**[ext1] vanno a cercar**[la]**[cat] per tutto il villaggio.

Dopo un po' davanti a **[loro]**[ext1,wolf] vedono una casa spaventosissima proprio quella dove si era nascosta **[la gatta Lucia]**[cat].

[La][cat] **[trovano]**[ext1,wolf] e **[la]**[cat] **[mangiano]**[ext1,wolf] poi **[si addormentano]**[ext1,wolf].

Infine **[il gatto Fufi]**[cat1] **[che]**[cat1] era **[l'amico di [Lucia]**[**cat**]**]**[cat1] tagliò la pancia a **[tutti]**[ext1,wolf].

[Fufi][cat1] prende **[Lucia]**[cat].

In cambio **[mette]**[cat1] dei sassi.

CE1, 599

[Un gatto di nome Alex][cat] gioca con un gomitolo e **[rompe]**[cat] i letti. La tv. I pupazzi. Il divano ecc. E **[butta]**[cat] giù il frigo ed **[esce]**[cat] di casa e **[rompe]**[cat] un'auto pure la scatola **[ha rotto]**[cat]. **[Alex]**[cat] ritorna e **[si fa sgridare]**[cat]. **[Ha fatto]**[cat] la cacca fuori e pure la pipì. **[Va]**[cat] anche al mercato. Insomma **[Alex]**[cat] è passeggione: **[è andato]**[cat] in Cina. Giappone. Africa. America. Sud America. Australia ecc. **[Va]**[cat] in una scuola due giorni dopo **[va]**[cat] all'università e poi **[distrugge]**[cat] tutta tutta tutta tutta tutta tutta tutta tutta tutta l'università.

CE1, 602

C'era una volta **[una strega]**[witch] **[che]**[witch] aveva **[un gatto]**[cat]. **[La strega]**[witch] si chiamava **[Viola]**[witch] e **[il gatto]**[cat] si chiamava **[Emma]**[cat].

[Emma][cat] un giorno combinò un pasticcio **[fece rovesciare]**[cat] il pentolone di **[Viola]**[witch] e **[le]**[cat] **[ha detto]**[witch]: "Cosa **[hai combinato]**[cat]!?" e **[la]**[cat] **[mise]**[witch] in castigo per due settimane.

Però un giorno **[Emma]**[cat] fece gli occhioni dolci e **[la strega Viola]**[witch] **[la]**[cat] tolse dal castigo.

CE1, 611

C'era una volta **[una strega]**[witch] con **[il [suo]**[**witch**] **gatto]**[cat].

Un giorno **[il gatto]**[cat] vide **[la strega]**[witch] fare una pozione, e **[il gatto]**[cat] si spaventò da morire, e **[il gatto]**[cat] nella **[sua]**[cat] testa disse « Che cos' è, **[ho paura]**[cat] **[non voglio più stare]**[cat] qui un minuto di più » allora **[scappò]**[cat] e **[la strega]**[witch] il giorno dopo si svegliò e **[il gatto]**[cat] non c'era più e **[la strega]**[witch] si preoccupò e con la **[sua]**[witch] scopa **[lo]**[cat] **[andò]**[witch] a cercare. Alla fine **[lo]**[cat] **[trovò]**[witch] e vissero felici e contenti.

CE1, 627

C'era una volta **[un gatto carino]**[cat] **[che]**[cat] si chiamava **[Ghiacciolina]**[cat] **[era]**[cat] bella e buona **[era]**[cat] buona con tutti e **[ascoltava]**[cat] tutti, e **[Ghiacciolina]**[cat] era una mangiona e il **[suo]**[cat] mangiare preferito erano i croccantini e il frutto preferito è la fragola.

[Lei][cat] era anche bellissima quando si vestiva.

E **[questo gatto]**[cat] è bianco con pochissimi pallini marroni.

CE1, 629

C'era una volta **[un robot]**[robot] **[che]**[robot] andava in giro per la città. Un giorno sotto casa **[del robot]**[robot] andò a vivere **[il lupo più cattivo della città]**[wolf]. **[Il robot]**[robot] disse: « Cosa **[ci fai]**[wolf] qui sotto casa **[mia]**[robot] » « **[sono venuto]**[wolf] per distrugger**[ti]**[robot]! Tanto con la casa di legno con un

soffio la casa cade giù. » E **[il robot]**[robot] dice “**[vogliamo diventare]**[robot,wolf] amici?” “ok” dice **[il lupo]**[wolf]. Alla fine **[il lupo]**[wolf] e **[il robot]**[robot] diventano amici.

CE1, 63

C’era una volta **[un gatto abbandonato]**[cat] **[che]**[cat] passeggiando incontrò **[un robot]**[robot] "ciao, come **[ti chiami]**[cat]?" disse **[il robot]**[robot] "**[Fagiolino]**[cat]. E **[tu]**[robot]?" "**[io]**[robot] mi chiamo **[Pietro il robottino]**[robot] ma **[tu]**[cat] sei abbandonato come **[me]**[robot]?" "sì anch'**[io]**[cat] ho perso **[la famiglia]**[ext1] però basta parlare di queste cose sennò **[mi metto]**[cat] a piangere. **[Andiamo]**[cat,robot] a prendere un gelato **[Fagiolino]**[cat]?" “sì grazie ma perché **[sei venuto]**[robot] qua?" "perché nel mondo dei robot **[mi]**[robot] prendono in giro" "perché?" "**[Mi]**[robot] dicono che **[sono]**[robot] una frana nel calcio però questo è vero va beh". Così **[arrivati]**[cat,robot] "limone" "e io fragola" "ma che buono" "adesso **[andiamo]**[cat,robot] al parco a divertirci un pochino poi **[andiamo]**[cat,robot] a casa" "non è vero scusa" "perché" "quello che ti è successo a **[te]**[robot]" "adesso dove **[dormiamo]**[cat,robot]?" "in una capanna" "va bene" "però adesso **[andiamo]**[cat,robot] al parco poi **[andiamo]**[cat,robot] a dormire **[andiamo]**[cat,robot] sulla altalena poi sullo scivolo, sul dondolo" "che bello, adesso **[ho sonno]**[cat] **[andiamo]**[cat,robot] a dormire nella capanna" disse **[il gatto]**[cat] "va bene **[io]**[robot] ne conosco una" "**[andiamo]**[cat,robot] perché si sta facendo buio." "buona notte" disse **[il robot]**[robot].

CE1, 65

C’era una volta **[una strega]**[witch] e **[volava]**[witch] in cielo con la scopa e un giorno **[atterrò]**[witch] in un bosco e **[incontrò]**[witch] **[un gatto tutto ferito]**[cat] e **[lo]**[cat] **[portò]**[witch] a casa **[sua]**[witch] e **[gli]**[cat] **[ha dato]**[witch] delle ali e poi **[uscirono]**[cat,witch] e **[il gatto]**[cat] non riusciva neanche a volare e **[la strega]**[witch] **[lo]**[cat] portò con **[lei]**[witch] perché **[non sapeva]**[cat] ancora volare e quindi **[provavano]**[cat,witch] e **[provavano]**[cat,witch] ma **[non riusciva]**[cat] ancora volare e a un certo punto sbucò una luce forte, forte e spuntavano delle ali a misura di gatto perché quelle erano troppo grandi, erano per adulti e quindi **[riuscì]**[cat] a volare e poi **[gli]**[cat] **[regalò]**[witch] una scopa e **[volavano insieme]**[cat,witch].

CE1, 653

C’era una volta una casa che dentro ci abitava **[il lupo]**[wolf] **[era]**[wolf] molto famoso perché **[era]**[wolf] un gran costruttore di robot e un giorno **[costruì]**[wolf] **[un robot gigante]**[robot] arrivò **[la polizia]**[ext1] e **[lo]**[wolf] **[arrestarono]**[ext1] **[lo]**[wolf] **[mettono]**[ext1] in galera per 3 mesi e dopo **[lo]**[wolf] **[lasciano andare]**[ext1] e il morale della storia è di essere buoni costruttori.

CE1, 673

C’era una volta **[un gatto randagio]**[cat] con **[i [suoi]**[cat] **cuccioli]**[cat1,cat2] e dopo un po' arriva **[un lupo]**[wolf] e **[mangiò]**[wolf] **[tutti quanti]**[cat,cat1,cat2] poi **[se ne andò]**[wolf] e **[va]**[wolf] alla cascata a bere l' acqua poi **[un gattino]**[cat1] si è salvato perché **[si è nascosto]**[cat1] sotto un albero poi **[ha trovato]**[cat1]

[un cucciolo di lupo][wolf1] e **[giocarono]**[cat1,wolf1] tutto il giorno con **[il cucciolo di lupo]**[wolf1] poi arriva **[la mamma lupo]**[wolf2] e **[il papà lupo]**[wolf3] e **[il gattino]**[cat1] scappò dopo un po' arriva **[la [sua]**[**cat1**] **mamma]**[cat].

CE1, 68

C'era una volta **[un lupo]**[wolf] **[che]**[wolf] passeggiava poi **[incontrò]**[wolf] **[un robot]**[robot], e insieme **[facevano]**[robot,wolf] gli scherzi come quando **[qualcuno]**[ext1] apre la porta della macchina, **[gli]**[ext1] va giù l'acqua e **[sono]**[robot,wolf] amici di scherzi per sempre, e continua fino a 70 anni. **[Sono]**[robot,wolf] migliori amici, poi **[si divertono]**[robot,wolf] a lanciare i palloncini di acqua, e **[sono]**[robot,wolf] amici del cuore.

CE1, 698

{[La strega][witch] e **[il gatto]**[cat]}

C'erano una volta **[una strega]**[witch] e **[un gatto]**[cat]. **[La strega]**[witch] aveva i vestiti strappati e **[il gatto con il pelo nero]**[cat]. **[La strega]**[witch] ha incontrato **[il gatto]**[cat]. **[L']**[cat]**[ha preso]**[witch] da terra e **[l']**[cat]**[ha portato]**[witch] nella **[sua]**[witch] capanna. **[Gli]**[cat] **[ha dato]**[witch] da mangiare e poi **[l']**[cat]**[ha portato]**[witch] a fare un giro sulla **[sua]**[witch] scopa volante. **[Il gatto]**[cat] si è divertito. **[Non aveva mai visto]**[cat] una cosa del genere. **[Il gatto]**[cat] e **[la strega]**[witch] diventarono amici e vissero felici e contenti.

CE1, 707

C'era una volta **[una gatta]**[cat] **[il proprietario]**[ext1] era **[una principessa]**[ext1] **[che]**[ext1] si chiamava **[Cenerentola]**[ext1]. **[Avevano]**[cat,ext1] un bel castello e una volta voleva uscire **[la gatta]**[cat] perché era arrivato **[il principe]**[ext2] e molto furba una notte **[aprì]**[cat] la finestra e **[scappò]**[cat] dopo di mattina prima che **[si svegliò]**[ext1] **[Cenerentola]**[ext1] si accorse che non c'era più **[la gatta]**[cat] e **[stava piangendo]**[ext1] tutto il giorno però **[andò]**[ext1] a cercar**[la]**[cat] e **[la]**[cat] **[trovò]**[ext1] ferita e **[la]**[cat] **[portò]**[ext1] subito al castello e **[si accorse]**[ext1] che **[non respirava]**[cat]. Al pronto animali **[le]**[ext1] dissero « **[Vuoi salutar[la]**[**cat**]**]**[ext1] per l' ultima volta? » sì perché **[sta]**[cat] per morire.

CE1, 710

C'era una volta **[una strega]**[witch] **[che]**[witch] aveva molta fame. Un giorno **[la strega]**[witch] vide **[un gatto]**[cat] **[che]**[cat] passeggiava nel bosco. **[La strega]**[witch] subito quando **[lo]**[cat] **[vide]**[witch] **[le]**[witch] venne fame. **[La strega]**[witch] però non sapeva come prender**[lo]**[cat]. Adesso **[la strega]**[witch] aveva molta più fame. Dopo un po' **[alla strega]**[witch] venne una idea. **[Si travestì]**[witch], **[prese]**[witch] un sacco per mettere giù **[il gatto]**[cat] e **[partì]**[witch] alla ricerca **[del gatto]**[cat]. Però **[il gatto]**[cat] si fece furbo. Mentre **[la strega]**[witch] era girata, **[il gatto]**[cat] prese il sacco dalle mani **[della strega]**[witch]. **[La strega]**[witch] rimase lì per sempre perché **[non è riuscita]**[witch] a liberarsi. Fine

CE1, 728

C’era una volta **[un robot]**^{robot} **[che]**^{robot} viveva nel villaggio di Robot city dove vivevano **[tanti tanti robot]**^{robot1}. **[Robot gioia]**^{robot1}, **[Robot calciatore]**^{robot1}, **[Robot burlone]**^{robot1}, **[Robot maldestro]**^{robot1} e **[tanti altri]**^{robot1}. Ma a un certo punto era ora di andare a calcio poi allora quando **[tornò]**^{robot} a casa **[cenò]**^{robot}, poi **[giocò]**^{robot} con la playstation, poi **[andò]**^{robot} a letto.

CE1, 736

Un giorno **[il gatto]**^{cat} decise di fare una passeggiata e **[incontrò]**^{cat} **[un cane]**^{ext1} **[che]**^{ext1} era a spasso anche **[lui]**^{ext1}. **[Il cane]**^{ext1} voleva mangiar**[lo]**^{cat}, allora **[il gatto]**^{cat} iniziò a correre. Ma **[il cane]**^{ext1} **[lo]**^{cat} prese e **[lo]**^{cat} **[mangiò]**^{ext1}. **[Il cane]**^{ext1} ormai pieno tornò a casa, ma all' improvviso saltò fuori **[un robot]**^{robot}.

CE1, 739

C’era una volta in una fiera **[un gatto]**^{cat}. Ogni volta che **[si sedeva]**^{cat} **[diventava]**^{cat} **[un lupo terrificante]**^{cat}, **[rompeva]**^{cat} e **[tagliava]**^{cat} tutto quel che **[vedeva]**^{cat}. Un giorno **[il gatto]**^{cat} andò in una foresta e **[si esercitò]**^{cat} fino a 30 anni **[ritornò]**^{cat} **[era]**^{cat} alto ma alto ma più alto del normale e veloce come un cobra, **[mangiò]**^{cat} tutto il mondo.

CE1, 74

{[La strega]^{witch} vuole mangiare **[il gatto]**^{cat}}

Un giorno **[una strega]**^{witch} voleva mangiare **[un gatto]**^{cat} ma **[il felino]**^{cat} voleva scappare però **[non poteva scappare]**^{cat} però **[era furbo]**^{cat} allora **[voleva mangiare]**^{cat} le sbarre di cioccolato.

CE1, 75

{[La strega]^{witch} e **[lo stregone]**^{ext1} e **[il gatto]**^{cat}}

C’era una volta **[una strega]**^{witch} e **[un gatto]**^{cat}.

Un giorno **[la strega]**^{witch} e **[il gatto]**^{cat} andarono sulla scopa a fare un giro e a un tratto venne **[lo stregone]**^{ext1}, **[era arrabbiato]**^{ext1} molto e per dispetto **[trasformò]**^{ext1} **[il gatto]**^{cat} in **[topo]**^{cat}. **[La strega]**^{witch} **[che]**^{witch} era intelligente trasformò ancora **[il topo]**^{cat} in **[gatto]**^{cat}, e alla fine **[ritornarono]**^{cat,witch} a casa, **[la strega]**^{witch} vede che fuori è buio e quindi **[il gatto]**^{cat} e **[la strega]**^{witch} andarono a dormire. Ma **[lo stregone]**^{ext1} era ancora arrabbiato.

CE1, 78

{[La strega]witch e **[il gatto]**cat}

C’era una volta **[una strega]**witch e **[un gatto]**cat, **[erano]**cat,witch molto amici e **[si volevano bene]**cat,witch. Un giorno si ammalò **[il gatto]**cat e **[non vuole più]**cat volare, **[aspettava]**cat e **[aspettavano]**cat,witch ma non **[fu guarito]**cat finché **[guarì]**cat. **[Esclamò]**witch "**[è guarito]**cat, **[è guarito]**cat" e **[fanno]**cat,witch un altro volo e **[la strega]**witch era felice di volare ancora "e **[tu]**cat come stai **[Briciolina]**cat?" "bene" "**[sono contenta]**witch **[Briciolina]**cat". "**[Ritornati]**cat,witch a casa **[mangiamo]**cat,witch?" **[Briciolina]**cat rispose "sì" dopo **[gli]**witch **[dice]**cat "**[ci prepariamo]**cat,witch" **[rispose]**witch "sì" "**[andiamo]**cat,witch a letto" "sì" ma **[non volle]**cat dormire. **[Prese]**cat la scopa magica e **[volò via]**cat **[dalla strega]**witch e a un certo punto **[si perse]**cat e incominciò a piovere e **[la strega]**witch si svegliò e **[lo]**cat **[vede]**witch volare in cielo e **[lo]**cat **[rincorse]**witch ma **[non si fermò]**cat e a un certo punto **[lo]**cat **[prese]**witch e **[tornò]**witch a casa e vissero felici e contenti.

CE1, 8

C’era una volta **[un robot]**robot **[che]**robot era atterrato sulla terra per fare una missione.

[Un gattino]cat andò a vedere e **[si spaventò]**cat. **[Quel gattino]**cat vide il razzo **[del robot]**robot ed **[entrò]**cat, **[vide]**cat tantissimi computer spaziali e **[vide]**cat un motore elettrico poi a un certo punto **[il robot]**robot entrò **[il gatto]**cat si nascose e **[vide]**cat **[il robot]**robot **[che]**robot scriveva a una base spaziale. **[Il gatto]**cat corse a dirlo a tutti **[i [suoi]cat amici]**ext1. **[Andarono]**ext1 anche **[loro]**ext1 a vedere **[il robot]**robot e **[li]**ext1 **[colse]**robot di sorpresa. **[I gattini]**ext1 scapparono dopo un po' **[ritornarono]**ext1 e **[non c’era più]**robot.

CE1, 81

{Gli incantesimi di **[Mella]**witch}

C’era una volta **[una strega di nome Mella]**witch. Dovete sapere che **[era]**witch da sola e **[voleva]**witch tanto un amico. Un giorno per strada **[trovò]**witch **[un gatto di nome Luna]**cat e con un incantesimo che **[lo]**cat trasforma in **[un uomo]**cat **[che]**cat **[le]**witch fece compagnia.

