

La découverte et la compréhension des profils d'apprenants : classification semi-supervisée et acquisition d'une langue seconde

THÈSE

présentée et soutenue publiquement le 16 octobre 2019

pour l'obtention du

Doctorat de l'Université Vincennes - Saint-Denis

(mention intelligence artificielle et sciences du langage)

par

Marie Durand

<i>Président :</i>	Khaldoun Zreik	Professeur Université Vincennes - Saint-Denis – Paris 8
<i>Rapporteurs :</i>	Barbara Hemforth	Professeur Université Paris Diderot – Paris 7
	Amedeo Napoli	Directeur de recherche CNRS, Université de Lorraine
<i>Examinatrice :</i>	Rebekah Rast	Professeur, Université Américaine de Paris
<i>Directrices :</i>	Isis Truck	Professeur, Université Vincennes - Saint-Denis – Paris 8
	Marzena Watorek	Professeur, Université Vincennes - Saint-Denis – Paris 8

Table des matières

Table des figures

vii

Introduction

Chapitre 1

Conceptions théoriques de l'acquisition d'une langue étrangère

- 1.1 Diversité des approches : une dichotomie? 5
 - 1.1.1 Le fonctionnalisme et la *learner variety approach* 7
- 1.2 Les débuts de l'apprentissage d'une langue étrangère et la morphologie flexionnelle 8
- 1.3 Première exposition à une nouvelle langue et input 11
 - 1.3.1 Apprentissages explicite et implicite 12
 - 1.3.2 Manipulation de l'input 13
- 1.4 La notion de profil d'apprenant 16

Chapitre 2

Projet VILLA

- 2.1 Objectifs et motivation du projet VILLA 19
- 2.2 Participants 22
- 2.3 Comparaison des langues en présence dans l'étude 22
- 2.4 Les caractéristiques de l'input 24
 - 2.4.1 Les séances d'enseignements : *form based* et *meaning based* 25
 - 2.4.2 La transparence et la fréquence des items lexicaux 26
- 2.5 Les tests de langue 26
 - 2.5.1 Niveau phonologique : *Phoneme Discrimination* 27
 - 2.5.2 Niveau lexical : *Word Recognition* 28

2.5.3	Niveau morphologique : <i>Gramaticality Judgment</i> et <i>Oral Question Answer</i>	28
2.5.4	Niveau morphosyntaxique : <i>Picture Verification</i> et <i>Sentence Imitation</i>	29
2.5.5	Observations critiques	31

Chapitre 3

Approches existantes

3.1	Gestion de données comportementales	33
3.1.1	Fuzzification	34
3.1.2	Opérations	36
3.1.3	Défuzzification	37
3.2	L'apprentissage non supervisé : le <i>clustering</i>	38
3.2.1	Notions générales	39
3.2.2	Mesure de distance ou similarité	41
3.2.3	Méthodes classiques	43
3.2.4	Méthodes floues	50
3.3	L'intégration de connaissances	52
3.3.1	La classification semi supervisée	52

Chapitre 4

Modélisation et base de données

4.1	Sources d'influence sur le parcours acquisitionnel dans le projet VILLA	64
4.1.1	La langue maternelle	64
4.1.2	Le type d'enseignement	66
4.1.3	Les caractéristiques de l'input et le type de tâche	69
4.2	Des données à la notion de représentation	72
4.2.1	La base de données	72
4.2.2	Représentation multicritère de données comportementales	75

Chapitre 5

Algorithme de classification semi supervisée spécifique au contexte

5.1	Pré-traitement des données	81
5.1.1	Vers quelles stratégies?	82

5.1.2	Transformations en sous-ensembles flous	86
5.1.3	Manipulations d'expressions linguistiques de jugements de performances	91
5.1.4	<i>Cannot-link</i> : qu'est-ce que la comparabilité?	95
5.1.5	Transformation des données	96
5.2	Les étapes de l'algorithme	100
5.3	Validation de l'algorithme	111

Chapitre 6

Application

6.1	Premier échantillon : Capacité de traitement en LC	114
6.1.1	Création des cores par l'utilisation de l'ensemble C	115
6.1.2	Raffinement des cores : création des clusters	128
6.2	Deuxième échantillon : Quel transfert du traitement vers la production de la morphologie en LC?	135
6.2.1	Création des cores par l'utilisation de l'ensemble C pour le deuxième échantillon	137
6.2.2	De l'intérêt d'une intégration des outliers	147

Conclusion et perspectives

Bibliographie	157
----------------------	------------

Annexe A

Meaning based et form based illustration

Table des figures

2.1	Tests de langue administrés au cours du projet VILLA. Les tests surlignés sont ceux utilisés dans le chapitre 6 d'application de l'algorithme	27
3.1	Spectre des couleurs visibles théorique	35
3.2	Spectre des couleurs visibles naturel	35
3.3	Découpage en SEFs du spectre des couleurs visibles	35
3.4	Différents problèmes d'apprentissage : (a) Supervisé; (b) Partiellement labellisé; (c) Partiellement contraint; (d) Non supervisé	38
3.5	Exemple de classification des méthodes de partitionnement de données	40
3.6	Dendrogramme d'une classification hiérarchique. En rouge les méthodes descendantes et en bleu les méthodes ascendantes	44
3.7	Agglomération hiérarchique.	45
3.8	Illustration du déroulement de l'algorithme des K-moyennes en quatre itérations ([Tan, 2006])	46
3.9	[Tan, 2006] où Eps correspond au rayon choisi	47
3.10	[Wowczko, 2013] <i>Reachability plot</i>	48
3.11	Carte topologique à deux dimensions avec un voisinage d'ordre 1 autour du neurone c. Voisinage carré (rouge) avec 8 voisins et hexagonal (bleu) avec 4 voisins ([Allab et al., 2011])	49
3.12	[Mallapragada et al., 2008] Illustration du clustering basé sur contraintes. (a) Projection en 2 dimensions des vraies classes d'appartenance des objets de la base de données "Diff300" tiré de [Basu et al., 2004b] après réduction de dimension (b) partition après déroulement de la méthode des K-moyennes pour 3 clusters sans utilisation de contraintes (c) & (d) deux partitions obtenues du même ensemble après sélection aléatoire et utilisation de 100 contraintes par paires. La statistique F_1 est un indice de qualité de la partition obtenue [Basu et al., 2004a]	54
3.13	Les différents types d'intégration dans le clustering semi-supervisé. Dans l'ordre, l'intégration de contraintes dans l'algorithme A prédéfini, l'intégration de contraintes dans la définition de la proximité, avant l'application de l'algorithme A quelconque et enfin l'intégration contrôlée par l'algorithme de clustering quelconque A . Tirée de [Sublemontier, 2012].	56
3.14	Illustration de l'importance des objectifs de l'utilisateur dans le processus de partitionnement : Clustering avec $k = 2$ avec trois contraintes en dur. Les contraintes <i>must-link</i> sont représentées par des barres continues et la contrainte <i>cannot-link</i> par une barre en pointillée.	60
4.1	Schéma en étoile représentant la BDD du projet VILLA	73

4.2	Échantillon en 3 dimensions de l'hypercube représentant la BDD du projet VILLA	74
4.3	Représentation de l'apprenant P en 2-tuple linguistique flou 3D.	77
4.4	Sommets et apex du tetrahedron du pattern P	79
5.1	Définition des SEFs S_1, S_2 et $S_3 \in F(X)$ sur l'ensemble de référence X	88
5.2	Définition des SEFs S_1 et $S_2 \in F(X)$ sur l'ensemble de référence X avec μ la valeur milieu théorique et $\epsilon = 0.5$	89
5.3	Résumé des GSMs existants, tiré de [Truck et Abchir, 2014]	94
5.4	Organisation et temps des passations des différents tests lors du projet VILLA	98
5.5	Vue d'ensemble des différentes étapes de l'algorithme.	110
6.1	Influence relative des dimensions de la modélisation dans le processus de création des cores : échantillon 1	116
6.2	Répartition des apprenants entre les différents cores, <i>outliers</i> non inclus	117
6.3	Tableau récapitulatif de la répartition des apprenants dans les trois SEFs (S_1, S_2, S_3) de l'ensemble d'appartenance en fonction de l'attribut considéré (partie 1 pour les niveaux linguistiques, partie 2 pour les caractéristiques de l'input) et selon les différents cores. Échantillon 1	120
6.4	Résultats des apprenants du core 3 à la tâche GJ en fonction du genre de l'item et du cas sollicité. Échantillon 1.	122
6.5	Moyenne et écart type des individus appartenant au core majoritaire en fonction de leur LM pour les 10 tirages aléatoires effectués	126
6.6	Constitution moyenne du core (5) "double" par LM.	127
6.7	Résultats de la méthode iVAT pour les cores 1 à 5 ; la correspondance entre le degré de gris et la distance numérique entre deux objets est légendée sur la droite de chaque core. Échantillon 1129	
6.8	Indice D obtenu pour le core 1 ; fonction NbClust() du logiciel R	131
6.9	Indice D obtenu pour le core 3	132
6.10	Partition des apprenants (référéncés avec leur identificateur)du core 3 pour $c = 2$ avec l'algorithme des K-moyennes	132
6.11	Partition des apprenants (référéncés avec leur identificateur) du core 3 pour $c = 3$ avec l'algorithme des K-moyennes	132
6.12	Partition obtenue après raffinement des cores en clusters. Échantillon 1	134
6.13	Score de comparabilité des outliers temporaires avec les différents <i>cores</i> et clusters.	135
6.14	Influence relative des différents attributs dans la création de l'ensemble de contraintes C . Échantillon 2	138
6.15	Part relative de chaque core et des outliers. Échantillon 2	139
6.16	SEF d'appartenance des apprenants de chaque core pour les 6 dimensions du deuxième échantillon (morpho_p_I : morphologie, production de l'instrumental; morpho_p_N : morphologie, production du nominatif; ms_p_SO : morphosyntaxe, production des phrases SVO ; ms_p_OS : morphosyntaxe, production des phrases OVS ; transf : transfert).	140
6.17	LMs des apprenants du core 1. Échantillon 2	145

6.18	LMs des apprenants du core 2. Échantillon 2	145
6.19	LMs des apprenants des outliers. Échantillon 2	146
6.20	Résultat de la méthode iVAT appliquée aux cores 1 à 2. Échantillon 2	147
6.21	Partition finale. Échantillon 2	148
6.22	Répartition des LMs des apprenants dans la partition finale. Échantillon 2	150
A.1	Différence entre les supports de cours <i>meaning based</i> (première image) et <i>form based</i> (deuxième image). Le genre des entités est ici souligné dans le cours form based.	171

Introduction

Parmi les questionnements des psycholinguistes se trouve la problématique de l'acquisition d'une langue seconde et de la *compréhension fine des mécanismes sous-jacents à l'apprentissage d'une nouvelle langue par un apprenant adulte*. L'apprentissage automatique, contrairement à la programmation traditionnelle, permet de concevoir des programmes agiles de classification, guidés et entraînés par les données et "écrits" par d'autres programmes.

Cette thèse a pour ambition de provoquer un dialogue fécond entre l'apprentissage automatique et la psycholinguistique dans le cas de l'apprentissage d'une langue seconde. Plus précisément, elle prétend permettre, d'une part, corroborer des résultats issus de la recherche en acquisition des langues grâce à de vrais jeux de données (i.e non artificiels) et, d'autre part, de mettre en évidence des mécanismes (ou à tout le moins, émettre des hypothèses) dans l'apprentissage d'une langue seconde, non triviaux pour les psycholinguistes.

Les techniques de classification non supervisée, ou *clustering*, permettent de créer des sous-ensembles homogènes d'objets à l'intérieur d'un plus grand ensemble, et ce sans spécifier de catégories prédéfinies dans lesquelles classer ces objets. Ces techniques permettent de faire émerger des données des patterns, structures *a priori* invisibles mais reposant sur des critères quantitatifs et/ou qualitatifs. Ces critères sont appelés dimensions et décrivent les objets de la base de données. A partir de cette description de chaque objet, un algorithme de classification non supervisé peut évaluer la distance entre chacun et ainsi regrouper ceux qui se ressemblent le plus, et séparer ceux qui s'accordent le moins.

La multitude des approches existantes provient de l'objectif final du clustering qui consiste en l'obtention d'une partition des données utile à l'utilisateur. Elle doit apporter une information nouvelle tout en s'accordant avec les connaissances externes au jeu de données. Lorsqu'il s'agit de partitionner des données touchant à un domaine scientifique particulier, il est en effet rare qu'il n'y ait aucune attente des spécialistes du domaine quant à la classification finale. Ces deux attentes bien que non contradictoires sont parfois délicates à satisfaire. La création de la partition finale repose en effet entièrement sur des comparaisons quantitatives des objets considérés et sur des regroupements fondés sur une fonction objective à minimiser. Ce processus est donc un processus aveugle de création, utile pour éviter les biais de confirmation mais s'achevant parfois sur un résultat obscur pour l'utilisateur, voire contradictoire avec des connaissances *a priori* et externes au jeu de données.

Le développement de la classification semi supervisée, ou clustering sous contrainte, est directement lié au besoin grandissant de trouver les moyens d'utiliser l'expertise existant sur un jeu de données. Bien qu'il soit tout à fait possible qu'un algorithme de classification non supervisée trouve une partition en accord avec les connaissances du domaine ou les attentes des utilisateurs, ces méthodes échouent pour beaucoup des cas les plus intéressants, et souvent les plus complexes, lorsqu'une expertise humaine est encore nécessaire.

Notre travail fait suite à une problématique relevant du domaine des sciences du langage et plus précisément de la caractérisation du processus d'acquisition d'une langue étrangère. Le projet *Varieties of*

Initial Learners in Language Acquisition : Controlled classroom input and elementary forms of linguistic organisation (VILLA) est un projet ANR européen dont l'objectif est d'étudier les stades initiaux d'acquisition du polonais par des apprenants de 5 langues maternelles (LMs) différentes (français, anglais, allemand, néerlandais et italien). Les chercheurs à l'origine du projet se sont ainsi intéressés à plusieurs problématiques, au cœur du domaine de la recherche en acquisition des langues (RAL) :

- L'observation de l'évolution de l'apprentissage du niveau initial de connaissances à 14 heures d'exposition à la nouvelle langue.
- L'étude du traitement de l'*input* (toute exposition, auditive ou visuelle, à la langue étrangère) sur différents niveaux linguistiques (perception, compréhension, analyse grammaticale et production) en relation avec les caractéristiques de ce dernier.
- Le rôle des connaissances initiales dans le traitement de l'*input* : le rôle des caractéristiques typologiques propres de la langue source (dimension translinguistique) ainsi que le rôle des principes universels spécifiques au langage et à la communication.

Ce projet s'inscrit dans une approche fonctionnaliste dont l'étude porte avant tout sur le procédé de mise en relation forme-fonction qu'effectue l'apprenant et l'influence qu'exerce l'*input* et ses différentes caractéristiques sur ce procédé. Cette approche théorique est centrée sur l'étude de l'apprenant et de la manière dont celui-ci concilie l'*input* à ses représentations et conceptions internes de la langue cible, son lecte. Ainsi toute performance de l'apprenant en langue cible est une fenêtre indirecte sur son lecte. L'analyse de ces performances par l'identification des stratégies de réponse mises en place dans des tâches de production en langue cible est ainsi la clé vers la caractérisation de l'évolution du processus d'acquisition de la langue étrangère par l'apprenant.

Les moyens mis en place par les chercheurs du projet VILLA pour collecter des données doivent permettre, ou non, d'observer les effets sur l'acquisition des différentes variables que sont la langue maternelle, le type d'enseignement et les caractéristiques de l'*input*. Une batterie de tests de langues a été administrée aux apprenants à la fin de chaque session d'enseignement. Ils recouvrent les différents niveaux d'analyse linguistique qui permettent la caractérisation du stade acquisitionnel d'un apprenant, à savoir la phonologie, le lexique, la morphologie, la morphosyntaxe et les capacités discursives.

Dans cette thèse nous nous intéressons à la découverte du profil d'un apprenant d'une langue étrangère, à travers une approche empirique et contextualisée de la caractérisation de son parcours acquisitionnel. La notion de profil d'apprenant est traditionnellement fondée sur des mesures de motivation et d'aptitude en langue étrangère. Ces mesures fournissent des données psychométriques généralement recueillies préalablement à l'expérience-même d'acquisition de la langue. Elles relèvent de capacités cognitives et linguistiques, mais ne donnent aucune indication sur les performances réelles des apprenants en langue cible. L'originalité de ce travail est de retravailler le concept de profil d'apprenant d'une langue étrangère en prenant le parti de la nécessité de la création des profils sur une base empirique constituée des productions des apprenants, tout au long du processus d'acquisition. Le profil n'est pas créé à partir des données *a priori* mais à partir des catégories construites au fur et à mesure de l'exposition à l'*input* et des interactions en langue cible.

Les techniques de classification semi supervisée rendent possible cette création, en utilisant les données de la base de données du projet VILLA, et en y intégrant les connaissances issues du domaine de la RAL. L'application d'un tel procédé à un jeu de données réel qualifiant un procédé aussi complexe que l'apprentissage d'une langue, est à la fois utile et complexe. L'intention est de fournir les bases pour la création d'un système dynamique et évolutif de "diagnostic" du profil d'un apprenant, possiblement en tant que module intégré à un système d'enseignement numérique d'une langue étrangère.

Les problématiques soulevées par ce travail relèvent donc d'une approche interdisciplinaire :

-
- Est-il possible de caractériser un apprenant dans toute la complexité de son processus d'acquisition ?
 - Quel accès à son interlangue procurent les données issues des tests en langue cible ?
 - Comment le modéliser formellement en vue d'une manipulation de ce modèle à des fins computationnelles ?
 - A partir d'un tel modèle, dans quelle mesure est-il possible de comparer deux apprenants ?
 - Comment intégrer l'expertise d'un chercheur en acquisition dans un algorithme de partitionnement d'un ensemble d'apprenant ?

Ces questions sont finalement des préalables à la véritable problématique qui se centre autour de l'utilité d'un tel partitionnement, pour la découverte des mécanismes sous-jacents à l'acquisition d'une langue chez des apprenants adultes et son potentiel explicatif de la variabilité constatée dans leurs performances.

La thèse est structurée autour de trois problématiques, à savoir la caractérisation et la modélisation du parcours acquisitionnel des apprenants et leurs comparaisons en vue d'un regroupement en profils. Le chapitre 1 présente les approches théoriques existantes dans l'étude de l'acquisition d'une langue étrangère et les notions clés dont nous aurons besoin pour l'analyse des données. Le chapitre 2 décrit le projet duquel provient la base de données dont nous disposons. Ces deux chapitres permettent de prendre connaissance des motivations théoriques et du contexte pratique du recueil de données que nous souhaitons analyser. Cette connaissance apporte une expertise sur les mécanismes à l'œuvre lors de l'acquisition d'une langue étrangère chez un apprenant débutant et éclaire sur les problématiques actuelles en recherche en acquisition des langues. Le chapitre 3 offre un aperçu de la gestion et de l'analyse de données comportementales par l'utilisation de la théorie des sous-ensembles flous et présente les différentes techniques de classification existantes. Il se centre principalement sur le domaine de la classification semi supervisée, permettant l'intégration de connaissances externes au jeu de données dans le processus de partitionnement par l'utilisation de contraintes.

Dans un second temps nous souhaitons résumer les différentes analyses existantes sur le jeu de données du projet VILLA. Les différentes études s'interrogeant sur les variables influant le parcours acquisitionnel des apprenants sont résumées dans la première partie du chapitre 4. Elles nous permettent d'obtenir une caractérisation globale d'un apprenant. Cette caractérisation nous pousse à l'élaboration d'une modélisation formelle de l'apprenant dont nous présentons plusieurs pistes en deuxième partie de ce même chapitre.

Le chapitre 5 est consacré au développement de la mesure de comparabilité entre deux apprenants d'un point de vue acquisitionniste et à son implémentation dans un algorithme globale de partitionnement en vue de l'obtention de profils d'apprenants. La notion de comparabilité entre apprenant est définie en terme de stratégies de réponses à une tâche en langue cible et permet la création d'un ensemble de contraintes qui seront prises en compte tout au long de la classification.

Enfin, le chapitre 6 permet la validation de l'algorithme par son utilisation sur deux échantillons de la base de données du projet VILLA. Cette validation passe par l'interprétabilité de la partition obtenue en accord avec la recherche en acquisition des langues.

Chapitre 1

Conceptions théoriques de l'acquisition d'une langue étrangère

Sommaire

1.1	Diversité des approches : une dichotomie ?	5
1.1.1	Le fonctionnalisme et la <i>learner variety approach</i>	7
1.2	Les débuts de l'apprentissage d'une langue étrangère et la morphologie flexionnelle	8
1.3	Première exposition à une nouvelle langue et input	11
1.3.1	Apprentissages explicite et implicite	12
1.3.2	Manipulation de l'input	13
1.4	La notion de profil d'apprenant	16

La recherche en acquisition des langues (RAL) recouvre l'étude de l'apprentissage d'une langue seconde (L2) ou langue étrangère (LE) c'est à dire de toute langue autre que la langue maternelle (LM). La RAL est un domaine d'étude d'une grande richesse en approches théoriques, approches développées pour l'étude et la caractérisation du processus d'acquisition d'une L2. Ceci s'explique en partie par le fait qu'elles prennent source dans les théories sur l'acquisition de la LM. Ainsi un certain transfert des conceptions sur l'apprentissage de la LM ont été appliquées avec plus ou moins de succès à l'apprentissage des L2s. Une autre grande partie de l'explication de cette richesse réside dans les différentes manières d'envisager la langue, conceptions différentes entraînant des théories multiples sur son acquisition. Dans cette section nous proposons de présenter brièvement les différentes théories existantes dans la littérature par le biais de la classification proposée par Mac Whinney ([MacWhinney, 2009],[Macwhinney, 2010]) et reprise par Hulstijn ([Hulstijn, 2015]). Ces auteurs identifient deux paradigmes principaux à savoir la pensée des générativistes, prônant l'existence d'une grammaire universelle (GU), et les émergentistes. Cette opposition reflète un débat plus général dans les sciences humaines et sociales concernant l'innéisme des comportements humains. Quelle est la part de l'inné et quelle est la part de l'acquis? Voilà le grand débat qui a animé et anime toujours le champ de la RAL depuis plusieurs décennies. Malgré tout, les recherches les plus récentes tentent de dépasser ce débat purement théorique, en se concentrant sur la validité des approches méthodologiques et des données ainsi produites.

1.1 Diversité des approches : une dichotomie ?

L'un des principaux thèmes de notre travail est l'*input*. Son rôle est crucial pour les études sur l'acquisition d'une langue seconde, car il représente intuitivement la principale source d'information permettant aux

apprenants de structurer et de perfectionner leur L2 en développement. Bien que cette notion a été problématisée et longuement discutée dans un certain nombre d'ouvrages ([Krashen, 1980, Krashen, 1992, Gass, 1997, Carroll, 1999, Carroll, 2001, E. Carroll, 2005, Piske et Young-Scholten, 2008]), nous définissons ici l'input comme n'importe quelle partie de la langue cible à laquelle les apprenants sont exposés, lorsque cette exposition est contextualisée, quel que soit le format de sa présentation (écrit ou audio). Même si toute la littérature en RAL se construit autour de la contribution de l'input à l'acquisition L2, la façon dont cette contribution est mise en œuvre et ses caractéristiques qui influent le plus sur le succès de l'acquisition, font encore l'objet d'un débat animé.

Les tenants de la théorie de la grammaire universelle (GU), introduite dans sa version initiale en 1957 par Chomsky ([Chomsky et Lightfoot, 1957]), postulent l'existence d'une compétence linguistique innée, relativement indépendante des autres fonctions cognitives, configurée minimalement par l'exposition de l'enfant à l'input de sa langue maternelle (LM), langue de son entourage. Leur principal argument repose sur la pauvreté des stimulus que reçoit l'enfant dans sa LM comparé à la complexité grammaticale de cette dernière. Comment pourrait-il alors être capable de maîtriser la LM dans un temps relativement réduit, en produisant des phrases qu'il n'a jamais entendu, l'input étant imparfait et incomplet? Partant de ce postulat, les générativistes adoptent une démarche formelle pour la caractérisation des propriétés de la GU, et des contraintes à l'œuvre lors de sa "reconfiguration" pour l'apprentissage d'une langue étrangère (L2).

Cette notion de reconfiguration a créé des dissensions parmi les partisans de la GU. Si l'apprentissage d'une L2 repose sur la GU comment se fait-il que les résultats des apprenants soient si variés? L'inégalité de réussite dans la maîtrise d'une ou plusieurs langues étrangères est incompatible avec l'idée d'un accès complet et illimité à la GU. Dès lors plusieurs pistes de réponses ont été données par les générativistes.

Comme White et White ([White et White, 2003]) le résumant, les hypothèses des générativistes quant au processus d'acquisition d'une L2 diffèrent autour de la question du contenu de la GU aux tous premiers stades d'acquisition d'une L2. D'une part, certains auteurs arguent que l'interlangue de l'apprenant contient les différentes catégories lexicales de la LM, mais se caractérise par l'absence des catégories fonctionnelles. Leurs arguments se fondent sur l'absence ou la quasi absence de morphologie ou de mots fonctionnels dans les productions des apprenants débutants. C'est le cas de l'hypothèse des arbres minimaux et de la *modulated structure-building hypothesis*. D'autre part, une autre partie de la littérature avance l'hypothèse d'une *full competence*, autrement dit, d'une interlangue au tout début d'acquisition pourvue de toutes les catégories lexicales et fonctionnelles, possiblement transférées de la LM.

Dans les deux cas, le processus d'acquisition suppose une construction *vs.* un réajustement des catégories syntaxiques et de leurs caractéristiques morphosyntaxiques associées en L2. Également, la théorie de la GU se solidarise autour de la question de la pauvreté de l'*input* et de son rôle minimal dans l'acquisition.

L'importance du rôle de l'*input* est une des raisons de la diversification des paradigmes en RAL à partir des années 80. Face aux générativistes défendant une faculté innée, spécifique à l'Homme, et indépendante des autres processus cognitifs, une multitude d'approches voient le jour, ayant toute en commun, entre autres, le rejet de l'existence d'un dispositif inné sous la forme de la GU, partisans d'un rôle plus ou moins déterminant du contexte d'acquisition, de l'*input* et de ses propriétés.

Ces approches regroupées sous le nom de "théorie de l'émergence" permettent de rendre compte de la variabilité des grammaires intermédiaires constatées chez les apprenants, et souhaitent étudier les mécanismes biologiques et statistiques desquels émerge la structure linguistique. En effet, ces approches ont une visée intégrative de l'acquisition d'une langue aux autres processus cognitifs (comme la mémoire, la catégorisation, la perception) et aux variables environnementales, sociologiques, développementales, etc. Ainsi, les mécanismes impliqués dans l'apprentissage d'une langue ne sont pas fondamentalement différents des mécanismes impliqués dans d'autres activités cognitives. Ses partisans suivent la tradition empirico-formelle des sciences et mettent le recueil de données et leurs analyses au cœur du procédé d'interprétation et validation théorique. Cette théorie générale regroupe des approches théoriques et méthodologiques très variées comme le connexionnisme, les modèles constructionnistes, les modèles statistiques *etc.* Pour un

recensement plus détaillé voir [MacWhinney, 2009] et [Rast, 2017].