CE1, 84

{[Il robot fortunato]robot}

Tanto tempo fa avevamo **[un robot]**robot **[che]**robot si chiamava **[robot - Andrea]**robot e **[usciva]**robot ogni giorno.

E **[voleva]**robot un gatto e un giorno fortunato **[trovò]**robot **[un gatto nero e bianco]**cat.

E ogni giorno **[giocava]**robot con **[lui]**cat. **[Andava]**robot al parco, giardino della robo - nonna.

Un giorno **[scappò]**cat e **[il robot - Andrea]**robot era triste e dopo due settimane **[lo]**cat **[ritrovò]**robot.

E **[lo]**[cat] **[chiamò]**[robot] **[coccolone]**[cat].

CE1, 91

{[La strega][witch] da brutta a bella}

C'era una volta **[una strega]**[witch], ed **[era]**[witch] così brutta che ogni volta che **[qualcuno]**[ext1] **[la]**[witch] vedeva **[scappava]**[ext1] a gambe levate.

Dopo un po' però **[era stufa]**[witch] e quindi **[preparò]**[witch] una pozione per diventare più bella. Quando **[ebbe finito]**[witch] **[si sciacquò]**[witch] le mani con quella pozione, **[si toccò]**[witch] la faccia e luuuum! **[Era diventata]**[witch] bellissima. Adesso nessuno **[l']**[witch]avrebbe presa in giro, **[era]**[witch] contentissima.

CE2, 1098

C'era una volta **[un mago]**[ext1] **[che]**[ext1] era **[il custode di una foresta incantata]**[ext1] e con **[il [suo]**[ext1] **unicorno]**[ext2], **[proteggeva]**[ext1] la foresta dalle streghe. Ma **[la più malvagia delle streghe]**[witch] **[che]**[witch] si chiamava **[Amelia]**[witch] con **[il [suo]**[witch] **corvo Gennaro]**[ext3], aveva lanciato un incantesimo alla foresta che faceva diventare **[tutti gli abitanti]**[ext4] di pietra compreso **[il mago]**[ext1]: **[che]**[ext1] era **[mago Merlino]**[ext1].

Ma **[mago Merlino]**[ext1] chiese aiuto **[agli abitanti non ancora pietrificati]**[ext5,ext6,ext7,ext8] di proteggere la foresta. Così **[unicorni]**[ext5], **[troll muschiosi]**[ext6], **[folletti]**[ext7] e **[gnomi]**[ext8] fecero un piano: **[i folletti]**[ext7] distraevano **[la strega]**[witch] mentre **[gli gnomi]**[ext8] portavano le pietre magiche al calderone per salvare la foresta e **[gli unicorni]**[ext5] aiutavano **[i troll muschiosi]**[ext6] a scavare le trappole e a ricoprirle di muschio. Alla fine **[salvarono]**[ext1,ext5,ext6,ext7,ext8] la foresta e vissero tutti felici e contenti!

CE2, 1099

C'era una volta **[una strega]**[witch]. Insieme **[al [suo]**[witch] **fratello]**[ext1] stavano andando a scuola. **[Loro due]**[ext1,witch] crescendo sempre di più diventarono ragazzi e **[decisero]**[ext1,witch] insieme di costruire **[un robot chiamato Roboti]**[robot]. **[I due]**[ext1,witch] diventati adulti decisero di andare in due case diverse dove **[ognuno]**[None,None] avrebbe continuato la **[sua]**[ext1,witch] passione. Poco dopo essersi trasferiti **[si accorsero]**[ext1,witch] che **[Roboti]**[robot] era **[uno]**[robot] e **[loro]**[ext1,witch] due così nacque il problema di chi se **[lo]**[robot] sarebbe portato a casa. **[Il fratello della [strega]**[witch]**]**[ext1] disse "**[io]**[ext1] sono **[uno scienziato]**[ext1] e quindi me **[lo]**[robot] **[ricostruirò]**[ext1] **[da solo]**[ext1], tieni**[te]**[witch]**[lo]**[robot] **[tu]**[witch] **[Roboti]**[robot]." Poco dopo **[il fratello della [strega]**[witch]**]**[ext1] se ne andò. **[La strega]**[witch] decise di buttare **[Roboti]**[robot] perché ora **[il [suo]**[witch] **fratello]**[ext1] non c'era più e quindi visto che **[lo]**[robot] **[avevano costruito]**[ext1,witch] insieme ora che **[lui]**[ext1] che **[lui]**[ext1] non c'era **[non era più]**[robot] un bel ricordo. Con il passare del tempo **[la strega]**[witch] decise di prendersi **[un gatto]**[cat] e [**il [suo]**[witch] **fratello]**[ext1] **[un cane]**[ext2]. Un giorno **[i due animali]**[cat,ext2] andarono **[uno]**[cat] nella casa della

[strega]witch e **[l'altro]**ext2 nell' altra casa con un biglietto nella bocca. Quando **[i due]**ext1,witch decisero di leggere il biglietto di scuse si commossero allora **[tutti e due]**ext1,witch iniziarono a costruirsi i regali a vicenda e **[i [loro]ext1,witch animali]**cat,ext2 li portarono da una casa all' altra. Poi visti tutti i regali **[decisero]**ext1,witch di chiamare **[i [loro]ext1,witch animali]**cat,ext2 con i nomi che **[si davano]**ext1,witch da piccoli, **[il gatto Felicetto]**cat perché quando **[la strega]**witch aveva 3 anni **[il [suo]witch fratellino]**ext1 9 e **[lei]**witch si confondeva con i nomi e alla fine è venuto fuori **[Felicetto]**cat poi **[il cane Violetta]**ext2 perché **[la strega]**witch si chiama **[Viola]**witch. **[I due fratelli]**ext1,witch decisero di incontrarsi al tramonto vicino a un lago. Lì **[fecero pace]**ext1,witch e **[costruirono]**ext1,witch **[un altro robot]**ext3, **[Feliola]**ext3, erano i loro nome mischiati, poi **[si erano trasferiti]**ext1,witch in una casa molto più grande. Lì **[continuarono]**ext1,witch le **[loro]**ext1,witch passioni e vissero felici e contenti.

CE2, 1100

{[Il topo]ext1 e **[il gatto]**cat }

C'era una volta **[un gatto di nome Bianchino]**cat **[che]**cat non mangiava da un anno. Allora **[vide]**cat **[un topo]**ext1 **[che]**ext1 si era addormentato e **[decise]**cat di preparare una trappola ma **[il topo]**ext1 si stava svegliando. Allora **[si svegliò]**ext1 e **[lo]**cat **[vide]**ext1 che **[stava preparando]**cat una trappola, allora **[si mise]**ext1 a correre. **[Il gatto]**cat vide che **[se ne stava andando]**ext1 allora **[si mise]**cat ad inseguir**[lo]**ext1 e alla fine **[il gatto]**cat a forza di correre iniziò a rallentare e **[non riuscì più]**cat a correre e **[il topo]**ext1 era così veloce che **[riusciva]**ext1 ancora a correre. **[Il topo]**ext1 riuscì a scappare e **[il gatto]**cat non riuscì a prender**[lo]**ext1. Il giorno dopo **[si incontrarono]**cat,ext1 e **[il gatto]**cat si avvicinò **[al topo]**ext1 perché **[gli]**ext1 voleva chiedere scusa per il giorno prima ma **[il topo]**ext1 aveva ancora paura **[del gatto]**cat, allora **[si nascose]**ext1 dietro un cespuglio e **[il gatto]**cat **[gli]**ext1 chiese se **[voleva diventare]**ext1 **[suo]**cat amico. **[Il topo]**ext1 uscì dal cespuglio e **[disse]**ext1 di sì. **[Diventarono]**cat,ext1 amici, **[visitarono]**cat,ext1 tutto il mondo insieme.

CE2, 1101

C'era una volta **[un ragazzo]**ext1 **[che]**ext1 aveva **[una sorellina]**ext2. **[Loro]**ext1,ext2 un giorno uscirono a pattinare sul ghiaccio. **[La sorellina]**ext2 non sapeva pattinare sul ghiaccio mentre **[il fratello più grande]**ext1 si. Un giorno **[lui]**ext1 cadde sul ghiaccio e il ghiaccio si ruppe e passarono molti anni ma **[lui]**ext1 non invecchiò e un bel giorno **[l'uomo sulla luna]**ext3 **[lo]**ext1 prese e **[lo]**ext1 **[portò]**ext3 in un posto speciale, e così **[il ragazzo]**ext1 creò una leggenda cioè **[aiutava]**ext1 **[Babbo Natale]**ext5, **[la Fatina dei dentini]**ext6, **[il Coniglio di pasqua]**ext7 e **[l'uomo che porta i bei sogni ai bambini]**ext8. Però c'era anche **[l'uomo nero]**ext4!!! **[Che]**ext4 **[portava]**ext4 ai bambini gli incubi e quindi ogni notte **[loro]**ext1,ext4 combattevano, ma **[l' uomo nero]**ext4 voleva sempre la vittoria. Quando arrivò **[il ragazzo]**ext1, **[Babbo Natale]**ext5 **[gli]**ext1 diede il compito di portare felicità ai bambini. E **[quel ragazzo]**ext1 si chiamò **[Black Jack]**ext1. **[Il robot]**robot era il gioco preferito dei bambini. E un bel giorno **[loro]**ext1 vinsero e da quel momento si vollero chiamare **[le 5 leggende]**ext1,ext5,ext6,ext7,ext8 e così vissero per sempre felici e contenti.

CE2, 1102

C’era una volta **[una strega]**witch **[che]**witch si sentiva sola e **[non aveva]**witch amici perché **[aveva cambiato]**witch casa. Lì dove abitava **[lei]**witch c’erano poche persone però un giorno mentre **[puliva]**witch in mansarda si sentiva miagolare. **[La strega incuriosita]**witch andò su convinta, **[aveva trovato]**witch **[un gattino]**cat **[che]**cat cercava da mangiare ed **[era assetato]**cat. **[Decise]**witch di tener**[se]**witch**[lo]**cat. **[Uscì]**witch per comprar**[gli]**cat qualcosa da mangiare e per comprare la lettiera, le ciotole per l' acqua e il cibo. Quando **[tornò]**witch a casa **[trovò]**witch **[il gatto]**cat **[che]**cat dormiva. **[Decise]**witch di lasciar**[lo]**cat stare. Quando **[si svegliò]**cat **[gli]**cat **[diede]**witch da mangiare, **[era]**cat magro. Dopo un paio di giorni **[si riprese]**cat. Adesso **[lei]**witch aveva **[un amico]**cat **[che]**cat **[le]**witch facesse compagnia. Da quel giorno **[non si sentiva]**witch più sola. **[Il gattino]**cat era proprio piccolo, **[lo]**cat **[chiamò]**witch **[Chicco]**cat. Anche **[Chicco]**cat si sentiva meglio da quando **[aveva]**cat **[una padrona]**witch. Da quel momento **[la strega]**witch e **[Chicco]**cat si sentivano meglio. **[Chicco]**cat si era trovato **[un nuovo giocattolo]**robot, era **[un robot]**robot. Giocavano sempre insieme. **[La strega]**witch era contenta di vedere **[Chicco]**cat divertirsi con **[il robot di nome Robin]**robot. **[Erano]**cat,robot talmente amici che **[non si volevano]**cat,robot lasciare. Un giorno **[Robin]**robot portò **[un topolino]**ext1 a **[Chicco]**cat, da una parte **[era contento]**cat dall'altra no, perché **[aveva paura]**cat che **[Robin]**robot **[lo]**cat lasciasse solo, **[Robin]**robot si avvicinò a **[Chicco]**cat e **[gli]**cat **[chiese]**robot cosa **[aveva]**cat. Allora **[Chicco]**cat si mise a raccontare. **[Robin]**robot **[gli]**cat disse che non sarebbe stato così, **[Chicco]**cat si sentiva felice. Vissero tutti felici e contenti. Fine

CE2, 1103

C’era una volta **[una strega]**witch **[che]**witch aveva **[un gatto]**cat **[che]**cat però a volte **[trattava]**witch male, tipo quando **[era arrabbiata]**witch magari non con **[lui]**cat ma **[con altre streghe]**ext1 **[gli]**cat **[diceva]**witch "**[vai]**cat fuori" oppure "**[stam[mi]witch]**cat lontano" e **[lui]**cat si sentiva un po' triste in questi casi, ma al contrario altre volte **[andavano]**cat,witch molto d'accordo, però era molto raro. Un giorno **[la strega]**witch era così arrabbiata con **[la sua amica strega]**ext1 che appena **[il gattino]**cat si avvicinò a **[lei]**witch **[la strega]**witch cominciò a urlare "**[vai]**cat via, **[sono già arrabbiata]**witch per conto **[mio]**witch, **[vai]**cat fuori" a quel punto **[il gattino]**cat andò fuori in giardino e **[si mise]**cat a piangere perché **[la strega]**witch era stata molto cattiva con **[lui]**cat. Prima **[il gattino]**cat ci pensò un po' su poi **[se ne andò]**cat. Dopo un po' **[la strega]**witch era pentita quindi **[uscì]**witch a cercar**[lo]**cat ma **[il gattino]**cat non c’era più, allora **[la strega]**witch si rannicchiò sul divano e **[si mise]**witch a piangere. Dopo una settimana **[il gattino]**cat tornò e **[la strega]**witch **[gli]**cat promise di non trattar**[lo]**cat più così, allora **[il gatto]**cat decise di restare e da lì **[il gatto]**cat e **[la strega]**witch andarono d'amore e d' accordo. Però arrivò un giorno che **[il povero gatto]**cat finì il **[suo]**cat ciclo vitale e **[la strega]**witch siccome **[non voleva lasciar[lo]cat]**witch **[lo]**cat **[mise]**witch in una bara e poi **[mise]**witch la bara in giardino in modo tale che potesse metter**[lo]**cat e toglier**[lo]**cat, come e quando **[voleva]**witch, così era come **[se fosse]**cat ancora **[vivo]**cat come prima. Un giorno passo di lì **[uno stregone]**ext1 che di lavoro faceva il prete e vedendo lì quella bara **[pensò]**ext1 di andare a vede come stava **[la strega]**witch visto che era **[una [sua]ext1 amica]**witch. **[La strega]**witch disse "tutto bene ma ora che **[il [mio]witch gatto]**cat è morto **[mi sento]**witch un po' sola" **[lo stregone]**ext1 **[le]**witch disse che **[poteva restare]**ext1 lì **[lui]**ext1 a

far**[le]**witch compagnia. **[La strega]**witch disse "va bene ma per quanto?" e **[lui]**ext1 risponde "anche tutta la vita se **[vuoi]**witch" da lì vissero tutti felici e contenti.

CE2, 1104

C'era una volta **[un uomo]**ext1 **[che]**ext1 aveva **[una moglie]**ext2 e **[un bambino di nome Giacomo e di 2 anni]**ext3. Un bel giorno di sole **[i genitori di [Giacomo]ext3]**ext1,ext2 decisero di portar**[lo]**ext3 allo zoo.

[Giacomo]ext3 ne fu molto felice perché **[aveva visto]**ext3 molti animali: i delfini, i pappagalli ecc. Giunse a un punto che **[i genitori]**ext1,ext2 si distrassero un secondo e **[il bambino]**ext3 venne morso da **[un lupo]**wolf con **[un padrone]**ext4. **[Il padrone]**ext4 si scusò subito e **[fece amicizia]**ext4 dopo che **[gli]**ext4 **[avevano detto]**ext1,ext2 che il giorno dopo era il **[suo]**ext3 compleanno. Insieme **[decisero]**ext1,ext2,ext4 di regalar**[gli]**ext3 **[un robot da 5.000€]**robot. **[Il bambino]**ext3 quando **[lo]**robot **[vide]**ext3 ne fu molto felice e **[ci]**robot **[giocò]**ext3 ogni ora.

Cresciuto **[giocò]**ext3 sempre di più.

CE2, 1106

C'era una volta **[un robot di nome Tomas]**robot **[che]**robot doveva andare a una festa dove c'erano **[tanti amici di [Tomas]robot]**ext1. **[Arrivato]**robot alla festa **[si mette]**robot a ballare ma **[non trova]**robot **[i [suoi]robot amici]**ext1 allora **[torna]**robot a casa scontento e **[si mette]**robot a piangere. Al momento appare **[un unicorno]**ext2 che era **[l'unicorno consola tutti]**ext2 e si chiamava **[Arcobaleno]**ext2 allora **[si mise]**ext2 a consolare **[Tomas]**robot. Allora **[l'unicorno]**ext2 dice a **[Tomas]**robot "**[prova]**robot ad andare lì con **[me]**ext2 e **[tutti]**ext1 **[li]**ext1 farò apparire" allora **[Tomas]**robot accetta. Allora **[arrivano]**ext2,robot lì e **[Arcobaleno]**ext2 dice "**[amici di [Tomas]robot]**ext1" e **[gli amici]**ext1 "bu bu buuuuu!" e **[Arcobaleno]**ext2 "**[eccoli]**ext1" e **[Tomas]**robot "ie ie ie!" Allora **[tutti]**ext1,robot si misero a ballare però **[Arcobaleno]**ext2 scompare e **[robot]**robot si sente triste però **[pensa]**robot che ci ha **[i [suoi]robot amici]**ext1 e allora **[è contento]**ext1 ma **[l'unicorno]**ext2 ancora di più perché **[sa]**ext2 che senza di **[me]**ext2 **[si diverte]**robot ancora di più. Allora **[robot]**robot si ricorda per tutta la vita di **[Arcobaleno]**ext2. Vissero così felici e contenti ma **[pensa]**robot che **[l'unicorno]**ext2 si sia impegnato tanto tanto. **[Robot]**robot e **[l'unicorno]**ext2 sono felicissimi.

CE2, 1107

Un giorno **[la mamma di [un cane]ext2]**ext1 muore e **[la sua nuova mamma]**cat è **[un gatto]**cat e **[il cane]**ext2 un giorno si perde, poi arriva **[un bambino di nome Lucas]**ext3 e **[porta]**ext3 **[il cane di nome Bella]**ext2 a casa di **[Lucas]**ext3. **[Lucas]**ext3 passa tanti anni con **[Bella]**ext2 finché **[si perde]**ext2. Con il tempo **[Bella]**ext2 si abitua a stare con **[un puma]**ext4. **[Bella]**ext2 ha cresciuto **[il puma]**ext4 finché **[si trova]**ext2 faccia a faccia con **[dei lupi]**wolf. **[I lupi]**wolf feriscono **[Bella]**ext2 finché arriva **[il puma]**ext4 **[che]**ext4 uccide **[i lupi]**wolf, poi **[il puma]**ext4 si prede cura di **[Bella]**ext2 fin quando **[si devono]**ext2,ext4

separare di nuovo perché **[Bella]**ext2 deve tornare da **[Lucas]**ext3 che ormai è da due anni che **[cerca]**ext2 di tornare a casa da **[Lucas]**ext3 ma alla fine **[perde]**ext2 il collare e nessun lo trova più. Poi passano ancora quattro anni poi **[Bella]**ext2 si perde nel bosco per un anno poi **[un signore di nome Antonio]**ext5 lo trova e **[chiama]**ext5 un po' di numeri di telefono finché **[trova]**ext5 quello di **[Lucas]**ext3, poi **[Lucas]**ext3 va nel bosco. Quando **[trova]**ext3 **[Bella]**ext2 **[gli]**ext3 corre incontro e **[vede]**ext3 che **[è ferita]**ext2 e **[Bella]**ext2 e **[il puma]**ext4 **[che]**ext2,ext4 erano migliori amici **[si devono separare]**ext2,ext4. **[Il puma]**ext4 va verso il bosco e **[Bella]**ext2 va verso **[Lucas]**ext3 poi **[Lucas]**ext3 porta **[Bella]**ext2 dal veterinario perché **[vede]**ext3 che **[è ferita]**ext2 poi quando **[trova]**ext3 la macchina vanno a casa di **[Lucas]**ext3, **[trova]**ext2 **[mamma gatta]**cat e **[si ricorda]**ext2 di **[lei]**cat che **[l']**ext2allattava e **[piange]**ext2 dall' emozione.

CE2, 1108

C'era una volta **[una strega di nome Orazia]**witch, **[[suo]witch marito Lupo Lucio]**wolf e **[i due figli di nome Strega Dina e Lupo Nello]**ext1,ext2. Come scuola **[i bambini]**ext1,ext2 frequentano quella di magie. **[I bambini]**ext1,ext2 andando a scuola si fecero degli amici. Allora **[i bambini]**ext1,ext2 chiesero **[alla mamma]**witch, "**[possiamo]**ext1,ext2 invitare **[la strega Lara]**ext3 e **[lo stregone Ronny]**ext4 ad uno stregalupo party?" "Si certo!" "Ok grazie **[mamma]**witch! Allora lo chiediamo **[a loro]**ext3,ext4." Il giorno dopo **[vanno]**ext1,ext2 a scuola e **[chiedono]**ext1,ext2 ai **[loro]**ext1,ext2 **[due amici]**ext3,ext4: "**[volete venire]**ext3,ext4 al **[nostro]**ext1,ext2 stregalupo party?" "Ok va bene!" Però intanto **[li]**ext1,ext2 aveva sentiti anche **[Gatto Matto]**cat, quindi **[si offese]**cat dicendo: "perché a **[me]**cat **[non [mi]cat invitate]**ext1?" **[I due fratelli]**ext1,ext2 risposero: "**[tu]**cat non ci stai troppo simpatico", quindi da lì iniziò il litigio: "**[siete]**ext1,ext2 cattivi" disse **[Gatto matto]**cat, "**[smettila]**cat, **[noi]**ext1,ext2 non siamo obbligati" dissero **[i due fratelli]**ext1,ext2. Dopo arrivò **[la mamma dei [due fratelli]ext1,ext2]**witch e **[disse]**witch: "cosa sta succedendo?" "**[Mamma]**witch, non **[la pianta]**cat!" "**[Bambini]**ext1,ext2 ci sta anche **[lui]**cat!" "Uffi! Vabbè ok allora **[invitiamo]**ext1,ext2 anche **[lui]**cat!" **[Tutti]**cat,ext1,ext2,ext3,ext4 contenti allora **[andarono]**cat,ext1,ext2,ext3,ext4 al party stregalupo. E vissero tutti felici e contenti.

CE2, 1427

{[Il gatto]cat e **[la strega birichina]**witch}

C'era una volta **[una strega]**witch. Un giorno **[andò]**witch nel negozio degli animali nella città di Stregandia, e **[comprò]**witch **[un gatto]**cat.

Quando **[lo]**cat **[portò]**witch a casa **[gli]**cat **[diede]**witch da mangiare. Dopo un po' **[la strega]**witch fece una pozione con: zampe di ragno, cibo di serpente e infine occhi di zombie. Quella pozione serviva per far parlare **[il gatto]**cat!

[Il gatto]cat visto che **[non sapeva]**cat che era per **[lui]**cat, la buttò giù e si allagò tutta la casa. **[La strega]**witch preparò un'altra pozione con: cervello di zombie, lumache bavose e zampe di gallina. Infine **[glie]**catla **[diede]**witch e vissero tutti felici e contenti.

CE2, 1428

{[Il gatto furbino]cat e **[il lupo birichino]**wolf}

C’era una volta **[un gatto]**cat **[che]**cat amava andare in piscina, dopo un po' arrivò **[il lupo dispettoso]**wolf e **[disse]**wolf **[al gatto]**cat: - Ei **[tu]**cat togliti, voglio andarci **[io]**wolf in piscina!- **[Il gatto furbo]**cat allora andò sott' acqua e **[tolse]**cat il tappo e allora **[tornò]**cat su e **[disse]**cat **[al lupo]**wolf: - Prego **[accomodati]**wolf! - **[Il lupo]**wolf allora andò in piscina senza sapere che non c’era il tappo. Quindi **[il lupo]**wolf entrò e **[venne risucchiato]**wolf. Due ore dopo **[il lupo]**wolf uscì da un tombino e **[il gatto]**cat scoppiò dalle risate! **[Il lupo]**wolf allora disse: - Non è divertente, se **[ti]**cat **[acchiappo]**wolf non sai che cosa **[ti]**cat **[farò]**wolf! - **[Il gatto]**cat allora scappò ma prima che **[il lupo]**wolf uscisse dal tombino **[il gatto]**cat fece uno scherzetto **[al lupo]**wolf, quindi **[fece]**cat il modo di far**[lo]**wolf incastrare. **[Il lupo]**wolf provò a liberarsi ma **[non ci riuscì]**wolf a liberarsi.

[Il gatto]cat ridendo **[gli]**wolf disse: -**[Qualcuno]**wolf mi sa che è un po' grassottello! -

[Il lupo]wolf dopo un po' si liberò e **[cominciò]**wolf a rincorrere **[il gatto]**cat. **[Il gatto]**cat scappò a gambe levate, ma per fortuna **[il gatto]**cat aveva un nascondiglio segreto e **[fece]**cat una statua uguale **[a lui]**cat e **[gliela]**wolf **[diede]**cat e **[il lupo]**wolf allora la prese e **[se la portò]**wolf nel bosco, **[se la mangiò]**wolf e **[si ruppe]**wolf tutti i denti e **[il gatto]**cat rise e visse felice e contento.

CE2, 1429

{[La strega]witch e **[il gatto]**cat}

C’era una volta **[una strega molto bella]**witch, **[questa strega]**witch si chiamava **[Perlacchia]**witch. **[La strega Perlacchia]**witch era molto simpatica e vivace, una volta **[stava saltellando]**witch in un bosco e **[aveva visto]**witch **[un gattino molto carino]**cat. **[Il gattino]**cat si chiamava **[Furia]**cat. **[Furia]**cat era tutto ferito. Visto che **[la strega Perlacchia]**witch era anche molto generosa **[lo]**cat **[portò]**witch nella torre con **[lei]**witch. Una volta **[la strega Perlacchia]**witch aveva saputo che **[il padre]**ext1 e **[la madre]**ext2 di **[due buoni e birichini bambini]**ext3,ext4 si erano arrabbiati quindi di sera **[lei]**witch e **[Furia]**cat andarono **[dai bambini]**ext3,ext4 che erano **[un maschio]**ext3 e **[una femmina]**ext4, **[il maschio quello più piccolo]**ext3 **[che]**ext3 si chiamava **[Diego]**ext3 di 14 anni e **[la femmina di 20]**ext4 **[che]**ext4 si chiamava **[Samantha]**ext4. **[Loro]**ext3,ext4 volevano mangiare, ma **[i genitori]**ext1,ext2 non volevano che **[loro]**ext3,ext4 mangiavano e quindi **[se ne stavano]**ext3,ext4 a letto con i cellulari poi **[Samantha]**ext4 aprì la porta perché **[voleva fare]**ext4 i compiti anche se erano le 3:00 di notte. Dopo un po' scese giù anche **[Diego]**ext3 e **[videro]**ext3,ext4 **[la strega Perlacchia]**witch e anche **[il gattino Furia]**cat. **[La strega Perlacchia]**witch **[ai bambini]**ext3,ext4 **[gli]**ext3,ext4 disse: - **[Non avete]**ext3,ext4 paura **[io]**witch sono buona, è che **[avevo visto]**witch **[questo gattino]**cat pero ' **[io]**witch ieri mi sono sposata con **[uno stregone]**ext5 **[che]**ext5 si chiama **[Christian]**ext5 e **[avevamo fatto]**ext5,witch anche **[un figlio]**ext6 cioé **[io]**witch **[lo]**ext6 avevo fatto e con **[[mio]witch figlio]**ext6 **[non posso badare]**witch **[al gattino]**witch e quindi **[avevo]**witch in mente di dar**[lo]**cat a voi. **[Io]**witch **[l']**catavevo chiamato **[Furia]**cat ma adesso tocca a **[voi]**ext3,ext4 decidere il nome. **[Diego]**ext3 e **[Samantha]**ext4 prendete**[ne]**cat cura **[io]**witch adesso devo andare ciao **[ci rivedremo]**ext3,ext4,witch presto e speriamo che **[vivrete]**ext3,ext4 felici e contenti. Buona notte e **[non dimenticatevi]**ext3,ext4 di far**[lo]**cat guarire. -

CE2, 1430

{[Il gatto][cat] e **[il lupo arrabbiato]**[wolf]}

C’era una volta **[un gatto]**[cat] **[che]**[cat] stava passeggiando nel bosco. All'improvviso **[vide]**[cat] un cespuglio con fragole grandi e piccole. **[Il gatto]**[cat] corse velocissimo per prendere quella fragola gigante e poi... **[Un lupo]**[wolf] gliela rubò. - **[Quel lupo]**[wolf] mi ha rubato la **[mia]**[cat] fragola che mi piaceva tanto. - disse **[il gatto]**[cat]. **[Il lupo]**[wolf] si mangiò tutta la fragola in un boccone. - **[Signor lupo]**[wolf] quella era la **[mia]**[cat] fragola! - disse **[il gatto]**[cat]. - **[Piccolo gatto]**[cat] **[mi]**[wolf] stai prendendo in giro. Questa è la **[mia]**[wolf] fragola non la **[tua]**[cat]! - rispose **[il lupo]**[wolf]. **[Il gatto]**[cat] andò a casa triste e lo **[disse]**[cat] a **[[sua]**[**cat**] **madre]**[ext1]. **[La mamma [del gatto]**[**cat**]**]**[ext1] dopo aver**[glie]**[ext1lo] detto si infuriò e **[andò]**[ext1] **[dal lupo]**[wolf]. - Ascolta **[signor lupo]**[wolf] ha preso la fragola a **[[mio]**[**ext1**] **figlio]**[cat] - disse **[la mamma [del gatto]**[**cat**]**]**[ext1] - Non **[mi]**[wolf] fai paura **[signora]**[ext1] **[io]**[wolf] sono più forte di **[lei]**[ext1] - disse **[il lupo]**[wolf] - Allora perché **[ho]**[ext1] una paletta in mano? **[Ti]**[wolf] **[devo fare male]**[ext1]! - disse **[la mamma infuriata]**[ext1] - Aiuto **[sono buono]**[wolf]! - disse **[il lupo]**[wolf].

CE2, 1432

{[Il gatto][cat] e **[il lupo sordo]**[wolf]}

C’era una volta **[un gatto]**[cat] e **[un lupo]**[wolf]. Tutto a un tratto **[il gatto]**[cat] urlò per sbaglio dentro l'orecchio **[del lupo]**[wolf] e **[diventò]**[wolf] sordo poi c’era **[una strega malvagia]**[witch] **[che]**[witch] per pranzo mangiava i gatti ma grazie all'urlo **[del gatto]**[cat] **[la strega]**[witch] **[lo]**[cat] sente! E quindi **[la strega]**[witch] arco **[il gatto]**[cat] da mangiare ma **[il gatto]**[cat] era molto furbo e **[il lupo poverino]**[wolf] non sentiva niente! E **[si chiedeva]**[wolf] "cosa **[mi succede]**[wolf] perché **[non sento]**[wolf] niente? Forse nessuno parla chi sa? Forse sono **[io]**[wolf] sordo? Beh **[spero]**[wolf] di no." **[Il gatto preoccupato]**[cat] per **[il lupo]**[wolf] cercò di curar**[lo]**[wolf] ma a tutto un tratto uscì fuori dal bosco **[la strega]**[witch] **[che]**[witch] voleva prendere **[il gatto]**[cat] con la **[sua]**[witch] brutta mira da vecchia perse **[il lupo]**[wolf] ma **[il lupo]**[wolf] non poteva parlare. Non solo **[non sentiva]**[wolf] ma **[non poteva neanche parlare]**[wolf] e **[la gatta]**[cat] rideva perché **[la strega]**[witch] continuava a sbagliare la mira. Ma **[la strega]**[witch] aveva sbagliato magia e la **[buttò]**[witch] contro **[il lupo]**[wolf] e quando gliel'ha lanciata **[lui]**[wolf] sentiva e **[riusciva]**[wolf] a parlare e alla fine **[il lupo]**[wolf] si vendicò e **[costringe]**[wolf] a vendicare pure **[il gatto]**[cat]. Ma **[il gatto]**[cat] era d'accordo di vendicarsi e insieme **[hanno ucciso]**[cat,wolf] **[la strega cattiva]**[witch] e vissero quasi tutti felici e contenti **[pensavano]**[cat,wolf] ma dopo 2 anni **[la strega]**[witch] risorge ma buona e con la **[sua]**[witch] magia crea **[un robot]**[robot] **[che]**[robot] proteggeva e **[cura]**[robot] **[il lupo]**[wolf] si chiamava **[K.2.O]**[robot.] e vissero ora tutti felici e contenti.

CE2, 1433

{[La strega][witch] e **[il gatto dispettoso]**[cat]}

C’era una volta **[una strega]**witch **[che]**witch era sempre felice. Un giorno **[era andata]**witch a fare una passeggiata e **[vide]**witch **[un gatto ferito]**cat. **[La strega]**witch pensò « **[Povero gattino]**cat! **[Lo]**cat **[porterò]**witch con **[me]**witch a casa, così **[potrò curar[lo]cat]**witch! » **[La strega]**witch **[lo]**cat portò a casa **[il gatto]**cat e con la **[sua]**witch magia **[lo]**cat **[curò]**witch in un battibaleno. **[La strega]**witch stava preparando la cena quando ad un certo punto, **[sentì]**witch un rumore. « Badaboom! » **[La strega]**witch disse « Che cosa è successo! » **[Si girò]**witch e **[vide]**witch la **[sua]**witch candela preferita per terra. « Chi è stato! » **[disse]**witch. **[Il gatto]**cat con gli occhi indicò il **[suo]**cat muso, **[la strega]**witch non capì e **[disse]**witch « Sei stato **[tu]**cat? » **[Il gatto]**cat disse sì con la testa. **[La strega]**witch al posto di essere arrabbiata era felice e **[disse]**witch « Non fa niente! **[Ti]**cat **[perdono]**witch, poi quella candela ormai valeva solo e 1,00! » Dopo un'ora **[sentì]**witch un altro rumore un altro e poi un altro ancora ma **[la strega]**witch era sempre felice! Ma **[il gatto]**cat continuò sempre ad essere dispettoso finché **[la strega]**witch non si arrabbiò per la prima volta nella **[sua]**witch vita piena di vasi rotti, casa infuocata e candele per terra!

CE2, 1434

C’era una volta **[un gatto]**cat con **[la [sua]cat padrona]**ext1 **[che]**ext1 si chiamava **[Francesca]**ext1. **[Il gatto]**cat uscì di casa, **[andò]**cat nel bosco e **[si sentiva ormai spacciato]**cat ma **[lui]**cat andò avanti pensando che qualcuno **[lo]**cat possa aiutare. **[Vide]**cat **[un lupo tutto affamato]**wolf **[che]**wolf **[lo]**cat rincorse e allora **[lui]**cat scappò ma **[il lupo]**wolf fece finta di ferirsi, **[il gatto]**cat **[lo]**wolf aiutò e **[il lupo]**wolf per poco non **[lo]**cat prese e **[il gatto]**cat scappò, **[si arrampicò]**cat su un albero. **[Il lupo]**wolf andò via ma era una finta, **[il gatto]**cat scese e **[il lupo]**wolf subito **[lo]**cat rincorse. **[Il gatto]**cat ritrovò le **[sue]**cat impronte dell’andata e **[ritornò]**cat in città e subito **[il cacciatore]**ext2 sparò **[al lupo]**wolf e **[il gatto]**cat tornò sano e salvo a casa.