Avec cette diversification des approches naît la diversification des questions de recherches qui visent désormais à identifier et évaluer l'importance de différentes variables comme partie intégrante d'un système complexe duquel émerge l'apprentissage d'une langue étrangère, par exemple les propriétés de distribution statistique de l'*input*, la question du stade initial de l'apprentissage, l'hypothèse de la période critique, l'importance de la LM et des autres LEs, le contexte d'acquisition/enseignement, le bilinguisme, *etc.* Ainsi, les approches regroupées sous la théorie de l'émergence possèdent malgré tout des différences fondamentales. La variété de base (*basic variety*, BV) observée par Klein et Perdue ([Klein et Perdue, 1997]), bien que définitivement non générativiste, n'accorde pas le même rôle à l'*input* que la théorie *usage based* (basée sur l'usage). Pour cette dernière, les propriétés de distribution statistiques de l'*input* sont mises au centre du processus d'acquisition, tandis que les tenants de l'approche du lecte des apprenants considère ceux-ci comme des communicants expérimentés, possédant une représentation interne du fonctionnement des langues relativement indépendante de la langue cible en cours d'acquisition et de leur langue maternelle. L'apprenant ne procède donc pas seulement à l'analyse des propriétés de distribution de l'*input* mais y transpose ses connaissances méta-linguistiques. En d'autres mots, au lieu d'adopter une vision entièrement *bottom up* (du bas vers le haut) de l'acquisition, l'approche du lecte des apprenants prend le parti d'une interaction entre l'*input* et les représentations internes de l'apprenant sur le fonctionnement d'une langue.

1.1.1 Le fonctionnalisme et la *learner variety approach*

En RAL, dans le texte fondateur de Corder ([Corder, 1967]), la variété des parcours et réussites des apprenants est centrale. Plus spécifiquement, l'auteur postule que les erreurs des apprenants sont le reflet direct des règles internes que l'apprenant a construites, règles portant sur le fonctionnement de la langue cible (LC), c'est-à-dire la langue en cours d'apprentissage. Ce système de règles de l'apprenant constitue un système cohérent et systématique, bien que celui-ci ne soit pas le système de règle de la LC. La propriété de systématisme des règles appliquées par l'apprenant en LC peut être inférée à partir des erreurs systématiques (par opposition à des erreurs ponctuels) commises par l'apprenant. En effet, quand l'apprenant commet plusieurs fois la même erreur, c'est qu'elle est représentative d'un postulat erroné sur le fonctionnement de la LC. Ce système linguistique est idiosyncrasique, c'est à dire propre à chaque apprenant, mais cohérent. Corder postule également que par nature ce système est instable, l'apprentissage étant en cours, et est donc mis à jour à travers l'exposition de l'apprenant à la LC, par l'enseignement (en classe), ou non (bain linguistique). L'instabilité du système en question soulève l'intérêt d'une description pas à pas du parcours acquisitionnel des apprenants. Dans les années 80, une série de projets longitudinaux voient le jour, ayant pour but la description des variétés de productions des apprenants (cf. projets ZISA, Meisel & al. ([Meisel et al., 1981]), projet P-Mol, Klein et Dittmar ([Klein et Dittmar, 1979]), projet ESF, Perdue ([Perdue, 1993])). Ces projets, ayant tous pour objet le recueil de données de productions d'un panel d'apprenants en LC, adoptent des approches théoriques différentes mais cherchent tous à rendre compte du postulat proposé par Corder, à savoir dégager les règles qui régissent les productions des apprenants et leur évolution au fur et à mesure que ceux-ci progressent dans leur maîtrise de la L2.

Parmi eux, Le projet ESF (Perdue, 1993 [Perdue, 1993]), que nous détaillerons en section suivante, s'intéresse à l'instabilité du dialecte idiosyncrasique de l'apprenant et cherche à définir ses états successifs à travers le suivi longitudinal d'immigrés de différentes langues maternelles, poursuivant l'acquisition de la langue de leur différent pays d'accueil. Cette étude translinguistique permet de dégager les universaux de l'apprentissage d'une langue en milieu non guidé (hors de la classe de langue) par des adultes mais également d'identifier l'influence des structures propres de chaque LCs et LMs.

De ces recherches émerge le concept de "lecte des apprenants" (*learner variety approach*), reprenant le principe d'un système idiosyncrasique mais cohérent de Corder, et le développant en identifiant les deux ensembles de facteurs permettant son évolution :

- Un ensemble de facteurs internes, le moteur de l'acquisition (*factors 'pushing' acquisition*), à savoir la motivation, portée par un besoin communicatif;
- un ensemble de facteurs externes, façonnant (*'shaping'*) l'acquisition, à savoir l'*input*, c'est à dire l'exposition à la LC.

L'interaction entre ces deux ensembles fournit à l'apprenant à la fois l'envie/besoin et les moyens de complexifier les structures de ses discours en LC.

L'importance de la motivation n'est pas un élément nouveau mais les auteurs l'abordent du point de vue linguistique en mettant en exergue le besoin communicatif. Cependant, l'étude ESF repose sur des apprenants immigrés cherchant à s'intégrer tant économiquement que socialement dans leur pays d'accueil le besoin communicatif est extrêmement présent, tandis que notre travail se base sur des données d'acquisition d'apprenants en milieu guidé. Comme nous le verrons en section 1.3, l'étude des apprenants *ab initio* en milieu guidé est un domaine encore peu exploré.

L'approche du lecte des apprenants, dans laquelle nous situons ce travail, s'intéresse donc aux questions de l'implémentation et de l'utilisation des connaissances linguistiques. De visée descriptive dans un premier temps, les fonctionnalistes s'attachent à l'étude de la mise en relation par l'apprenant de la forme d'un énoncé avec sa fonction. Les contraintes pesant sur une production d'un apprenant sont considérées comme d'origines multiples et non seulement linguistiques. La mise en relation d'un énoncé, à quel que niveau linguistique que ce soit (morphème, syntactique, discursif), avec un sens, un objectif communicatif, est central à cette approche. Les relations sémantiques et les informations pragmatiques doivent être reliés à des formes syntaxiques pour construire un discours. Nous proposons de décrire les mécanismes permettant à l'apprenant débutant d'effectuer cette mise en relation forme-fonction en LC en présentant les observations issues du projet ESF dans la section suivante.

1.2 Les débuts de l'apprentissage d'une langue étrangère et la morphologie flexionnelle

Au fur et à mesure qu'un apprenant est exposé à des entrées linguistiques dans la LC (*input*) et qu'il est soumis à des sollicitations en LC, il commence à émettre des hypothèses sur son fonctionnement. C'est ce qu'on nomme le lecte de l'apprenant. Le lecte de l'apprenant constitue un système cohérent mais instable. Sa perméabilité permet son évolution sous l'effet de différentes contraintes (notamment l'*input*). Il est donc intéressant d'essayer de retracer cette évolution au cours de l'acquisition d'une LE. Dans un premier temps l'évolution de l'interlangue d'un apprenant a été considérée comme un continuum entre le stade initial d'exposition à l'*input* jusqu'à un stade avancé, potentiellement final, sous réserve de l'acceptation qu'une langue puisse être maîtrisée par un apprenant comme un natif. Ce continuum présuppose une acquisition de la LE assimilable à celle de la LM, ainsi qu'une approche linguistique de la LE par l'apprenant. Les recherches en RAL issues du fonctionnalisme s'attachent à décrire le parcours acquisitionnel de l'apprenant en termes d'étapes qualitativement distinctes, notamment en ce qui concerne les débuts de l'apprentissage d'une LE.

Klein & Perdue ([Klein et Perdue, 1992]) analysent les données longitudinales recueillies au cours du projet ESF ([Perdue, 1993]) et observent que lors des débuts de l'exposition à une nouvelle langue, l'apprenant analyse celle-ci non pas en termes de règles morphosyntaxiques mais selon des principes extra linguistiques. Un apprenant à l'acquisition naturelle non guidée et donc non explicitement exposé à la grammaire de la LC (comme en classe de langue) possède un certain nombre de principes organisationnels discursif régissant la structure de ses productions. En effet, le lecte d'un apprenant débutant semble dénué de grammaire et ne constituer qu'"une collection chaotique de mots". Pourtant Klein & Perdue dégagent trois grands types de principes universels d'organisation du discours chez des apprenants effectuant une tâche verbale complexe. Afin de résoudre une telle tâche l'apprenant doit être à même de fournir de l'information, de satisfaire un but communicatif, en réponse à une question. La tâche verbale complexe à laquelle les auteurs se réfèrent

est la description de scènes d'un film. Cette tâche était effectuée par les participants au projet à différents intervalles, à partir de leur arrivée dans leur pays d'accueil respectif. Le traitement minimal, ou prototypique ([Watorek, 1996]), d'une tâche linguistiquement complexe constitue une interaction entre la résolution de la tâche (réponse à la *quaestio*) et les moyens linguistiques disponibles. Si ceux-ci peuvent être insuffisants pour une réponse grammaticalement appropriée, l'apprenant débutant parvient à satisfaire au but communicatif par la manipulation de l'ordre des mots de la phrase, placés de telle sorte qu'ils répondent à une organisation de l'information universelle. Le choix et l'organisation de l'information dans un discours en fonction du niveau syntaxique, sémantique et pragmatique sont régis par les principes informationnels :

- Controller comes first ("Le contrôleur vient en premier")
- Focus comes last ("le focus vient en dernier")

La notion de topique, définie selon l'approche syntaxique que nous adoptons, évoque les informations déjà connues par l'interlocuteur ou jugées comme telles par le locuteur (cf. "état mental de l'interlocuteur", [Lahousse, 2003]). Autrement dit il s'agit de l'élément sur lequel porte le discours. Sa connaissance préalable par l'interlocuteur est la condition *sine qua none* de l'intégration de nouvelles informations dans l'énoncé. Or, selon le modèle de la *quaestio* de Klein et von Stutterheim ([Klein et Von Stutterheim, 1987]), un énoncé est exprimé pour répondre à une question, implicite ou explicite, qui va ainsi conditionner la structure globale d'un discours, mais également sa structure interne à un niveau plus micro. Ainsi un texte cohérent est constitué d'un topique, induit par la *quaestio*, et suivie de son pendant, l'information (souvent) nouvelle produite, le focus, répondant à la *quaestio*. Cette structure Topique-Focus au niveau d'un énoncé est extrêmement saillant chez les apprenants débutants.

Les constituants d'une phrase, bien que dénuée de morphologie, remplissent un rôle, si non syntaxique, du moins informatif. Par exemple, on ne peut gratifier l'apprenant de l'attribution du statut de sujet à un item non marqué, pas plus que d'objet ou de verbe. Cependant, un item lexical peut se comporter dans les productions de l'apprenant comme un noyau verbal autour duquel se construit la phrase. Les éléments placés avant ce noyau ont le rôle d'agent de la phrase, et ceux placés après ont le rôle de patient, lorsqu'ils existent. En effet, les auteurs définissent comme structure syntaxique de base fréquemment utilisés par les apprenants débutant la structure "NP-V-(NP)", assimilée à une organisation "Agent-Action(-Patient)". Cette organisation "par défaut" de la phrase est également celle que l'apprenant va utiliser dans son analyse de l'input en LC, menant potentiellement à des difficultés de compréhension dans le cadre d'une langue à l'ordre des mots flexibles.

Cet état de compétence est nommé par les auteurs comme la variété de base (*basic variety*, BV). Elle est donc caractérisée par l'absence de morphologie nominale (cas, genre, nombre) et verbale. L'organisation du discours en réponse à des tâches verbales complexes se structurent autour de principes organisationnels universels, le lexique utilisé est faible bien que composé de noyaux de verbe (non conjugués) servant de pivot à l'organisation syntaxique. Cet état de l'interlangue tend à se fossiliser ([Selinker, 1972]) chez les apprenants du projet ESF. C'est pourquoi les auteurs s'y sont particulièrement intéressés. Il s'agirait en effet d'un état d'acquisition suffisamment efficace du point de vue communicatif pour que les participants du projet tendent à ne plus progresser dans la maîtrise de la LC, malgré l'évidente inadéquation entre leurs productions et l'input reçu.

Ainsi, au stade de la variété de base, un apprenant peut communiquer de manière "simple, versatile et efficace" en L2 ([Klein et Perdue, 1997], p.3).

C'est de ce constat que Klein et Perdue se placent pour se questionner sur la complexité potentielle que nous offrent les langues, notamment à travers la morphologie flexionnelle et/ou lexicale. S'interroger sur la nécessaire complexité de la langue c'est également s'interroger sur ce qui pousse les apprenants à dépasser la variété de base et à progresser dans leur parcours acquisitionnel. Celle-ci établie une référence dans les débuts de l'apprentissage d'une langue étrangère car elle constitue un palier de stabilisation probable et attesté pour un tiers des apprenants ayant participé au projet ESF. Cependant, le contexte d'énonciation ainsi que la possible mise en contradiction de certains de ces principes organisationnels de l'information dans

le discours vont plus ou moins altérer leur application dans la production de l'apprenant. C'est en effet de la mise en compétition de ces principes et de leur possibilité de contradiction que va naître le besoin pour l'apprenant de dépasser le stade de la variété de base. Ainsi, l'apprenant va au cours de son acquisition se détacher de ces principes grâce à l'acquisition de la morphologie, lui permettant ainsi de marquer la fonction d'un élément de l'énoncé de manière (morpho-)syntaxique et non plus pragmatique. L'apprenant rentre ainsi dans le stade post-basique.

Parallèlement, il serait possible d'imaginer un état d'acquisition encore antérieur à la variété de base. C'est à partir de cette hypothèse que Perdue ([Perdue, 1996]) tente de caractériser ce qu'il nomme la *pre-BV*. Avant même l'exposition à la LC, l'apprenant possède des connaissances utiles pour son acquisition :

- un savoir encyclopédique
- la capacité à percevoir son environnement et un appareil articulatoire fonctionnel
- l'habileté à découper un flux de parole et à attribuer du sens aux segments

C'est en accord avec ces prémisses que l'auteur en conclut que l'apprenant débute son acquisition d'une LE avec des connaissances sur les catégories cognitives qui sont généralement grammaticalement exprimées dans les langues (temporalité, spatialité, mais aussi agentivité, etc.), mais aussi avec des connaissances ou des procédés internalisés d'organisation de l'information dans différents types de discours (constat plus évident chez l'adulte).

Cet étape d'acquisition se différencie de la BV par l'absence de noyau verbal dans la construction syntaxique. Il se situe avant l'organisation "NP-V(-NP)" constaté en BV et n'est constitué que d'items nominaux ("*Noun based utterance organisation*"). Ainsi, les items utilisés par l'apprenant peuvent remplir le rôle grammatical d'un verbe ou pronom personnel alternativement et sont des mots spécifiques à l'apprenant les utilisant. Il s'agit en fait de découpe du flux de parole en LC ("*phonetic matrix*"), extrait du contexte, et réutilisé en bloc non analysé (souvent constitué d'un déterminant et d'un item nominal regroupé). Ces items sont arrangés sémantiquement selon le pattern Topic-Focus lors des productions de l'apprenant qui sont donc généralement constitués de deux éléments.

Ainsi, les étapes initiales de l'apprentissage d'une LE sont majoritairement dominés par des principes universels dits *language neutral* (neutre vis-à-vis de la langue), l'influence des spécificités de la langue maternelle de l'apprenant et de la langue cible se manifestant plus tardivement dans l'acquisition. La mise en concurrence des principes universaux évoqués dans certaines situations de production amène l'apprenant à réaliser la limite d'expressivité qu'ils permettent et à revoir son lecte, afin de les dépasser et de complexifier et enrichir ses moyens linguistiques ([Klein et Perdue, 1997]), sortant ainsi de la *basic variety* et entrant dans la *post-BV*, aussi appelée la "*finite utterance organisation*" (organisation basée sur les formes fléchies), caractérisée par la sensibilité de l'apprenant au système morphosyntaxique de la LC, et visible dans ses productions.

En milieu guidé, la continuité observée entre ces différents stades d'organisation de l'interlangue doit être remise en cause. En effet, l'enseignement des règles morphosyntaxique, mises en exergue en classe de langue, est complètement absent d'un input naturel tel que reçu par les immigrés du projet ESF. De plus, les tâches auxquelles l'apprenant devait participer étaient des tâches verbales complexes de production. Ces tâches ne permettent pas d'avoir accès à un éventuel traitement de la grammaire de la LC effectué par les apprenants dans leur analyse de l'input. Ainsi, la manipulation du contexte des premières heures de l'acquisition d'une langue étrangère permet l'examen de plusieurs hypothèses. Le premier aspect intéressant d'une recherche en environnement contrôlé (et donc forcément moins écologique) est la possible mise en relation directe des résultats des apprenants avec les caractéristiques de l'input.

1.3 Première exposition à une nouvelle langue et input

Des tâches linguistiques ciblées permettent l'évaluation des capacités de l'apprenant à traiter la LC (perception, compréhension), capacités généralement antérieures à leur équivalent en production. En effet, l'élaboration de la notion de la variété de base repose sur l'analyse de réponses à des tâches linguistiquement complexes (description d'image, de séquence de film), or, dans ces tâches le but communicatif est naturellement prioritaire à la forme de l'énoncé. Ainsi, elles ne permettent pas d'observer dans les stades initiaux d'acquisition, l'étendue des connaissances et hypothèses de l'apprenant sur le fonctionnement de la LC. Pourtant, par la manipulation de l'input, il est possible de mettre en évidence une sensibilité précoce de l'apprenant à la morphologie d'une langue étrangère. Par exemple, Dimroth ([Dimroth, 2006]) démontre que l'association son-sens est possible après quelques heures seulement d'exposition si la fréquence d'un pattern est associée à une mise en exergue (*highlighting*) gestuelle.

L'apprenant s'appuie donc sur le contexte d'énonciation de l'input pour le caractériser. Quelles autres sources d'information l'apprenant utilise-t-il au tout début de l'acquisition? Quelles sont les "bagages", linguistiques et autres, dont l'apprenant dispose avant même toute exposition à la LC, et quelles stratégies d'analyse de l'input en découle-t-il? Ce sont les questions que Rast ([Rast, 2008]) énoncent dans son ouvrage sur l'input et les premières heures d'exposition à une LE. Ces questions recouvrent la notion du *prior system* (système antérieur) évoqué par Giacobe ([Giacobe, 1992]). Dans un premier temps les auteurs de la RAL n'attribuait à ce système que la fonction de filtre de la L1 sur toutes LE potentielles (théorie du transfert). Pourtant cette approche de filtre ne permettait pas de rendre compte de la créativité des apprenants dans leurs productions en LC ([Sharwood-Smith, 1986]). En effet, ce n'est pas tant les caractéristiques de la L1 mais les différences typologiques entre la LM et la LC, et plus encore la différence typologique perçue par l'apprenant, qui va conditionner le transfert de structures de la LM à la LC. De plus, la théorie du transfert suppose un transfert de structures de la LM telles que définies par des linguistes, et non par des apprenants. Plus encore, ce filtre ne concerne pas seulement la LM puisque de nombreux travaux décrivent l'influence des autres LEs de l'apprenant dans son approche de la LC. Hendriks et Prodeau ([Hendriks et Prodeau, 1999]) démontrent qu'un français ayant appris l'allemand à un niveau avancé place de manière appropriée le verbe en deuxième position (V2) en néerlandais dès les débuts de son acquisition (40h) contrairement à un français ayant appris l'anglais, différence expliquée par le fait que le V2 existe également en allemand et non en anglais. Rast, toujours dans son ouvrage, va plus loin et analyse les performances d'apprenants français du polonais dans une tâche de traduction. Ces apprenants français ont soit appris (en contexte institutionnel) le russe, l'allemand, une autre langue romane (italien ou espagnol), soit une langue romane et une langue germanique (hors anglais). L'auteur retient deux facteurs majeurs de la réussite à cette tâche au cours des 8h d'exposition au polonais. Premièrement, les mots les plus correctement traduits sont les mots à la plus grande transparence avec le français. Deuxièmement, les apprenants ayant appris le russe ou l'allemand traduisent plus facilement des mots dont la transparence est élevée avec leur L2. De plus, en l'absence de connaissance sur le système flexionnel riche de la LC, ces apprenants s'appuient tout de même sur leur connaissance du système flexionnel de leur L2 pour la recherche d'indices morphologiques dans les mots à traduire. Ainsi, bien que le système polonais est doté d'un système morphologique plus complexe que l'allemand ou le russe, les apprenants opèrent une stratégie d'analyse de la fin de mots afin de catégoriser les mots en polonais (adverbe, nom etc.) ou d'effectuer un accord en genre. L'auteur conclut ainsi que des apprenants à "L1 ou L2 à fortes flexions seraient plus sensibles aux marqueurs morphologiques que ceux n'ayant aucune connaissance de ce type de langue" ([Rast, 2008], p.99).

La L1 ne constitue donc pas le seul élément dont dispose l'apprenant pour appréhender une LE. Au-delà de ses connaissances linguistiques d'autres langues, celles-ci ont également potentiellement enrichie ses connaissances métalinguistiques sur le fonctionnement d'une langue, comme en atteste la recherche d'indices grammaticaux en fin de mot par certains apprenants. Cette stratégie déjà présente chez les apprenants allemands du projet sur lequel se base cette thèse, n'apparaît d'ailleurs que plus tardivement chez les apprenants d'autres LMs comme nous le verrons par la suite. Les connaissances méta-linguistiques

constituent une grande part des connaissances utiles d'un apprenant face à la résolution d'une tâche dans une nouvelle LE. Ces connaissances méta-linguistiques vont influencer sur les stratégies mises en place par l'apprenant pour apprendre une LE, au-delà de la distance typologique (actuelle ou perçue) entre la LM de l'apprenant et ses éventuelles L2s avec la LC.

1.3.1 Apprentissages explicite et implicite

Pour mettre en évidence le rôle de ces connaissances, les tâches *word order test* et *grammaticality judgment task* ont été choisies pour leur nature hautement méta-linguistique par Rast dans son recueil de données. La tâche *grammaticality judgment* a également été utilisée par Kachinske et collaborateurs ([Kachinske et al., 2015]) afin de mettre en évidence le rôle de l'apprentissage implicite dans l'acquisition d'une LE artificielle. L'apprentissage implicite est simplement défini par les auteurs comme l'apprentissage non intentionnel d'un item par un apprenant sans que celui-ci en soit conscient, ou par Ellis comme "*the acquisition of knowledge about the underlying structure of a complex stimulus environment by a process which takes place naturally, simply, and without conscious operations*" ("l'acquisition de connaissance sur la structure sous jacente d'un stimulus complexe par un processus naturel, simple et sans opérations conscientes." [Ellis, 1994], p. 1). L'usage d'une langue artificielle comme objet d'acquisition enlève une certaine validité écologique à l'étude mais permet de s'affranchir de la complexité d'une langue naturelle et ne garder comme caractéristique que celle d'intérêt pour l'étude. Dans cette étude, la structure d'intérêt concerne l'ordre des mots d'une phrase.

Le débat sur un apprentissage implicite d'une L2 concerne principalement la notion (et la mesure) de "conscience". Afin de vérifier la nature implicite d'une règle, les chercheurs du domaine utilisent plusieurs procédés. Certains demandent explicitement aux participants d'exprimer la règle qui leur a permis de résoudre la tâche après la réalisation de celle-ci ou demandent aux participants de "penser à voix haute" lors de sa réalisation. Une mesure directe consiste à évaluer l'apprentissage effectif d'un participant sur une structure syntaxique tandis que lors de la phrase d'apprentissage celui-ci est explicitement orienté sur le sens de l'input reçu. Kachinske et collaborateurs suppose ainsi la présentation de l'input comme modificateur de la conscience de l'apprenant des formes de la LC, ce qui pose directement la question de la méthode d'enseignement des règles morphosyntaxiques d'une LE, question reprise dans le projet VILLA (*form focused vs. meaning based*, cf. chapitre 2).

Ainsi, la manipulation de la présentation de l'input effectuée par Kachinske et collaborateurs, explicite ou non vis-à-vis de la structure morphosyntaxique d'intérêt, permet de s'intéresser à ce que la méthode d'enseignement a comme influence sur la rétention de cette structure.

La règle que les participants devaient induire de la phase d'entraînement consistant en l'écoute de 132 phrases était que le déterminant d'un nom doit être placé en deuxième position. Trois catégories de phrases étaient ainsi construites : Nom-Déterminant ; Adjectif-Nom-Déterminant ; Adjectif-Déterminant-Adjectif-Nom. Les adjectifs, noms et autres items des phrases étaient dans la LM des participants, l'anglais. Seules les deux premières catégories étaient présentes dans la phase d'entraînement, les trois dans la phase de test, afin de tester un transfert éventuel de la règle sur un nouveau type de phrase. Les deux groupes, implicite *vs.* explicite (*incidental vs. intentional*), recevaient des consignes différentes. Un groupe contrôle sans entraînement préalable était également testé. Il était demandé aux participants du groupe implicite de faire attention au sens des phrases lors de l'entraînement, alors qu'à l'autre groupe avait été demandé de trouver la/les règle(s) régissant la grammaticalité des phrases présentées. L'objectif étant d'observer si l'attention portée sur la forme *vs.* sur le fond entraîne une différence. Les individus du groupe implicite et explicite ont obtenus de meilleur résultat que le groupe contrôle pour les phrases de type Nom-Déterminant, suggérant un apprentissage de la règle morphosyntaxique dans sa version la moins complexe après seulement 132 phrases d'input dans la langue artificielle.

Ce résultat concerne directement la détection, incidente ou non, de régularité dans l'input. Cette capacité est très étudiée en RAL, originellement par Saffran et collaborateurs ([Saffran et al., 1996]), sous le terme d'apprentissage statistique défini comme "une approche considérant l'acquisition d'une LE comme l'absorption passive et non sélective des régularités statistiques de l'environnement" ([Schmidt, 2010], p.7). Les résultats de ce domaine pose la question plus globale de l'importance de l'input et de ses caractéristiques dans l'acquisition d'une LE. Dans l'étude de Rast déjà mentionnée, l'effet de la transparence des items ou d'une structure morphologique (avec la LM ou les autres L2s) sur leur rétention a été abordée. La notion d'apprentissage statistique statue sur l'intérêt de la fréquence d'une structure dans l'input. L'exposition répétée aux régularités du langage permet l'émergence du système de la LC, non sous forme de règles, mais sous forme de probabilités contingentes. Cette théorie possède un fort pouvoir explicatif et est actuellement à l'origine des meilleurs modèles de traitement automatique du langage naturel (réseaux de neurones), lorsqu'une quantité suffisante d'input (pour le modèle) est disponible.

1.3.2 Manipulation de l'input

Nous avons déjà évoqué le fait que le lecte était un système grammatical, cohérent et dynamique car instable. Le lecte de l'apprenant est en effet issu du traitement de l'input par l'apprenant ainsi que des tâches communicatives auxquelles il prend part. Il est ainsi légitime de se poser la question de l'influence des caractéristiques de l'input sur la création du lecte d'un apprenant et donc sur son acquisition. Que ce soit en termes d'étendue, de qualité, de spécificité, ou de quantité, il existe de nombreuses études sur le lien entre les modalités de l'input et les productions des apprenants.

En ce qui concerne la quantité globale d'input et son effet sur l'acquisition, il y a une dizaine d'années le constat de Flege ([Flege, 2009]) est que l'input en L2 n'a jamais été vraiment mesuré, mais seulement estimé sur la base des auto-évaluations des apprenants étudiés. Ce problème méthodologique est directement lié à la difficulté de contrôle pouvant être exercé sur l'input de l'apprenant, impossibilité même dans le cas d'un apprentissage en milieu non guidé. Ce constat a été à l'origine du regain d'intérêt pour les études sur les premières heures d'exposition à la LC ([Gullberg et al., 2010]; [Myles, 2012]; [Rast, 2008]; [Shoemaker et Rast, 2013]; [Geertje van Bergen et al., 2014]). Partir du stade initial d'acquisition et fournir l'input dans un environnement contrôlé sont les deux prémisses nécessaires à la mise en perspective directe entre input et gain d'acquisition. Ainsi, les études sur les débuts de l'apprentissage d'une langue offrent un aperçu direct des liens entre quantité et qualités d'input.

VanPatten ([VanPatten, 1996, Van Patten, 2004]) définit deux principes fondateurs du traitement de l'input par des apprenants :

1. La primauté du sens : les apprenants traitent l'input à la recherche d'un sens avant la recherche d'une forme.
2. Le nom en premier : les apprenants ont tendance à attribuer au premier item nominal d'une phrase le statut d'agent.

Ces principes (déclinés en sous principes) recourent les travaux de Klein et Perdue déjà évoqués avec la variété de base et soulignent, dans le cadre de l'étude d'apprenant *ab initio*, l'importance du sens sur la forme. Ces différentes recherches sont autant d'arguments à une approche *meaning based* du traitement de l'input, comme d'une approche par défaut. La conséquence didactique supposée serait donc de privilégier un enseignement également *meaning based*. En effet, pour revenir à l'étude de Kachinske et collaborateurs, la seule différence constatée entre le groupe implicite et le groupe explicite se situe sur un résultat légèrement meilleur des apprenants du groupe explicite sur les phrases de structure plus complexe Adjectif-Nom-Déterminant. Aucun des deux groupes ne montrait de meilleur résultat sur les phrases de la catégorie n'ayant

pas été présentes dans l'input (Adjectif-Déterminant-Adjectif-Nom). Toutefois, la manipulation de l'attention de l'apprenant sur le sens *vs.* la forme de l'input n'apporte pas d'information directe sur l'influence du mode d'enseignement/présentation de la LC.

Wong ([Wong, 2004]) énonce ainsi qu'une structure grammaticale ne pourra être acquise que par l'intermédiaire de contenu propositionnel dont l'attribution d'un sens nécessite une structure grammaticale. La méthode d'enseignement *processing instruction* (enseignement par traitement, PI) est ainsi étudiée par plusieurs chercheurs regroupant leur résultat dans le livre de VanPatten ([VanPatten, 2004]), et comparée à d'autres méthodes d'enseignement comme *Focus on Form*, *Focus on Forms*, et *Focus on Meaning* ([Collentine, 2004]). La méthode PI se caractérise par un enseignement purement basé sur le sens et se rapproche en cela de la méthode *Focus on Meaning* assimilable à l'approche communicative dans sa version la plus radicale. L'approche *Focus on Forms* se rapproche de l'enseignement grammatical dit traditionnel, qui ne fait pas une priorité de la compétence communicative des apprenants et du besoin de sens et de contexte pour l'apprentissage d'une forme. L'approche *Focus on Form* a été créée par Long ([Long, 1991], puis grandement étudiée en vue de trouver un équilibre entre les deux méthodes dernièrement énoncées. L'objectif théorique de cette approche est de privilégier les activités communicatives et d'utiliser les occasions incidentes au cours de ces activités pour expliciter un point de grammaire en contexte. Ainsi l'explicitation grammaticale devient utile au but communicatif de la tâche.

Cette approche a émergé notamment du constat que l'approche purement basée sur le sens pouvait ne pas permettre d'atteindre une compétence en LE proche de celle d'un natif ([Genesee, 1987, Swain, 1985]), mais s'est vu conforter par les nombreuses études des résultats positifs obtenus auprès des apprenants l'ayant expérimentée (par exemple [Long et Doughty, 2011, Williams, 2001, Tian, 2011]). La méthode consiste en des activités communicatives où de l'information linguistique est fournie ou élicitée par le professeur en réponse à une production de l'apprenant ou directement via une reprise grammaticalement corrigée de l'énoncé de l'apprenant.

Par exemple, l'étude menée par Loewen ([Loewen, 2005]) a pour objectif l'évaluation de la rétention de formes ayant été mise en avant lors de tâches communicatives. Pour cela Loewen a extrait un certain nombre de ces formes travaillées pour tester leur acquisition par 118 apprenants de l'anglais sur deux périodes de temps, quelques jours suivant leurs occurrences et deux semaines après. Les apprenants venaient de Corée, de Chine, du Japon et de Taiwan et possédaient un niveau en anglais considéré par l'auteur comme allant de bas à intermédiaire. Trois types de tests ont été développés afin de tester l'acquisition produite par les différents types de *focus* opérés en classe : un test de définition pour les précisions fournies sur le sens d'un mot ; un test de jugement de grammaticalité et de correction pour les élicitations des formes incorrectes produites par l'apprenant ; un test de prononciation pour les items incorrects produits par l'apprenant. Chaque apprenant était testé individuellement et uniquement sur les formes pour lesquelles celui-ci avait soit reçu un *feedback* pendant les activités communicatives, soit demandé une information linguistique. Si l'on considère les résultats des apprenants quel que soit le test (et donc le type de focus) considéré, les formes travaillées incidemment lors d'épisodes communicatifs ont été restituées correctement à hauteur de 50% quelques jours après et à hauteur de 40% deux semaines plus tard. Cette étude va au delà de la visée descriptive d'autres études dans le sens où elle permet de tester les apprenants directement sur les formes individuellement travaillées, c'est à dire où un échange a eu lieu entre l'apprenant et le professeur sur une forme spécifique et donc où l'apprenant a fait montre de l'inadéquation de son utilisation avec son utilisation appropriée. Dans le projet sur lequel se base cette thèse, deux méthodes d'enseignement étaient employées. La première était entièrement *meaning focused* tandis que la deuxième contenait des mises en évidences de formes (visuellement et dans l'input du professeur) ainsi que des techniques d'élicitation et de reprises corrigées d'énoncés des apprenants. Cependant il est important de souligner que les apprenants de l'étude de Loewen étaient à un stade bien plus avancé d'acquisition que celui de nos participants et également que l'étude de Loewen a directement lié la sensibilité des apprenants au travail incident des formes à ce qu'il nomme le *successful uptake*, c'est à dire à l'intégration réussie et directe de l'apprenant de la correction ou information

du professeur dans son discours au cours de la tâche communicative.

De plus, une revue de la littérature couvrant les années 1980 à 1998 n'a pas relevé de différence de gain en termes d'acquisition entre les approches *Focus on Forms* et *Focus on Form* ([Norris et Ortega, 2000]). Une autre plus récente ([Ellis, 2016]) considère même que plus qu'une approche, le terme *Focus on Form* désigne aujourd'hui un ensemble d'activités ou procédures en classe. Cependant, ce qui nous intéresse particulièrement est l'impulsion théorique à l'origine de cette méthode d'enseignement qui se situe autour de l'hypothèse *Noticing* et de l'hypothèse interactionniste.

L'hypothèse interactionniste initiée par Krashen ([Krashen, 1980]) avec l'*input hypothesis* mais établie par Long ([Long, 1983]) et consolidée par Ellis ([Ellis, 2009]) reprend l'idée que l'input doit avant tout véhiculer du sens et que l'interaction en LC est primordiale, notamment lorsque celle-ci incite l'apprenant à modifier ses productions. Ainsi, la problématique de l'acquisition en LE se déplace sur les facteurs poussant l'apprenant à modifier ses productions, et la solution se centre toujours autour de l'input. La question est donc : comment faire en sorte que l'input reçu permette la modification des productions des apprenants, et donc sous-entend sa représentation du fonctionnement de la LC ?

Schmidt ([Schmidt, 2010]) considère que l'acquisition doit passer par l'attention à des formes spécifiques de la langue pour leur rétention. L'approche réfère ainsi à la mise en relation d'une forme et de sa fonction. Afin d'autoriser la rétention de formes spécifiques, il est nécessaire de mettre en place les éléments nécessaires au "*noticing*" de ces formes dans l'input par l'apprenant. En effet, Schmidt, dans l'étude de sa propre acquisition du portugais ([Schmidt, 1990]) souligne qu'il ne maîtrisait pas certaines des formes fréquentes dans l'input reçu. C'est seulement après avoir consciemment remarqué ces formes que Schmidt a pu les retenir et les réutiliser. Ce constat est à l'origine de l'hypothèse *noticing*. Ainsi, la distinction entre l'input, le matériel en LC auquel l'apprenant est exposé, et l'intake est devenue un sujet d'étude central.

En effet, le fait de présenter un item à l'apprenant ne garantit pas son intégration (*what comes in*, [Corder, 1967], p.165). L'input n'est ainsi que ce que l'apprenant voit ou entend en LC. L'intake est également considéré comme un processus plus qu'une connaissance, assimilable à la notion de compréhension. Ainsi, l'intake correspond à une représentation correcte du fonctionnement de la LC directement issue de l'exposition à l'input. Hatch ([Hatch, 1983]) souligne qu'un élément en LC auquel est exposé l'apprenant n'est véritablement de l'input que si l'apprenant essaye (à minima) de le traiter. Cela implique que l'apprenant doit y prêter attention.

L'action de "noter" une forme, un item, *etc.* dans l'input est un processus défini par Schmidt comme "un terme technique limité à l'enregistrement conscient de structure spécifique de la langue" ([Schmidt, 2010], p.5). On peut ainsi l'opposer directement à l'apprentissage implicite qui est surtout caractérisé par l'inconscience d'un tel processus. Cependant, Schmidt souligne ainsi que "faire attention" délibérément à certaines formes est indispensable pour l'acquisition, spécifiquement lorsque ces formes ne sont pas saillantes, mais également que l'apprentissage implicite ne résulte pas en connaissance implicite. Il distingue également le fait de "noter" du fait de "comprendre", ce dernier supposant l'abstraction d'une règle générale. L'hypothèse *noticing* n'apparaît donc pas en contradiction avec l'approche de l'apprentissage statistique mais plutôt comme un complément de celle-ci.

Le fait est que dans la littérature en RAL se côtoient les études de nombreuses caractéristiques de l'input comme facilitatrices, ou inversement, de l'acquisition. Mais la problématique vers laquelle l'hypothèse *noticing* nous pousse semble être tout simplement la saillance de l'input comme facteur de l'intake par l'apprenant. La saillance caractérise des parties d'un stimulus qui se démarquent du reste (*stand out*, [Cintrón-Valentín et Ellis, 2016]) et qui sont donc plus à même d'être intégrées par l'apprenant, et apprises.

Slobin ([Slobin, 1985]) indique, pour l'acquisition de la L1, que la notion de saillance s'applique en premier lieu à la perception phonétique, que ce soit pour l'extraction d'un bloc de langage (*chunk*), ou d'une syllabe. Il indique également qu'au niveau du mot, la partie première et la fin du mot sont saillantes. Rast et Dommergues ([Rast et Dommergues, 2003, Rast, 2008]) rajoutent qu'en acquisition L2 la distance phonémique entre la L1 de l'apprenant et la LC rend les caractéristiques topologiquement éloignées plus difficiles à répéter par l'apprenant. La transparence lexicale, comme nous l'avons déjà évoqué, a également un effet sur la rétention de l'apprenant et l'utilisation, appropriée ou non, des structures transparentes. Au niveau syntaxique, les positions initiale et finale d'un item facilitent également leur prise en compte par l'apprenant. Dans l'étude de Kachinske et collaborateurs déjà évoquée, les apprenants, du groupe implicite comme du groupe explicite, obtenaient de meilleurs résultats en jugement grammatical pour la structure morphosyntaxique (déterminant en deuxième position) la moins complexe de l'input (Nom-Déterminant *vs.* Adjectif-Déterminant-Nom). Ce résultat peut être attribué au résultat de Rast et Dommergues sur la saillance d'un item en fonction de sa position dans la phrase. Plus récemment, Rast et collaborateurs ([Rast et al., 2018]) se sont intéressés aux différences de saillance entre différentes marques morphologiques situées en fin de mot. Ainsi, pour un même emplacement, les différences d'utilisation appropriée d'un suffixe marquant le statut de l'objet ou du sujet dans la phrase sont attribuables à des caractéristiques phonologiques du suffixe, mais également à leur régularité dans l'input, et à leur faible niveau d'ambiguïté fonctionnelle (un suffixe pour une fonction, pour plus de détails sur cette étude voir le chapitre 4 section 4.1.3). Ainsi la saillance concerne plus que la perception de la langue (distance phonémique et transparence) mais est également impactée par la complexité syntaxique, et par son système grammatical, l'association forme-fonction.

Ainsi, le lien entre *noticing* et l'acquisition n'est pas directement observé, la saillance de l'input ne garantissant pas son appropriation par l'apprenant. Ce processus serait donc nécessaire mais non suffisant. La question est donc de savoir quels sont les mécanismes et les stratégies associées qui sont en place dès les stades initiaux, et quel type ou étape d'intake en résulte ([Rast, 2010]). Aux toutes premières heures d'exposition à une LE, l'apprenant utilise-t-il les mécanismes d'apprentissage statistique pour le développement de connaissance pattern-spécifique (via l'extraction de chunks par exemple); les caractéristiques de l'input lui permettront-elles de dépasser l'exposition passive à celui ci pour "noter" certaines formes de l'input; l'apprenant arrivera-t-il à généraliser certains patterns en règle et à l'appliquer à des nouveaux items?

1.4 La notion de profil d'apprenant

Toutes les variables que nous venons d'évoquer sont autant de dimensions pertinentes dans l'appropriation d'une LE, et de ce fait il est intéressant de les manipuler pour observer, et si possible quantifier (ou plus généralement décrire), leur effet sur l'acquisition. Le recensement de toutes ces variables a pour but originel l'explication d'un constat sans appel, à savoir les variations dans le niveau d'acquisition des apprenants d'une LE, et la faible proportion de ces apprenants à atteindre un "niveau natif".

Afin de caractériser un phénomène, et d'expliquer la variation de ses manifestations dans une population, la recherche en sciences sociales est traditionnellement abordée de deux points de vue. Dans le premier, la recherche se centre autour de la découverte de tendances générales. Il s'agit de la modélisation d'un phénomène complexe ramené à un ensemble de variables interagissant entre elles au travers des procédures expérimentales sur des échantillons de la population d'intérêt. Un exemple de cette approche sera développé dans le chapitre 2 décrivant le projet VILLA, à l'origine de la base de données utilisée dans cette thèse.

L'autre approche est l'étude de l'individu, du cas d'étude. Elle permet une description plus fine des comportements et est souvent une étude longitudinale, difficile à mettre en place dans le cas d'échantillon important de la population. Cette technique d'étude est très intéressante pour la découverte des variables d'intérêt dans la caractérisation d'un phénomène, ainsi que pour le développement d'intuitions et la

construction d'hypothèses sur l'influence de ces variables. Seulement, les constats observés sur un individu ne sont pas généralisables à toute une population et ne possèdent donc pas un grand potentiel d'abstraction.

C'est pourquoi le chercheur Véronique argue que les deux objectifs de ces techniques, loin d'être contradictoires, sont complémentaires : "la dialectique de l'individuel et du social est un topos de la recherche en sciences sociales" ([Véronique, 1994], p.3). Cette dialectique amène naturellement à rechercher une synthèse des deux extrêmes que l'on trouve dans la notion de profil. Le profil, au sens de la découverte de régularités dans des sous groupes de la population, possède un fort potentiel explicatif dans les variations de comportement d'une population face à un même phénomène, dans les mêmes conditions de manifestation. Cependant, la construction de profil d'apprenant d'une LE est problématique.

Dans la littérature, la notion de profil d'apprenant est en effet un statut individuel de l'apprenant construit indépendamment de son acquisition. Carroll ([Carroll et Sapon, 1955]) cherche à mesurer une "aptitude en langue étrangère". Cette aptitude doit être prédictive de l'acquisition d'une L2, procède donc d'un état interne initial de l'apprenant, *a priori* de toute exposition, et regroupe enfin un ensemble de capacités psycholinguistiques mesurables, d'habiletés générales dans l'apprentissage d'une langue. Cette vision de Carroll est donc une vision figée de l'aptitude d'un apprenant à apprendre une langue étrangère, basée sur un score individuel unique quel que soit la LC considérée. Le *Modern Language Aptitude Test* (MLAT) développé par Carroll ([Carroll, 1965], [Carroll, 1981], [Carroll, 1990] et [Carroll et Sapon, 1955]) cherche à mesurer cette aptitude en LE. Il teste un ensemble de 4 composantes :

1. La capacité de codage phonémique
2. la sensibilité grammaticale
3. la capacité d'induction linguistique
4. la mémoire associative

Le modèle sur lequel se base ce test a été largement étendu et revisité depuis, et encore récemment, par le Modèle LCDH ([Sparks et Ganschow, 2001]), la Théorie du CANAL-F ([Grigornko et al., 2000]), ou encore le Modèle Hi-LAB pour apprenant avancé ([Doughty et al., 2010]). Pour une liste exhaustive des modèles d'extension du MLAT voir [Wen et al., 2017]. Cependant, bien que ces auteurs apportent des éclaircissements et des raffinements, ils restent cantonnés à une référence au modèle original proposé par Carroll, composé de quatre grandes dimensions fortement corrélées au futur niveau d'acquisition de l'apprenant.

Skehan ([Skehan, 2002], [Skehan, 2013] et [Skehan, 2016]) rajoute une dimension dynamique à cette vision, en prônant l'importance de relier ces aptitudes individuelles à des étapes de l'acquisition. Il revendique alors l'importance de chercher à expliquer les parcours acquisitionnel des apprenants plutôt qu'à seulement chercher à calculer un score prédictif immuable. Malheureusement, cette revendication reste largement théorique et spéculative du fait du peu d'opérationnalisation proposée par l'auteur pour cette réconciliation entre état acquisitionnel et aptitude en LE.

Robinson ([Robinson, 2005], [Robinson, 2007], [Robinson, 2012]) considère également l'aptitude en LE à travers un modèle à 4 habiletés mais sous l'angle du type de sollicitation de l'apprenant. Dans cette perspective Robinson argue que la tâche communicative à laquelle est soumis l'apprenant va conditionner l'habileté utilisée par l'apprenant pour répondre à cette tâche. Il souligne ainsi l'interaction entre les capacités d'apprentissage de l'apprenant et le contexte dans lequel l'apprenant va mobiliser ses capacités.

Malgré tout, tous ces modèles relèvent d'habiletés générales et fondamentales et sont issues d'une perspective différentielle où la construction d'un profil de l'apprenant se fera hors de tout cadre d'acquisition d'une LE. Or comme l'exprime Véronique :

"[...] il ne s'agit pas de partir de catégories *a priori* mais plutôt d'observer la gestion *hic et nunc* des situations exolingues.[Les profils d'apprenants] ne constituent pas des données a priori mais des catégories construites au fil des interactions." ([Véronique, 1994], p.9).

La littérature en RAL et en didactique manque cruellement de références sur la création d'un profil d'apprenant à partir de réelles données d'apprentissage, c'est à dire de données de l'apprenant issues de ses expériences de traitement et de production en LC. De plus, la vision figée d'une "aptitude en LE" ne procède pas de la description et la compréhension du processus d'acquisition d'une LE mais plutôt de l'établissement docimologique de l'avenir immuable d'un individu dans l'apprentissage d'une nouvelle langue. Or, en plus du fait que l'acquisition d'une nouvelle langue est un processus dynamique, la *learner variety approach*, dans laquelle nous nous plaçons, s'intéresse davantage à la caractérisation du lecte de l'apprenant plutôt qu'à son placement en terme de niveau d'acquisition. Ainsi, dans cette thèse, l'objectif consiste en la création de profils d'apprenant basés sur les données de réponses à des tests de langue, sollicitant soit une production en LC soit un traitement en LC de la part de l'apprenant. Ces profils seront identifiables en termes de stratégies de réponses à une tâche spécifique, et mis en parallèles avec les différentes étapes d'acquisition d'une LE observées dans d'autres études issues de la RAL.

Chapitre 2

Projet VILLA

Sommaire

2.1	Objectifs et motivation du projet VILLA	19
2.2	Participants	22
2.3	Comparaison des langues en présence dans l'étude	22
2.4	Les caractéristiques de l'input	24
2.4.1	Les séances d'enseignements : <i>form based</i> et <i>meaning based</i>	25
2.4.2	La transparence et la fréquence des items lexicaux	26
2.5	Les tests de langue	26
2.5.1	Niveau phonologique : <i>Phoneme Discrimination</i>	27
2.5.2	Niveau lexical : <i>Word Recognition</i>	28
2.5.3	Niveau morphologique : <i>Gramaticality Judgment</i> et <i>Oral Question Answer</i>	28
2.5.4	Niveau morphosyntaxique : <i>Picture Verification</i> et <i>Sentence Imitation</i>	29
2.5.5	Observations critiques	31

Le projet VILLA « *Varieties of Initial Learners in Language Acquisition : Controlled classroom input and elementary forms of linguistic organisation* » est un projet européen financé par une subvention Open Research Area en France (ANR), en Allemagne (DFG) et aux Pays-Bas (NOW) pour une durée de trois ans (2011-2014). L'équipe italienne a obtenu une subvention PRIN pour la réalisation du projet, et l'équipe anglaise à York a reçu un financement additionnel par le British Royal Academy.

2.1 Objectifs et motivation du projet VILLA

Le projet VILLA porte sur les premiers stades d'acquisition d'une langue étrangère, en l'occurrence le polonais, dans des conditions contrôlées (une salle de classe filmée) avec une attention particulière portée sur le traitement de l'*input* auquel sont exposés les apprenants, locuteurs de 5 langues typologiquement différentes (l'allemand, l'anglais, l'italien, le français et le néerlandais) [Dimroth et al., 2013]. L'*input* est défini dans ce projet comme l'"ensemble de matériau linguistique auquel l'apprenant est exposé" ([Carroll, 1999], [Carroll, 2001], [Gass, 1997]). Les trois objectifs principaux, formulés à partir de l'état de connaissance et des interrogations actuelles en RAL, sont :

- L'observation de l'évolution de l'apprentissage du niveau initial de connaissances à 14 heures d'exposition à la nouvelle langue.
Bien que l'acquisition d'une L2 soit un phénomène très étudié depuis plusieurs décennies, peu

d'attention a été portée par les chercheurs aux tous premiers stades d'acquisition d'une LE. Les travaux existants concernent principalement l'étude des premiers moments d'exposition à une langue artificielle, ou n'offrent une analyse basée que sur une période d'observation très restreinte. En effet, les contraintes méthodologiques liées à ce type d'étude rendent difficile une observation prolongée, qui pourtant permet l'observation de l'évolution de l'acquisition et de la consolidation des mécanismes à l'œuvre. L'étude menée par Rast ([Rast, 2008]) est pionnière dans ce domaine étant donné la plus longue période d'observation. Elle constitue une base pour l'élaboration du projet VILLA.

- L'étude du traitement de l'*input* sur différents niveaux linguistiques (perception, compréhension, analyse grammaticale et production) en relation avec ses caractéristiques.
En réaction aux hypothèses générativistes ne considérant pas l'input comme un facteur déterminant dans l'acquisition d'une L2, les théories *usage-based* se centrent sur l'importance de la distribution statistique des propriétés de l'input et son influence sur leur traitement par l'apprenant ([Robinson et Ellis, 2008]). Ainsi, l'input auquel les apprenants ont été exposés dans le projet VILLA est un input scripté en amont (contenu des cours), et contrôlé en aval (caméra et micros), afin d'être à même d'observer les possibles liens entre distribution de l'input et acquisition, en traitement comme en production.

- Le rôle des connaissances initiales dans le traitement de l'*input* : le rôle des caractéristiques typologiques propres de la langue source (dimension translinguistique) ainsi que le rôle des principes universels spécifiques au langage et à la communication.

Ce projet s'inscrit dans la tradition fonctionnaliste des travaux en acquisition des langues secondes qui ont mis en évidence le rôle de la langue source dans la construction initiale des connaissances de l'apprenant. Les apprenants adultes abordent l'acquisition d'une nouvelle langue avec des connaissances antérieures qui vont conditionner l'acquisition d'une L2. Premièrement, la langue maternelle des apprenants contribue à l'élaboration des représentations de ceux-ci sur le fonctionnement de la LC (son lecte), notamment à travers la distance typologique qu'ils peuvent percevoir entre LM et LC (cf. 1). Deuxièmement, au tout début d'acquisition, il a été montré (cf. 1) que les apprenants reposent sur des principes universels dits "*language neutral*" [Klein, 2001] pour traiter et utiliser la LC. Le projet VILLA étudie des apprenants du polonais de 5 LM différentes à des fins d'observation d'effets différenciés en fonction de la typologie de la LM de l'apprenant, mais également d'observation de comportements similaires du fait de l'utilisation des principes "*language neutral*".

En résumé, le projet VILLA vise à recueillir des données longitudinales sur une période allant d'un état initial de connaissances en polonais à un état de quatorze heures d'exposition à l'input. Les données recueillies tout au long de cette période d'observation, sous formes d'enregistrement des séances d'enseignements et de résultats à des tests linguistiques en LC, permettront de répondre à cette question :

1. Quelles sont les capacités mises en œuvre par les apprenants en perception, compréhension et production dans une nouvelle langue cible après les 14 premières heures d'exposition ? Quels processus sont activés avant même que ces apprenants arrivent à produire de façon autonome dans la langue cible ?

Le choix d'étudier des apprenants de 5 LM différentes pour la même LC permet d'intégrer la dimension translinguistique au projet et de répondre aux questions suivantes :

2. Dans ces processus, quel est le rôle de la langue source (langue maternelle des apprenants) ainsi que d'autres langues étrangères apprises auparavant ? Quelles autres connaissances indépendantes des langues source et cible sont mobilisées par les apprenants (principes discursifs, structure informationnelle

etc.) pour s'approprier la nouvelle langue? Quelles connaissances extralinguistiques (connaissances du monde) contribuent au processus d'appropriation de la nouvelle langue?