CE2, 1436

C’era una volta **[un robot]**robot **[che]**robot camminò verso il bosco poi tra i cespugli **[il lupo]**wolf vede **[il robot]**robot poi **[il robot]**robot sentì un rumore au au poi **[il lupo]**wolf si fece vedere e **[dice]**wolf "ha ha ha adesso **[tu]**robot vieni con **[me]**wolf." "No" "perchè" "no perchè **[tu]**wolf sei molto cattivo" "non è vero" "sì che è vero" "**[io]**wolf **[ti]**robot voglio invitare a mangiare la pizza" "non è vero" parla **[il lupo]**wolf "**[stai]**robot zitto di no **[io]**wolf sai che **[ti]**robot faccio " "cosa" "**[ti]**robot **[spengo]**wolf" "no **[non farlo]**wolf per favore **[ti]**wolf **[supplico]**robot per favore" e **[lo]**robot **[spegne]**wolf e si sentì il rumore quando **[si spegne]**robot sp sp sp **[il lupo]**wolf riparla "ha ha ha" **[il robot]**robot si riaccese e **[disse]**robot "**[lupo]**wolf smettila di trattare male le persone **[tu]**wolf vuoi essere un cattivo si o no" "no va bene **[non faro più]**wolf del male **[te]**robot lo **[supplico]**wolf" "va bene **[siamo amici]**robot,wolf" "ochei **[andiamo]**robot,wolf nel bosco" e vissero tutti felici e contenti. Fine

CE2, 1437

{[Il robot]robot e **[la strega]**witch}

C’era una volta **[una strega]**witch **[che]**witch faceva il compleanno, ma **[pensava]**witch che **[il [suo]witch amico robot]**robot l'avesse dimenticato. **[Il robot]**robot invece non se l'è dimenticato, **[voleva]**robot fare una sorpresa **[alla [sua]robot amica]**witch allora **[si mise]**robot d' accordo con **[la famiglia della [sua]witch amica]**ext1,ext2,ext3,ext4 e allora **[la mamma della [strega]witch]**ext2 "si va bene allora" dice **[la mamma]**ext2. - Dico **[io]**ext2 penso che la torta la farà **[la sorella maggiore]**ext3 e i preparativi li farà **[il papà]**ext4 invece **[io]**ext2 farò la cena e **[tu]**robot **[robot]**robot prendi i palloncini neri. Allora a lavoro bene **[abbiamo]**ext2,ext3,ext4,robot tutto **[vai]**robot a prendere **[la strega]**witch... - **[La strega]**witch era triste allora **[il robot]**robot dice - **[Vieni]**witch c' è una sorpresa per **[te]**witch. - **[La strega]**witch e **[il robot]**robot andarono a casa **[della strega]**witch. **[Lei]**witch dice **[al robot]**robot - Allora **[non ti sei dimenticato]**robot del **[mio]**witch compleanno. - **[Il robot]**robot dice - Adesso **[andiamo]**robot,witch c' è una festa per **[te]**witch dai **[andiamo]**robot,witch a giocare. -

fine

CE2, 1438

{[La strega]witch e **[il lupo]**wolf**}**

C’era una volta **[una strega molto ma molto strana]**witch, non è **[una strega come le altre cattiva]**witch, ma **[era]**witch buona e generosa.

Un giorno **[incontrò]**witch **[un lupo]**wolf **[che]**wolf era in una fattoria, ma **[il contadino della fattoria]**ext1 era andato in città a fare la spesa quindi **[il lupo]**wolf si stava avvicinando a mangiare le bestie, **[la strega]**witch era un po' incapace a fare gli incantesimi e quindi con la **[sua]**witch scopa **[arrivò]**witch prima **[del lupo]**wolf. **[Il lupo]**wolf **[che]**wolf aveva fame corse all' entrata sul retro ed **[entrò]**wolf ma **[la strega]**witch provò a fare qualche incantesimo, il primo incantesimo fu far diventare **[il lupo]**wolf una gallina, poi **[riprovò]**witch e **[il lupo]**wolf divenne **[un piccione]**wolf e quindi **[scappò]**wolf. **[La strega]**witch aveva un'intenzione che era di far diventare **[il lupo]**wolf da cattivo a buono. Ma era un problema, **[la strega]**witch non sapeva fare incantesimi!

Allora la sera **[lesse]**witch tutti i libri di magia e il giorno dopo a cavallo della **[sua]**witch scopa **[trovò]**witch **[il lupo in forma di piccione]**wolf, **[la strega]**witch volò subito da **[lui]**wolf e **[lo]**wolf **[fece diventare]**witch **[un lupo buono]**wolf. E da quel punto **[fu]**wolf sempre buono, e **[divenne]**wolf vegetariano. E vissero tutti felici e contenti.

THE END

CE2, 1440

C’era una volta **[un robot]**robot e **[il [suo]robot amico gatto di nome Caddy]**cat. Un giorno **[stavano giocando]**cat,robot in giardino a palla. Dopo aver giocato, **[fecero]**cat,robot merenda. **[Il gatto]**cat mangiò un pesce e **[il robot]**robot una fetta di torta alle fragole.

Dopo la merenda **[tornarono]**cat,robot ognuno a casa **[sua]**cat,robot e **[si misero]**cat,robot nel letto a dormire.

Il giorno dopo una volta svegli **[si incontrarono]**[cat,robot] di nuovo e **[andarono]**[cat,robot] a giocare a calcio e dopo anche a ping pong. **[Entrambi]**[cat,robot] sono bravissimi in tutti gli sport. Mentre **[il gatto]**[cat] e **[il robot]**[robot] erano al parco, all 'improvviso arrivò **[un lupo feroce]**[wolf] **[che]**[wolf] voleva mangiarse**[li]**[cat,robot].

[Il gatto][cat] e **[il robot]**[robot] erano molto coraggiosi e **[scapparono]**[cat,robot], e da lontano **[spararono]**[cat,robot] **[al lupo]**[wolf], **[che]**[wolf] morì. Da quel giorno **[rimasero]**[cat,robot] amici per sempre.

CE2, 1441

{[Il gatto][cat] e **[il lupo cattivo]**[wolf]}

C'era una volta **[un gatto]**[cat] **[che]**[cat] si divertiva con **[i [suoi]**[cat] **amici]**[cat1] a giocare con la palla, a un certo punto arrivò **[un lupo cattivo]**[wolf] **[che]**[wolf] **[gli]**[cat,cat1] prese la palla e **[disse]**[wolf]: " - questo giardino è **[mio]**[wolf], **[andate]**[cat,cat1] via!" **[Il gatto]**[cat] e **[i [suoi]**[cat] **amici]**[cat1] scapparono via. Un giorno dopo **[il gatto]**[cat] e **[i [suoi]**[cat] **amici]**[cat1] ritornarono in giardino a giocare, arrivò ancora **[il lupo]**[wolf] ed **[era triste]**[wolf] perché **[non aveva]**[wolf] nessun amico, allora **[il gatto]**[cat] e **[i [suoi]**[cat] **amici]**[cat1] **[lo]**[wolf] chiamarono a giocare tutti insieme. Arrivarono **[un branco di lupi feroci]**[wolf1], ed erano **[i fratelli [del lupo]**[wolf]**]**[wolf1] e **[dissero]**[wolf1]: " - **[prendiamo]**[wolf1] **[questi gatti]**[cat,cat1]", **[il lupo]**[wolf] disse: "- no, sono **[i [miei]**[wolf] **amici]**[cat,cat1]!" E **[i [suoi]**[wolf] **fratelli]**[wolf1] dissero **[ai gatti]**[cat,cat1]: "- **[possiamo]**[wolf1] giocare anche **[noi]**[wolf1]?" **[I gatti]**[cat,cat1] dissero: "- certo!" E così diventarono **[tutti]**[cat,cat1,wolf,wolf1] amici.

CE2, 1442

{[La strega maldestra][witch] e **[il lupo]**[wolf]}

C'era una volta **[una strega]**[witch] **[che]**[witch] voleva fare delle pozioni ma neanche **[riusciva]**[witch] a fare 1 pozione. C'era **[il lupo]**[wolf] **[che]**[wolf] si vantava sempre diceva "**[io]**[wolf] sono il più bello, **[sono]**[wolf] il più bravo" con il **[suo]**[wolf] skateboard volante. C'era in gioco una gara, chi arrivava primo era il più veloce, il secondo il più intelligente e il terzo era il più lento. **[La strega]**[witch] **[che]**[witch] era più furba **[del lupo]**[wolf] inventò una pozione che **[la]**[witch] faceva far diventare più veloce. Dopo un giorno siamo alla gara, ci sono **[la strega]**[witch], **[il lupo]**[wolf] e **[il maiale]**[ext1] e c'erano tre gare, la prima la gara di oro, la gara di argento e la gara di bronzo, poi quella di oro la prima era tarocca. C'è anche **[il direttore]**[ext2] **[che]**[ext2] parlava di tutte le gare al **[suo]**[ext2] tre via la prima gara inizierà e così fu per tutte le altre. La prima inizia, **[il lupo]**[wolf] è in vantaggio ma poi con una velocità massima **[la strega]**[witch] arriva in vantaggio e per non far vincere **[il lupo]**[wolf] per sbaglio **[la strega]**[witch] lanciò una puzzetta che puzzava molto! **[Il lupo]**[wolf] **[che]**[wolf] sentiva qualcosa cadde per terra e quindi **[era eliminato]**[wolf] ma c'era ancora **[il maiale]**[ext1] **[che]**[ext1] non aveva nessun mezzo di trasporto e quindi **[rallentava]**[ext1] e **[il maiale]**[ext1] correva e **[aveva vinto]**[ext1]. Siamo alla seconda gara e si parte, **[il maiale]**[ext1] è primo e **[il lupo]**[wolf] è secondo, manca davvero poco e **[il maiale]**[ext1] è ancora primo ma **[il lupo]**[wolf] ha pareggio con **[il maiale]**[ext1] ma la seconda partita finisce in pareggio. E manca ancora l'ultima gara che vale la coppa delle grandi orecchie. Si parte e **[il lupo]**[wolf] è primo e manca

davvero poco, **[il maiale]**[ext1] è ancora primo e **[il maiale]**[ext1] vince! E **[il maiale]**[ext1] vince ancora la coppa delle grandi orecchie e vissero tutti felici tranne **[il lupo]**[wolf] e **[la strega]**[witch].

THE END

CE2, 1444

{[La strega pasticciona][witch] e **[il robot]**[robot]}

Un giorno **[la strega pasticciona]**[witch] voleva costruire **[un robot]**[robot] cercando delle ricette: - Vel poziones, vel robotos, fungus elettronicus **[fammi]**[witch] uscire un robotus! - **[esclamò]**[witch] cercando sempre delle ricette sul sito robotuselettronico.it. Ne **[cercava]**[witch] tantissime ma niente da fare; nel frattempo arrivarono **[i bambini]**[ext1] e dissero: - Ciao **[strega]**[witch], **[abbiamo portato]**[ext1] le mele elettronicus - **[la strega]**[witch] rispose: - Che? Le mele ma chi **[ve]**[ext1] l' ha detto? **[Fatemi]**[witch] pensare, ecco cosa **[mi]**[witch] mancava le mele! - rispose **[la strega]**[witch] e **[iniziò]**[witch] di nuovo: - Vel poziones vel robotus melus elettronicus fammi uscire un robotus - ma niente da fare ancora: - Ma come è possibile sono ricette tarocche! - " Intanto **[i bambini dispiaciuti]**[ext1] se ne andavano nel negozio giocattoli. C'erano tante cose, tutto tranne **[il robot]**[robot]. **[Chiesero]**[ext1] e per la grazia del cielo **[uno]**[robot] era rimasto, però **[costava]**[robot] 100.000 lire **[avevano]**[ext1] solo 99.999 lire ma ne **[trovarono]**[ext1] una lira a terra, per poco cadeva nel tombino. Ritornando **[alla strega disperata]**[witch] **[piangeva]**[witch] ogni settimana ma un giorno arrivò un pacco **[dai bambini]**[ext1], era **[il robot]**[robot] finalmente, **[il robot piccolino]**[robot]. Una settimana dopo **[il robot]**[robot] cadde nell' acqua. **[La strega ancora più disperata]**[witch] portò **[il robot]**[robot] **[dall' elettricista robottista]**[ext2] ma **[rimase]**[robot] sempre lo stesso. **[Il robot]**[robot] si riprese: - ciao **[strega]**[witch], **[io]**[robot] sto bene! - **[La strega]**[witch] disse: - oh mio dio finalmente ora **[posso fare]**[witch] i piatti più buoni del mondo! –

THE END

CE2, 1515

{[Il gatto][cat] e **[il robot]**[robot]}

C'era una volta **[un gatto]**[cat] **[che]**[cat] era solo a giocare, **[passeggiava]**[cat], e **[passeggiava]**[cat] ma **[non trovò]**[cat] nessun amico.

Ma un giorno passeggiando sulla collina **[vede]**[cat] **[un robot]**[robot] **[che]**[robot] si divertiva a correre su e giù dalla collina e allora **[il gatto]**[cat] **[gli]**[robot] disse: " - Ei **[tu]**[robot] vuoi fare amicizia con **[me]**[cat]? "

[Il robot][robot] **[gli]**[cat] rispose di sì.

Allora **[il gatto]**[cat] **[gli]**[robot] disse **[al robot]**[robot] "come **[ti chiami]**[robot]?" E **[il robot]**[robot] gli rispose "**[io]**[robot] mi chiamo **[Max]**[robot]."

Dopo **[il robot Max]**[robot] **[gli]**[cat] disse "invece come ti chiami **[tu]**[cat]?"

[Il gatto][cat] **[gli]**[robot] rispose "**[io]**[cat] mi chiamo **[Luna]**[cat] e sono **[una gatta]**[cat] sempre sola a passeggiare e **[cerco]**[cat] sempre di trovare almeno un amico."

[Il robot][robot] **[gli]**[cat] disse "ci sono **[io]**[robot] adesso sarò il **[tuo]**[cat] migliore amico."

"Adesso **[devo andare]**[robot] ciao **[Luna]**[cat]." "Ciao **[Max]**[robot] alla prossima".

CE2, 1516

C'era una volta **[una strega]**[witch]. **[Era]**[witch] tutta sola e **[voleva]**[witch] un animaletto. Un giorno **[ha visto]**[witch] **[un gatto nero come il carbone]**[cat] allora dice **[la strega]**[witch] " vuoi essere **[il [mio]**[**witch**] **gatto]**[cat]? " " Ma certo " e passarono il tempo insieme e sono tutti felici e contenti.

CE2, 1517

Titolo {**[il brutto gatto]**[cat]}

C'era una volta **[un gatto]**[cat] **[che]**[cat] si chiamava **[Sprin]**[cat]. **[Aveva]**[cat] una famiglia però un giorno **[Sprin]**[cat] e **[la [sua]**[**cat**] **padrona]**[ext1] erano andati a fare una passeggiata però **[Sprin]**[cat] dopo un po' non trova **[la [sua]**[**cat**] **padrona]**[ext1] e allora **[s'era messo]**[cat] a camminare e **[trovò]**[cat] **[un robot in mille pezzi]**[robot] e allora **[lo]**[robot] **[montò]**[cat] e **[lo]**[robot] **[chiamò]**[cat] **[Igienico]**[robot]. **[Sprin]**[cat] e **[Igienico]**[robot] diventarono migliori amici e un giorno **[Sprin]**[cat] e **[Igienico]**[robot] erano andati a fare una passeggiata e **[s'erano persi]**[cat,robot] però **[Sprin]**[cat] aveva incontrato **[un cane]**[ext2] **[che]**[ext2] **[lo]**[cat] stava per mordere però **[Igienico]**[robot] con il suo GPS salvò **[Sprin]**[cat] e **[s'erano incontrati]**[cat,robot] e erano felici e contenti.

Fine.

CE2, 1653

C'era una volta **[un bambino]**[ext1] **[che]**[ext1] aveva **[un gatto]**[cat] **[che]**[cat] si chiamava **[Tommy]**[cat]. Un giorno **[Tommy]**[cat] scappò, ad un certo punto da lontano **[intravide]**[cat] un palazzo. **[Lui]**[cat] si avvicinò, **[entrò]**[cat] e **[vide]**[cat] **[un strega di nome Buonafaccia]**[witch]. **[Buonafaccia]**[witch] **[gli]**[cat] disse: - **[vieni]**[cat!] - **[Tommy]**[cat] **[gli]**[witch] andò in braccio e **[la strega]**[witch] **[lo]**[cat] portò in una sala immensa tutta vuota. **[Buonafaccia]**[witch] **[gli]**[cat] disse: - è qui che **[vivrai]**[cat] però visto che la stanza è vuota **[dobbiamo andare]**[cat,witch] a fare shopping! - **[Buonafaccia]**[witch] però non aveva uno spicciolo allora **[andò]**[witch] al pentolone, **[prese]**[witch] della calze bucate, le **[immerse]**[witch] e si trasformarono in banconote. **[Tommy]**[cat] e **[Buonafaccia]**[witch] ai negozi presero: una lettiera, 1 piumino, 2 coperte, 5 quadri, 1 cuscino, 10 kg di mangime per i gatti e 1 bistecca. **[Tommy]**[cat] e **[Buonafaccia]**[witch] allestirono la sala vuota e **[divennero]**[cat,witch] amici. Il titolo della storia è =

{La nuova vita di **[Tommy]**[cat] }

CE2, 1654

{[I tre fratelli][ext1,ext2,ext3]}

C’era **[un robot]**[robot] **[che]**[robot] era amico di **[un bambino]**[ext1] **[che]**[ext1] si chiamava **[Andrea]**[ext1] e un giorno **[si incontrarono]**[ext1,robot] sul balcone perché **[un bullo]**[ext2] **[lo]**[robot] picchiò e **[lanciò]**[ext2] **[il robot]**[robot] e **[cadde]**[robot] sul balcone e **[il bambino]**[ext1] disse "ma chi sei **[tu]**[robot]?" "sono **[un robot]**[robot]" "perché **[sei ridotto]**[robot] così?" disse **[il bambino]**[ext1] e **[il robot]**[robot] rispose "perché **[mi]**[robot] hanno bullizzato" e **[il bambino]**[ext1] disse "ah **[ho capito]**[ext1]" **[il robot]**[robot] disse "**[vuoi]**[ext1] essere **[mio]**[robot] amico?" e **[il bambino]**[ext1] rispose "perché **[tu]**[robot] non hai amici?" "no" disse **[il robot]**[robot], "perché nessuno **[mi]**[robot] vuole" "ah ok. **[Ci vediamo]**[ext1,robot] sabato al parco." Sabato **[vanno]**[ext1,robot] al parco e **[il robot]**[robot] vede **[i due bulletti]**[ext2,ext3] e disse "**[torniamo]**[ext1,robot] a casa veloce." Ma **[loro]**[ext1,ext2,ext3] erano fratelli. E **[il robot]**[robot] si chiamava **[Diego]**[robot] ma **[loro]**[ext1,ext2,ext3] non sapevano che **[erano]**[ext1,ext2,ext3] fratelli, poi **[la mamma]**[ext4] **[gli]**[ext1] comprò **[un gatto]**[cat] **[che]**[cat] **[chiamò]**[ext1] **[Tatino]**[cat] **[che]**[cat] era molto bravo e favoloso.