Les apprenants du projet ont reçu un cours de polonais basé sur l'approche communicative assuré par un enseignant de langue maternelle polonaise. En effet, les thématiques enseignées étaient contextualisées par des situations de communications. Le script des séances d'enseignement était défini à l'avance et resté le même en terme de qualité et de quantité d'input pour les différents groupes d'apprenants. Le choix des items et le nombre de fois où l'apprenant y sera exposé a été prédéfini pour chacun d'entre eux. Ainsi, les caractéristiques de ces items ont pu être établies, leur fréquence et leur transparence, et utilisées en fonction dans la création des tests linguistiques. Également, les cours étaient monolingues, c'est à dire qu'il n'y avait pas de recours à la LM des apprenants. Cependant, la mise en contexte, l'appui des illustrations et la communication extra linguistique potentielle lors d'interactions en classe sont autant d'aides pour l'apprenant. Ces deux choix méthodologiques permettent d'étudier la question globale de l'effet de l'input sur l'acquisition :

3. Quel est le rôle même de l'*input* (fréquence et transparence lexicale des items présents dans l'*input*)? Quel est le rôle des indices paralinguistiques tels que gestes, phénomènes interactionnels, contexte, etc.?

En dehors de cette ligne de conduite générale en matière de méthodologie de l'enseignement, dans chaque pays (excepté en Allemagne), deux groupes distincts d'apprenants (14 à 20 apprenants adultes selon le pays) représentant le même profil (âge, niveau d'instruction, type d'études, etc.) ont été soumis à deux types de cours de langue polonaise différenciés par le degré d'explicitation métalinguistique des contenus des leçons. Ainsi, un groupe a reçu un enseignement basé sur le sens (*meaning based input*) tandis que l'autre, un enseignement davantage axé sur la forme (*form based input*), le contenu des cours restant le même. En Allemagne, le groupe d'apprenants ayant reçu le cours basé sur le sens était un groupe d'enfants, c'est pourquoi il n'est pas utilisé dans la présente analyse. La mise en place de deux méthodologies d'enseignement a pour objectif de répondre à :

4. Quel est l'effet sur le traitement des deux différentes manières de présenter la nouvelle langue (cours axés sur le sens vs cours axés sur la forme)?

Cette question formulée à partir des problématiques issues du domaine de la didactique des langues rejoint directement une autre question plus orientée sur les activités internes de l'apprenant :

5. Au fur et à mesure que l'acquisition avance, l'attention que les apprenants portent sur les aspects sémantiques laisse-t-elle la place à l'attention portée davantage sur les aspects structuraux, ou leur attention est-elle portée sur les deux aspects en même temps, en évoluant graduellement?

En effet, si l'attention de l'apprenant aux aspects sémantiques *vs.* structuraux de la LC est différenciée, il est intéressant d'observer si une présentation de l'input orientée sur les uns *vs.* sur les autres aspects influe sur cette attention.

Le corpus VILLA offre une documentation complète des séances de l'enseignement d'une langue étrangère, des acquis des apprenants et de leurs profils individuels. Il donne ainsi la possibilité d'examiner avec précision des séquences didactiques en relation avec les comportements des apprenants, leurs interactions mutuelles et avec l'enseignant. L'analyse de ces séquences didactiques et du déroulement des cours réalisés avec deux démarches différentes (fondée sur le sens et fondée sur la forme), mise en parallèle avec les résultats des tests linguistiques et ceux des tests psychométriques, permet d'affiner la réflexion autour des questions concernant l'apport de la recherche en acquisition pour la didactique des langues.

2.2 Participants

Les apprenants sont des étudiants universitaires, âgés de 18 à 25 ans. Les étudiants en cursus sciences du langage, psychologie ou langues n'ont pas été retenus pour le projet. Également, les participants recrutés pour l'étude n'ont jamais appris ni été en contact avec le polonais ou aucune autre langue slave. Dans tous les pays sauf l'Angleterre, la langue vivante 1 (LV1) des participants est l'anglais. Les autres L2 acquises peuvent varier d'un pays à l'autre en fonction de la politique linguistique du pays. En France, tous les participants ont comme LV2 une langue romane (espagnol, italien, portugais) et les étudiants ayant fait du latin sont exclus. Ainsi, les apprenants francophones n'ont jamais été confrontés à un système casuel. En Allemagne et aux Pays-Bas, les LV2 des participants correspondent pour l'essentiel aux langues romanes également avec quelques exceptions. De même en Italie pour la LV2, et les participants ont tous appris l'anglais en LV1. Pour ce qui est de l'Angleterre, les apprenants possèdent comme LV1 le français, l'allemand ou l'espagnol. Quatre apprenants anglophones n'ont jamais appris une LE au cours de leur scolarisation, l'étude d'une LE n'étant pas obligatoire. Le polonais constitue donc pour eux leur première langue étrangère. Même si dans chaque pays 20 sujets ont été recrutés pour l'étude, les divers abandons ont entraîné un nombre variable d'apprenants d'un groupe à l'autre. Le tableau 2.1 résume le nombre de participants par groupe.

Langues sources	<i>form based</i> input Apprenants adultes	<i>meaning based</i> input Apprenants adultes	<i>meaning based</i> input Apprenants enfants
Français	17	19	
Italien	15	14	
Anglais	17	18	
Néerlandais	20	20	
Allemands		20	19

TABLE 2.1 – Répartition des participants dans les différents groupes d'enseignement

2.3 Comparaison des langues en présence dans l'étude

Afin d'étudier le traitement de l'input aux stades initiaux d'acquisition d'une nouvelle langue, il est important de connaître l'ensemble de l'input auquel les apprenants peuvent être exposés. L'accès qu'ils peuvent avoir à la LC à travers les médias divers et variés échappe au contrôle des chercheurs. C'est pourquoi une langue peu diffusée en Europe occidentale comme le polonais est un bon candidat pour la LC du projet VILLA. Par ailleurs, le polonais présente des caractéristiques typologiques intéressantes par rapport aux langues sources des apprenants. Un bref résumé des structures évoquées pour chaque langue est visible en tableau 2.2.

	Langue cible			LMs		
	Polonais	Néerlandais	Anglais	Français	Allemands	Italiens
Liberté dans l'ordre des mots	+	+	-	-	+	(+)
Pro-drop	+	-	-	-	-	+
Nombre d'accords Sujet-Verbe	6	3	2	5	4	6
Nombre de cas	7	-	-	-	4	-
Genre	3	2	-	2	3	2

TABLE 2.2 – Différences morpho-syntaxique majeures entre LMs et LC du projet VILLA

Le polonais est une langue faisant partie du groupe occidental des langues slaves dotées d'une morphologie verbale (4 conjugaisons) et d'une morphologie nominale (5 déclinaisons) très riches. L'accord est marqué

d'une part entre le sujet et le verbe où on atteste un marquage systématique de la personne et du nombre. Au passé, la désinence indique également le genre du sujet. Le polonais possède trois genres : le féminin, le masculin et le neutre. Le nom s'accorde en genre, en nombre et en cas avec l'adjectif, et avec certains numéraux. Cette riche morphologie va de pair non seulement avec une organisation 'pragmatique' des constituants mais aussi avec l'absence du pronom sujet (sujet nul ou *pro drop*) sauf dans des contextes de contraste. On note également une absence de tout article sauf du démonstratif.

Ainsi, la morphologie nominale du polonais permet de marquer le genre et le nombre, mais également la fonction syntaxique d'un constituant de la phrase, l'ordre des mots étant relativement libre. Ce sont les morphèmes casuels qui indiquent s'il s'agit du sujet ou de l'objet dans une phrase. Enfin, le marquage casuel du nom varie en fonction de la préposition qui l'accompagne. Par exemple, la préposition « *za* » (derrière) implique que le nom qui la suit soit à l'instrumental et la préposition « *w* » (dans) déclenche l'emploi de l'accusatif.

Par rapport à ces caractéristiques de la LC, les langues maternelles des apprenants du projet VILLA s'éloignent du polonais à un degré variable. La comparaison entre le polonais d'une part, et le français, l'italien, l'allemand, le néerlandais et l'anglais d'autre part, résumée dans le tableau ci-dessous permet de voir les relations de proximité/distance entre ces langues.

En ce qui concerne l'ordre des mots, les langues maternelles des apprenants présentent des caractéristiques différentes. En allemand et en néerlandais, l'ordre des mots est régi par la règle syntaxique « V2 » selon laquelle le verbe fini doit occuper la deuxième position dans la phrase, la position pré-verbale doit être remplie par un constituant qui n'est pas obligatoirement le sujet. Ainsi, sans être complètement libre, l'ordre des mots n'est pas obligatoirement Sujet-Verbe-Objet (SVO). Le français et l'anglais sont traditionnellement considérés comme langue à l'ordre SVO. Cependant, en anglais, cet ordre est plus rigide qu'en français qui accepte le renversement de l'ordre SV dans le cas d'utilisation de verbes intransitifs (par exemple, « Et au fond de la rue arrive un tram »), ainsi que l'ordre SOV lorsque l'objet est exprimé par des pronoms clitiques (par exemple « elle le voit »). En revanche, l'anglais ne connaît pas d'exception de l'ordre SVO. L'italien, en principe une langue à l'ordre SVO, autorise les variations liées à la structure informationnelle. Ainsi, l'ordre syntaxique SVO en italien est neutre mais d'autres ordres sont possibles au niveau de l'énoncé et motivés par les aspects pragmatiques.

"Puisque l'ordre de la phrase est grammaticalement très contrôlé en français alors qu'en italien, la structure formelle est « altérée » dans un but pragmatique en dépendant du contexte et du besoin communicatif, il importe de distinguer des degrés de rigidité de l'ordre des mots et de ne pas considérer que la structure SVO a la même fonction dans les deux langues" ([Augendre, 2008]).

De plus, l'italien est la seule langue parmi les langues maternelles des apprenants du projet VILLA à sujet nul (langue *pro-drop*), comme le polonais.

En ce qui concerne la morphologie verbale, elle est riche dans toutes les langues sources ainsi que dans la langue cible même si on trouve des différences. Nous n'entrons pas dans les détails de la comparaison interlinguistique de la flexion verbale car elle n'a pas fait l'objet d'analyse systématique dans le projet VILLA. Nous nous attardons, en revanche, sur les différences entre les langues quant à la morphologie nominale qui a été la plus étudiée.

Comme le polonais, les noms allemands ont un genre inhérent (masculin, féminin, ou neutre) et sont marqués en fonction du nombre (singulier et pluriel). Cependant, l'allemand ne possède que quatre cas (nominatif, génitif, datif et accusatif) contrairement au polonais qui en possède 7 avec l'instrumental, locatif et le vocatif en plus. Le marquage casuel concerne essentiellement les noms masculins et neutres et leur forme dépend de la classe du nom. Les articles, les pronoms et les adjectifs s'accordent avec les noms qu'ils accompagnent en genre, en nombre et en cas. Les cas en allemand sont principalement marqués sur les

déterminants. Le marquage casuel sur le nom même est moins manifeste. Aucune autre langue maternelle des apprenants du projet VILLA n'est dotée du système casuel, y compris le néerlandais qui partage d'autres propriétés avec l'allemand.

Cependant, l'accord en genre et en nombre peut être plus ou moins complexe selon les langues considérées. Ainsi, en ce qui concerne le français, le genre peut être marqué de différentes manières. L'une d'entre elles est la voyelle finale « -e » marquant le genre féminin qui, selon le cas, a des conséquences sur la prononciation. Par exemple, dans le nom « étudiant » au féminin l'ajout de « -e » déclenche la prononciation du [t]. Cela dit, il existe en français un grand nombre de noms terminés par « -e » et invariables en genre (« journaliste »/ « article ») ainsi que des noms où l'ajout de « -e » n'affecte pas la prononciation (ami/amie). D'autres phénomènes sont aussi présents tels que la double consonne comme dans le nom « chien » qui devient « chienne » au féminin. Les noms en français sont également marqués en nombre (singulier et pluriel). Généralement, c'est le « -s » final qui marque le pluriel à l'écrit et qui est habituellement silencieux en français parlé (« étudiant(s) »). Certaines désinences telles que -aux [o], peuvent également marquer le pluriel (par exemple « le cheval » - « les chevaux »), ce qui affecte la prononciation et est perceptible à l'oral. Le genre et le nombre sont signalés par les déterminants qui varient en genre et en nombre.

En italien le marquage du genre et du nombre est plus clair qu'en français et est perceptible à l'oral. De façon générale, au singulier les noms féminins terminent par « -a » et les noms masculins par « -o ». Au pluriel, les noms féminins terminent généralement par « -e » et masculins par « -i ». Ceci est toujours le cas pour le genre naturel défini par le référent animé sexué tandis que le genre grammatical peut avoir d'autres terminaisons comme « -e » dans « una rete/un réseau » (féminin) ou « un lampone/une framboise » (masculin). Les noms s'accordent en genre et en nombre avec les adjectifs et avec les déterminants.

L'anglais ne possède que le genre naturel qui concerne seulement les noms référant aux entités animées et sexuées. Des entités non animées ne possèdent pas de genre (ex. « a table » *vs.* « a woman »). Les articles ne signalent donc pas le genre (« a girl » *vs.* « a boy »). De façon générale, la marque du pluriel est « -s » et l'article ne porte pas l'information sur le nombre dans la mesure où il n'accompagne que les noms au singulier. L'adjectif en anglais est invariable et ne s'accorde ni en genre ni en nombre avec le nom. L'anglais se caractérise donc par un système plus simple en ce qui concerne la structure interne du syntagme nominal.

L'input auquel les apprenants du projet VILLA ont été exposé correspond à une situation propre à l'acquisition guidée. Les apprenants ont accès à une variété de polonais structurée selon des contraintes relevant du contrôle total des contenus d'enseignement. En effet, l'enseignant a reçu des consignes très précises sur le nombre d'occurrence de différents items à utiliser à chaque séance ainsi que sur les items qu'il ne devait jamais employer pendant les cours. Ce type d'input diffère de celui auquel un apprenant pourrait avoir accès en milieu naturel (par exemple lors d'un séjour en Pologne), propre à l'acquisition non guidée. Cependant, l'input fourni a été suffisamment riche à la fois en lexique et en formes grammaticales (notamment la morphologie nominale), afin de donner la possibilité aux apprenants débutant de la matière pour le traitement du nouveau système linguistique.

2.4 Les caractéristiques de l'input

Les participants aux projets VILLA ont été soumis à un *input* en LC très contrôlé. Le contenu des séances d'enseignement a été planifié avant le début du projet. Ainsi les apprenants de toutes les LMs du projet ont reçu le même enseignement en polonais, dispensé par le même enseignant. Les thèmes abordés et le nombre d'occurrences des différents lexèmes étaient établis d'emblée. Les thèmes abordés incluaient des sujets de vie courante tels que la présentation de soi, la famille, la nourriture, le déplacement, etc. Le choix des paradigmes linguistiques enseignés se justifie par ces thèmes. Ainsi, le système flexionnelle de la LC

est au cœur de l'enseignement du projet VILLA. Il a été privilégié parce qu'il constitue le contraste le plus important entre les systèmes linguistiques des langues sources et de la langue cible. La totalité des cours (14 heures) a été enregistrée par deux caméras dont l'une dirigée sur les apprenants et l'autre sur l'enseignant. Ces enregistrements ont pour objectif premier d'assurer le contrôle total de l'input et de ce qui s'est passé pendant les séances. Le support principal des cours étaient les powerpoints conçus à l'avance et accompagnés pour certaines séances de dialogues enregistrés au préalable, en relation avec le contenu du cours. Afin que l'input reçu par les apprenants soit parfaitement connu et contrôlé, tant en contenu qu'en quantité, les apprenants avaient pour consignes de ne pas travailler le polonais en dehors des séances d'enseignements et de ne pas consulter d'autres sources d'input dans cette langue.

La période d'enseignement du polonais s'est étendue sur 10 jours (deux semaines hors weekend). Une session d'enseignement journalière à proprement parler dure 90 minutes avec une courte pause à mi chemin, excepté le dernier jour où la session a duré 30 minutes. L'ensemble des cours représente donc un total de 14h d'enseignement. Les sessions d'enseignement se composent d'un exposé du professeur, de questions-réponses entre le professeur et l'ensemble de la classe, entre le professeur et un élève en particulier, et de jeux de rôles sollicitant un dialogue entre élèves.

2.4.1 Les séances d'enseignements : *form based* et *meaning based*

Dans chaque pays participant, excepté l'Allemagne, deux groupes d'apprenants d'environ 20 adultes ont été formés (*cf.* tableau 2.1). Le premier groupe a reçu un enseignement dit *meaning based*, basé sur le sens, l'autre groupe a reçu un enseignement dit *form based*, avec une explicitation de la forme, notamment par l'intermédiaire de signaux visuels présents dans les slides de présentation du cours (*cf.* annexe A).

L'importance de la focalisation sur la forme de la LC dans l'enseignement est une question qui fait encore débat en acquisition des langues secondes. Plusieurs degrés et différents moyens d'explicitation des formes enseignées peuvent être envisagés. Ainsi, la dichotomie *Meaning-focused* et *Form-focused instruction* (MFI et FFI) de Doughty et Williams (1998) ([Doughty et Williams, 1998]) comporte elle-même des ramifications. L'approche MFI centrée sur l'acquisition implicite du savoir suppose un engagement de la part de l'apprenant dans des tâches communicatives, et donc d'une acquisition incidente. Cette approche propose à l'enseignant deux types de méthodologies pour la gestion de l'enseignement de la forme ([Long, 1998]; [Doughty, 1991]). L'enseignant peut décider à l'avance des formes travaillées en cours (*Focus on Forms*), ou il peut profiter des tentatives de communications des apprenants (*Focus on Form*) ([Ellis et al., 2002]; [Doughty et Williams, 1998]).

L'approche FFI est l'approche dont le travail des formes morphologiques de la LC suppose un travail d'appropriation explicite par les apprenants. Les différentes formes enseignées sont contextualisées, et les apprenants s'entraînent à les décliner en période de classe. Le métalangage, par le biais de la langue maternelle des apprenants, peut être présent.

Les deux démarches adoptées dans le projet VILLA ne se positionnent pas clairement vis-à-vis de ces approches. Les deux différences majeures entre l'approche *meaning based* et l'approche *form based* sont, pour l'enseignement *form based* :

- L'explicitation visuelle (par symbole) des liens forme-fonction dans les supports de cours.
- Les corrections explicites de l'enseignant sur les tentatives de communications des apprenants.

Les deux cours étaient monolingues et donc ne permettaient pas un recours à la langue maternelle des apprenants. Le déroulement des leçons se centrait sur le message communicatif et il n'y avait pas d'entraînement explicite sur les déclinaisons.

2.4.2 La transparence et la fréquence des items lexicaux

L'*input* reçu par les apprenants lors du projet VILLA était entièrement planifié à l'avance, mais également enregistré lors des sessions d'enseignement, à des fins de vérification. Ainsi, en terme d'étendue du lexique et de fréquence d'exposition des items, le projet VILLA offre la possibilité d'une analyse de corrélation entre les performances des apprenants et l'*input* réel reçu. Comme nous l'avons vu dans le chapitre 1, la majeure partie des courants théoriques en RAL admettent que l'*input* est nécessaire à l'acquisition. Néanmoins, les procédés par lesquelles l'*input* influence le lecte de l'apprenant sont une des principales problématiques actuelles en RAL.

Le rôle de la fréquence de différents items sur la manière dont les apprenants construisent l'association « forme-fonction » est étudiée dans le projet VILLA grâce au contrôle des items enseignés et leur classement dans deux catégories : items fréquents *vs.* items non fréquents. Un item est considéré comme fréquent lorsqu'il est entendu par l'apprenant au moins 20 fois dans l'*input*. Ce seuil de fréquence a été étudié dans des travaux antérieurs ([Goldschneider et DeKeyser, 2001]) et correspond au nombre d'occurrences à partir duquel le traitement de l'item par l'apprenant est modifié. L'objectif est d'observer si et dans quelle mesure un item fréquent est analysé plus rapidement par l'apprenant en traitement et en production. Différents niveaux de traitement sont considérés. Au niveau phonologique, il s'agit d'étudier la discrimination des phonèmes et l'extraction d'un item dans une phrase. Au niveau morphosyntaxique, le projet VILLA étudie la mise en place du système casuel en relation avec l'ordre des mots, à la fois dans des tâches de traitement et dans des tâches de production ciblée et semi-guidée. Finalement, au niveau discursif, les productions semi-guidées permettent d'analyser l'interaction entre la structure informationnelle et la structure phrastique lors du processus de construction du discours sollicité par une tâche communicative complexe.

Le même objectif s'applique à la deuxième caractéristique d'intérêt de l'*input* dans le projet VILLA qu'est la transparence. En effet, la transparence d'un item faciliterait sa perception, plus généralement son traitement, et donc sa restitution ([Rast, 2008]). Cette caractéristique repose sur le constat global que l'apprenant mobilise ses connaissances linguistiques antérieures dans le traitement d'une nouvelle langue. Ainsi, la transparence des items lexicaux auxquels les apprenants ont été exposés a été contrôlée et manipulée afin d'observer des effets de cette variable dans le traitement et la production des apprenants. En effet, les mots présentés dans les séances successives font partie de deux catégories, mots transparents et mots opaques. Ce classement a été fait à la suite d'un test de transparence effectué avant la préparation du cours de polonais auprès des locuteurs natifs des 5 langues sources considérées. Ces locuteurs natifs n'ont jamais été exposés au polonais avant ce test et ont dû traduire dans leurs langues respectives les mots polonais en se basant sur une familiarité potentielle avec les mots de leur langue source. Ce test a permis de constituer la catégorie des noms transparents avec les items que les locuteurs des 5 langues ont pu traduire correctement en polonais sans aucune connaissance préalable de cette langue. À titre d'exemple, un item comme « francuz/français » est considéré comme transparent dans la mesure où il a été correctement traduit par au moins 50% de locuteurs natifs des 5 langues. En revanche, un mot polonais comme « lekarz/médecin » est considéré comme opaque.

L'impact des caractéristiques des langues sources et cibles en présence, du type de l'enseignement (meaning et form based) et des caractéristiques de l'*input* (fréquence et transparence) sur le traitement de l'*input* en L2 a été testé dans le projet VILLA à travers une batterie de tests présentée dans la section suivante.

2.5 Les tests de langue

En plus des enregistrements audios et vidéos permettant de recueillir les productions orales semi-spontanées des apprenants lors des exercices en classes, une batterie de tests en langue polonaise ont été

administrés aux apprenants à la fin de chaque session d'enseignement. Certains de ces tests de langues ont été administrés plusieurs fois afin de suivre l'évolution des réponses des apprenants au cours de la période d'observation. Ces tests permettent de tester les apprenants sur les différents niveaux linguistiques impliqués dans l'acquisition d'une nouvelle langue tels que la phonologie, le lexique, la morphologie et la morphosyntaxe, et la construction du discours. Ces différents tests sont présentés dans la figure 2.1.

Tests de langue	Période de passation des tests, avant toute exposition (0), ou après une session d'enseignement (1-9)										
	0	1	2	3	4	5	6	7	8	9	10
Phoneme Discrimination	X			X				X			
Lexical Decision	X	X			X		X		X		
Word Recognition	X					X					X
Grammaticality Judgement (morphologie nominale)				X				X			
Oral question-answer task (morphologie nominale)				X				X			
Sentence Puzzle (ordre des mots)						X			X		
Picture Verification (morphosyntaxe)							X				X
Sentence Imitation (morphosyntaxe)							X				X
Grammaticality Judgement (accord sujet verbe)											X
Cloze test (pronoms personnels)											X
Elicited production: route direction (discours)											X
Elicited production: film retelling (discours)											X

FIGURE 2.1 – Tests de langue administrés au cours du projet VILLA. Les tests surlignés sont ceux utilisés dans le chapitre 6 d'application de l'algorithme

Les tests surlignés dans le tableau sont centraux pour ce travail et leurs résultats ont été utilisés dans le chapitre 6. L'ensemble de ces tests permet de décrire le parcours acquisitionnel des apprenants en phonologie, lexique, morphologie et morphosyntaxe en LC. La structure des tests est importante pour comprendre ce que reflètent les résultats obtenus par les apprenants étudiés. Il s'agit de tâches ciblées testant des paradigmes linguistiques précis. De plus la structure de ces tâches nous renseigne sur la façon dont leur analyse doit être conduite, et participe ainsi de la construction de l'algorithme de clustering semi supervisé utilisé pour la création de profils d'apprenants. Nous décrivons ici leur structure, et verrons en chapitre 5 la manière dont ces informations sont utilisées pour la construction de l'algorithme.

2.5.1 Niveau phonologique : *Phoneme Discrimination*

La tâche *Phoneme Discrimination* (PD) vise à examiner la sensibilité perceptuelle en L2 au tout début d'acquisition et à évaluer un éventuel impact des LMs des apprenants et du type de l'input sur la discrimination des nouveaux sons. La tâche de phonologie PD exige de l'apprenant de reconnaître des phonèmes du polonais qui n'existent pas dans sa LM. La distance phonémique (différenciation d'un phonème non présent dans la LM ou d'un phénomène dont le cluster de phonème est non présent dans la LM) des

différentes langues maternelles des apprenants avec le polonais est élevée ([Rast, 2010]).

Le polonais est une langue comprenant un grand panel de fricative, dont un grand nombre qui n'existent pas dans les langues maternelles respectives des apprenants. Des études antérieures sur des apprenants anglais ont montré que leur capacité de discriminations des fricatives du polonais (alveopalatal *vs.* retroflex) est quasi nulle au début de l'apprentissage, mais qu'elle s'améliorait avec un entraînement ciblé ([Lisker, 2001]; [McGuire, 2007]). Une étude sur les apprenants français de notre base de données a montré des résultats similaires après seulement quelques heures d'exposition ([Shoemaker, 2014]).

Dans cette tâche, deux phonèmes constituant des paires minimales sont entendus à la suite. L'apprenant doit alors dire si les deux sont identiques ou différents en pressant la touche de clavier appropriée sur l'ordinateur. Le stimuli est un son construit sous la forme Consonne-Voyelle (CV). La consonne est choisie parmi un ensemble de 6 consonnes sibilantes du répertoire polonais (/s/, /z/, /ɕ/, /ʒ/, /ʂ/, /ʐ/), et la voyelle est toujours /a/. Les 6 sons ainsi créés ont été combinés de toutes les manières possibles, et dans les deux ordres d'écoute, pour créer les 30 stimulus finaux, dont 6 paires de sons identiques.

Ce test a été administré aux apprenants trois fois au cours du projet VILLA. La première fois à 0 heure d'exposition au polonais, c'est à dire avant le début des sessions d'enseignement, et la deuxième et troisième fois après 4 heures et demie et 10 heures et demie d'exposition respectivement.

2.5.2 Niveau lexical : *Word Recognition*

Le test *Word Recognition* (WR) programmé sur E-prime vise à comprendre comment et à quel moment de l'apprentissage, les items lexicaux sont reconnus dans un flux sonore. Dans la tâche WR, les apprenants entendaient une phrase en polonais suivie d'un mot isolé, et devaient décider si le mot isolé était contenu ou non dans la phrase précédemment entendue. La tâche a été administrée trois fois au cours des sessions d'enseignement. La première fois avant le début du projet, après 0 heure d'exposition, et les fois suivantes après 7 heures et demie, et 13 heures et demie d'exposition à l'*input*.

Les mots à reconnaître dans les phrases successives de ce test font parti des quatre catégories, transparents/fréquents, opaques/fréquents, transparents/absents, opaques/absents dans l'*input*. Ceci permet de mesurer l'impact de la transparence et de la fréquence sur la capacité d'extraction d'un item. Par ailleurs ce test a été utilisé dans une étude pilote ([Shoemaker et Rast, 2013]). Ces auteurs ont remarqué que la position initiale et finale du mot dans la phrase facilite clairement l'extraction indépendamment de l'une des quatre catégories. C'est pourquoi dans le test de VILLA, seule la position médiane a été utilisée parce que celle-ci permet de mesurer véritablement la capacité à extraire un item du flux sonore.

2.5.3 Niveau morphologique : *Gramaticality Judgment* et *Oral Question Answer*

Les deux tests *Gramaticality Judgment* (GJ) et *Oral Question Answer* (OQA) visent à évaluer les compétences des apprenants en morphologie nominale du polonais. La tâche GJ nous permet de statuer sur les capacités de jugement d'un apprenant quant à la reconnaissance des morphèmes casuels corrects dans une phrase simple, tandis que la tâche OQA sollicite de l'apprenant une production originale et offre une idée de la production de désinences appropriées au contexte. Les deux tâches sollicitent les connaissances de l'apprenant, en traitement et en production, portant sur l'opposition nominatif (féminin et masculin) *vs.* instrumental (féminin et masculin).

La tâche GJ dont la première version a été élaborée dans l'étude de Rast et col. ([Rast et al., 2014]) consiste en la présentation sonore d'une phrase en polonais à l'apprenant qui doit l'écouter puis décider (en appuyant sur la touche appropriée) si cette phrase est grammaticale (correcte) ou non. Le test comporte 64

phrases dont 32 correctement marquées. Les phrases sont construites de trois mots sur le schéma "Nom propre + copule 'être/być' (est/jest) + nom désignant une nationalité" ou "Nom propre + copule 'être/być' (est/jest) + nom désignant une profession". Dans les phrases grammaticalement correctes, le nom de profession ou de nationalité est marqué à l'instrumental en polonais. Les phrases non grammaticales sont des phrases où le nom de nationalité ou de profession est dans la forme du nominatif au lieu de l'instrumental. Ainsi, dans les phrases grammaticales, la désinence de l'instrumental masculin *-em* ou celle de l'instrumental féminin *-ą* est utilisée tandis que dans les phrases non grammaticales les noms au nominatif masculin et féminin portent respectivement les désinences *-Ø* et *-ka*.

Albert jest fotografem / Albert est photographe	Intrumental Masc
*Tomasz jest fotograf / Tomasz est photographe	Nominatif Masc
Patryk jest lekarzem / Patryk est médecin	Instrumental Masc
*Dawid jest lekarz / Dawid est médecin	Nominatif Masc
Daniela jest artystką / Daniela est artiste	Intrumental Fém
*Anna jest artystka / Anna est artiste	Nominatif Fém
Monika jest tłumaczką / Monika est traductrice	Instrumental Fém
*Joanna jest tłumaczka / Monika est traductrice	Nominatif Fém

Exemple 2.5.1.

La tâche OQA est une tâche de question-réponse où la forme de la question conditionne le marquage casuel que l'apprenant devra produire lors de sa réponse à l'oral. Chaque participant devait répondre à 32 questions pour une passation du test. Un stimuli visuel comportant un icône féminin ou masculin est d'abord présenté, puis l'apprenant entend une question, et enfin il voit apparaître sur l'écran une image représentant une nationalité ou une profession. La question posée peut être de deux types. Le premier type "*Kto to jest ?*" (Qui est ce ?) induit une réponse de type *To jest* (C'est) + un complément nominal marqué au nominatif masculin ou féminin. Le deuxième type de question "*Kim on/ona jest ?*" (Qui est-il/elle ?) induit une réponse de type *On/Ona jest* (Il/Elle est) + un complément nominal marqué à l'instrumental féminin ou masculin.

Les deux tâches étaient administrées deux fois, après 4h30 d'exposition au polonais (T1), et après 10h30 d'exposition (T2), afin de pouvoir analyser l'évolution des performances des apprenants à ces deux tâches. Ces deux tâches testent le même paradigme linguistique (Nominatif *vs.* Instrumental) en traitement et en production. Leur mise en parallèle est donc particulièrement intéressante pour l'étude d'un transfert entre capacité de traitement et capacité de production en LC.

2.5.4 Niveau morphosyntaxique : *Picture Verification* et *Sentence Imitation*

Afin d'observer l'acquisition du système morphosyntaxique de la LC, deux tests ont été utilisés : celui de *Picture Verification* (PV) et de *Sentence Imitation* (SI). En effet, en polonais, l'ordre des mots est relativement libre. Bien que l'ordre des mots canoniques d'une phrase soit de type Sujet-Verbe-Objet (SVO), le contexte d'énonciation peut amener un locuteur de polonais à organiser la phrase selon d'autres ordres qui se justifient par des contraintes pragmatiques. Ainsi, les ordres dits marqués OVS, SOV et OSV sont possibles dans certains contextes bien précis, souvent liés à la focalisation sur un des constituants de l'énoncé. Afin de marquer le statut d'un constituant dans la phrase, on utilise alors le marquage casuel. Le nominatif est utilisé pour marquer le sujet de la phrase tandis que l'objet est accordé à l'accusatif. C'est l'acquisition de cette opposition nominatif *vs.* accusatif qui est étudiée grâce aux tâches PV et SI (cf. exemple 2.5.2).

Tłumaczke pozdrawia artystka / L'artiste salue le traducteur	OVS
Artystka pozdrawia tłumaczke / L'artiste salue le traducteur	SVO
Tłumaczka pozdrawia artystke / Le traducteur salue l'artiste	SVO
Studentke pcha nauczycielka / Le professeur pousse l'étudiant	OVS
Nauczycielka pcha studentke / Le professeur pousse l'étudiant	SVO
Studentka pcha nauczycielke / L'étudiant pousse le professeur	SVO

Exemple 2.5.2.

Dans le test PV, l'apprenant entend deux fois une courte phrase en polonais et voit deux images simultanément. Il doit ensuite décider à quelle image correspond la phrase entendue. Les deux images représentent deux actants animés humains, un frère (*brat*) et une sœur (*siostra*), qui effectuent une action exprimée par des verbes transitifs tels qu'appeler, tirer, saluer. Tantôt l'actant "frère", tantôt l'actant "sœur" est agent de l'action. Si le frère est agent de l'action, il sera donc le sujet de la phrase, et sera marqué au nominatif tandis que le patient "sœur" sera marqué à l'accusatif, et inversement. L'apprenant doit réagir à 24 phrases tests qui sont originalement de trois types. Elles sont construites sur le modèle SVO, OVS ou OSV. Dans les premières analyses, les chercheurs du projet VILLA ont observé que le verbe n'a pas d'impact sur le traitement de la phrase et que c'est la position du SN par rapport au verbe qui joue le rôle prépondérant. Ce qui attire son attention c'est l'ordre SO ou OS. Il n'y a donc que des différences mineures entre OSV et OVS que l'on pourrait analyser de façon plus qualitative, afin de voir si le type de verbe impacte la compréhension, mais ces travaux sont encore en cours. Ainsi, pour notre étude, seule la position du sujet et de l'objet est prise en compte. Les 24 phrases du tests peuvent donc être regroupées en deux catégories : SO *vs.* OS.

L'ordre des mots SVO est un ordre neutre et correspond au principe de l'ordre sémantique très fort "Agent Action Patient". Ce principe est connu d'Aristote comme principe logique :

prius essere quam operari (d'abord l'être alors action)

prius actio deinde passio (d'abord agent puis patient)

Ce principe "*language neutral*" est privilégié par des apprenants débutants et est sous-jacent à leurs productions (cf. chapitre 1). Lorsque l'ordre syntaxique SVO et l'ordre sémantique "Agent Action Patient" coïncident dans une phrase soumise à l'apprenant et qu'il la traite correctement, il est impossible de savoir quel principe est à l'œuvre dans son lecte (cf. chapitre 1). Si l'apprenant traite correctement les phrases de type SO, on ne peut pas savoir s'il est sensible au système flexionnel de la LC. Il est possible qu'il applique ici le principe sémantique en assignant le rôle de l'agent au premier constituant de la phrase sans que son attention soit portée sur la désinence du nom. En revanche, si l'apprenant commence à être sensible au système flexionnel de la LC, il peut donner des réponses correctes également aux phrases de type OS. Ce test a été administré deux fois durant la période d'observation, après 9 heures d'enseignement (T1) et après 13 heures et demie (T2).

La tâche SI vise également à analyser le traitement du marquage casuel Nominatif *vs.* Accusatif en relation avec l'ordre des mots. Les apprenants entendent une phrase et doivent la répéter après une courte pause durant laquelle ils sont amenés à dessiner une figure géométrique qui s'affiche sur un écran d'ordinateur. L'activité de dessiner constitue un distracteur et permet d'éviter à ce que l'apprenant s'appuie juste sur sa mémoire à court-terme pour répéter la phrase. L'hypothèse sous-tendant ce type de tâche de répétition est qu'un apprenant ne peut répéter correctement une structure grammaticale que si celle-ci a été intégrée dans son lecte d'apprenant. A contrario, si la structure grammaticale à répéter est au-delà de la complexité de la connaissance grammaticale de l'apprenant, celui-ci échouera à la répéter correctement. L'étude des compétences implicites en L2 via ce type de tâche a été largement validée dans la littérature des sciences du langage. En effet, il a été démontré que les résultats à ce type de tâche en L2 dépassent sensiblement ceux qui seraient attendu si l'apprenant ne reposait que sur sa mémoire à court terme pour la réaliser [Naiman, 1974].

De plus, McDade et al. ([McDade et al., 1982]) ont soulevé le fait que lorsque l'action de répétition n'est pas immédiate après l'écoute de la phrase mais décalée dans le temps, alors seules les phrases comprises par l'apprenant sont correctement répétées. Ces conclusions avaient également été formulées par Fraser et al. ([Fraser et al., 1963]) qui ont utilisé le laps de temps entre l'écoute et la répétition pour rajouter une tâche distractive consistant à dessiner une figure géométrique, comme pour le projet VILLA. Ainsi, les tâches de répétition impliquent plusieurs procédés cognitifs, à savoir le décodage de la phrase, son interprétation et sa (re)production, procédés allant au delà du simple *parroting* (répéter comme un perroquet).

Cependant, des doutes peuvent être émis quant à la qualité de ce type de tâche dans l'évaluation des capacités de production des apprenants. Comme Vinther le formule, "il est difficile d'établir si l'échec de la répétition d'une phrase provient d'une compréhension insuffisante ou d'un manque de capacité productive" ([Vinther, 2002], p.62). En effet, de nombreuses différences existent entre les productions spontanées d'un apprenant et ses résultats à une tâche de répétition. Ainsi, Connell et Myles-Zitzer ([Connell et Myles-Zitzer, 1982]) concluent que ce type de tâche n'est pas un bon indicateur des productions non sollicitées.

2.5.5 Observations critiques

Le design expérimental du projet VILLA permet l'étude d'un même paradigme linguistique par plusieurs tests en LC. Le premier paradigme est celui de la déclinaison au nominatif *vs.* accusatif d'un constituant de la phrase. Le traitement par l'apprenant de ce paradigme est étudié à travers la tâche PV, tandis que la tâche SI permet l'étude de la production de ce paradigme. Le second paradigme est celui de la déclinaison du nominatif *vs.* de l'instrumental des items nominaux d'une phrase. Les capacités de traitement de ce paradigme par les apprenants sont étudiés par leurs résultats à la tâche GJ. Les capacités de production de ce paradigme sont quant à elles étudiées à travers les résultats des apprenants à la tâche OQA. Ainsi, une mise en parallèle des résultats obtenus aux deux couples de tâches PV-SI puis GJ-OQA permet l'observation des capacités de transfert du traitement d'un paradigme linguistique vers sa production en LC.

Cependant, les deux couples de tâches ne sont pas équivalents. Tout d'abord, la tâche PV ne comporte que deux items nominaux pour refléter l'opposition nominatif accusatif, tandis que la tâche SI en incorpore huit. La variété des items utilisés dans la tâche SI peut entraîner une difficulté supplémentaire dans la production du paradigme sollicité. De plus, comme nous l'avons déjà évoqué, la tâche SI est une tâche de répétition de phrase, tandis que la tâche OQA est une tâche de question-réponse où l'apprenant émet une production originale. Ces deux types de tâches n'entraînent pas le même type de résultat, et sont donc difficilement comparables. Ainsi, lors de la comparaison des capacités de transfert du traitement d'un paradigme linguistique vers sa production, il convient de connaître ces différences dans la construction des tâches utilisées.

Chapitre 3

Approches existantes

Sommaire

3.1	Gestion de données comportementales	33
3.1.1	Fuzzification	34
3.1.2	Opérations	36
3.1.3	Défuzzification	37
3.2	L'apprentissage non supervisé : le <i>clustering</i>	38
3.2.1	Notions générales	39
3.2.2	Mesure de distance ou similarité	41
3.2.3	Méthodes classiques	43
3.2.4	Méthodes floues	50
3.3	L'intégration de connaissances	52
3.3.1	La classification semi supervisée	52

Les chapitres précédents nous ont permis de présenter les notions théoriques en acquisition des langues secondes essentielles à la compréhension des données du projet VILLA, ainsi que le contexte et la méthodologie du recueil de ces données. Nous présentons dans ce chapitre les outils informatiques nécessaires pour leurs analyses en vue de la création de profils d'apprenant, à partir de la Base De Données (BDD) à notre disposition.

3.1 Gestion de données comportementales

Un problème peut souvent être résolu de plusieurs manières lorsqu'il n'est pas un problème abstrait mais bien concret. Le choix et l'application d'une manière est appelée mise en place d'une stratégie. Au cours de son existence, un être humain va être amené à changer de stratégie de résolution pour des problèmes similaires. Ce constat pourrait être considéré comme une description de ce que représente l'apprentissage. La transition d'un choix stratégique à un autre n'est pas discret ni irréversible. Il existe souvent une phase de confusion, précédant l'apprentissage. Pour certains auteurs, l'état de confusion est un nécessaire à tout apprentissage, lorsque l'individu se rend compte qu'il y a inadéquation entre ce qu'il sait du monde et ce qui est (ou plus précisément perçoit). Ainsi, même si pour un problème donné l'individu va appliquer une stratégie identifiable et pour le problème similaire suivant une autre stratégie va être adoptée, rien ne nous assure que l'individu ne va revenir à l'ancienne stratégie pour le prochain. Soit qu'il teste encore l'efficacité relative de chacune des stratégies, soit qu'il identifie mal les caractéristiques des problèmes et échoue à reconnaître la similarité entre eux.

De par notre problématique, nous souhaitons identifier la stratégie (bonne ou mauvaise) de réponse d'un apprenant à une question d'un test (cf. chapitre 2). Si la réponse à cette question est bonne, l'apprenant obtient un point, si elle est erronée, il obtient 0. La note obtenue à un test représente l'agrégation de ces points, ramenée à l'ensemble $[0; 1]$. Ainsi, la note obtenue ne reflète pas directement le choix stratégique de l'apprenant. L'agrégation des résultats en une note unique, bien que pratique, introduit de l'imprécision quant aux choix stratégiques de l'apprenant, et donc quant à la représentativité de sa note pour son niveau d'acquisition en LC.

Dans la littérature, il existe plusieurs formalismes de la gestion de l'imprécision dans l'information fournie par les données. Ces approches sont issues des premiers travaux menés par Zadeh ([Zadeh, 1965a]) sur les sous-ensembles flous. Dans cette section, nous proposons une synthèse des travaux de Zadeh et des formalismes pertinents pour notre problématique.

3.1.1 Fuzzification

La notion de *sous-ensemble flou* (SEF) permet d'établir qu'un objet appartient de manière graduelle à une classe d'élément. Ainsi cet objet peut appartenir de manière plus ou moins forte à plusieurs classes différentes. Cette notion dérive de la théorie des ensembles classiques. On associe ainsi à un objet une fonction d'appartenance f à un ensemble de classes X , qui répartit le degré d'appartenance de l'objet aux différents termes composant X . Les termes composants X sont appelés sous-ensembles flous.

Définition 1. Un sous ensemble flou A de l'ensemble X est caractérisé par une fonction d'appartenance $f_A(x)$ telle que $\forall x \in X, f_A(x) \in [0; 1]$

Un sous ensemble classique est en fait un cas particulier d'un sous ensemble flou où f_A ne peut être égal qu'à 0, dans le cas où l'objet considéré n'appartient pas du tout à A , ou 1, si il lui appartient totalement. Prenons l'exemple des couleurs. Les couleurs se décomposent sur une échelle numérique selon leur longueur d'onde. On peut connaître la longueur d'onde exacte d'un rayonnement lumineux mais est-il facile de définir avec précision de quelle couleur il s'agit ? Un rayonnement d'une longueur d'onde de 625 nanomètres est il plus orange ou rouge ? La logique floue nous permet d'affirmer qu'il est les deux à la fois, et en même temps aucun des deux. Pour se représenter l'appartenance multiple de ce rayonnement lumineux nous considérons le spectre des couleurs visibles situé entre 400 et 780 nanomètres. La figure 3.1 représente le découpage officiel des longueurs d'ondes en couleurs perceptibles, la couleur rouge commençant à 630 nm, un rayonnement de 625 nm est considéré orange. Ce découpage nous apparaît comme non naturel et ne ressemble pas aux couleurs que nous observons tous les jours. La question de nuance est importante. Si l'on regarde la figure 3.2, la représentation de l'ensemble des couleurs nous paraît plus naturelle. Sur cette figure est également indiqué notre exemple de rayonnement lumineux correspondant à une longueur d'onde de 625 nm. Nous nous posons maintenant la question de sa caractérisation. La longueur d'onde est une mesure utile mais ne nous permet pas de raisonner sur une base quotidienne. D'un autre côté certains pourraient arguer que le découpage officiel du spectre des couleurs ne permet pas de saisir la nuance de ce rayonnement lumineux, que d'aucuns pourrait qualifier de rouge orangé. Si l'on considère l'ensemble X des couleurs comme constitué des sous-ensembles flous violet, bleu, vert, jaune, orange, rouge, on obtient la figure 3.3. Grâce à la fuzzification du spectre visible des couleurs, on observe que un rayonnement lumineux x de longueur d'onde de 625 nm se caractérise par $f_{orange}(x) = 0.8$ et $f_{rouge}(x) = 0.2$.

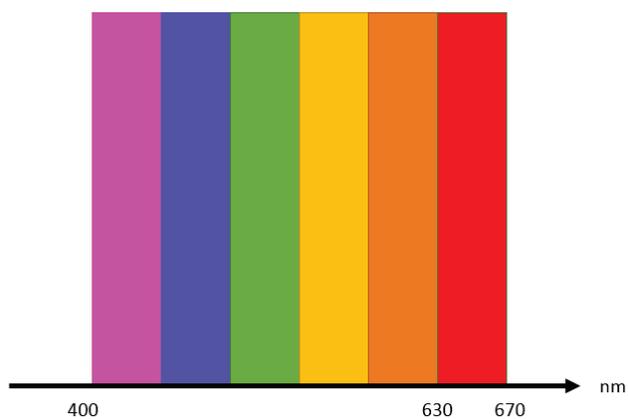


FIGURE 3.1 – Spectre des couleurs visibles théorique

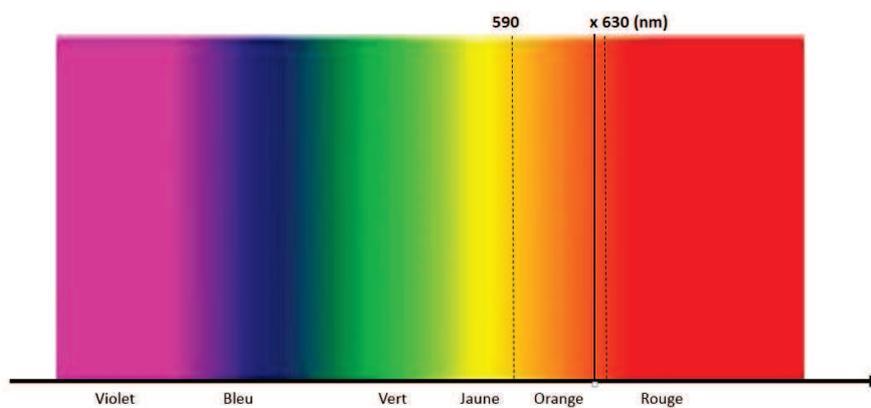


FIGURE 3.2 – Spectre des couleurs visibles naturel

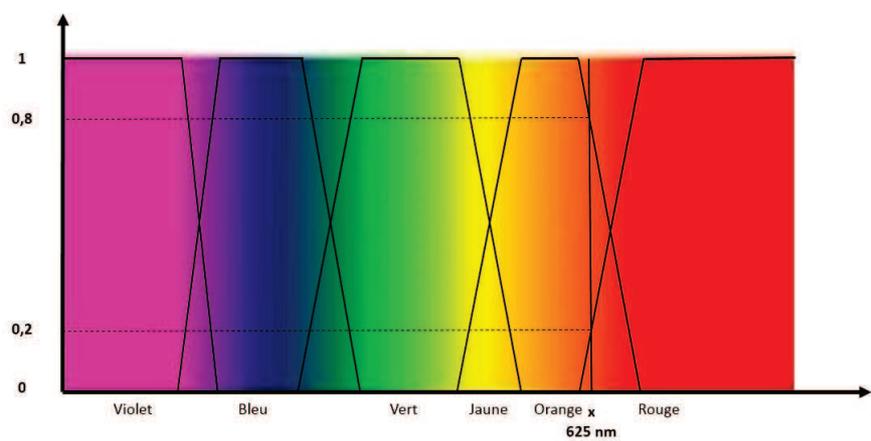


FIGURE 3.3 – Découpage en SEFs du spectre des couleurs visibles

Les SEFs présentés dans l'exemple sont d'un type particulier, il existe en effet plusieurs types de sous ensembles flous. Le type le plus important décrit dans la théorie des possibilités par Zadeh en 1999 ([Zadeh, 1999], et celui que nous considérons ici, correspond aux sous-ensembles flous normalisés. On catégorise les SEFs selon ses caractéristiques, support, hauteur et noyau.

Définition 2. Caractéristiques d'un SEF :

- le support : $supp(A) = \{x \in X / f_A(x) \neq 0\}$
- la hauteur : $h(A) = \max_{x \in X} (f_A(x))$
- le noyau : $noy(A) = \{x \in X / f_A(x) = 1\}$

Le support correspond à tous les éléments x qui appartiennent "au moins un peu" à A et le noyau tous les éléments qui appartiennent "entièrement" à A . Il n'est pas obligatoire qu'un SEF catégorise des éléments de manière certaine, ainsi son noyau peut être un ensemble vide, et la hauteur devient alors un qualificatif important de ce SEF.

3.1.2 Opérations

Il existe plusieurs opérations possible sur deux SEFs, opérations dérivées de celles sur les ensembles classiques :

- Intersection
- Union
- Complémentarité

Définition 3. Soit A et B deux SEFs :

Intersection : $A \cap B \Rightarrow f_{A \cap B}(x) = \min(f_A(x), f_B(x))$

Union : $A \cup B \Rightarrow f_{A \cup B}(x) = \max(f_A(x), f_B(x))$

Complémentaire : $A^C = \{x \in X / f_{A^C}(x) = 1 - f_A(x)\}$ avec $A^C \cup A \neq X$ et $A^C \cap A \neq \emptyset$

Le résultat de ces différentes opérations est lui-même un SEF. Le min et le max définis ici comme opérateurs d'intersection d'union peuvent être remplacés par d'autres opérateurs qui satisfont les propriétés énoncées en définition 4.

Définition 4. Soit \star un opérateur :

- commutatif $\Rightarrow \star(x, y) = \star(y, x)$
- associatif $\Rightarrow \star(x, \star(y, z)) = \star(\star(x, y), z)$
- monotone $\Rightarrow \star(x, y) \leq \star(z, t)$ si $x \leq z$ et $y \leq t$
- qui a 1 pour élément neutre : $\iff \star(x, 1) = x$

min et max sont ainsi des opérateurs appartenant respectivement à la famille générique des t-normes et t-conormes. Une t-norme est un opérateur d'intersection par une norme triangulaire et une t-conorme est un opérateur d'union par une conorme triangulaire. Ces deux familles d'opérateurs respectent les propriétés présentées en définition 4 et sont formalisées en définition 5.

TABLE 3.1 – Exemples d'opérateurs t-norme et t-conorme

t-norme	t-conorme
$\min(x, y)$	$\max(x, y)$
$\max(0, x + y - 1)$	$\min(1, x + y)$
$1 - \min(((1-x)^p + (1-y)^p)^{1/p}, 1)$	$\min((x^p + y^p)^{1/p}, 1)$

Définition 5. Une t-norme est une fonction $\tau: [0; 1] \times [0; 1] \rightarrow [0; 1]$ définie par

$$A \cap^\tau B \Rightarrow f_{A \cap^\tau B}(x) = \tau(f_A(x), f_B(x))$$

Une t-conorme est une fonction $\perp: [0; 1] \times [0; 1] \rightarrow [0; 1]$ définie par

$$A \cup^\perp B \Rightarrow f_{A \cup^\perp B}(x) = \perp(f_A(x), f_B(x))$$

Dans la littérature, on peut trouver plusieurs opérateurs appartenant à ces deux familles et nous en relevons quelques-uns dans le tableau 3.1.

Quel que soit le choix d'un opérateur d'union et d'intersection, il se doit d'être dual afin de préserver les lois de logique propositionnelle de De Morgan, à savoir

$$\text{non}(A \text{ ou } B) = (\text{non } A) \text{ et } (\text{non } B)$$

$$\text{non}(A \text{ et } B) = (\text{non } A) \text{ ou } (\text{non } B)$$

Définition 6. La dualité d'une t-norme et d'une t-conorme vérifie :

- $1 - \tau(x, y) = \perp(1 - x, 1 - y)$
- $1 - \perp(x, y) = \tau(1 - x, 1 - y)$

Les différentes opérations décrites dans cette section permettent de manipuler les SEFs et d'ainsi raisonner dessus. Néanmoins les valeurs d'entrées sont souvent numériques, il convient alors d'être capable, une fois la manipulation des données terminée, de défuzzifier le résultat obtenu afin de pouvoir retourner une valeur définie sur un ensemble numérique équivalent à celui des données en entrée, exploitable par le système.

3.1.3 Défuzzification

La dernière étape pour opérationnaliser un système flou est la défuzzification. Cette étape permet de trouver la classe d'appartenance représentant au mieux le résultat obtenu après les divers opérations effectuées sur les SEFs. Au sein des méthodes les plus utilisées, on peut citer la méthode des maximas (MoM) et la méthode du centre de gravité (*Center of Gravity*, CoG).