CE2, 1655

C’era una volta **[una strega molto cattiva e molto potente]**[witch] e con **[lei]**[witch] c' è anche **[un gatto nero]**[cat]. **[Questa strega]**[witch] e **[questo gatto]**[cat] vivevano in un pianeta molto lontano dalla terra perché almeno quando **[avrà dominato]**[witch] la terra **[non morirà]**[witch]. Ma **[la strega]**[witch] non sapeva che sulla terra c’era **[una squadra di supereroi molto forti]**[ext1] **[che]**[ext1] risolvevano tutti i problemi che facevano **[i cattivi]**[ext2]. **[I supereroi]**[ext1] un giorno hanno saputo di **[questa strega malvagia]**[witch] e **[hanno saputo]**[ext1] pure dove **[viveva]**[witch] però **[non hanno saputo]**[ext1] di **[questo gatto nero]**[cat] **[che]**[cat] portava molta sfortuna. Il giorno dopo **[i supereroi]**[ext1] partono con una navicella spaziale per andare in questo pianeta molto lontano poi dopo 100 giorni **[sono arrivati]**[ext1] in questo pianeta, però si era fatto notte e **[hanno avuto]**[ext1] molta fortuna perché **[hanno trovato]**[ext1] una casa abbandonata. Dopo la notte **[si sono messi]**[ext1] a camminare molto molto, dopo un'altra notte **[hanno trovato]**[ext1] la casa **[della strega malvagia]**[witch]. Piano piano **[sono entrati]**[ext1] nella casa **[della strega]**[witch] **[che]**[witch] stava dormendo e cosi **[non [l']**[witch]**hanno fatta svegliare]**[ext1] con una magia poi **[sono tornati]**[ext1] sulla terra e **[[l']**[witch]**hanno messa]**[ext1] in prigione. La magia durava per almeno 20 giorni il tempo di studiar**[la]**[witch]. Dopo i 20 giorni **[si è svegliata]**[witch] ed **[era impazzita]**[witch] e con la magia **[è uscita]**[witch] dalla prigione però **[i supereroi]**[ext1] sapevano la **[sua]**[witch] parte debole e **[la]**[witch] **[continuavano]**[ext1] a colpire su quella parte però **[era incavolata]**[witch] e con la magia **[li]**[ext1] **[spinge]**[witch] **[tutti]**[ext1] via poi **[il più grosso]**[ext3] con il **[suo]**[ext3] martello **[gli]**[witchelo] tira addosso e **[stava]**[witch] per morire però **[non morì]**[witch]. Ma **[era indebolita]**[witch] e poi **[la]**[witch] **[catturano]**[ext1] e **[le]**[witch] **[dicono]**[ext1] cosa vuole ma **[lei]**[witch] non vuole dir**[glie]**[ext1lo] e così **[la]**[witch] **[torturano]**[ext1] ed **[era molto ferita]**[witch] ma **[riuscì]**[witch] a scappare ma **[la]**[witch] **[continuavano]**[ext1] a colpire nella parte debole e **[morì]**[witch]. **[Tutti i cittadini]**[ext4] uscirono di casa e **[dicono]**[ext4] "**[voi]**[ext1] siete i più forti **[vincete]**[ext1] sempre" poi arrivò **[il gatto nero]**[cat] **[che]**[cat] faceva cadere i pali e **[uccide]**[cat] metà della città e **[i supereroi]**[ext1] con un colpo **[lo]**[cat] **[uccidono]**[ext1] e vissero tutti felici e contenti.

CE2, 1663

{[Il gattino povero]cat}

C’era una volta **[un gattino selvaggio]**cat **[che]**cat viveva in un villaggio. **[Questo gattino]**cat era senzatetto e **[tutto il villaggio]**ext1 aveva paura di **[lui]**cat perché **[la gente]**ext1 pensava che se ti **[vede]**cat ti **[fa]**cat male, ma **[lui]**cat non faceva male! **[Questo gattino]**cat era povero e quando pioveva **[graffiava]**cat le porte per dire aiuto, ma nessuno **[gli]**cat aprì la porta quindi **[si bagnava]**cat tutto. Un giorno ha grandinato e **[il gatto]**cat si faceva male allora **[andò]**cat in una casa dove una volta **[lui]**cat e **[la [sua]cat padrona]**ext2 vivevano finché un giorno **[la padrona]**ext2 per le vacanze andò in Puglia e **[lasciò]**ext2 **[il gattino]**cat là a casa però **[la padrona]**ext2 non ci è più e quindi **[il gattino]**cat se ne andò da quella casa. Così **[il gattino]**cat entrò in quella casa e **[vide]**cat che era tutto cambiato e che era un pentolone gigante e c’erano delle cose disgustose tipo: occhi di zombie, lingua di serpente, zampe di ranocchie insetti morti... Poi c’era una stanza con **[una signora]**witch, **[quella signora]**witch si svegliò e **[vide]**witch **[il gatto]**cat.. **[scappa]**cat ma **[la strega]**witch **[lo]**cat prese e **[disse]**witch « ma **[che bel gattino]**cat **[vorrei]**witch adottar**[lo]**cat » **[il gattino]**cat era felice penso "sì che bello **[un altro padrona]**witch sì, sì, sì..." così **[la strega]**witch adottò **[il gattino]**cat e ora **[dice]**witch "è l'ora di **[darti]**cat un nome. **[Ti]**cat **[chiamerò]**witch: Lusio, Cerentol, Simpaticone, Ciocochino, Ray" **[il gattino]**cat "miau" quindi **[lo]**cat **[chiamò]**witch **[Ray]**cat e vissero felici e contenti.

CE2, 1664

C’era una volta **[una strega molto malvagia]**witch. Un giorno **[uscì]**witch per andare a comprare 4 ragni per la pozione. E **[vide]**witch **[un robot molto felice]**robot e **[sentì]**witch che oggi era il **[suo]**robot compleanno e **[le venne in mente]**witch un' altra pozione e invece di prendere 4 ragni **[prese]**witch 10 funghi velenosissimi e poi **[andò]**witch a casa e **[si travestì]**witch da **[una nonna innocente]**witch e poi **[andò]**witch **[dal robot]**robot e **[gli]**robot **[chiese]**witch « come ti chiami **[bel robottino]**robot? » e **[il robot]**robot rispose « **[io]**robot mi chiamo **[Billi]**robot e **[tu]**witch nonnina? » E **[la nonnina]**witch rispose « ha ha ha **[io]**witch non sono **[una nonnina]**witch sono **[una strega]**witch » e **[il robot Billi]**robot dice « aiuto qualcuno **[mi]**robot aiuti » poi **[la strega]**witch prese **[il robot Billi]**robot e **[lo]**robot **[fece diventare]**witch cattivo molto cattivo e **[gli]**robot **[fece]**witch scordare il **[suo]**robot compleanno poi l' incantesimo svanì e poi tutti felici e contenti tranne **[la strega]**witch. Fine.

CE2, 1665

{[Il robot]robot e **[il lupo]**wolf}

C’era una volta **[un robot molto ma molto intelligente, giocherellone, divertente e studioso]**robot. **[Il robot]**robot fabbricava altri giocattoli di metallo come: macchinine, altri robot e aeroplani telecomandati. **[Lui]**robot viveva in una casa rossa molto carina costruita da **[lui]**robot con resistenti mattoni di cemento.

Un giorno arrivò **[un branco di lupi]**[wolf]. **[Questi lupi]**[wolf] erano molto invidiosi di **[lui]**[robot], della **[sua]**[robot] casa e delle **[sue]**[robot] invenzioni. **[Il robot]**[robot] dal **[suo]**[robot] carattere giocherellone voleva sempre giocare con **[i lupi]**[wolf], ma invece **[i lupi]**[wolf] **[lo]**[robot] ignoravano.

[I lupi][wolf] stavano escogitando un piano per distruggere le **[sue]**[robot] invenzioni, la **[sua]**[robot] casa e se **[ci riuscivano]**[wolf] **[volevano]**[wolf] distruggere anche **[lui]**[robot].

Il giorno in cui accadde **[il robot]**[robot] aveva capito tutto e **[li]**[wolf] **[uccise]**[robot] perché sennò prima o poi **[l']**[robot]**[avrebbero ammazzato]**[wolf].

CE2, 1673

Tanto tempo fa c'era **[una strega]**[witch] **[che]**[witch] era stufa di stare da sola. Un giorno però **[la strega]**[witch] sentì dei rumori, **[era]**[witch] contentissima perché **[pensava]**[witch] che era qualcuno ma quando **[si voltò]**[witch] **[vide]**[witch] **[un dolcissimo gattino]**[cat], **[decise]**[witch] che era meglio stare con **[il gatto]**[cat] che da sola. **[Prese]**[witch] in braccio **[il gatto]**[cat] **[che]**[cat] era talmente magro perché **[non aveva mangiato]**[cat] così **[gli]**[cat] **[offrì]**[witch] subito un po' di carne e latte. **[La strega]**[witch] di accorse che **[il gatto]**[cat] si annoiava allora **[prese]**[witch] **[il micio]**[cat] e **[lo]**[cat] **[portò]**[witch] in negozio per far**[gli]**[cat] scegliere un gioco ma quando **[il gatto]**[cat] rispose miao, **[la strega]**[witch] non sentì miao ma **[sentì]**[witch]: " - Sì **[voglio]**[cat] questo non quello! " **[La strega]**[witch] si spaventò perché **[riusciva]**[witch] a capire cosa dicevano i gatti però **[non si sbalordì]**[witch] tanto perché è **[una strega]**[witch] e le streghe sono magiche. Così **[chiese]**[witch] **[al gatto]**[cat]: " - Come **[ti vuoi]**[cat] chiamare? " **[Il gatto]**[cat] rispose: " - **[Io]**[cat] mi voglio chiamare **[Stella]**[cat]! " **[La strega]**[witch] rispose: " - Bene **[Stella]**[cat] cosa vorresti fare? " Così **[Stella]**[cat] disse: " - **[Vorrei]**[cat] tornare a casa per **[farci fare]**[cat,witch] tante coccole." **[I due]**[cat,witch] si fecero un mare di coccole e **[fecero]**[cat,witch] una promessa: che era di restare per sempre amici.

CE2, 1676

{[Il gatto][cat] e **[la strega]**[witch]**}**

Un giorno **[dei bambini]**[ext1] quando **[stavano]**[ext1] giocando fuori in giardino avevano visto **[un gatto nero]**[cat], allora **[avevano pensato]**[ext1] che era di **[una strega]**[witch].

Quella sera **[i bambini]**[ext1] erano andati a letto e **[hanno dormito]**[ext1] ma **[la bambina]**[ext2] **[che]**[ext2] si chiamava **[Lisa]**[ext2] era **[una bambina molto simpatica]**[ext2] e **[aveva]**[ext2] i capelli lunghi e sempre la coda e **[aveva]**[ext2] anche i capelli color biondi e gli occhi verdi.

[Il gatto][cat] ha il pelo lungo e gli occhi verdi, **[lui]**[cat] **[il gatto]**[cat] è davvero di **[una strega]**[witch].

[Quella strega][witch] aveva il naso lungo e un brufolo, **[aveva]**[witch] gli occhi verdi, i capelli neri, il cappello viola e le unghie lunghe.

[La strega][witch] viene di notte per fare uno scherzo **[ai bambini]**[ext1] perché **[lei]**[witch] adora i bambini. **[I bambini]**[ext1] si svegliarono e **[videro]**[ext1] **[una strega spaventosa]**[witch], allora **[fecero]**[ext1,witch] amicizia e **[lei]**[witch] ha preparato la cena e come cena c'era... Fagioli, scarafaggi, cavolo verza ecc.

[I bambini][ext1] e **[la strega]**[witch] sono andati a letto ma **[Gianni il bambino che è allergico ai gatti]**[ext3], e in camera **[loro]**[ext1] c'era **[un gatto]**[cat] **[che]**[cat] stava facendo le fusa.

E alla fine **[tutti]**[cat,ext1,ext2,ext3,witch] diventarono amici.

CE2, 1677

{[Il gatto][cat] e **[la strega]**[witch]}

C'era una volta **[un gatto]**[cat] **[che]**[cat] si era perso e passò di lì **[una strega]**[witch] con una grande scopa e **[trovò]**[witch] **[il gatto]**[cat], allora **[aspettò]**[witch] un po' però **[la mamma [del gatto]**[cat]**]**[cat1] non si fece vedere, allora **[la strega]**[witch] si portò dietro **[il gatto]**[cat] a casa. **[La strega]**[witch] **[gli]**[cat] diede da mangiare, anche **[la strega]**[witch] mangiò.

Dopo aver mangiato **[sparecchiarono]**[cat,witch] però con la magia **[della strega]**[witch], allora i piatti si muovevano e andarono nel lavandino, si accese l'acqua e intanto **[loro]**[cat,witch] si addormentarono.

Il giorno dopo **[fecero]**[cat,witch] colazione e una bella passeggiata.

[Fecero][cat,witch] un disegno, un bagnetto e una cena, **[mangiarono]**[cat,witch] pollo e arrosticini.

[Si fecero][cat,witch] una bella dormitina insieme nelle coperte tutte calde anche se pizzicavano un po'.

Così vissero felici e contenti.

CE2, 760

C'era una volta **[una strega]**[witch]: magica, con il naso lungo perché **[diceva]**[witch] le bugie, **[aveva]**[witch] la **[sua]**[witch] scopa magica.

[Lei][witch] quando **[era]**[witch] a casa aveva **[un lupo]**[wolf] **[che]**[wolf] faceva gli scherzi alcuni buoni e alcuni cattivi.

Il giorno dopo **[il lupo]**[wolf] voleva fare uno scherzo **[alla strega]**[witch], lo scherzo era che **[lasciava]**[wolf] un palloncino sulla sedia, quando **[lei]**[witch] si era seduta, e quando **[aveva scoperto]**[witch] che era stato **[il lupo]**[wolf], **[la strega arrabbiata]**[witch] **[gli]**[wolf] fa uno scherzo cattivo, che ha voluto fare un incantesimo **[al lupo]**[wolf] **[che]**[wolf] non poteva più dire le bugie. Il giorno dopo **[il lupo]**[wolf] ha chiesto scusa per lo scherzo che **[aveva fatto]**[wolf] anche **[alla strega]**[witch], e vissero felici e contenti.

Fine.

CE2, 761

C'era una volta **[una strega]**[witch] con **[il [suo]**[witch] **gatto]**[cat], un giorno **[il [suo]**[witch] **gatto]**[cat] era scappato. **[Non poteva fare]**[witch] magie, la scopa non andava più e allora un giorno **[si arrabbiò]**[witch] tanto anzi tantissimo e **[fa]**[witch] una magia anche se il giorno prima **[poteva farle]**[witch] quindi **[si**

arrabbiò]witch ancora di più, dopo **[il gatto]**cat ritornò e **[gli]**witch era passata l'arrabbiatura. La scopa ripartì, si fece un giro con **[il gatto]**cat ed era anche **[lui]**cat era molto felice.

Però anche **[lui]**cat i giorni che **[la strega]**witch era arrabbiata **[lui]**cat era arrabbiato nello stesso giorno. **[La strega]**witch aveva scoperto che **[avevano]**cat,witch la telepatia, **[erano felicissimi]**cat,witch poi si scopre che **[il gatto]**cat era stato rubato da **[un'altra strega]**witch1 **[che]**witch1 era **[[sua]witch sorella]]**witch1 che **[lei]**witch non si era arrabbiata perché c'era **[il gatto]**cat vicino e **[la sorella]**witch1 però sì perché **[il gatto]**cat poteva parlare quindi **[si arrabbiò]**witch tantissimo. Dopo **[il gatto]**cat fa riunire **[le due streghe]]**witch,witch1 e **[il gatto]**cat era di **[tutte e due]**witch,witch1.

CE2, 762

{[La strega]witch e **[il robot]**robot}

C'era una volta, **[un robot]**robot **[che]**robot inventò **[una strega]**witch.

Allora **[questo robot]**robot era molto socievole, ma siccome **[il [suo]robot inventore]**ext1 **[l']**robotha costruito per uccidere, allora tutti avevano paura di **[lui]**robot. **[Lui]**robot voleva una amica magari che facesse un incantesimo per far**[lo]**robot diventare umano. **[Lui]**robot pensò: « **[potrei inventarmela]**robot una? e magari costruirla. » **[Continuò]**robot « **[inventerò]**robot una fata? No come **[potrei farla volare]**robot? Mhmm

[Inventerò]robot, un "genio della lampada" no, è un maschio, ora ci sono **[inventerò]**robot **[una strega]**witch. » E così dopo **[aver pensato]**robot a lungo, **[inventò]**robot **[una strega]**witch.

Il giorno dopo **[si mise]**robot al lavoro. Ci è voluta una settimana, ma il duro lavoro ripaga sempre, **[il nostro robot]**robot è riuscito a inventare **[una strega]**witch.

[La strega]witch appena creata chiese subito: "Come **[mi chiamo]**witch? Quanti anni **[ho]**witch? Ce l'**[ho]**witch un lavoro ?..." allora **[il robot]**robot **[gli]**witch disse: "**[Tu]**witch ti chiami **[Endora]**witch, **[hai]**witch 35 anni, e il **[tuo]**witch lavoro è realizzare i sogni degli altri !..." **[Endora]**witch: "Che bello **[tu]**robot hai un sogno? **[Io]**witch lo posso realizzare, comunque un'ultima domanda, chi è **[[mio]witch papà]**robot, chi è **[[mia]witch mamma]**ext4, ce l'**[ho]**witch una sorella?" Lui: "**[Io]**robot sono **[[tuo]witch papà]**robot, **[[tua]witch mamma]**ext4 è in cucina e se **[vuoi]**witch una sorella o un fratello lo **[avrai]**witch." **[Endora]**witch: "Maaa, **[tu]**robot ce l'hai un desiderio?" **[Lui]**robot: "Sì, **[ti]**witch **[ho creato]**robot apposta!" **[Endora]**witch: "**[Scommetto]**witch che **[vuoi diventare]**robot umano." **[Lui]**robot: "Si proprio cosi." E allora **[diventò]**robot umano e nessuno ebbe più paura di **[lui]**robot.

CE2, 764

{[Il lupo]wolf e **[la strega]**witch}

C'era una volta **[un lupo]**wolf **[che]**wolf era affamato, **[aveva]**wolf dei denti molto affilati e orecchie grandi.