La méthode MoM consiste à prendre l'abscisse de la moyenne des points maximaux (en hauteur) de la fonction d'appartenance du SEF résultat. Cette méthode permet de n'utiliser que les valeurs les plus représentatives du SEF résultat et de gérer le cas où plusieurs valeurs possèdent la même hauteur maximale.

La méthode CoG consiste à prendre l'abscisse du centre de gravité du SEF résultat. Cette méthode est considérée comme plus efficace dans le cas d'un SEF résultat étalé sur l'ensemble de définition, car elle prend en compte l'ensemble du SEF réponse et non seulement les valeurs à forte hauteur.

La logique floue nous permet de nous affranchir de seuils brusques entre classes d'appartenance et d'ainsi identifier de manière plus précise les changements de comportement adoptés par les apprenants. Une fois la stratégie d'un apprenant en réponse à une tâche identifiée, nous sommes à même de le comparer avec d'autres apprenants. La caractérisation correcte d'un objet est en effet un préalable essentiel à sa comparaison avec d'autres objets. Nous souhaitons dans ce travail créer des profils d'apprenants en nous basant sur leurs données d'apprentissage d'une LC. Créer des profils revient à créer des sous groupes d'apprenants, des sous groupes d'apprenants similaires en intra et dissimilaires en inter. La similarité entre deux objets, menant à leur regroupement, ne peut s'établir qu'à travers un procédé de comparaison suffisamment sensible pour capter les différences, et suffisamment souple pour identifier les similarités.

3.2 L'apprentissage non supervisé : le *clustering*

L'objectif de la classification non supervisée ou *clustering* est de découvrir des sous-ensembles homogènes d'observations ou *clusters* parmi un plus grand ensemble de données ([Jain et Dubes, 1988]). Ces regroupements doivent être utiles pour l'utilisateur, où l'utilité est définie en fonction des objectifs de l'analyse des données. La notion d'utilité étant ancrée dans le contexte de l'analyse, il existe une multitude d'algorithmes de clustering.

Les observations ou encore objets partagent des propriétés communes appelées attributs qui vont être à la base de leurs regroupements. Dans le cas de l'apprentissage non supervisé, les propriétés des clusters ne sont pas connues à l'avance. Par propriétés de clusters on pourra comprendre leurs nombres, leurs tailles, leurs distributions, leurs formes ... C'est pourquoi dans le cas d'un partitionnement supervisé on parlera d'un problème de classification des nouvelles entrées, alors qu'en clustering la structure des données reste à découvrir. Il existe bien sûr des cas intermédiaires à ces deux extrêmes (cf. figure 3.4). En effet bien que les algorithmes de classification entièrement non supervisée soient essentiels du fait de leur généralisation et donc applications possibles à n'importe quelle situation, en contexte réel, il est rare qu'aucune connaissances sur les données ne soient disponibles, ne serait-ce que parce qu'elles ont été collectées, et donc choisies pour leur valeur potentiellement informative. Ces cas intermédiaires seront examinés en section 3.3.1.

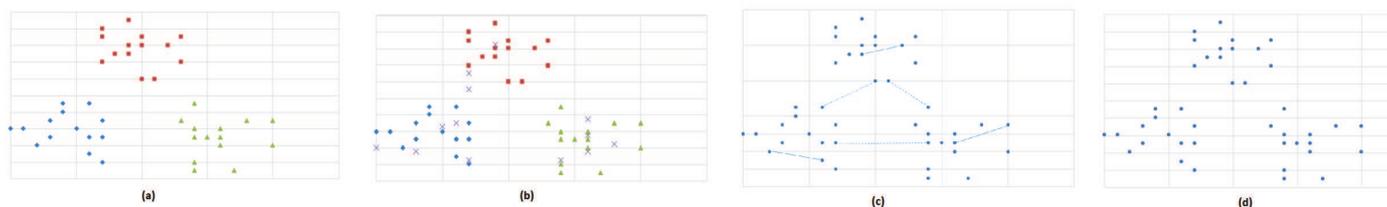


FIGURE 3.4 – Différents problèmes d'apprentissage : (a) Supervisé ; (b) Partiellement labellisé ; (c) Partiellement contraint ; (d) Non supervisé

Afin de découvrir ces clusters, deux grandes familles de procédés sont utilisées. La première appelée *model based* se basent sur des distributions probabilistes des observations ([Banfield et Raftery, 1993]). La deuxième grande famille qui nous intéresse tout particulièrement dans cette thèse repose quant à elle sur des notions de représentations et distances géométriques ([Berkhin, 2002]).

Parallèlement à cette propriété distinctive entre algorithmes de clustering, on peut également les différencier selon leur méthode de découpage de l'ensemble de données. La première méthode de découpage classique est dite hiérarchique. Les méthodes hiérarchiques ascendantes partent d'un découpage des données en singleton pour aller vers des clusters réunissant de plus en plus d'objets. Les méthodes hiérarchiques dites descendantes à contrario partent de l'ensemble entier de données pour aller vers un découpage de plus en plus fin. L'autre méthode est dite de partitionnement et produit directement une classification en un nombre de clusters K

défini au préalable. Cette distinction un peu datée permet néanmoins une première approche assez didactique du domaine.

Le cœur de tout algorithme de clustering repose dans sa manière de comparer les objets. Il existe de nombreuses manières d'établir la proximité ou distance entre deux observations. Cette prolifération de mesure de distance donne lieu sur un même ensemble de données à autant de classifications différentes de cet ensemble ([Jain et al., 1999a]). Son adaptation au problème considéré est donc un des grands enjeux de la formulation d'une classification et fera l'objet dans cette thèse d'un intérêt tout particulier.

Des milliers d'algorithmes de clustering existent dans la littérature, dans une grande variété de domaines d'applications, allant de la biologie à l'économie en passant par l'environnement. Cet état de fait rend extrêmement difficile une approche exhaustive de présentation des approches publiées. Dans tous les cas, le processus de clustering dépend, pour [Jain et Dubes, 1988], de la réponse à un certain nombre de questions :

1. Qu'est-ce qu'un cluster ?
2. Quelles attributs/caractéristiques doivent être utilisés ?
3. Les données doivent-elles être normalisées ?
4. Les données contiennent-elles des *outliers* (valeurs anormales)
5. Comment définit-on la similarité entre deux objets ?
6. Combien de clusters sont présents dans les données ?
7. Quelle méthode de clustering doit être utilisée ?
8. Un regroupement est-il naturellement présent dans les données ?
9. Les clusters et la partition finale obtenue sont-ils valides ?

Cet état de l'art essaiera malgré tout de présenter le domaine de la classification non supervisée en fonction des trois piliers de construction d'un algorithme de clustering : le choix d'une fonction objective, des modèles génératifs probabilistes et des heuristiques ([Jain et Dubes, 1988]).

3.2.1 Notions générales

La classification d'un ensemble Z de données s'obtient par son découpage en k sous-ensembles de clusters (groupes ou classes). Pour l'instant nous assumons que k est connu, par exemple par connaissances à priori sur l'ensemble de données, et cette question fera l'objet d'une présentation dans notre chapitre sur la description de notre algorithme.

On pourra noter dans cette état de l'art qu'il existe un package R nommé `NbClust()` ([Charrad et al., 2014]) regroupant 30 indices du nombre optimal de clusters pour un ensemble de données. Malgré cela, les indices sont souvent contradictoires, et même si la règle de la majorité peut être appliquée, elle ne constitue pas forcément le choix le plus légitime.

Les techniques de clustering peuvent être appliquées à des données quantitatives (numériques), qualitatives (catégoriques ou nominales), symboliques ou un mélange des trois. Dans ce chapitre et de manière plus générale dans cette thèse nous nous intéresserons particulièrement aux données de types numériques, étant celles avec lesquels nous devons travailler. Les données sont généralement des observations multicritères sur des mêmes objets. Chaque observation consiste donc en d mesures de variables, regroupées dans un vecteur ligne de dimension d , $z_i = [z_{1i}, z_{2i}, \dots, z_{di}]$, $z_i \in R_n$. Un nombre n d'observations est formalisé par $Z = \{z_i, i = 1, 2, \dots, n\}$, et est représenté par une matrice $Z_{d \times n}$:

$$Z_{d \times n} = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1n} \\ z_{21} & z_{22} & \dots & z_{2n} \\ \dots & \dots & \dots & \dots \\ z_{d1} & \dots & \dots & z_{dn} \end{pmatrix} \quad (3.1)$$

Dans notre cas, les n vecteurs lignes représenteront nos apprenants et leurs résultats aux d dimensions psycholinguistiques. Un cluster c peut être défini de diverses manières selon Jain et Dubes ([Jain et Dubes, 1988]) et Blashfield et Aldenderfer ([Blashfield et Aldenderfer, 1988]) :

- Approche théorique : "Un cluster est un ensemble d'entités qui sont similaires et dont les entités de différents clusters ne sont pas similaires."
- Approche géométrique : "Un cluster est une agrégation de points dans l'espace tel que la distance entre deux points dans un même cluster est inférieure à la distance entre n'importe quel point appartenant au cluster et n'importe quel point n'y appartenant pas."
- Approche ensembliste : "Des clusters peuvent être décrits comme des régions connectées dans un espace multidimensionnel contenant une densité relativement forte de points, séparées entre elles de régions contenant une densité relativement faible de points."

Une partition P sur Z en k classes est définie par $P = \{c_j, j = 1, 2, \dots, K\}$ avec $c_j \in P$:

1. $c_k \neq \emptyset$ pour $k = 1, \dots, K$
2. $\cup_{k=1}^K c_k = Z$
3. $c_l \cap c_k = \emptyset$

Les contraintes 2 et 3 peuvent être relâchées dans un grand nombre de cas. En particulier, la relaxation de la contrainte 3 donne lieu à un partitionnement flou de l'ensemble Z , où un objet peut appartenir à un ou plusieurs clusters. Cette particularité fait l'objet d'un approfondissement en section 3.2.4.

Le partitionnement de cet ensemble est généralement formulé comme un problème d'optimisation et différents critères d'optimisation ont été formulés dans la littérature. Globalement ces critères d'optimisation sont basés sur l'idée d'un rapport homogénéité intracluster sur hétérogénéité intercluster. Les algorithmes heuristiques de partitionnement essaient d'optimiser ce critère à travers la minimisation d'une fonction objective (section 3.2.3.2). Du fait de la nature non convexe de ces fonctions objectives, souvent un minimum local seul est trouvé et non un optimum global.

Type d'algorithme de classification non supervisée

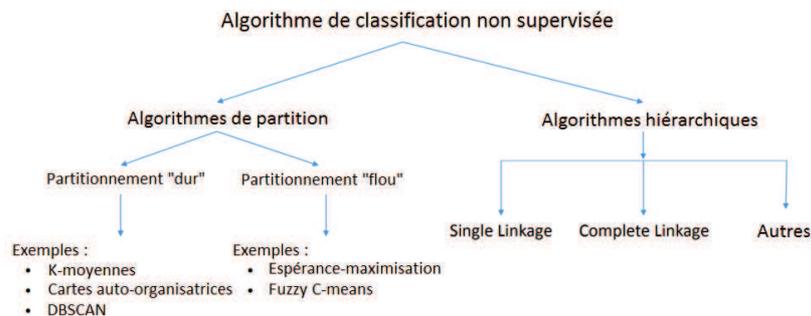


FIGURE 3.5 – Exemple de classification des méthodes de partitionnement de données

Nous présentons ici quelques méthodes de partitionnement, afin d'essayer de montrer le large panel de solutions à la classification non supervisée (cf. figure 3.5). En premier lieu nous abordons les classiques du clustering avec les méthodes hiérarchiques (section 3.2.3), reposant entièrement sur la notion de distance, puis nous introduisons la notion de fonction objective et d'optimisation avec l'algorithme des K-moyennes (section 3.2.3.2). Sa dérivée en version floue (section 3.2.4.1) et sa généralisation probabiliste (section 3.2.4.2) sont également présentées. Puis nous présentons deux familles d'algorithmes de clustering affranchis d'une fonction objective au sens strict du terme, avec la présentation du clustering basé sur densité (section 3.2.3.3), et l'utilisation des réseaux de neurones (section 3.2.3.4). Mais d'abord, nous revenons sur la notion de distance, particulièrement importante pour notre problématique.

3.2.2 Mesure de distance ou similarité

La performance de nombreux algorithmes d'apprentissage et de *data mining* reposent drastiquement sur l'obtention d'une mesure de distance adaptée à l'ensemble de données. Ce que l'on demande en fait à une bonne mesure de distance est de représenter correctement les relations entre les données. Ce problème est particulièrement central dans le domaine de la classification non supervisée puisque dès lors que peu de connaissances existent sur la forme de la partition finale, il n'existe pas de vraie réponse au processus de clustering. Par exemple si l'on doit regrouper en classes différentes des ouvrages, que trois algorithmes sont utilisés et que l'un d'entre eux renvoie un regroupement par auteur, l'autre par domaine, et le troisième par époque, lequel a raison ? Pire encore, si l'on choisit de classer les ouvrages par domaine, que faire d'un livre en psychologie-sociale, à cheval entre deux domaines ? Et d'ailleurs comment indiquer à un algorithme de clustering notre volonté de les classer par domaine ? Ces deux dernières questions seront traitées dans les sections 3.2.4 et 3.3.1 respectivement. Pour ce qui est de la première, commençons par comprendre comment est construit une mesure de distance ou similarité.

La notion de distance est à la base même de la définition d'un espace métrique. On appelle distance sur un ensemble E une application d de $E * E$ dans l'ensemble R^+ telle que, quel que soit z appartenant à E ([Verley,]):

1. $\forall z_i, z_j \in Z, d(z_i, z_j) = 0 \iff z_i = z_j$
2. $\forall z_i, z_j \in Z, d(z_i, z_j) = d(z_j, z_i)$
3. $\forall z_i, z_j, z_k \in Z, d(z_i, z_j) \leq d(z_i, z_k) + d(z_k, z_j)$

Une classe très importante d'espaces métriques est représentée par les espaces vectoriels normés et sont ceux ressemblant le plus aux espaces numériques habituels. On y définit la notion de distance entre deux éléments par la norme de leur différence : $d(x, y) = \|x - y\|$ La distance entre deux objets avec un certain nombre d'attributs est typiquement définie par une combinaison des distances de leurs attributs individuels. Nous notons ici trois normes communes pour x et y possédant deux attributs :

$$d_1(x, y) = |x_1 - y_1| + |x_2 - y_2| = \|x - y\|_1$$

$$d_2(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} = \|x - y\|_2$$

$$d_3(x, y) = \sup(|x_1 - y_1|, |x_2 - y_2|)$$

Nous pouvons généraliser la deuxième distance, nommée distance euclidienne, pour des éléments à n dimensions comme suit :

$$d(x, y) = \sqrt{\sum_{k=1}^n (x_k - y_k)^2} \quad (3.2)$$

Cette distance euclidienne appartient elle-même aux distances de Minkowski (Tan et al., 2006) :

$$d(x, y) = \left(\sqrt[r]{\sum_{k=1}^n |x_k - y_k|^r} \right)^{1/r} \quad (3.3)$$

La dernière propriété des distances référencée s'appelle l'inégalité triangulaire, et, si elle est relâchée, il ne s'agit plus d'une distance à proprement parler, car on sort de l'espace métrique, et les deux propriétés restantes caractérisent alors ce qu'on nomme plus globalement une mesure de dissimilarité.

Par opposition à celle-ci on définit les mesures de similarité, qui sont généralement leur transposé et possèdent les mêmes propriétés, et on regroupe ces deux notions sous le terme de mesure de proximité. Le tableau 3.2 définit les relations possibles entre distance et similarité en fonction du type de mesures des attributs de nos données (nominal *vs.* ordinal *vs.* numérique ou métrique d'intervalle). Dans ce tableau x et y sont deux objets à un seul attribut du type défini par le tableau.

Type d'attribut	Dissimilarité	Similarité
Nominal	$d = \begin{cases} 0, & \text{if } x = y \\ 1, & \text{if } x \neq y \end{cases}$	$s = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{if } x \neq y \end{cases}$
Ordinal	$d = x - y / (n - 1)$	$s = 1 - d$
Numérique ou métrique d'intervalle	$d = x - y $	$s = -d, s = \frac{1}{1 + d}, s = e^{-d}, s = 1 - \frac{d - \min_d}{\max_d - \min_d}$

TABLE 3.2 – Rapport entre mesure de similarité et de dissimilarité selon le type de l'attribut considéré, tableau tiré de [Tan, 2006], où n est le nombre (entier) de valeurs dans l'ensemble de définition de l'attribut

Nous ne pouvons pas lister ici toutes les mesures de similarités, étant donné l'énorme quantité disponible. Par exemple pour les mesures binaires, Choi et collaborateurs ([Choi et al., 2010]) en recensent soixante-seize, alors que Cha et collaborateurs ([Cha, 2007]) en dénombrent 45 pour les densités de probabilité. Nous proposons juste de comprendre comment se construit une mesure de similarité dans l'objectif de comparaison de deux objets multidimensionnels.

Plusieurs définitions de la similarité sont utilisées en clustering, dépendant de l'objectif du clustering. Si nous revenons à notre ensemble de données représenté par une matrice Z de dimensions n éléments \times d attributs (cf. 3.2.1), une classe populaire de fonctions de similarité consiste en des mesures linéaires de type $(x_i - x_j)^T M (x_i - x_j)$ où x_i et x_j sont deux items que nous souhaitons comparer et appartenant à Z et M une matrice définie positive, c'est-à-dire une matrice symétrique dont toutes les valeurs propres sont non négatives (égales à 0 ou positives).

Le problème de la distance euclidienne présentée dans l'équation 3.2 est qu'elle ne permet pas d'ordonner l'importance des différents attributs de nos objets. Pour remédier à cela nous pouvons associer à chaque attribut un poids $w_k \geq 0$, $1 \leq k \leq m$ dans le calcul de la distance, en fonction de son importance :

$$d_w(x_i, x_j) = \left(\sum_{k=1}^m w_k (x_{ik} - x_{jk})^2 \right)^{1/2} \quad (3.4)$$

où m est le nombre d'attribut caractérisant un objet.

Cette configuration nous permet de traiter chaque attribut de manière indépendante, mais ne rend pas compte des interactions entre eux. C'est pourquoi a été introduit une représentation matricielle de la mesure de distance :

$$d_A(x_i, x_j) = \left((x_i - x_j)^T A (x_i - x_j) \right)^{1/2} \quad (3.5)$$

Cette notation permet d'ailleurs de passer d'une mesure de distance globale à une multiplicité de distances locales sur un même ensemble de données, captant leur aspect hétérogène. Cette approche comme nous le

verrons par la suite a été démontrée comme très efficace.

Par une mesure de distance appliquée à l'ensemble Z on obtient une matrice de similarité $S^{n \times n}$ dont l'élément s_{ij} représente le coefficient de similarité entre l'élément x_i et l'élément x_j . C'est sur cette matrice de similarité (vs. dissimilarité) que vont s'appliquer nombre de processus de partitionnement (par exemple l'algorithme des K-moyennes, les algorithmes utilisant l'astuce du noyau comme les *Support Vector Machine*, et plus globalement tous les algorithmes reposant sur la notion de voisinage). Une connaissance à priori des données pourra guider le choix (manuel ou automatique) ou la création (manuelle ou par apprentissage) d'une mesure de distance appropriée au contexte.

3.2.3 Méthodes classiques

Nous définissons une méthodes comme classique lorsque la partition finale résultant d'une méthode possède les caractéristiques suivantes :

- Si C_1 et C_2 sont deux clusters de la partition finale, alors $C_1 \cap C_2 = \emptyset$.
- o_i un objet de l'ensemble de données partitionné appartient à un unique cluster.

Ces méthodes de clustering sont à mettre en opposition avec celles dites floues, sur lesquelles nous reviendrons en section 3.2.4.

3.2.3.1 Méthode de classification hiérarchique

Les méthodes de classification hiérarchique produisent un arbre ou dendrogramme mettant en œuvre une succession de partitions. Les méthodes descendantes divisent ces partitions successives jusqu'à l'obtention d'une partition à clusters singletons. Chaque cluster de la partition courante est divisé en deux à chaque étape (itérative) de l'algorithme pour l'obtention de la partition $n + 1$. Les méthodes ascendantes fusionnent les clusters singletons jusqu'à l'obtention d'un seul cluster regroupant toutes les données contenues dans Z (cf. figure 3.6).

En techniques ascendantes ou agglomératives, qui sont les plus utilisées, il existe plusieurs manières de procéder à la fusion de deux clusters. Les deux les plus connues sont les suivantes :

1. La méthode de regroupement *single-linkage* (dite du plus proche voisin) groupe entre eux les deux clusters possédant chacun un objet dont la distance entre eux est la plus petite.
2. La méthode *complete-linkage* regroupe les deux clusters dont la distance maximum entre deux objets est la plus petite.

On peut également citer la méthode *average-linkage*, centroid ([Lance et Williams, 1967]) ou de Ward. La méthode de *single-linkage* est dite efficace pour capter les clusters de formes non elliptiques mais sont sensibles aux outliers et au bruit. La méthode *complete-linkage* est moins sensible aux outliers du fait qu'elle ne les intègre à la partition que tardivement car ils représentent la distance maximum la plus forte ([Tan, 2006]).

En résumé, étant donné une matrice de dissimilarité $D_{n \times n}$, les étapes d'un algorithme de clustering hiérarchique agglomératif sont :

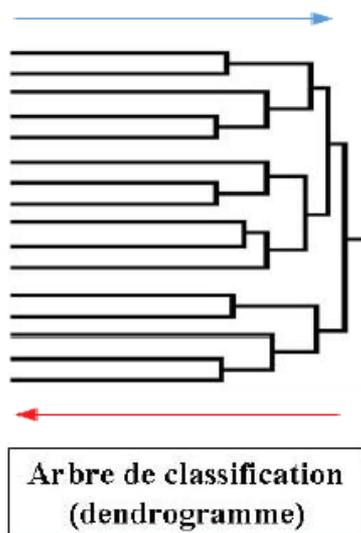


FIGURE 3.6 – Dendrogramme d’une classification hiérarchique. En rouge les méthodes descendantes et en bleu les méthodes ascendantes

Algorithme 3.1 Algorithme de classification hiérarchique ascendante

- 1: Créer n clusters, chacun contenant seulement un objet.
 - 2: Chercher dans la matrice de dissimilarité D la paire d’objets la plus similaire, notée d_{rs} pour les objets r et s .
 - 3: Combiner les deux objets r et s dans un nouveau cluster (rs) et réduire le nombre de cluster de 1 en supprimant les lignes et colonnes représentant r et s . Mettre à jour D avec les dissimilarités entre le nouvel objet (rs) et tous les autres objets.
 - 4: Répéter les étapes 2 et 3 ($n - 1$) fois, c’est à dire jusqu’à ce que tous les objets ne forment qu’un seul cluster. A chaque itération, garder une trace des clusters fusionnés et leur valeur de dissimilarité associée.
-

Ce genre d’algorithme est surtout utilisé afin de pouvoir créer une taxonomie ou une ontologie des données. Le clustering hiérarchique renseigne en effet sur les liens entre clusters. De par sa nature, les différents niveaux d’agglomération peuvent être assimilés à une succession de recouvrements entre clusters, permettant ainsi d’établir également une relation (hiérarchique) entre clusters (cf. figure 3.7). Cependant les algorithmes hiérarchiques sont coûteux en termes de capacité spatiale et de temps. Comme dit précédemment, ils sont aussi sensibles au « bruit » contenu dans les données et aux outliers.

Pour un état de l’art récent centré sur l’application des méthodes de classification hiérarchique en *data mining* et ses diverses implémentations voir [Murtagh et Contreras, 2017].

3.2.3.2 Méthode de partitionnement direct

Par opposition aux méthodes hiérarchiques, les méthodes de partitionnement trouvent tous les clusters de la partition de manière simultanée et n’imposent pas une structure hiérarchique des données. Ceci est rendu possible par la présence d’une fonction objective globale, c’est-à-dire un objectif global à atteindre, et non plus une vision fragmentaire de la partition à optimiser. En ce qui concerne les techniques probabilistes de clustering (cf. section sur EM), cette fonction objective peut s’apparenter à minimiser la somme du carré des écarts (SCE) entre le paramétrage actuel du modèle et les données ; pour l’algorithme des K-moyennes elle consistera en une minimisation de la distance entre les items et le prototype de leur cluster respectif.

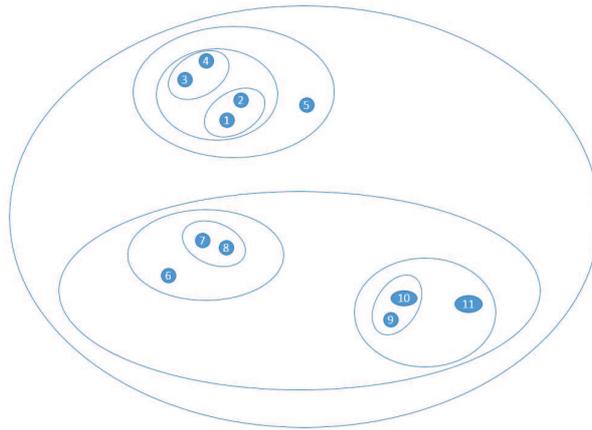


FIGURE 3.7 – Agglomération hiérarchique.

Plus généralement, cette fonction objective décrit l'état interne de la partition, en fonction de l'homogénéité intracluster relativement à l'hétérogénéité intercluster. Minimiser la fonction objective revient dès lors à maximiser ces deux derniers éléments.

Tout comme les méthodes *complete link* et *single link* pour les méthodes hiérarchiques, l'algorithme des K-moyennes est la plus populaire et la plus simple des méthodes de partitionnement direct. Bien que sa découverte ("ses découvertes", cf. [Jain, 2010]) remonte à plus de 70 ans, il reste un des algorithmes les plus utilisés, et surtout dérivés. Nous décrivons ici la structure basique de l'algorithme des K-moyennes.

Premièrement il faut déterminer un nombre pré-spécifié k (nombre de clusters souhaité) de centroïdes (ou prototypes). Chaque point de l'ensemble de données est ensuite assigné au centroïde le plus proche de lui, et la collection de points assignée à un centroïde forme ainsi un cluster. L'ensemble des centroïdes est ensuite mis à jour, de manière à ce que chaque centroïde représente le centre de densité de son cluster. L'assignation des points à un centroïde puis la mise à jour des centroïdes sont ensuite répétées jusqu'à ce que l'assignement des points ne change plus le statut de la partition, cela revenant à arrêter lorsqu'il y a convergence.

Algorithme 3.2 Algorithme des K-moyennes basique

Selectionner k points comme centroïdes initiaux.

repeat

Créer k cluster en assignant chaque objet au centroïde le plus proche.

Recalculer le centroïde de chaque cluster.

until Les centroïdes ne changent plus.

La figure 3.8 illustre le déroulement de l'algorithme des K-moyennes. Les centroïdes (illustrés par le symbole « + ») sont d'abord situés au centre de la masse de points la plus dense. Puis, par les itérations successives, ils se déplacent vers le centre des trois masses visuellement identifiables dans l'ensemble en deux dimensions des points. La condition d'arrêt du clustering des K-moyennes est souvent remplacée par une marge d'erreur statuant l'arrêt si le changement des centroïdes est inférieur à ce seuil d'erreur. Dans cet exemple, le calcul des centroïdes repose sur la moyenne des positions pour leurs clusters respectifs.