Un giorno **[il lupo]**^{wolf} incontrò **[una strega]**^{witch} e **[le]**^{witch} **[disse]**^{wolf}: "**[Voglio mangiar[ti]**$^{\mathbf{witch}}$**]**^{wolf}!" e **[la strega]**^{witch} **[gli]**^{wolf} rispose: "**[Ci devi soltanto provare]**^{wolf}!" **[Il lupo]**^{wolf} disse: "Eh sennò cosa succede?" **[La strega]**^{witch} disse: "Sennò **[ti]**^{wolf} **[faccio]**^{witch} a piccoli pezzettini!" **[Il lupo]**^{wolf} disse: "Come **[fai]**^{witch}?" **[La strega]**^{witch} disse: "Con la **[mia]**^{witch} bacchetta! Mi basterà pronunciare questa formula, bum bidi bum, bum bum far ei pezzet bum bum bidi bum!" E **[la strega]**^{witch} la pronunciò, in un batter d' occhio **[il lupo]**^{wolf} iniziò a dividersi in piccoli pezzettini e **[la strega]**^{witch} cantò vittoria, peccato che l'incantesimo durava soltanto 15 secondi e **[il lupo]**^{wolf} ritornò come prima, **[il lupo]**^{wolf} diventò furioso e **[le]**^{witch} **[tirò]**^{wolf} una manata con i **[suoi]**^{wolf} artigli che **[la]**^{witch} stende e **[le]**^{witch} **[lasciò]**^{wolf} tutti i segni degli artigli, ma **[la strega]**^{witch} non si arrende e **[combatterono]**^{witch,wolf} fino alla fine, un po' di tempo dopo **[tutti e due]**^{witch,wolf} finirono a terra e **[non avevano]**^{witch,wolf} più forze per combattere e allora **[si arrendono]**^{witch,wolf}, ma **[il lupo]**^{wolf} **[che]**^{wolf} era più furbo **[della strega]**^{witch} disse una bugia e quando **[si alzano]**^{witch,wolf} **[il lupo]**^{wolf} tirò una manata **[alla strega]**^{witch} e **[la]**^{witch} **[buttò]**^{wolf} giù dalla montagna e **[il lupo]**^{wolf} **[che]**^{wolf} non aveva più forze, si alza e **[entrò]**^{wolf} nella casa **[della strega]**^{witch} e **[dormì]**^{wolf}. Il giorno dopo **[il lupo]**^{wolf} morì perché il veleno che c'era sulla scrivania **[gli]**^{wolf} cadde addosso perché di notte c'era stato il temporale.

CE2, 765

C'era una volta **[un gatto]**^{cat} **[che]**^{cat} si era perso in un bosco. **[Il gatto]**^{cat} era molto spaventato, così **[preso]**^{cat} dalla paura **[scappò]**^{cat}. Il giorno dopo **[il gatto]**^{cat} vide una cascina. In quella cascina ci abitava **[un robot]**^{robot}. **[Quel robot]**^{robot} si chiamava **[Axel 2]**^{robot}. **[Axel]**^{robot} era solo ma quando **[ha sentito]**^{robot} miagolare il **[suo]**^{robot} cuore era esploso dalla gioia. **[Aprì]**^{robot} la porta e **[vide]**^{robot} **[una gatta]**^{cat}. **[La gatta]**^{cat} non aveva un nome e quindi **[Axel]**^{robot} **[l']**^{cat} aveva chiamata **[Jachen]**^{cat}. **[Jachen]**^{cat} era disidratata quindi **[Axel]**^{robot} andò in cantina. **[Lui]**^{robot} aveva una scorta di scatolette di tonno. **[Jachen]**^{cat} **[lo]**^{robot} seguì e **[lo]**^{robot} **[vide]**^{cat}. **[Jachen]**^{cat} con un balzo **[gli]**^{robot} saltò addosso. **[Axel]**^{robot} si spaventò ma **[Jachen]**^{cat} si calmò subito. **[Il robot]**^{robot} **[gli]**^{cat} ha detto: "**[hai]**^{cat} tanta fame?" **[Jachen]**^{cat} miagolò ma **[Axel]**^{robot} non **[la]**^{cat} capiva ma **[gli]**^{cat} **[diede]**^{robot} da mangiare e alla fine **[lo]**^{cat} **[allevò]**^{robot}. Alla fine quando **[Axel]**^{robot} morì **[lasciò]**^{robot} **[Jachen]**^{cat} a **[[suo]**$^{\mathbf{robot}}$ **figlio]**^{ext1}.

Fine.

CE2, 767

C'era una volta **[un gatto]**^{cat}, **[il gatto]**^{cat} era felice con il **[suo]**^{cat} gomitolo, un giorno arrivò **[una strega]**^{witch} e **[gli]**^{cat} **[rubò]**^{witch} il **[suo]**^{cat} gomitolo.

Ma **[il [suo]**$^{\mathbf{cat}}$ **padrone]**^{ext1} **[gli]**^{cat} comprò una palla, ma **[la strega]**^{witch} era pronta a fare un altro colpo e il giorno dopo **[colpì]**^{witch} però **[il gatto]**^{cat} si aggrappò **[alla strega]**^{witch}.

Poi **[la strega]**^{witch} tornò a casa e **[il gatto]**^{cat} prende le **[sue]**^{cat} cose, arrivò anche **[il padrone]**^{ext1} e **[la]**^{witch} **[buttò]**^{ext1} nel **[suo]**^{ext1} camino e fu così che morì **[la strega]**^{witch}.

CE2, 768

C’era una volta **[un gatto]**[cat] **[che]**[cat] girava in un bosco. **[Il gatto]**[cat] era con tante macchie nere, **[è]**[cat] molto tenero, ma anche furbo, **[Un lupo]**[wolf] **[lo]**[cat] vide e **[disse]**[wolf] "uh! **[Che bel bocconcino]**[cat]", **[il gatto]**[cat] non **[lo]**[wolf] vide quindi **[il lupo]**[wolf] **[lo]**[cat] può mangiare facilmente. **[Il gatto]**[cat] **[lo]**[wolf] vide all'ultimo quindi è iniziato uno scontro, **[vanno avanti]**[cat,wolf] per ore finché **[non si addormentavano]**[cat,wolf]. La mattina dopo **[il gatto]**[cat] si svegliò prima **[del lupo]**[wolf] e **[il gatto]**[cat] scappò. **[Tornò]**[cat] a casa **[dalla mamma]**[cat1], ma **[il lupo]**[wolf] con il **[suo]**[wolf] super fiuto riuscì a ritrovar**[lo]**[cat], **[il gatto]**[cat] e **[la [sua]**[cat] **mamma]**[cat1] vanno **[dalla padrona]**[ext1] e **[la mamma gatta]**[cat1] **[gli]**[ext1] disse " c'è **[un lupo]**[wolf] " gridando e **[la padrona]**[ext1] prese una scopa e **[lo]**[wolf] bastona, dopo **[gridano]**[cat,cat1,ext1] "**[il lupo]**[wolf] è morto!"

Fine della storia.

CE2, 769

{[Il gatto costruttore][cat]}

C’era una volta **[un gatto costruttore]**[cat], **[quel gatto]**[cat] era così bravo a costruire che un giorno **[ha costruito]**[cat] **[un robot]**[robot]. **[Quel robot]**[robot], però a **[quel robot]**[robot] serviva una card con sopra scritto il codice. **[Il gatto]**[cat] si ricordava che l'aveva conservata però **[non sapeva]**[cat] dove l'aveva messa. Il giorno dopo **[il gatto]**[cat] si trovò in una camera piena di card, **[era riuscito]**[cat] a trovarla però la card diceva che **[il robot]**[robot] doveva essere chiamato **[SCP O 96]**[robot] **[il gatto]**[cat] chiamò **[il robot]**[robot] così però **[il robot]**[robot] era diventato **[un SCP O 96]**[robot]. **[Il gatto]**[cat] **[lo]**[robot] mise in una cella così **[lo]**[robot] **[studiava]**[cat] e così vissero tutti felici e contenti tranne **[il mostro SCP O 96]**[robot]

CE2, 770

C’era una volta **[un gatto]**[cat] e **[un robot]**[robot], **[erano]**[cat,robot] due amici perché **[giocavano]**[cat,robot] sempre insieme.

[Facevano][cat,robot] sempre la lotta, il più svantaggiato era **[il robot]**[robot] perché **[era]**[robot] il più lento mentre **[il gatto]**[cat] era velocissimo; **[giocavano]**[cat,robot] sempre a pallone e come sempre è più forte **[il gatto]**[cat] ma nel gioco della matematica era più bravo **[il robot]**[robot] perché è più intelligente e **[il gatto]**[cat] manco di 1 % e **[il robot]**[robot] 1.000.000.

Ma direi che in questi giochi è più bravo **[il gatto]**[cat] ecco come **[sono]**[cat,robot] amici.

Poi **[il robot]**[robot] si stancò perché era più bravo **[il gatto]**[cat], **[il gatto]**[cat] decise di non andarsene così **[diventò]**[cat] più debole e **[il robot]**[robot] **[gli]**[cat] dice che era uno scherzo dopo un po' di tempo perché **[era diventato forte]**[robot] anche **[lui]**[robot] e il detto è:

sempre di mettercela tutta, e **[saranno]**[cat,robot] sempre amici.

CE2, 771

C’era una volta **[un gatto di nome Birba]**[cat]. **[Birba]**[cat] un giorno scappò, e **[andò]**[cat] nel bosco e **[incontrò]**[cat] **[un lupo di nome Bot]**[wolf], e **[disse]**[wolf] "**[tu]**[cat] non puoi passare". E **[Birba]**[cat] disse "perché." **[Bot]**[wolf] disse "perché **[sei piccola]**[cat]". **[Birba]**[cat] disse che **[cerca]**[cat] solo una casa e **[Bot]**[wolf] disse "**[puoi]**[cat] venire nella **[mia]**[wolf]", **[Birba]**[cat] andò a casa di **[Bot]**[wolf]. **[Disse]**[cat] che era bellissima

e dopo **[si misero]**[cat,wolf] a guardare la tv e **[Bot]**[wolf] e **[Birba]**[cat] mentre **[guardavano]**[cat,wolf] la tv mangiavano i popcorn e vissero tutti felici e contenti per il momento.

CE2, 773

C’era una volta **[un gatto]**[cat] **[che]**[cat] gironzolava nei boschi, poi comparve una casa in mezzo agli alberi, **[il gatto]**[cat] entrò e **[vide]**[cat] **[una strega]**[witch] **[che]**[witch] stava facendo la pozione per diventare tutto quello che esisteva. Poi il giorno dopo **[il gatto]**[cat] ritornò là **[dalla strega]**[witch]. **[Il gatto tutto silenzioso]**[cat] saltò addosso **[alla strega]**[witch], poi **[la strega]**[witch] prese **[il gatto]**[cat] e **[gli]**[cat] **[lanciò]**[witch] la pozione e diventò **[un leone]**[cat], **[tigre]**[cat], **[squalo]**[cat], **[serpente]**[cat], **[robot]**[cat], **[umano]**[cat] infine **[un Godzilla]**[cat] e **[spaccò]**[cat] la casa **[della strega]**[witch]. Poi **[Godzilla]**[cat] prese la pozione e la lanciò **[alla strega]**[witch] e **[la strega]**[witch] diventò una formica e **[Godzilla]**[cat] **[la]**[witch] schiacciò, e **[la]**[witch] **[mette]**[cat] dentro il pentolone e **[la strega]**[witch] morì e **[Godzilla]**[cat] si tirò la pozione addosso e ritornò **[gatto]**[cat].

CE2, 813

C’era una volta **[un gatto]**[cat] **[che]**[cat] andava in giro e **[andava]**[cat] in una casa ma non c' è nessuno nessuno allora **[uscì]**[cat] nella notte e **[cammina]**[cat] nel bosco e **[vede]**[cat] **[un lupo]**[wolf] e **[il lupo]**[wolf] dice " **[non sono cattivo]**[wolf] " **[il gatto]**[cat] dice " **[mi sono perso]**[cat], **[mi]**[cat] **[puoi aiutare]**[wolf] e portare a casa? " **[il lupo]**[wolf] dice " vediamo, **[non so]**[wolf] la **[tua]**[cat] casa dov' è pero **[proviamo]**[cat,wolf] a vedere " **[si incamminano]**[cat,wolf] e **[trovano]**[cat,wolf] una casa arancione ma **[il gatto]**[cat] è addormentato. **[Il lupo]**[wolf] **[lo]**[cat] scuote e **[si sveglia]**[cat], dice **[il lupo]**[wolf] " questa è casa **[tua]**[None] " **[il gatto]**[cat] dice " non è la **[mia]**[cat] casa " **[il lupo]**[wolf] " di là, è quella la " **[il gatto]**[cat] dice di sì " **[andiamo]**[cat,wolf] " **[apre]**[cat] la porta " ciao **[mamma]**[ext1], ciao **[papà]**[ext2] " **[il gatto]**[cat] dice " **[possiamo tener[lo]**[wolf]**]**[cat,ext1,ext2] **[il lupo]**[wolf]? " **[la mamma]**[ext1] dice di sì.

CE2, 814

[Un robot di nome Leon][robot] stava camminando tranquillo e **[vede]**[robot] **[un lupo]**[wolf] e **[il lupo]**[wolf] **[lo]**[robot] vede. **[Il lupo di nome Bull]**[wolf] **[lo]**[robot] insegue, **[Leon]**[robot] corre e **[corre]**[robot] **[vede]**[robot] una casa e **[ci entra]**[robot]. Non c’era nessuno. **[Si nasconde]**[robot], **[spacca]**[robot] la porta, **[Leon]**[robot] non si fa vedere e **[dice]**[robot]: "**[vai]**[wolf] via -" e **[il lupo]**[wolf] risponde: "- no mai **[dimmi]**[robot] dove **[sei]**[wolf]."

[Il robot][robot] esce dalla finestra e **[corre]**[robot].

[Vede][robot] **[una bambina]**[ext1] e **[la bambina]**[ext1] **[lo]**[robot] prende in braccio poi **[il lupo]**[wolf] **[lo]**[robot] vede e **[Leon]**[robot] dice "**[corri]**[ext1] c'è **[un lupo]**[wolf]" e **[la bambina]**[ext1] dice "dove" "dietro di **[te]**[ext1]".

Allora **[la bambina]**[ext1] corre veloce verso casa **[sua]**[ext1], **[arriva]**[ext1], **[entra]**[ext1], **[non vede]**[ext1] **[il lupo]**[wolf], **[si siede]**[ext1], **[fa scendere]**[ext1] **[il robot]**[robot].

"Come **[ti chiami]**[robot]?"

"**[Io]**[robot] mi chiamo **[Leon]**[robot], e **[tu]**[ext1]?"

"**[Io]**[ext1] mi chiamo **[Elisabetta]**[ext1]."

[Il lupo][wolf] apre la porta **[entra]**[wolf] e **[dice]**[wolf]: "Sorpresa" **[Leon]**[robot] e **[Elisabetta]**[ext1] dicono "**[scappiamo]**[ext1,robot]!"

Ma prima di scappare **[Bull]**[wolf] si addormenta allora **[Leon]**[robot] **[lo]**[wolf] porta fuori bum cadde, **[prende]**[robot] **[il lupo]**[wolf], **[lo]**[wolf] **[porta]**[robot] in discarica

poi **[il lupo]**[wolf] si sveglia, **[cerca]**[wolf] di uscire.

Poi **[lo]**[wolf] **[buttarono]**[ext1,robot], **[cerca]**[wolf] di uscire e **[muore]**[wolf].

E vissero fino alla morte.

CE2, 815

C'erano una volta

[un gatto][cat] e **[un robot]**[robot]. **[Il gatto]**[cat] si chiama **[Leonardo di cognome Davinci]**[cat]. E **[il robot]**[robot] si chiamava **[Daniele di cognome Ibraliu]**[robot].

Un giorno **[Leonardo]**[cat] stava andando in giro per la casa poi **[cade]**[cat] dentro una tipo di cantina e **[Leonardo]**[cat] si spaventò e **[vide]**[cat] **[un robot molto grande]**[robot], **[pensava]**[cat] di buttar**[lo]**[robot] per terra infatti l'**[ha fatto]**[cat]. Buttando**[lo]**[robot] **[si è acceso]**[robot]. **[Il robot]**[robot] si è alzato **[ha iniziato]**[robot] a parlare con **[il gatto]**[cat] e **[sono diventati]**[cat,robot] migliori amici.

CE2, 816

{[La strega][witch] e **[il gatto rubato]**[cat]}

In una casa in cima ad una montagna viveva **[una strega]**[witch] **[che]**[witch] si chiamava **[Mella]**[witch].

[Mella][witch] desiderava tanto un gatto allora **[provò]**[witch] a fare tantissimi incantesimi ma nessuno riuscì a fare apparire un gatto.

Un giorno **[alla strega Mella]**witch venne un'idea e **[disse]**witch: - al posto di fare un incantesimo **[ruberò]**witch un gatto ad un bambino." Quella sera **[la strega Mella]**witch andò nella città di fianco, alla montagna e la **[perlustrò]**witch tutta.

Infine **[arrivò]**witch in una casa, **[entrò]**witch, **[vide]**witch **[un gatto]**cat e **[lo]**cat **[rubò]**witch.

La mattina dopo **[Mella]**witch diede il nome **[al gatto]**cat e **[lo]**cat **[chiamò]**witch **[Lulù]**cat, dopo **[disse]**witch a **[Lulù]**cat: " - **[ti]**cat **[porteró]**witch con **[me]**witch sulla **[mia]**witch scopa e dopo **[ti]**cat faro vedere la **[mia]**witch sfera magica.- "

Dopo aver fatto il giro sulla scopa **[gli]**cat mostrò **[la strega]**witch la sfera magica e **[vide]**witch **[il bambino]**ext1 a cui aveva rubato **[il gatto]**cat molto triste.

[La strega Mella]witch riportò **[Lulù]**cat **[al bambino]**ext1.

La mattina dopo **[la strega]**witch era felice perché **[aveva restituito]**witch **[il gatto Lulù]**cat ma **[era]**witch un po' triste perché gli mancava di già **[Lulù]**cat.

CE2, 817

{La casa magica}

C'era una volta **[un gatto arancione]**cat **[che]**cat aveva paura dei tuoni allora venne chiamato **[Tuono]**cat. Un giorno **[venne buttato]**cat via **[dai [suoi]cat padroni]**ext1 e **[si ritrovò]**cat sotto una panchina, poi **[andò]**cat in giro e **[vide]**cat una casa grande e alta e **[ci volle entrare]**cat però la porta di casa era chiusa allora **[vide]**cat una finestra aperta, **[si arrampicò]**cat su un albero e **[riuscì]**cat ad entrare. **[Si guardò]**cat intorno, **[vide]**cat **[un robot]**robot **[che]**robot parlava e **[si muoveva]**robot. Prima **[si spaventò]**cat ma dopo **[capì]**cat che **[lui]**robot era buono ma che **[voleva]**robot solo proteggere la **[sua]**robot casa. **[Il gatto Tuono]**cat **[lo]**robot aiutò a difendere la casa e **[loro]**cat,robot diventarono amici per la pelle.