L'algorithme des K-moyennes converge toujours. Cela étant, il est possible que la convergence se fasse, en accord avec la fonction objective choisie, au niveau d'un minimum local. L'algorithme des K-moyennes est extrêmement conditionné par le choix premier de ses centroïdes, souvent choisis aléatoirement. Une des manières de pallier à cette sensibilité aux outliers, découlant directement de ce choix, est d'appliquer l'algorithme plusieurs fois sur le même ensemble de données, et d'ensuite choisir la partition avec le SCE minimum. Une autre solution est proposé par Arthur et Vassilvitskii ([Arthur et Vassilvitskii, 2007]) pour

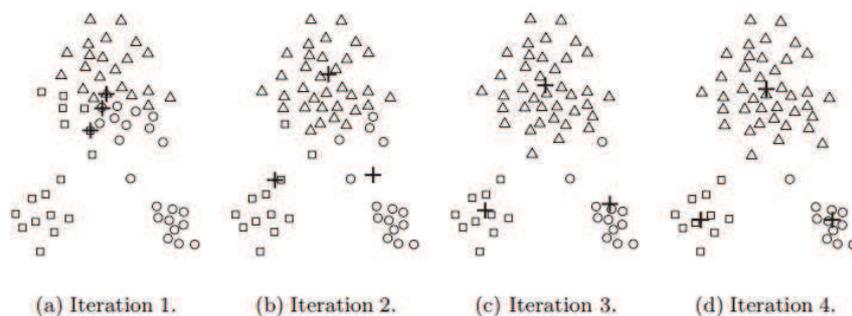


FIGURE 3.8 – Illustration du déroulement de l'algorithme des K-moyennes en quatre itérations ([Tan, 2006])

l'initialisation des centroïdes, donnant naissance à l'algorithme k-means++.

La sensibilité aux outliers présentée par cet algorithme ne constitue pas un problème lorsque les données sont pré-traitées pour en éliminer le bruit. Dans notre procédé, du fait de la transformation appliquée à nos données avant le passage d'un algorithme de clustering classique, les outliers attestés sont retirés avant l'application de l'algorithme de partitionnement.

3.2.3.3 Clustering fondé sur la densité et DBSCAN

Malgré la simplicité et la rapidité de l'algorithme des K-moyennes, le clustering fondé sur la densité (*Density-based clustering*) possède de nombreux avantages. Tout d'abord, il ne nécessite pas de choisir un nombre de clusters *a priori*. De plus, il permet la détection de clusters aux formes variées et non plus uniquement sphériques, et est résistant par rapport au bruit et aux outliers. La notion de densité est cette fois-ci au centre du problème, à la place de la notion de similarité ou distance. Ceci peut apparaître comme une simplification de fait, la densité possédant un nombre possible de définition apparemment moins élevé. La définition traditionnelle est l'approche center-based. C'est aussi celle utilisée dans l'algorithme DBSCAN ([Ester et al., 1996]) que nous utilisons ici comme illustration du partitionnement basé sur la densité. La densité d'un point spécifique est calculée par le nombre de points présents dans un rayon prédéfini autour du point sélectionné, incluant ce dernier. On peut dès lors différencier trois types de points en fonction de leurs densités : le point noyau, doté d'une forte densité et qui se trouve relativement au centre d'une zone peuplée, généralement dépassant un certain seuil de densité pré spécifié par l'utilisateur ; les points frontières, appartenant à une zone dense mais en représentant la frontière, car suivi d'une zone vide ; et les points outliers dont la densité est très faible et constitue un bruit plutôt qu'une information. Pour illustration voir figure 3.9.

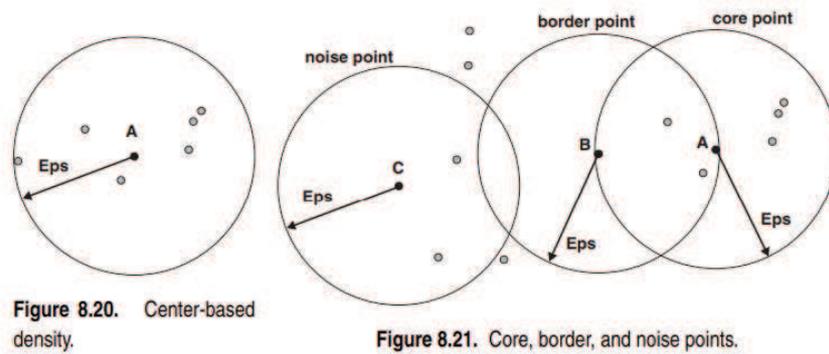


FIGURE 3.9 – [Tan, 2006] où Eps correspond au rayon choisi

Finalement une fois la densité de chaque point calculée, l'algorithme DBSCAN en lui-même est très simple : Éliminer les points outliers, regrouper les points noyaux qui sont suffisamment proches (en général en fonction du rayon utilisé pour le calcul de densité), rattacher les points frontières au point noyau le plus proche.

Cette définition de la densité (et l'algorithme DBSCAN qui en découle) est simple mais va donc entièrement reposer sur le choix du rayon, et éventuellement d'une densité seuil pour la définition des points noyaux (minimum de points nécessaire à la création d'un cluster, souvent noté MinPts). Ces paramétrages initiaux viennent en quelque sorte remplacer les problèmes de choix de nombre de clusters présents pour l'algorithme des K-moyennes et dérivés. De plus, le principe d'un rayon unique pour évaluer la densité de clusters de densités potentiellement différentes n'est pas adapté. Si le rayon est trop petit il risque de passer à côté des clusters à densité faible, et inversement si celui-ci est trop grand il fusionnera les clusters de densité élevée. Le problème de densité variable entre clusters d'une même partition est un des problèmes que nous aurons à gérer du fait de la nature de nos données, et de la recherche d'information effectuée. Il apparait en effet que les profils d'acquisition d'une LE, au cœur de notre étude, ne sont en aucun cas numériquement équilibrés parmi la population des apprenants.

L'algorithme OPTICS ([Ankerst et al., 1999]) résout ce problème en laissant flotter le rayon considéré. OPTICS applique les principes de DBSCAN mais l'algorithme est traité un nombre infini de rayon ϵ_i , où $0 \leq \epsilon_i \leq \epsilon$. Le résultat de cet algorithme n'est donc pas une partition mais un ordonnancement des points, accompagné d'un paramètre de « *reachability* » (accessibilité), représentant la structure hiérarchique interne des clusters. Cette structure peut s'observer à travers un « *reachability plot* », représentant les objets et leurs valeurs respectives de *reachability*. L'ordre de traitement des objets est présenté sur l'axe des abscisses et leur valeur de *reachability* sur l'axe des ordonnées. Des valeurs faibles de *reachability* indiquent une forte densité, et apparaissent sur le graphique sous forme de « vallées ». Ce graphique permet une visualisation rapide de la structure interne précédemment mentionnée, et d'ainsi pouvoir fixer le nombre minimum de points nécessaires à la création d'un cluster (représenté par la barre rouge sur la figure 3.10) .

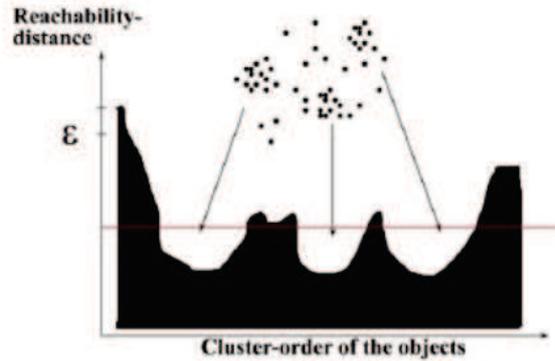


FIGURE 3.10 – [Wowczko, 2013] *Reachability plot*

Pour une comparaison détaillées des algorithmes DBSCAN et OPTICS voir [Shah et al., 2012].

Pour résumer, les méthodes de partitionnement possèdent trois caractéristiques communes dont chacune est sujette à des variantes et des optimisations diverses : 1) une première phase d’initialisation, 2) le ré-assignement des items à regrouper dans les clusters et 3) la mise à jour des paramètres des clusters de la partition. Bien que les propriétés statistiques, heuristiques et de calcul soient conditionnées par la réalisation de ces trois éléments, il existe d’autres facteurs influençant le déroulement du processus de regroupement, souvent contenus dans la nature même des données (données manquantes, à hautes dimensions, autres).

Étant donné la nature non convexe (admettant un optimum global et des optimums locaux) de la plupart des fonctions objectives utilisées dans les méthodes de partitionnement, un des risques majoritairement rencontrés en clustering est que la partition finale ne représente que l’atteinte d’un optimum local. Une des manières de pallier cette problématique est d’appliquer un grand nombre de fois l’algorithme de partitionnement avec à chaque fois une initialisation différente, pour ensuite retenir la partition dont l’erreur quadratique moyenne (ou SCE) est la plus faible. Une autre manière de faire est de choisir parmi les différents indices existants dans la littérature, afin de trouver une optimisation optimale. L’initialisation des méthodes de partitionnement direct est un enjeu majeur dans la littérature ([Fraley et Raftery, 1998], [Ji He et al., 2004], [Khan et Ahmad, 2004],...)

Comme dit précédemment, l’approche originale présentée dans ce travail possède un pré traitement particulier des données, considéré comme une étape à part entière du processus de regroupement, permettant d’éviter la plupart des problèmes d’initialisation des méthodes de regroupement par partitionnement.

3.2.3.4 Réseaux de neurones et cartes topologiques

Une des caractéristiques les plus évidentes des cerveaux des mammifères est leur organisation topographique. Des voisinages de cellules nerveuses (neurones) peuvent être activés par des stimuli provenant de régions de neurones voisines et vont à leur tour activer des cellules voisines. Ainsi les connexions établies entre cellules nerveuses suivent une carte topologique, et chaque cellule possède un voisinage. Les cartes topologiques de Kohonen (ou auto-organisatrice) s’inspirent de cette organisation spatiale afin de représenter des données multidimensionnelles, de manière à préserver leur topologie intrinsèque en les représentant par des vecteurs prototypes (*code vectors*) sur un espace de faible dimension qu’on appelle carte.

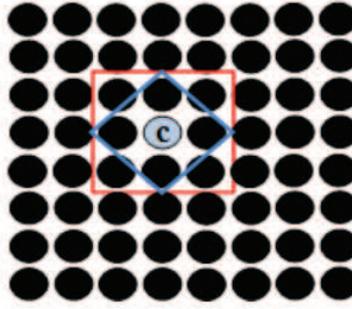


FIGURE 3.11 – Carte topologique à deux dimensions avec un voisinage d'ordre 1 autour du neurone c .
Voisinage carré (rouge) avec 8 voisins et hexagonal (bleu) avec 4 voisins ([Allab et al., 2011])

Chaque neurone c de la carte est connecté au reste de la carte par son voisinage. La carte est alimentée par d axones du même nombre que celui des attributs d'un élément x de notre ensemble de données. Le voisinage du neurone c sur la carte A se note $N_A(c)$ et définit ainsi un ordonnancement topologique des cellules. A chaque neurone c de A est associé un vecteur prototype w^c appartenant à un ensemble W tel que deux neurones voisins c et r soient associés à deux vecteurs w^c et w^r proche selon la mesure de distance choisie.

$$f(x_i) = \arg \min_{r \in C} d^T(x_i, w^r)$$

L'objectif de l'algorithme de Kohonen ([Kohonen, 1990]) est donc d'adapter la forme de A contenant m neurones à la distribution de l'ensemble de données d'entrée X à n éléments. La version en ligne de cet algorithme (proche des nuées dynamiques de [Diday, 1971]) possède deux étapes. La première est une étape d'affectation procédant grâce à une fonction f de projection d'un élément x_i de X dans la carte A , en lui associant un neurone c dont le vecteur prototype w^c est le plus proche de x_i :

$$d^T(x_i, w^{f(x_i)}) = \sum_{r \in C} K^T(\delta(r, f(x_i))) \|x_i - w^c\|^2$$

où $K^T(\delta(r, f(x_i)))$ est une fonction de voisinage autour du neurone choisi par $f(x_i)$ et T est le rayon de voisinage qui décroît au cours du temps; $\delta(r, f(x_i))$ est une distance entre deux neurones. Généralement on utilise $K^T(\delta(r, f(x_i))) = e^{-(\delta_{rc}/2T^2)}$.

Une fois les éléments x_i affecté à leur vecteur prototype, la deuxième étape de l'algorithme va mettre à jour ces derniers selon la formule suivante :

$$w^c = \frac{\sum_{r \in C} K_T(\delta_{cr}) X_r}{\sum_{r \in C} K_T(\delta_{cr} n_r)}$$

Ces deux étapes se succèdent de manière itérative jusqu'à ce que le recalcul des vecteurs prototypes reste stable.

L'un des principaux problèmes des algorithmes de partitionnement basé sur les réseaux de neurones à apprentissage compétitif est la stabilité. Moore ([Moore, 1988]) définit la stabilité d'un algorithme itératif comme le gage qu'une présentation infinie de données ne mènera qu'à un nombre fini de cluster, assurant ainsi une généralisation de la partition trouvée, malgré l'intégration de nouvelles données. Seulement les nouvelles données peuvent inclure des nouveaux patterns important sémantiquement et qui pourtant ne doivent pas écraser la connaissance précédemment apprise par l'algorithme de clustering. Ce problème est énoncé comme le dilemme stabilité-plasticité ([Carpenter et Grossberg, 1987]). Ces mêmes auteurs posant le dilemme proposent une solution sous la forme *Adaptative Resonance Theory* (ART), incluant la notion d'attente (expectations), permettant la création de nouveaux vecteurs prototypes, empêchant l'érosion des anciennes

associations input/vecteurs-prototypes. Pour plus de détails voir Grossberg (2013). Cette problématique est particulièrement intéressante dans le cas d’une utilisation continue d’un système de partitionnement, avec de nombreuses données arrivant dynamiquement sur un tel système. L’état d’avancement de nos travaux et la faible quantité de données actuellement disponible dans notre projet place la considération de ces aspects et donc des travaux y existants dans des perspectives d’utilisation.

3.2.4 Méthodes floues

La classification non supervisée floue est la synthèse entre le clustering et la théorie des ensembles flous ([Zadeh, 1965b], cf. section 3.1). Elle est supérieure aux méthodes dites « dures » lorsque les frontières entre clusters sont vagues et ambiguës ([Klir et Yuan, 1995]) et plus particulièrement lorsqu’elles se recouvrent. Le clustering flou permet la visualisation d’un tel recouvrement des données dans la matrice d’appartenance résultante d’une telle procédure. Ceci est rendu possible par l’assignation à chaque point de l’ensemble de données d’un indice d’appartenance à chacun des clusters de la partition ([Baraldi et Blonda, 1999a], [Baraldi et Blonda, 1999b]).

3.2.4.1 Fuzzy C-means

Le clustering flou consiste à relâcher la contrainte d’une appartenance unique et entière à un cluster pour un item donné. Pour tout individu x_i , i appartenant à N , on associe donc un vecteur U_i $1 \times K$ où u_{ij} représente le taux d’appartenance de x_i au cluster j , $0 \leq j \leq K$. Généralement cette appartenance est comprise entre 0 et 1. L’algorithme fuzzy C-means (FCM) développé par Bezdek ([Bezdek et al., 1984]) est une variation de l’algorithme des K-moyennes. Étant donné une matrice $Z_{d \times n}$:

$$Z_{d \times n} = \begin{pmatrix} z_{11} & z_{12} & \dots & z_{1n} \\ z_{21} & z_{22} & \dots & z_{2n} \\ \dots & \dots & \dots & \dots \\ z_{d1} & \dots & \dots & z_{dn} \end{pmatrix} \quad (3.6)$$

Qui contient n observations possédant d attributs, une matrice $U_{c \times n}$:

$$U_{c \times n} = \begin{pmatrix} \mu_{11} & \mu_{12} & \dots & \mu_{1n} \\ \mu_{21} & \mu_{22} & \dots & \mu_{2n} \\ \dots & \dots & \dots & \dots \\ \mu_{c1} & \dots & \dots & \mu_{cn} \end{pmatrix} \quad (3.7)$$

est une matrice de partition floue de Z où μ_{ik} correspond au degré d’appartenance de l’observation i au cluster k , et un vecteur V :

$$V_c = \begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_c \end{pmatrix} \quad (3.8)$$

contenant les prototypes de clusters (centroïdes), $V_i \in R^n$, on obtient la fonction objective suivante F ([Dunn, 1973]) à minimiser :

$$F(Z; U; V) = \sum_{i=1}^c \sum_{k=1}^n (\mu_{ik})^m \| z_k - v_i \|_A^2 \quad (3.9)$$

Où m est un paramètre déterminant le niveau de flou de la partition résultante, et $\| \cdot \|_A$ une norme matricielle (basiquement la distance euclidienne sur la matrice identité A).

Pour un $m > 1$ donné et $1 < c < N$, on minimise F par l’algorithme suivant :

Algorithme 3.3 Algorithme FCM**repeat**Calcul des prototypes de clusters. Pour l'initialisation choisir c points aléatoirement.

Calcul des distances pour tous les points avec tous les prototypes de clusters.

Mise à jour de la matrice de partitionnement U par le calcul des nouveaux degrés d'appartenance de tous les individus avec tous les clusters, selon la mesure de distance choisie.**until** convergence i.e. $\|U^p - U^{p-1}\| < \epsilon$

Pour plus de détails et autres algorithmes on pourra se référer à [Babuška, 2000].

3.2.4.2 Classification non supervisée fondée sur la distribution, modèle de mélange et l'algorithme espérance-maximisation

L'algorithme espérance-maximisation (ou *model-based clustering*, ou *mixture-resolving clustering*) introduit par [Dempster et al., 1977] (originellement développé dans les cas de manque de données) est une généralisation probabiliste de l'algorithme des K-moyennes. Dans ce dernier, nous recherchons les centroïdes optimaux pour représenter leur cluster respectif. Dans les méthodes de clustering à modèles de mélanges, l'idée est que les patterns à regrouper sont issus de la même distribution (assumée connue), et l'objectif est dès lors d'identifier les paramètres de chaque distribution en présence, et de déterminer leur nombre en trouvant la loi de mélange (souvent une combinaison linéaire des sous-distributions) sur l'ensemble des données. L'appartenance d'un point à un cluster s'exprimera donc sous la forme d'une probabilité, cela étant assimilable à un degré d'appartenance, d'où la classification de cet algorithme et ceux découlant dans les méthodes floues de partitionnement.

Soit p une loi mélange sur un espace X , une loi de probabilité s'exprimant comme une combinaison linéaire de plusieurs lois de probabilités p_1, p_2, \dots, p_k sur X . A chaque p_i (composante de la loi mélange) est donc associé un coefficient (ou proportion) π_i tels que, pour tout x_j appartenant à X :

$$p(x_j) = \sum_{k=1}^g \pi_k p_k(x_j)$$

La plupart des travaux existant dans ce domaine assume que les composants du mélange de densité suivent une distribution gaussienne, et cherchent donc à paramétrer correctement ces gaussiennes individuelles à travers leur algorithme. La loi mélange devient alors une loi multinormale :

$$p(\cdot; \alpha_k) = N(\mu_k, \Sigma_k)$$

avec $\alpha_k = (\mu_k, \Sigma_k)$, $\mu_k \in R^d$ désignant la moyenne de la composante k et $\Sigma_k \in R^{d \times d}$ la matrice de variance correspondante.

L'obtention du paramétrage de la loi mélange s'effectue traditionnellement par l'obtention d'un maximum satisfaisant de vraisemblance ([Jain et al., 1999a]).

$$L(\theta; x) = \sum_{i=1}^n \ln \left(\sum_{k=1}^g \pi_k p(x_i; \alpha_k) \right)$$

Cependant l'optimisation directe sur θ est difficile, la solution analytique n'est donc pas une approche crédible. L'algorithme d'espérance maximisation permet de maximiser le log vraisemblance de manière itérative, afin d'obtenir le θ qui maximisera la vraisemblance, et ce grâce à deux étapes-clés.

- L'étape E procède au calcul de l'espérance associée au modèle qui revient à calculer la probabilité que l'individu x_i ait été généré par la loi de distribution dont le paramétrage est issu de l'itération

précédente (θ estimé, initialisé souvent de manière aléatoire). Dans cette étape, θ est donc fixé. Cette étape correspond à l'assignement des items à leur centroïde pour l'algorithme des K-moyennes.

- La phase M procède ensuite au calcul du maximum de vraisemblance en se basant sur le l'espérance calculée à l'étape E, ce qui permet d'obtenir un nouveau paramètre θ . Cette étape correspond à la mise à jour des centroïdes dans l'algorithme des K-moyennes.
- L'algorithme s'arrête si le log de vraisemblance obtenue entre deux itérations est stationnaire.

Cette méthode n'étant pas analytique mais heuristique, l'algorithme EM tout comme l'algorithme des K-moyennes peut converger vers un maximum (respectivement minimum) local, résultat grandement influencé par le choix initial de θ (respectivement des centroïdes). Aussi le choix de la loi de distribution revient à fixer la forme des clusters obtenus, tout comme le choix de la mesure de distance conditionne leur forme dans l'algorithme des K-moyennes.

Bien que les méthodes mentionnées ici soient efficaces dans certaines limites, leurs performances reposent sur la définition précise de paramètres spécifiques au problème (la distribution des données, le nombre de clusters). De plus, la compréhensibilité de certains clusters obtenus n'est pas expertisée, ce qui en data mining, rend le savoir ainsi découvert inutilisable. Ce constat est valable pour les méthodes que nous venons de présenter, à savoir les méthodes hiérarchiques, de partitionnement, basées sur densité, topologiques, et floues (nous n'avons fait qu'effleurer la surface du problème, et aurions pu parler de *support vector machine*, de méthodes basées sur graphe etc.). Par conséquent il est nécessaire de valider chaque partition une fois obtenue, en aval, ce que nous verrons notamment en section 5.3, ou en amont (ou du moins de maximiser nos chances) en abattant toutes nos cartes, ce qui peut aussi se traduire par l'intégration de connaissances supplémentaires, préalablement disponibles, au processus de partitionnement lui-même, ce que nous abordons dès maintenant.

3.3 L'intégration de connaissances

Les approches guidées par les données, entièrement non supervisées, sont adaptées lorsque qu'aucune connaissance préalable n'est disponible. Mais, il est bien connu que le résultat de ces méthodes peut apparaître obscur aux experts du domaine, rendant ainsi inutilisable le produit de ces algorithmes. Aussi, inclure à la genèse de ces algorithmes un aspect plus théorique propre à la problématique sur laquelle ils vont être appliqués peut améliorer l'interprétabilité de leur résultat.

Deux types d'outils distincts de data mining sont disponibles dans la littérature. Des outils généraux, ne prenant pas en compte les aspects particuliers des phénomènes étudiés, comme des caractéristiques des variables du modèle, spatiales, temporelles, indépendances, etc. ; et des outils conçus pour un certain domaine, auquel ils sont fortement liés, et peuvent difficilement être adaptés à un autre. Par exemple en économie, les outils de data mining disponibles intègrent plus souvent des modèles économiques existants que des techniques d'intelligence artificielle pure comme le deep learning. Il se trouve que ces algorithmes, bien que plus lents à construire, sont nécessaires du fait de la clarté de leurs résultats aux yeux de l'utilisateur, et également parce qu'ils implémentent les connaissances disponibles du domaine en question, permettant un déroulement guidé (et non supervisé) du processus de partitionnement.

3.3.1 La classification semi supervisée

Le domaine du clustering ou les méthodes de classification non supervisée de données sont des outils permettant de regrouper en cluster des groupes similaires d'objets. Le clustering repose essentiellement sur un certain nombre de suppositions quant à la structure des données, le nombre de clusters, et la distribution des données. L'efficacité des méthodes de classification non supervisée dépend pour beaucoup de la correspondance entre ces suppositions et la réalité des données. Dans les faits, pour beaucoup de problèmes appliqués, une certaine quantité de connaissances antérieures du problème ou d'informations supplémentaires aux données sont à disposition. Ces connaissances, lorsqu'elles sont intégrées au processus de regroupement,

en augmentent considérablement les performances ([Wagstaff et al., 2001], [Xiong et De la Torre, 2013], [Basu et al., 2004b], [Li et al., 2008]).

Ces connaissances peuvent concerner la nature des similarités pouvant être observées entre les objets, être exprimées sous la forme d'un échantillon de données déjà classifiées, des préférences d'un utilisateur, ou encore porter sur les caractéristiques même des clusters. Les méthodes utilisant ce type de connaissance extérieure à l'ensemble brut de données se trouvent donc à mi-chemin entre la classification supervisée et la non supervisée et sont regroupées sous l'intitulé classification semi supervisée ou clustering sous contraintes. De plus les approches développées du domaine peuvent s'appliquer à une variété de types de données complexes. Basu et al. ([Basu et al., 2009]) recense un certain nombre d'applications à des données relationnelles ([Desjardins et al., 2007]), des données textuelles ([Oyama et Tanaka, 2008]; [Tung et al., 2008]), ou encore de surveillance vidéo ([Yan et al., 2008]).

Souvent, ces informations sont sous la forme d'un sous ensemble labellisé (pré classifié) d'un plus grand ensemble de données. Dans ce cas, le jeu de données $X = \{x_i\}$, avec $1 \leq i \leq n$, peut être divisé en deux parties $X_l = \{x_1, \dots, x_l\}$ pour lesquels les labels $Y_l = \{y_1, \dots, y_l\}$ sont fournis, et les points $X_u = \{x_{l+1}, \dots, x_{l+u}\}$ pour lesquels les classes ne sont pas connues. Ce cas est ce qui est appelé la classification semi-supervisée « classique » ([Chapelle et al., 2006]). Dans cette revue de la littérature, nous noterons l'existence de ces algorithmes, étant souvent à l'origine de nombreuses approches et variations de clustering, mais nous nous permettrons de ne pas être exhaustifs, du fait que notre problématique/contexte ne tombe pas dans ce cas classique. En effet les informations connexes à notre ensemble de données ne se présentent pas comme un échantillon dont la classe d'appartenance des items est connue.

Le développement du domaine plus général du clustering sous contrainte est directement lié au besoin grandissant de trouver les moyens d'utiliser les informations collatérales quand elles sont disponibles. Bien qu'il soit tout à fait possible qu'un algorithme de classification non supervisée trouve une partition en accord avec les connaissances du domaine ou les attentes des utilisateurs, pour beaucoup des cas les plus intéressants, et souvent les plus complexes, lorsqu'une expertise humaine est nécessaire, ces méthodes échouent. La classification semi supervisée tente alors d'automatiser au maximum l'acquisition et l'utilisation de connaissances additionnelles au jeu de données.