Un giorno però **[litigarono]**cat,robot per chi comandava la casa allora **[fecero]**cat,robot a sorteggio e venne fuori **[il gatto]**cat (**[Tuono]**cat) ma **[il gatto]**cat visto che **[era buono]**cat decise che comandavano **[tutti e due]**cat,robot allora **[festeggiarono]**cat,robot la **[loro]**cat,robot amicizia e da quel giorno **[il gatto]**cat e **[il robot]**robot vissero insieme per tutta la vita felici e contenti fin quando **[il gatto]**cat non diventò vecchio e **[si ammalò]**cat, però **[il robot]**robot quando **[il gatto]**cat morì **[lui]**robot per ricordarsi **[del gatto Tuono]**cat e tutta la **[sua]**cat generosità **[gli]**cat **[dedicò]**robot una festa soltanto per **[lui]**robot.

CE2, 818

C'era una volta **[un gatto di nome Giuliano]**cat, **[gli]**cat piaceva molto giocare e dormire.

[Giuliano]cat dorme solo quando è in casa da solo, invece quando c'è a casa **[la padrona]**ext1 **[gioca]**cat molto.

[Giuliano][cat] è un po' arancione, un po' rosso, un po' bianco e un po' marroncino.

A **[lui]**[cat] non **[gli]**[cat] piace l'acqua infatti quando piove **[non va]**[cat] sul terrazzo, quando **[va]**[cat] a bere **[cerca]**[cat] di non bagnarsi il muso e quando **[la padrona]**[ext1] **[gli]**[cat] deve lavare le zampe **[lui]**[cat] scappa sul letto!

Per **[lui]**[cat] il letto è la cosa più morbida che ci sia nella casa, e la cosa più dura sia il tavolo.

[Giuliano][cat] mangia i crocchini secchi, l'umido e **[beve]**[cat] l'acqua.

Un giorno, quando **[la padrona]**[ext1] tornò, in una mano aveva una nuovissima cuccia con dentro un cuscinetto verde molto comodo.

Invece nell'altra mano **[aveva]**[ext1] un topino peluche, una pallina morbida e un topo grande peluche. **[La padrona]**[ext1] fece entrare nella cuccia **[Giuliano]**[cat], quando **[era]**[cat] nella cuccia **[gli]**[cat] **[diede]**[ext1] i giochi e **[lui]**[cat] era molto contento. Il titolo della storia é:

[il gatto Giuliano][cat]

CE2, 819

[Il gatto scrittore][cat]

C'era una volta **[un gattino di nome Marino]**[cat], **[lui]**[cat] adora giocare con **[i [suoi]**[cat] **amici]**[ext1], a collezionare carte e specialmente fare disegni di animali inventati, il **[suo]**[cat] preferito è il pappaunicorno.

Un giorno **[andò]**[cat] nella città di Cucciolander e **[trovò]**[cat] la foresta incantata ed **[entrò]**[cat] ; **[si sedette]**[cat] e **[scrisse]**[cat] una poesia. Mentre **[era seduto]**[cat] **[si sdraiò]**[cat] e **[vide]**[cat] il cielo infinito, **[abbracciò]**[cat] il **[suo]**[cat] quadernino e **[scrisse]**[cat] due pagine, la **[sua]**[cat] poesia alla fine si chiamò Infinito, dopo aver finito di scrivere **[si rialzò]**[cat] e **[vide]**[cat] **[una cosa spaventosa]**[robot], ma **[lui]**[cat] non sapeva che era **[il robottino]**[robot], **[dei padroni]**[ext1] in ricerca **[del [loro]**[ext1] **gatto]**[cat1] **[lo]**[cat] **[presero]**[ext1] e **[presero]**[ext1] pure **[il robottino]**[robot] ma alla fine **[vinsero]**[cat,robot] e **[ritornarono]**[cat,robot] nel bosco. Per tutta la vita **[sono stati felici]**[cat,robot] e il 27 maggio **[gli]**[cat,robot] hanno dedicato una giornata.

CE2, 820

C'era una volta **[una strega di nome Saretta]**[witch]. **[La strega]**[witch] aveva **[un gatto di nome Gippetto]**[cat]. **[Gippetto]**[cat] combinava sempre guai. Un giorno **[la strega Saretta]**[witch] prese la scopa, **[chiama]**[witch] **[Gippetto]**[cat] **[saltarono]**[cat,witch] sulla scopa e **[partirono]**[cat,witch] **[Gippetto]**[cat] **[che]**[cat] combinava guai scese dalla scopa e **[dice]**[cat] "**[io]**[cat] scendo dalla scopa" **[scese]**[cat] e **[disse]**[cat] "il mondo è cosi bello!!!" Ma **[lui]**[cat] s'era portato i regali perché oggi era Halloween quindi quando **[la strega Saretta]**[witch] si era girata, **[aveva visto]**[witch] che i regali non c'erano quindi **[andò]**[witch] a cercare **[Gippetto]**[cat], **[lo]**[cat] **[trovò]**[witch], **[presero]**[cat,witch] i regali e **[andarono]**[cat,witch] a dare i regali a tutti i bambini. Da quel punto **[non fece più]**[cat] casini.

Il titolo è **[la strega]**witch e **[il gatto]**cat

CE2, 821

C'era una volta **[un robot]**robot **[che]**robot stava camminando. E **[incontra]**robot **[il [suo]robot amico lupo]**wolf e **[il robot]**robot **[lo]**wolf saluta. Dopo **[sono andati]**robot,wolf al parco e **[hanno giocato]**robot,wolf a calcio e poi **[sono andati]**robot,wolf in piazza a mangiare il gelato, e **[sono andati]**robot,wolf al negozio di giocattoli e **[hanno comprato]**robot,wolf una macchinina telecomandata e **[sono andati]**robot,wolf a camminare e poi **[sono andati]**robot,wolf a casa.

CE2, 822

{[La strega pasticcera]witch e **[il [suo]witch gatto Gattastro]**cat}

[La strega Sara abbastanza buona]witch aveva **[un gatto di nome Gattastro]**cat.

[La strega Sara]witch era puzzolente e sporca. **[Aveva]**witch un vestito tutto nero, un cappello nero appuntito, una scopa magica che aveva un pulsante che la faceva volare... Ma sapete la cosa più spettacolare di **[questa strega]**witch... È che non **[le]**witch piace fare le pozioni magiche... Ma **[le]**witch piace cucinare!

Per **[lei]**witch la cucina è il **[suo]**witch mondo, il **[suo]**witch sogno è diventare una pasticcera ma **[crede]**witch che non ha molte speranze per colpa del **[suo]**witch aspetto... Ma la cosa più bella di **[lei]**witch è che non si arrende, **[prova]**witch **[prova]**witch e **[prova]**witch ma niente.

Allora un giorno **[la strega]**witch ebbe un'idea, l'idea è trasformarsi in una donna, **[prese]**witch la bacchetta ed ecco fatto!

[Fece]witch delle cose buonissime, **[prese]**witch ancora la bacchetta e **[fece]**witch un negozio solo per **[lei]**witch. Era tutto ok il piano andava a gonfie vele. Arrivò la notte ed era l'ora di tornare a casa.

[La strega]witch era stanchissima, prese **[Gattastro]**cat e **[gli]**cat fece le coccole sulla **[loro]**cat,witch poltrona. Ma successe una cosa stranissima nel negozio di notte... Stava scomparendo tutto!

[La strega]witch si svegliò, **[prese]**witch la scopa e **[volò]**witch al negozio, **[prese]**witch subito la bacchetta magica e **[disse]**witch: " - Abra cadabra!- "

Menomale era tornato come prima tutto. **[La strega]**witch era felicissima, **[iniziò]**witch a comportarsi come una vera donna, **[era diventata]**witch profumata, pulita e molto bella.

[Era]witch più contenta che mai con **[il [suo]witch gatto Gattastro]**cat.

CE2, 823

{Le avventure di **[Ciao]**ext1 e di **[Leo]**robot}

Tanto tempo fa c'era **[un robot]**[robot] **[che]**[robot] era come il supereroe di mariatolis crismal e **[il supereroe]**[robot] si chiama **[Leo]**[robot]. **[Lui]**[robot] aveva **[un assistente]**[ext1] **[che]**[ext1] si chiama **[Ciao]**[ext1], **[Leo]**[robot]: " - **[Leo]**[robot] c'è **[pennuto Puré]**[ext2] **[Ciao]**[ext1] usa la balestra di fuoco veloce!"

[Ciao][ext1] tirò con la **[sua]**[ext1] balestra e **[usò]**[ext1] la freccia acida e **[Pennuto Puré]**[ext2] chiamò **[Stada]**[robot1] **[che]**[robot1] è **[un robot]**[robot1]. **[Suchi]**[ext3] e **[Coltism]**[ext4] **[loro]**[ext3,ext4] sono stati polverizzati dalla freccia acida poi **[usò]**[ext1] la freccia di **[Sdendermen]**[ext5] e **[Pennuto Puré]**[ext2] viene stato polverizzato e poi **[Ciao]**[ext1] divenne **[un eroe]**[ext1].

CE2, 824

{[Il robot][robot] e **[il lupo]**[wolf]:}

C'era una volta **[un robot di nome Yack]**[robot] **[che]**[robot] abitava in una foresta in una capanna però quando **[aveva fame]**[robot] **[andava]**[robot] nella foresta a cercare cibo ma quasi tutte le volte **[non trovava]**[robot] cibo ma poi **[incontrò]**[robot] **[un animale]**[wolf] **[non sapeva]**[robot] qua l' era perché **[lui]**[robot] non aveva mai visto **[un' animale]**[wolf] **[che]**[wolf] andava così veloce tra gli alberi. **[Aveva]**[wolf] le orecchie a punta, il naso rotondo, i denti affilati e la bocca lunga e **[camminava]**[wolf] a quattro zampe e **[aveva]**[wolf] una coda lunga e poi **[indicò]**[wolf] **[me]**[robot] con il **[suo]**[wolf] dito allora **[io]**[robot] rientrai nella **[mia]**[robot] capanna e **[quel coso]**[wolf] come **[lo]**[wolf] chiamava **[il robot]**[robot] scappò via velocemente nella **[sua]**[wolf] tana. Il giorno dopo **[il robot]**[robot] uscì dalla capanna e **[andò]**[robot] nella foresta e **[incontrò]**[robot] ancora **[quel coso]**[wolf], allora **[lui]**[robot] **[il robot]**[robot] si avvicinò a **[quel coso]**[wolf] e allora si avvicinò anche **[il lupo]**[wolf] e **[gli]**[robot] **[disse]**[wolf]: " - ciao **[robot]**[robot] " **[il robot]**[robot] rispose: " - ciao come ti chiami? " **[gli]**[wolf] chiese **[il robot]**[robot]. **[Il lupo]**[wolf] **[gli]**[robot] rispose " - **[io]**[wolf] mi chiamo **[lupo]**[wolf] " **[gli]**[robot] rispose **[il lupo]**[wolf] il allora **[non aveva ancora capito]**[robot] che **[era]**[wolf] un animale allora **[il robot]**[robot] **[gli]**[wolf] disse: " - **[io]**[robot] invece mi chiamo **[Yack]**[robot] " **[il lupo]**[wolf] **[gli]**[robot] rispose: " - che bel nome " e **[il robot]**[robot] **[lo]**[wolf] ringraziò.

CE2, 825

{**[Il gatto]**[cat] ed **[il lupo]**[wolf] "amore nascosto"}

C'erano una volta **[una gatta]**[cat] ed **[un lupo]**[wolf], **[la gatta]**[cat] si chiamava **[Sissi]**[cat] e **[il lupo Alexander]**[wolf], **[si erano appena conosciuti]**[cat,wolf] ma **[si amavano]**[cat,wolf] già e spesso **[erano insieme]**[cat,wolf] perché o **[i [loro]**[cat,wolf] **padroni]**[ext1] stavano insieme a parlare, o **[loro]**[cat,wolf] di notte scappavano e **[si incontravano]**[cat,wolf]. Un giorno **[erano stati]**[cat,wolf] talmente insieme **[la gatta]**[cat] diede la prima leccata **[al lupo]**[wolf], e la **[loro]**[cat,wolf] leccata e il nostro bacio. Il giorno dopo **[il lupo]**[wolf] andò **[dalla gatta]**[cat] e **[le]**[cat] **[chiese]**[wolf] se **[volevano scappare]**[cat,wolf] vie per sempre insieme, **[la gatta]**[cat] accettò e prima che si svegliassero **[i [loro]**[cat,wolf] **padroni]**[ext1] **[il lupo]**[wolf] disse di scappare, e **[scapparono]**[cat,wolf]. Dopo un anno che **[erano scappati]**[cat,wolf] **[la gatta]**[cat] scoprì di essere incinta, e lo disse **[al lupo]**[wolf] **[il lupo]**[wolf] era felicissimo di essere papà. I nove mesi passarono e **[la gatta]**[cat] diede alla luce **[due bellissimi gattini]**[cat1,cat2] **[una femmina]**[cat1] e **[un gattino maschio]**[cat2] e **[la gatta]**[cat] ed **[il lupo]**[wolf] furono molto felici e **[li]**[cat1,cat2] **[chiamarono]**[cat,wolf]

[Sasha][cat1] e visto che **[il gattino piccolo]**[cat2] era molto vivace **[lo]**[cat2] **[chiamarono]**[cat,wolf] **[Saetta]**[cat2]. **[I gattini]**[cat1,cat2] crebbero e **[la mamma]**[cat] e e **[il papà]**[wolf] morirono, e **[i gattini]**[cat1,cat2] ormai diventati adulti **[si fecero]**[cat1,cat2] **[le [loro]**[**cat1,cat2**] **famiglie]**[ext2] e anche loro fecero **[i [loro]**[**cat1,cat2**] **figli]**[ext3] e **[fecero]**[cat1,cat2] una bella vita.

CE2, 827

{[Un bravo gatto][cat] con **[un bel padrone]**[ext1]}

C'era una volta **[un bel gattino]**[cat] **[che]**[cat] si chiamava **[Gianni]**[cat].

[Gianni][cat] era **[un buon gatto]**[cat] **[ascoltava]**[cat] molto **[i [suoi]**[**cat**] **padroni]**[ext2].

[Gianni][cat] giocava con **[il suo padrone]**[ext1]. Un giorno **[i padroni di [Gianni]**[**cat**]**]**[ext2] partirono quindi **[Gianni]**[cat] non sapeva con chi giocare.

Un giorno passò di lì **[un bambino di nome Mario]**[ext3] e **[giocò]**[ext3] con **[Gianni]**[cat] finché non tornarono **[i padroni]**[ext2],

quando **[i padroni]**[ext2] tornarono fecero **[un bambino]**[ext4] **[che]**[ext4] si chiamava **[Gabriele]**[ext4]. Quando **[Gabriele]**[ext4] aveva otto anni **[giocavano]**[cat,ext4] insieme.

Poi però **[Gabriele]**[ext4] dovette andare a scuola e già in 2° elementare **[aveva]**[ext4] molto da studiare quindi **[Gianni]**[cat] fu molto triste di questa cosa. Poi **[Gabriele]**[ext4] andò alle medie, alle superiori e all' università e così **[finì]**[ext4] le scuole e **[si laureò]**[ext4] in chimica e **[fu]**[ext4] molto bravo.

[Gabriele][ext4] sposò **[una donna di nome Laura]**[ext5] e **[Gabriele]**[ext4] vissero con **[Gianni]**[cat] felici e contenti in una bellissima casa.

CE2, 828

C'era una volta **[un gattino di nome Nerino]**[cat] e un giorno **[la [sua]**[**cat**] **padroncina]**[ext1] **[gli]**[cat] mise un po' di croccantini e **[Nerino]**[cat] se li mangiò tutti e **[la [sua]**[**cat**] **padroncina]**[ext1] **[gli]**[cat] disse " ma **[avevi]**[cat] molta fame " poi **[si mise]**[cat] a dormire.

CE2, 829

{[Alex il gatto pasticcione][cat]}

[Alex][cat] era **[un gatto felice]**[cat] **[che]**[cat] abitava con **[la [sua]**[**cat**] **famiglia]**[ext1] però **[era]**[cat] anche maldestro. **[Combinava]**[cat] sempre dei pasticci però comunque **[la [sua]**[**cat**] **famiglia]**[ext1] **[gli]**[cat] voleva bene.

Un giorno **[il padrone di [Alex]**[**cat**]**]**[ext2] aveva dimenticato di dar**[gli]**[cat] la pappa che **[aveva lasciato]**[ext2] su un mobiletto. Allora **[Alex]**[cat] saltò e **[fece cadere]**[cat] tutto: bum! !! Un altro giorno **[Alex]**[cat] era andato a fare l' esame dei gatti più agili di Bollate. **[Alex]**[cat] pensò:- questa volta ce la

devo fare -. **[Alex]**[cat] fece anche amicizia con un gatto che **[lo]**[cat] aveva difeso. E quel gatto si chiamava Tigro. Era il turno di **[Alex]**[cat] però arrivò decimo su dieci perché aveva fatto cadere l'istruttrice però Tigro era arrivato primo e consolò **[Alex]**[cat] dicendo**[gli]**[cat:-] **[tu]**[cat] hai fatto quello che potevi - allora tornò a casa. Col passare degli anni **[Alex]**[cat] non diventò meno pasticcione però più affettuoso perché aveva un amico. Tigro.

CE2, 830

{[Il robot senza dio][robot]}

Un giorno c'era **[una famiglia di robot]**[ext1] e c'era **[un robot]**[robot] però **[della famiglia]**[ext1] **[che]**[robot] era sempre giù di morale non si sapeva perché, si stavano preoccupando **[i [suoi]**[robot] **genitori]**[ext2]. Allora **[i genitori]**[ext2] e **[il [suo]**[robot] **fratellino]**[ext3] cercano di scoprire cosa aveva ma **[lui]**[robot] non diceva niente, **[non mangiava]**[robot], **[stava]**[robot] sempre da solo chiuso nella **[sua]**[robot] stanza.

A scuola **[non rispondeva]**[robot], **[non diceva]**[robot] le interrogazioni quindi **[i genitori]**[ext2] **[lo]**[robot] portarono all' ospedale ma non capivano neanche **[i dottori]**[ext4] allora **[cercarono]**[ext4] di parlar**[gli]**[robot] ma **[lui]**[robot] scappava e **[usciva]**[robot] di casa e **[faceva]**[robot] passeggiate lunghissime ma **[lui]**[robot] non si stancava.

A chi **[lo]**[robot] incontrava e **[lo]**[robot] salutava o **[gli]**[robot] diceva qualcosa **[correva]**[robot] **[correva]**[robot] e **[non ritornava]**[robot] a casa. **[I [suoi]**[robot] **genitori]**[ext2] non **[lo]**[robot] trovavano ma **[erano abituati]**[ext2], **[saltava]**[robot] anche i giorni di scuola, **[le maestre]**[ext5] si preoccuparono e **[andarono]**[ext5] a casa **[sua]**[robot] e **[chiesero]**[ext5] **[alla mamma [del robot]**[robot]**]**[ext2,ext6] "dov'é **[Carlo]**[robot]?" "**[Non lo sappiamo]**[ext2]" allora **[lo]**[robot] **[andarono]**[ext2] a cercare lontano lontano. Alla fine **[lo]**[robot] **[trovarono]**[ext2] e **[scoprirono]**[ext2] che **[non aveva più]**[robot] dio e **[non riusciva]**[robot] a parlare, **[era]**[robot] giù di morale perché **[lo]**[robot] **[trattavano]**[ext2] come se fosse uno strano.

CE2, 834

{[Il gatto con tante emozioni][cat]!}

C'era una volta **[un gatto]**[cat] **[che]**[cat] si chiamava **[Milù]**[cat].

[Aveva][cat] **[una padrona bella e gentile]**[ext1] e un giorno però **[la padrona]**[ext1] non si sentì di curar**[lo]**[cat] e **[lo]**[cat] **[portò]**[ext1] a un gattile.

[Il gatto][cat] era molto legato **[alla padrona]**[ext1] e quindi **[piangeva]**[cat] tanto! Ma **[i signori del gattile]**[ext2] chiamavano sempre **[la padrona]**[ext1]. Finché un giorno **[la padrona]**[ext1] **[lo]**[cat] andò a prendere. Erano passati mesi e mesi che **[non [lo]**[cat] **vedeva]**[ext1]. Quindi **[il gatto]**[cat] si mise a piangere dalla gioia e anche **[la padrona]**[ext1]. **[Il gatto]**[cat] ritornò a casa e **[gli]**[cat] fecero una festa. **[Lui]**[cat] giocò con **[i [suoi]**[cat] **amici]**[ext3] e **[gli altri padroni]**[ext4] ridevano e ridevano.

[La padrona][ext1] non era stata mai così contenta e anche **[il piccolo gatto]**[cat] e così vissero felici e contenti insieme e **[non si lasciarono]**[cat,ext1] mai !

CE2, 835

{La gita al parco di **[Bull]**[robot] e **[Lucario]**[wolf]}

C’era una volta **[un robot di nome Bull]**[robot]. Un giorno **[Bull]**[robot] andò **[dal [suo]**[robot] **amico Lucario il lupo]**[wolf] e **[gli]**[wolf] **[disse]**[robot]: " - **[Andiamo]**[robot,wolf] al parco. - " **[Lucario]**[wolf] rispose: " - ma prima **[facciamo]**[robot,wolf] merenda. **[Ho]**[wolf] una fame da lupi. - "

[I due amici][robot,wolf] si incamminarono per andare al parco con un cesto con dentro del cibo: due panini col salame, due sacchetti di circuiti alla fragola e delle biglie.

Nel parco c’erano tanti giochi: uno scivolo di trenta metri, quattro altalene, sei trampolini, tre girandole e una pista per le biglie. **[Lucario]**[wolf] andò sulla girandola, sull' altalena e poi **[mangiò]**[wolf] il panino.

Invece **[Bull]**[robot] svuotò il sacchetto di circuiti alla fragola poi **[andò]**[robot] sullo scivolo e sull' altalena.

Alle ore 5:30 **[i due amici]**[robot,wolf] si stancarono e **[giocarono]**[robot,wolf] con le biglie; e alle ore 7:15 **[andarono]**[robot,wolf] a casa a mangiare. The end.

CE2, 836

{[Il gatto][cat] e **[la strega]**[witch]}

C’era una volta **[una strega]**[witch] **[che]**[witch] odiava molto i gatti. **[La strega]**[witch] viveva in una foresta incantata.

Ma un giorno **[un gatto]**[cat] entrò in casa **[sua]**[witch] e **[cercò]**[witch] di prender**[lo]**[cat] ma **[non ci riuscì]**[witch] ma **[il gatto]**[cat] aveva un potere solo.

[Aveva][cat] il potere di fare innamorare le streghe di **[lui]**[cat] perché **[aveva]**[cat] gli occhi troppo dolci.

Quindi **[il gatto]**[cat] visse con **[la strega]**[witch].

Ma dietro la porta c’era **[un'altra strega]**[witch1] **[che]**[witch1] aveva sentito tutto. Ma **[quella strega]**[witch1] adorava i gatti e così **[entrò]**[witch1] in casa e **[si rubò]**[witch1] **[il gatto]**[cat] e **[se [lo]**[cat] **porto]**[witch1] a casa.

[La strega][witch] pensa per due giorni e poi **[disse]**[witch] " - **[Io]**[witch] devo riprendermi **[quel gatto magico]**[cat]."

E così **[andò]**[witch] alla ricerca **[del gatto]**[cat]. **[Andò]**[witch] sulle montagne e **[non [lo]**[cat] **trovo]**[witch], **[andò]**[witch] nel deserto e **[non [lo]**[cat] **trovò]**[witch], poi **[andò]**[witch] nella foresta e **[incontrò]**[witch] una casa piena zeppa di gatti.

Un giorno di notte **[andò]**[witch] dentro alla casa e **[si riprese]**[witch] **[il gatto magico]**[cat] e vissero per sempre felici e contenti. E **[la strega]**[witch1] il giorno dopo quando **[non trovo più]**[witch1] **[il gatto]**[cat] **[si arrabbia]**[witch1] tantissimo.

CE2, 839

C’era una volta **[un lupo]**^{wolf} **[che]**^{wolf} ogni sera andava a caccia di conigli, di lepri, di cervi e altri animali di questo tipo.

Un giorno **[il lupo]**^{wolf} si addentrò nella foresta e **[vide]**^{wolf} **[una cosa]**^{robot} in lontananza. **[Arrivò]**^{wolf} di soppiatto e con un balzo **[saltò]**^{wolf} **[sulla preda]**^{robot}.

Solamente che **[la preda]**^{robot} non era un animale dei soliti che **[cacciava]**^{wolf}, anzi, non era proprio un animale.

[Era]^{robot} alto, di ferro, con la testa rotonda e sulla testa **[aveva]**^{robot} un'antenna.

Talmente **[era]**^{robot} duro che per dar**[gli]**^{robot} un morso **[il lupo]**^{wolf} si fece dondolare un dente.

[L'oggetto di metallo]^{robot} era **[un robot]**^{robot}.

[Il lupo]^{wolf} dopo aver dato un morso **[al robot]**^{robot} fu un po' stordito. Poi **[il lupo]**^{wolf} e **[il robot]**^{robot} fecero amicizia però **[il lupo]**^{wolf} aveva fame e **[il robot]**^{robot} senza dire una parola, **[andò]**^{robot} fitto fitto nei boschi e **[tornò]**^{robot} con un coniglio da mangiare

E vissero tutti felici e contenti.

{[Il lupo]^{wolf} e **[il robot]**^{robot}}

CE2, 841

{[Il gatto]^{cat} e **[la strega]**^{witch}}

C’era una volta **[un gatto]**^{cat} **[che]**^{cat} si chiama **[Chiara]**^{cat} ed **[era]**^{cat} molto brutta, **[non aveva]**^{cat} pelo, **[aveva]**^{cat} dei graffi sulla pancia, ma un giorno **[Chiara]**^{cat} incontra **[una strega]**^{witch}, e **[la strega]**^{witch} **[la]**^{cat} guarda negli occhi e **[il gatto]**^{cat} diventa bellissimo, **[aveva]**^{cat} gli occhi azzurri, **[aveva]**^{cat} il pelo tutto nero e la punta della coda bianca e da quel giorno se guardavi **[il gatto]**^{cat} negli occhi per 3 secondi diventava **[un gatto]**^{cat}. **[La strega]**^{witch} aveva delle caramelline che quando le **[mangiava]**^{witch} diventava **[un gatto]**^{witch} e alla notte andavano **[tutti i gatti]**^{cat1} che **[la strega]**^{witch} e **[Chiara]**^{cat} avevano fatti diventare **[gatti]**^{cat1} e **[andavano]**^{cat,witch} nelle case, **[[si facevano]**$^{\mathbf{cat1}}$ **accarezzare]**^{cat,witch} poi **[li]**^{cat1} **[guardavano]**^{cat,witch} negli occhi e diventano gatti.

Poi il giorno dopo **[Chiara]**^{cat} non si trova e **[viene catturata]**^{cat} e **[la]**^{cat} portano nel deserto e per 22 giorni **[Chiara]**^{cat} non si trova. **[La strega]**^{witch} si arrabbia e **[fa diventare]**^{witch} **[tutti i gatti]**^{cat1} **[delle streghe]**^{cat1} e **[volano]**^{cat1,witch} sulle scope e dopo 2 ore **[la]**^{cat} **[trovano]**^{cat1} e se ne vanno **[tutti]**^{cat,cat1,witch} insieme e **[festeggiano]**^{cat,cat1,witch} e vivono felici e contenti.

CE2, 842

{[Il lupo]^{wolf} e **[il robot]**^{robot}}

Un giorno **[un lupo mannaro]**[wolf] **[che]**[wolf] aveva molta fame **[andò]**[wolf] nel bosco a cercare una preda per poi gustarsela.

Però **[non trovo]**[wolf] niente, solo un cacciavite, **[non capiva]**[wolf] perché c'era un cacciavite nel bosco però **[sapeva]**[wolf] che **[aveva]**[wolf] una grande fame. Così **[tornò]**[wolf] a caccia ma **[non trovava]**[wolf] niente, **[aveva trovato]**[wolf] un pezzo di metallo e ancora **[non capiva]**[wolf] più niente, quindi **[si mise]**[wolf] a dormire.

La notte dopo **[il lupo]**[wolf] non voleva mangiare ma **[voleva investigare]**[wolf].

Così **[uscì]**[wolf] dalla tana e **[trovo]**[wolf] **[un robot]**[robot] **[che]**[robot] sembrava morto ma **[stava solo dormendo]**[robot] poi **[si sveglio]**[robot], **[guardò]**[robot] negli occhi e sia **[il robot]**[robot] e sia **[il lupo]**[wolf] urlarono a squarcia voce, **[il robot]**[robot] si mise a scappare a più non posso e **[il lupo]**[wolf] **[lo]**[robot] inseguì, fino a quando che **[il robot]**[robot] non sbatte la testa contro un albero.

[Il lupo][wolf] portò nella tana **[il robot]**[robot] e **[gli]**[robot] **[chiese]**[wolf] se il pezzo di metallo e il cacciavite erano **[suoi]**[robot], **[il robot]**[robot] **[gli]**[wolf] disse di sì.

Poi **[il robot]**[robot] si mise il pezzo di metallo sulla pancia e **[si aggiustò]**[robot] e **[il robot]**[robot] ringraziò **[il lupo]**[wolf] e **[se ne andò]**[robot] dando**[gli]**[wolf] una bistecca come regalo, e **[il lupo]**[wolf] se la mangio e poi **[andò]**[wolf] a dormire.

CE2, 843

{[Robi il robot][robot] scappa **[dal lupo]**[wolf]}

C'era una volta **[un robot di nome Robi]**[robot] ed **[era]**[robot] speciale, perché **[riusciva]**[robot] a vivere ed **[era]**[robot] molto furbo, così un giorno **[riuscì]**[robot] a ingannare **[un lupo]**[wolf].

Iniziò tutto nel bosco e **[un lupo]**[wolf] **[lo]**[robot] stava inseguendo, **[lui]**[robot] si arrampicò su un albero ma **[non riuscì]**[robot] a fermar**[lo]**[wolf] ma **[riuscì]**[robot] a fuggire, e **[riuscì]**[robot] ad andare nella **[sua]**[robot] casetta, una casetta normalissima ma con un sacco di funzioni in più.

Un giorno **[il lupo]**[wolf] bussò alla porta. **[Robi]**[robot] aprì la porta, **[il lupo]**[wolf] **[lo]**[robot] inseguì fino al terzo piano ma **[Robi]**[robot] si buttò fuori dalla finestra e **[si aggrappò]**[robot] alla finestra del secondo piano, **[il lupo]**[wolf] cercò di buttarsi ma **[soffriva]**[wolf] di vertigini. Intanto **[Robi]**[robot] senza farsi vedere entrò al secondo piano, **[prese]**[robot] una vasca di acqua freddissima e la **[mise]**[robot] fuori dalla porta. **[Fece]**[robot] anche in tempo a mangiare una brioche e dopo **[andò]**[robot] al terzo piano, **[spinse]**[robot] **[il lupo]**[wolf] giù dalla finestra e **[il lupo]**[wolf] finì dritto nell' acqua e **[scappò]**[wolf] ma **[non cercò mai più]**[wolf] di attaccare **[Robi]**[robot].

CE2, 844

{[Il gatto][cat] e **[il lupo]**[wolf]}

C'era una volta **[un piccolo gattino di nome Bibi]**[cat].

Un giorno **[il piccolo gattino]**[cat] uscì dalla sua casa e **[andó]**[cat] nel bosco e **[sentì]**[cat] qualcosa e era **[un lupo]**[wolf] ma **[voleva mangiar[lo]**[cat]**]**[wolf], ma **[il gatto]**[cat] corre corre e **[entrò]**[cat] in casa **[sua]**[cat]. **[Il lupo]**[wolf] buttò giù la porta, e **[il gatto]**[cat] corre ma **[si nasconde]**[cat]. **[Il lupo]**[wolf] **[lo]**[cat] vede e poi **[il gatto]**[cat] e **[il lupo]**[wolf] diventarono amici e **[fanno]**[cat,wolf] tutto insieme come: giocare, mangiare, correre, pescare, parlare, camminare.

CE2, 847

{[La strega][witch] e **[la gatta misteriosa]**[cat]**}**

C'era una volta **[un gatto]**[cat] **[che]**[cat] viveva nel bosco davanti a una casa abbandonata.

Dove c'era dentro **[una strega di nome Malefica]**[witch].

Un giorno **[Malefica]**[witch] decise di fare una torta ai mirtilli. Quindi **[decise]**[witch] di andare nel bosco vicino a casa sua a raccogliere un po' di frutta.

[Arrivati][witch] alla radura. **[Malefica]**[witch] incominciò a raccogliere le bacche per fare la torta. Dopo un po' **[ritorna]**[witch] a casa e inizia a preparare la torta. Una volta finita la **[mise]**[witch] sopra il davanzale della finestra per farla raffreddare. Poi **[incominciò]**[witch] a pulire e **[sentì]**[witch] un rumore. **[Si girò]**[witch] e **[vide]**[witch] **[un gattino]**[cat] **[che]**[cat] stava mangiando la torta. **[Era]**[cat] piccolo. Con due occhi blu enormi. Il pelo nero e **[era]**[cat] dolce come un coniglietto.

[Malefica][witch] **[lo]**[cat] prese e **[lo]**[cat] **[chiamo]**[witch] **[Pilù]**[cat].

[Pilù][cat] era freddo e piccolo quindi **[Malefica]**[witch] decise di adottar**[lo]**[cat] e da quel giorno **[Malefica]**[witch] si prenderà cura di **[Pilù]**[cat].

CE2, 850

{[La strega][witch] e **[il lupo]**[wolf]**}**

C'era **[una strega]**[witch] **[che]**[witch] stava pulendo la casa e il balcone però **[la strega]**[witch] aveva i poteri, però **[usava]**[witch] i poteri per spostare la scopa. Dopo **[la strega]**[witch] stava dormendo e poi arrivo **[un lupo]**[wolf] e **[aprì]**[wolf] la casa e **[la strega]**[witch] si svegliò e **[il lupo]**[wolf] stava dormendo in piedi e **[il lupo]**[wolf] dice: " - **[posso]**[wolf] fare un pisolino qui nel **[tuo]**[witch] letto?- "

E **[la strega]**[witch] dice: -si-.

- E allora **[io]**[wolf] lo facevo un pisolino –

[La strega][witch] dice se **[poteva andare]**[witch] a casa di **[lui]**[wolf].

E **[il lupo]**[wolf] dice di sì.

E allora **[lei]**[witch] andava via, nel frattempo **[raccoglieva]**[witch] delle ciliegie e le fragole.

Il giorno dopo **[lei]**witch aveva dormito vicino a un albero, **[aveva visto]**witch una porta arcobaleno e **[il lupo]**wolf nel frattempo stava dormendo tanto tanto tanto tempo. **[La strega]**witch aveva aperto quella porta arcobaleno, dentro c'era la casa **[del lupo]**wolf e **[lei]**witch dice **[al lupo]**wolf se **[potevano fare]**witch,wolf scambio di casa.

E **[il lupo]**wolf dice di sì.

Allora **[loro due]**witch,wolf smontavano la casa e **[scambiavano]**witch,wolf la casa poi **[il lupo]**wolf chiese se **[lo]**wolf **[sposava]**witch! E **[lei]**witch dice di sì.

CE2, 870

Un bel giorno, **[la strega Lina]**witch e **[il [suo]witch gatto nero Jordy]**cat, volevano provare a fare una pozione nuova, perciò **[la strega Lina]**witch andò a prendere tre barattoli e una ciotola piena d' acqua e la **[versò]**witch in uno dei tre barattoli, poi **[prese]**witch una sostanza viola e un' altra verde, le **[mischiò]**witch e uscì una polvere color blu e i capelli di **[strega Lina]**witch diventarono blu mentre **[al gatto Jordy]**cat andò la polvere su per il naso, allora **[la strega]**witch fece un'altra pozione per raffreddore da dare **[al gattino]**cat ma **[Jordi]**cat non accettò.

[La strega]witch allora disse: - O **[mangi]**cat questa oppure o **[tieni]**cat il raffreddore. Allora **[il gatto]**cat lo mangiò e il raffreddore sparì.