3.3.1.1 Les contraintes par paires

Les travaux initiaux du domaine présentent des algorithmes de clustering pouvant incorporer des contraintes portant sur des relations entre paires d'objets, ou étant capables d'apprendre des mesures de distances spécifique au problème considéré. Les contraintes par paire ont été proposées par Wagstaff et Cardie (2000). Ces contraintes sont de deux types : *must-link* ou littéralement « doivent être connectés » et *cannot-link* ou littéralement « ne peuvent être connectés ». Le terme connecté définit ici le fait d'appartenir au même cluster.

Ils ont de plus développé une variante des K-moyennes qui inclut ces connaissances dans un ensemble M , et en tire avantage. Si de telles contraintes sont connues, plutôt que de retourner une partition qui minimise au mieux la fonction objective générique utilisée, l'algorithme va adapter sa solution pour qu'elle s'accorde avec cet ensemble M . Ces contraintes sur objets ont plusieurs propriétés. La contrainte *must-link* est une relation d'équivalence, elle est donc symétrique, réflexive et transitive. La propriété de transitivité de ce type de contrainte va nous permettre d'inférer d'autres contraintes *must-link* entre d'autres paires d'objets n'appartenant pas à l'ensemble C initial ([Basu et al., 2004a]). Nous pouvons visualiser cette transitivité à partir de l'image d'un graphe. Chaque objet est représenté par un nœud et chaque contrainte *must-link* par une arête. Si on opère une fermeture transitive sur ce graphe, toutes les nouvelles arêtes ainsi obtenues représenteront autant de contraintes *must-link* que l'on pourra inclure dans M .

$$\text{Transitivité des contraintes } \textit{must-link}(x_i, x_k) \notin M \& (x_i, x_j) \in M \& (x_j, x_k) \in M \Rightarrow (x_i, x_k) \in M$$

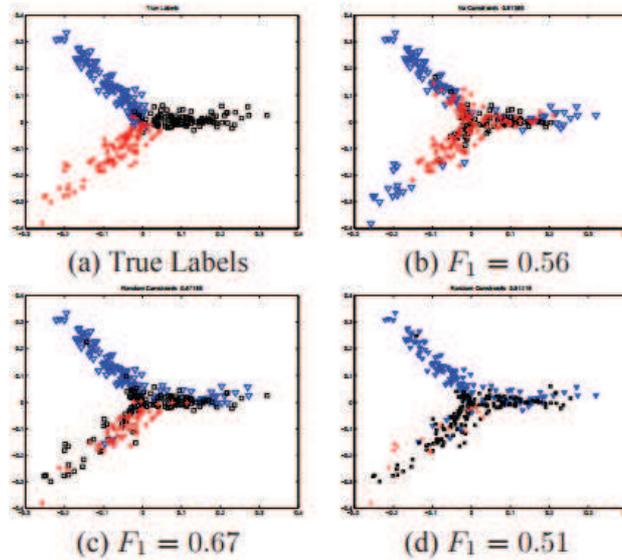


FIGURE 3.12 – [Mallapragada et al., 2008] Illustration du clustering basé sur contraintes. (a) Projection en 2 dimensions des vraies classes d'appartenance des objets de la base de données "Diff300" tiré de [Basu et al., 2004b] après réduction de dimension (b) partition après déroulement de la méthode des K-moyennes pour 3 clusters sans utilisation de contraintes (c) & (d) deux partitions obtenues du même ensemble après sélection aléatoire et utilisation de 100 contraintes par paires. La statistique F_1 est un indice de qualité de la partition obtenue [Basu et al., 2004a]

Par opposition, la contrainte *cannot-link* n'est pas une relation d'équivalence. La transitivité ne s'y applique pas ($M_{c(i,j)}$ et $M_{c(j,k)}$ n'implique pas $M_{c(i,k)}$). Des inférences sur des contraintes *cannot-link* peuvent tout de même se faire lorsque combinées à des contraintes *must-link*.

$$\text{Transitivité des contraintes } \text{cannot-link}(x_i, x_k) \notin C \& (x_i, x_j) \in C \& (x_j, x_k) \in M \Rightarrow (x_i, x_k) \in C$$

où M est l'ensemble des contraintes *must-link* et C l'ensemble des contraintes *cannot-link*.

L'utilisation de ces contraintes améliore sensiblement l'efficacité de l'algorithme des K-moyennes, comme on peut le visualiser en figure 3.12.

3.3.1.2 Acquisition des contraintes

Comme mentionné précédemment, il est possible qu'une partie de l'ensemble de données à traiter soit accompagné d'un plus petit ensemble, un échantillon, déjà labellisé. Ces objets pré-classifiés peuvent servir de *seeds*, c'est-à-dire servir de point de départ à certains algorithmes pour classifier le reste de l'ensemble. Cette méthode s'appelle le *seeding*. Comme vu dans certains des algorithmes précédemment mentionnés, l'initialisation est une étape critique d'un algorithme de partitionnement, et contraint souvent l'utilisateur à faire tourner plusieurs fois un même algorithme avec différentes initialisations pour contrebalancer ses effets. Il est possible d'utiliser cet échantillon pré-classifié afin d'obtenir un ensemble de contraintes portant sur le reste de l'ensemble de données pour lesquels aucune information n'est fournie à l'origine. Le voisinage de ces *seeds* sera considéré comme peuplé d'objets *must-link*, alors que les objets trop éloignés d'une *seed* se verront qualifiés d'une contrainte *cannot-link* avec la *seed* en considération, et par extension son voisinage proche. Basu et al. [Basu et al., 2002] propose deux versions modifiées de l'algorithme des K-moyennes utilisant une approche de *seeding* pour l'initialisation des centroïdes. Le voisinage d'une *seed* étant obtenu par clôture transitive (cf. section précédente) sur le graphe représentant les relations *must-link* découlant de l'échantillon

prélabellisé. Cette technique très limitée dépend entièrement de la taille de l'échantillon classifié fourni avec les données.

Ainsi les contraintes utilisées sont déterminantes dans la partition obtenue. Si certaines contraintes sont fausses ou « pauvre[s] en information » ([Davidson et al., 2006]), la partition risque d'être altérée en précision et par rapport aux volontés initiales de l'utilisateur ([Xiong et De la Torre, 2013]). Une solution possible pour l'acquisition de contraintes supplémentaires est la demande à un expert ou utilisateur des données. Cependant la spécification de ces contraintes par l'utilisateur, ou par étude approfondie du problème, peut être extrêmement lente et coûteuse. L'apprentissage actif de contraintes en classification semi supervisée est un domaine permettant de minimiser l'intervention humaine dans l'acquisition de l'ensemble des contraintes. Le problème se déplace ainsi sur la qualité informationnelle, indice permettant de réduire l'intervention humaine aux seules contraintes hautement informationnelles. La première méthode d'acquisition de contrainte automatique du domaine est connue sous le nom de *farthest first query* ([Basu et al., 2004a]). Elle est composée de deux phases : *Explore* et *Consolidate*. La phase d'exploration recherche de manière incrémentale un nombre de points (ou objets) k correspondant au nombre de cluster souhaité. Ces k points sont choisis de manière à tous appartenir à un cluster différent, information fournie par un « oracle » (souvent un utilisateur), avec un nombre de requête pour l'oracle prélimité. Ils forment ainsi le squelette du clustering à venir, et sont donc considéré comme hautement informatif sur le partitionnement. Les points restants, n'appartenant pas aux squelettes, sont ensuite choisis de manière aléatoire, puis comparés successivement avec les centroïdes, en commençant par le plus proche (obtenu par mesure de distance), jusqu'à ce qu'une contrainte *must-link* avec un des k points de la phase *Explore* soient trouvées, toujours par requête à l'oracle. Cette phase étant la phase de consolidation. Une version améliorée de cette procédure a été proposée par Mallapragada et al. ([Mallapragada et al., 2008]), qui au lieu de choisir un point de manière aléatoire lors de la phase *Consolidate*, utilise une fonction Min Max destinée à trouver le point q , de l'ensemble restant de points à traiter, dont l'appartenance à un cluster est la plus incertaine, i.e. dont l'indice de similarité s le plus grand avec un des points du squelette (i.e. appartenant à X_s), est le plus petit comparé aux autres points :

$$q = \arg \min_i P(X_s, x_i) = \arg \min_i \max_{x_j \in X_s} s_{ij}$$

On choisit ensuite un représentatif par cluster u_k , dont l'indice de similarité avec q est le plus grand, puis on les ordonne de manière à demander à l'oracle si q et u_k peuvent être reliés par une contrainte *must-link*. Ces méthodes sont malgré tout influencées par la phase d'exploration, qui reste une sélection aléatoire du squelette initial du clustering, et bien sûr requiert un utilisateur non naïf (au moins), voir un expert, comme oracle. D'autres algorithmes ont été proposés, s'affranchissant de cette sélection aléatoire en phase *Explore* ([Cai et al., 2016]), et prenant avantage du nombre de requêtes déjà formulées pour améliorer la performance du clustering ([Xiong et De la Torre, 2013]). Dans Vu et al. ([Vu et al., 2010]), les auteurs proposent une mesure d'utilité des contraintes *must-link* candidates, définie par la densité de population de la région dans laquelle se trouve la paire d'item concerné. Une contrainte *must-link* reliant deux items dans un environnement peuplé possède *a priori* moins de capacité discriminante. Afin de sélectionner les contraintes dont le poids informationnel est le plus élevé, Cai et collaborateurs proposent en plus d'une telle mesure de la pertinence d'une contrainte, de n'effectuer les requêtes à l'utilisateur que sur ensemble réduit des données non labellisés, ensemble choisi pour sa qualité informationnelle. L'intervention humaine comme nous l'avons déjà mentionné peut être incontournable pour des problèmes complexes, malgré tout dans ce cas de figure l'expert doit être capable d'établir une similarité/dissimilarité entre deux objets, de manière quasi instantanée, en interaction avec le processus de clustering. Une telle comparaison est parfois impossible, lorsque peu de connaissances *a priori* sur les objets sont connues, ou si les objets sont trop complexes. De plus, comparé à d'autres domaines d'apprentissage, le domaine de l'apprentissage de contraintes reste peu étudié ([Cai et al., 2016]).

3.3.1.3 Utilisation des contraintes

Une fois les contraintes obtenues, la question de leur implémentation dans le processus de partitionnement doit se poser. La plupart des méthodes en clustering semi supervisé tombent sous deux catégories générales : par contrainte *vs.* sur distance. Dans la figure 3.13, le premier schéma représente le clustering semi

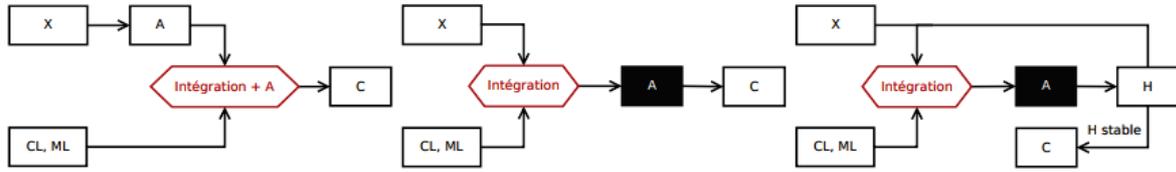


FIGURE 3.13 – Les différents types d’intégration dans le clustering semi-supervisé. Dans l’ordre, l’intégration de contraintes dans l’algorithme A prédéfini, l’intégration de contraintes dans la définition de la proximité, avant l’application de l’algorithme A quelconque et enfin l’intégration contrôlée par l’algorithme de clustering quelconque A . Tirée de [Sublemontier, 2012].

supervisé par contrainte, tandis que les deux autres schémas représentent celui basé sur l’apprentissage d’une métrique de distance. Les méthodes par contraintes utilisent les connaissances extérieures pour guider le partitionnement en empêchant la violation des contraintes de l’ensemble C ([Demiriz et al., 1999]; [Wagstaff et al., 2001]; [Basu et al., 2002]). Dans les approches basées sur distance, un algorithme de clustering classique est utilisé mais se base sur une fonction de distance paramétrée par apprentissage s’assurant ainsi que les contraintes *must-link* se traduisent effectivement par le regroupement des deux points considérés lors de l’application de cette fonction de distance ([Bilenko et al., 2004]; [Cohn et al., 2003]; [Klein et al., 2002]; [Xing et al., 2003]).

3.3.1.3.1 Modification de la mesure de similarité

Le besoin d’une mesure de distance appropriée en machine learning est d’une grande importance (cf. section 3.2.2), mais choisir manuellement une telle mesure, de manière à ce qu’elle capte correctement les relations entre objets d’un ensemble de données, peut paraître fastidieux. Cet état de fait a mené à l’émergence du domaine de l’apprentissage de mesure, qui vise à apprendre automatiquement une mesure de distance directement sur les données considérées ([Bellet et al., 2013]; [Kulis, 2013]). Ainsi la première approche pour inclure l’ensemble de contraintes au déroulement du processus de partitionnement consiste à s’intéresser à une mesure de distance (respectivement similarité) entre les items. Dans la littérature, cette approche se fait soit par la présence d’un ensemble de données pré-étiquetées, i.e. déjà classifiées, soit possédant un certain ensemble de contraintes. Idéalement cette mesure devra tendre pour deux éléments liés par une contrainte *must-link* vers 0 (respectivement 1) et pour deux éléments liés par contraintes *cannot-link* vers 1 (respectivement 0).

La prise en compte des contraintes par paires dans le processus de partitionnement se fera par le biais d’une fonction de coût (ou de perte) composée d’un paramètre alourdissant la distance calculée entre deux items dissimilaires, et allégeant (voir annihilant) la distance entre deux items similaires. Cette fonction de coût est intégrée de manière plus globale dans la fonction objective à minimiser par le processus de partitionnement. Le problème d’une bonne mesure de distance (ou similarité) en clustering est critique du fait de l’absence d’une bonne réponse pour la classification intrinsèque à l’ensemble de données (cf. section 3.2.2). Xing et al. ([Xing et al., 2003]) dans ses travaux pionniers sur l’apprentissage d’une mesure métrique de distance pose la question d’un apprentissage automatique d’une bonne mesure de distance par le biais des préférences utilisateurs exprimées sous la forme de contraintes (de similarité et de dissimilarité) par paires sur un sous-ensemble des données. Pour ce faire, il formalise le problème par la détection des attributs critiques des items pour leur regroupement *vs.* séparation. On peut rapprocher ces travaux de la littérature sur l’analyse en composantes principales ([Jolliffe, 1986]), avec la différence notable d’une utilisation directe des préférences utilisateurs, assurant ainsi une partition « utile », en accord avec les objectifs spécifiques au contexte. De plus, ce procédé assure la réalisation d’une mesure complète généralisable et pouvant dès lors être appliqué

à tout nouvel item intégré à l'ensemble de données, assurant une continuité entre ensemble d'entraînement et nouvel input, potentiellement dénué d'information supplémentaire donnée par l'utilisateur. Prenons un ensemble de points $\{x_i\}$, $1 \leq i \leq n$ appartenant à R^d , et ensemble S où (x_i, x_j) appartient à S si x_i et x_j sont similaires. On cherche une mesure de distance linéaire $d(x, y)$ de la forme :

$$d(x, y) = d_A(x, y) = \|x - y\|_A = \sqrt{(x - y)^T A (x - y)}$$

Où A doit être définie semi-positive (positive avec la valeur 0 autorisée) pour respecter les propriétés d'une mesure de distance (positivité et inégalité triangulaire). Si $A = I$, la matrice identité, alors notre mesure devient la distance euclidienne. Si on restreint A à être diagonale, cela revient en fait à appliquer un poids pour chacun des attributs, considérant alors un certain ordonnancement dans l'importance de chacun pour la comparaison de deux items. Le problème revient alors, d'après les auteurs, à minimiser la distance pour les paires d'items appartenant à S , et à la maximiser pour les paires d'items n'appartenant pas à S , appartenant à un ensemble D représentant les paires d'items dissimilaires, par le biais de A .

$$\min_A \quad \sum_{m(i,j) \in M} D_A^2(x_i, x_j) \quad (3.10)$$

$$\text{tel que} \quad \sum_{c(i,j) \in C} D_A(x_i, x_j) \geq 1 \quad (3.11)$$

$$A \geq 0 \quad (3.12)$$

$$(3.13)$$

où M est l'ensemble des contraintes *must-link* et C l'ensemble des contraintes *cannot-link*.

Cette optimisation, ce faisant par le biais de A qui est un paramètre linéaire, est donc convexe et par définition sa résolution n'engendrera pas le calcul d'un optimum local. C'est là un affranchissement considérable si l'on considère les autres algorithmes présentés dans cet état de l'art.

Dans le cas d'une matrice diagonale, le problème revient alors à minimiser la transformation g sur A .

$$g(A) = g(A_{11}, \dots, A_{nn}) = \sum_{(x_i, x_j) \in S} \|x_i - x_j\|_A - \left(\sum_{(x_i, x_j) \in D} \|x_i - x_j\|_A \right)$$

Ceci est réalisé facilement avec la méthode Newton-Raphson, méthode itérative d'approximation du zéro d'une fonction. Pour le cas d'une matrice A pleine, cette méthode est trop lourde et les auteurs proposent alors de maximiser la fonction g appliquée sur l'ensemble D par l'utilisation de l'algorithme (itératif) du gradient, et en utilisant des méthodes de résolutions de systèmes linéaires sur les deux contraintes 3.15 et 3.16.

$$\max_A \quad g(A) = \sum_{(x_i, x_j) \in D} \|x_i, x_j\|_A \quad (3.14)$$

$$\text{tel que} \quad f(A) = \sum_{(x_i, x_j) \in S} \|x_i, x_j\|_A^2 \leq 1 \quad (3.15)$$

$$A \geq 0 \quad (3.16)$$

L'apprentissage d'une mesure a été le sujet de nombreux travaux dans la littérature. La plupart d'entre eux se centrent sur une mesure de distance globale, on peut citer de manière non exhaustive ITML ([Davis et al., 2007]), LDML ([Guillaumin et al., 2009]), LMNN ([Weinberger et al., 2006]) ou DML_{eig} ([Ying et Li, 2012]).

Cette approche peut être inefficace dans le cas de données hétérogènes. En effet ce qui est discriminant pour une partie des attributs des items, ne l'est peut-être pas pour d'autres. Cette affirmation est supportée par Bohné et al. ([Bohné et al., 2018]), où les auteurs proposent l'apprentissage de métriques locales, plutôt qu'une globale. Cette approche n'est pas de leur fait et d'autres travaux abordent cette question. Notamment on peut citer une extension du LMNN aux mesures locales (MM-LMNN [Weinberger et Saul, 2009]), pour

laquelle chaque classe possède sa propre mesure de distance (à noter qu'il faut donc connaître à l'avance nos classes) et toutes ces mesures sont optimisées parallèlement de manière à optimiser un critère de classification global. Mais ces méthodes échouent à s'affranchir du besoin d'un échantillon des données déjà labellisés, informations souvent inexistantes et coûteuse à mettre en place dans de nombreuses applications, et information absolument manquante dans la problématique spécifique à cette thèse. La méthode de Bohné et collaborateurs, nommée Large Margin Local Metric Learning (LMLML), repose sur le principe des modèles de mélange (cf. section 3.2.4.2), et ne nécessite qu'un ensemble de contraintes par paires sur une partie des données. De plus leur méthode est décrite par ses auteurs comme une généralisation des travaux existants en apprentissage de mesures locales, du fait qu'elle n'active pas une mesure de distance en fonction de la région de discrimination, mais utilise des combinaisons pondérées de mesures. Au lieu de chercher à paramétrer une unique matrice A , on cherche à pondérer $K + 1$ matrice,

$$M_\theta(x_i, x_j) = \sum_{k=0}^K w_\theta^k(x_i, x_j) M_k$$

la matrice M_0 faisant office d'influence globale, dont le poids reste constant à travers les régions de l'ensemble de données.

A est donc remplacée par cette fonction matricielle M_θ dans l'apprentissage de la mesure de distance :

$$d^2(x_i, x_j, M_\theta) = (x_i - x_j)^T M_\theta(x_i, x_j)(x_i - x_j)$$

On la note s .

La multiplicité des matrices nous assurant de la localité de leur paramétrage. De plus le calcul des poids $w_\theta^k(x_i, x_j)$ s'effectue par l'addition des probabilités qu' x_i et x_j appartiennent à la gaussienne k du modèle de mélange (voir section 3.2.4.2). L'utilisation de cette méthode nous assure un partitionnement flou des données rendant le passage d'une région à une autre, et leurs poids associés, moins abrupte. Chaque mesure locale aura donc une forte influence dans une région très spécifique, influence allant doucement décroître aux abords de sa région, pour laisser place à une autre mesure locale, et s'appliquera donc proportionnellement à la probabilité pour les deux items en considération à appartenir à cette région.

La mesure de distance apprise s'intègre plus globalement dans une fonction objective à minimiser par l'intermédiaire d'une fonction de coût $l(y, s)$ où $y_{ij} = 1$ (respectivement -1) si x_i et x_j sont deux points similaires (respectivement dissimilaires), représentant l'ensemble des contraintes par paires connues et l est définie par :

$$l(y, s) = \max\left(0, 1 - \frac{y}{\gamma}(1 - s)\right)$$

où γ appartenant à $[0, 1]$ permet de paramétrer les marges.

Cette fonction de coût donne ainsi lieu à la fonction objective suivante :

$$\phi(M, \theta) = \frac{1}{|D|} \sum_{(i,j) \in D} l(y_{ij}, d^2(x_i, x_j, M_\theta)) = \lambda \sum_{k=0}^K \omega(M_k)$$

où D est le sous ensemble de données pour l'apprentissage de la mesure, et ω est une fonction convexe régulatrice empêchant les problèmes de surapprentissage (ou surajustement) aux données d'entraînement utilisées D .

Cette approche prenant en compte l'aspect instable des caractéristiques à considérer comme importante dans le processus de discrimination et généralisant les différentes méthodes du domaine à ce jour dans la littérature, rencontre tout de même des limitations. Tout d'abord le choix du nombre K de régions à identifier doit rester petit, le temps de calcul étant linéairement lié à sa taille, ce qui ne poserait pas de problème étant donné la faible quantité de données dont nous disposons. Un autre des problèmes directement lié à notre contexte est la nature des contraintes que nous obtenons. En effet du fait de la modélisation utilisée, nous nous centrons sur une approche et une construction de la comparabilité entre items, par détection des individus dissimilaires. Nous n'obtenons donc comme information supplémentaire que des contraintes

de type *cannot-link*. L'implémentation d'une telle solution sur nos données semble donc compromise. Une telle contrainte a déjà été étudiée par Oyama & Tanaka ([Oyama et Tanaka, 2008]). De plus les auteurs montrent que l'utilisation unique des contraintes *cannot-link* dans l'apprentissage d'une mesure globale rend cet apprentissage plus simple, transformant la formule d'optimisation énoncée 3.10 en :

$$\min_A \quad \frac{1}{2} \|A\|_F^2 \quad (3.17)$$

$$\text{tel que } D_A^2(x_i, x_j) \geq 1 \quad \forall c(i, j) \in C \quad (3.18)$$

$$A \geq 0 \quad (3.19)$$

où $\|\cdot\|_F$ est la norme Frobenius et C l'ensemble des contraintes *cannot-link*.

Ce problème est ainsi formalisé par les auteurs comme un problème de programmation quadratique convexe. Malgré tout, les techniques d'intégration des connaissances supplémentaires par apprentissage d'une mesure de distance ne peuvent pas s'appliquer à notre ensemble de données. En effet toutes ces techniques reposent sur un apprentissage effectué sur un échantillon des données à partitionner. Étant donné le faible nombre d'individus (faible au vu des demandes computationnelles, mais élevé au vu de la littérature en acquisition des langues), nous ne pouvons nous permettre de tirer un échantillon à des fins d'apprentissage d'une telle mesure. Ces approches ne sont effet pas conçues pour des domaines où le recueil de données est encore limité et coûteux, mais pour des domaines propulsés par l'ère du numérique. Ces différents algorithmes d'ailleurs généralement rivalisent de performances, où un taux de précision de 98% sera considéré comme une réelle avancée par rapport à un taux de précision de 97%. De tels objectifs ne sont pas encore considérés à notre état actuel d'avancement.

3.3.1.3.2 Modification de l'algorithme par implémentation de contraintes

Les algorithmes heuristiques classiques peuvent rapidement trouver une solution de partitionnement et s'adaptent à des ensembles de données très grands. Par contre ils ne garantissent pas que toutes les contraintes soient satisfaites, ou plus généralement la qualité de la solution trouvée. Une approche déclarative et exacte offre une meilleure compréhension des données, ce qui est incontournable pour de petits ensembles de données de grandes valeurs car prenant un temps considérable à être collectés. Selon Bilenko et al. ([Bilenko et al., 2004]), l'autre approche pour l'implémentation de contraintes externes dans un algorithme de clustering est dite *constrained-based*, où les contraintes guident l'algorithme de clustering plutôt que d'être utilisées dans la mesure de distance.

Le premier travail dans ce domaine propose une version modifié de COBWEB ([Wagstaff et Cardie, 2000]) permettant de respecter strictement les contraintes connues. Il a été suivi par une version modifiée de l'algorithme des K-moyennes, appelé *COP-Kmeans* ([Wagstaff et al., 2001], 3.4), dont l'ajout principal se trouve être une fonction *violate - constraints* qui s'assure à chaque itération de respecter l'ensemble C de contraintes (*must-link* et *cannot-link*), tout en cherchant à minimiser la fonction objective, qui dans le cas de l'algorithme des K-moyennes se trouve être l'erreur de quantification vectorielle liée à la distance entre un item et son centroïde.

Algorithme 3.4 COP-Kmeans (ensemble de données X , nombre de cluster k , ensemble des contraintes *must-link* M , et *cannot-link* C)

- 1: Soit μ_1, \dots, μ_k les k centres initiaux de cluster (centroïdes)
 - 2: Pour chaque $x_i \in X$, lui assigner son plus proche cluster c tel que *violate - constraints*(x_i, c, M, C) retourne faux. Si aucun cluster ne satisfait cette condition, échec.
 - 3: Mettre à jour chaque centroïde μ_i par calcul de la moyenne des chaque instances $x_j \in \mu_i$.
 - 4: Répéter les étapes (2) et (3) jusqu'à convergence
 - 5: Retourner $\{\mu_1, \dots, \mu_k\}$
-

Il est important de noter qu'ici il est possible qu'aucune partition ne satisfasse toutes les contraintes, ou que les contraintes puissent permettre de minimiser la fonction objective. C'est tout là l'intérêt et le problème

Algorithme 3.5 *violate – constraints*(x_i, c, M, C)

- 1: Pour chaque $m(i, j) \in M$: Si $x_j \notin c$, retourner vrai
 - 2: Pour chaque $c(i, j) \in C$: Si $x_j \in c$, retourner vrai
 - 3: Retourner faux
-

des algorithmes basé sur contraintes : trouver une partition qui n’aurait pas été trouvée par un algorithme de clustering classique, et en même temps s’assurer que la partition obtenue soit réellement améliorée à travers l’ensemble de contrainte du point de vue de l’utilisateur.

Dans [Basu et al., 2009] ce problème est illustré en figure 3.14.

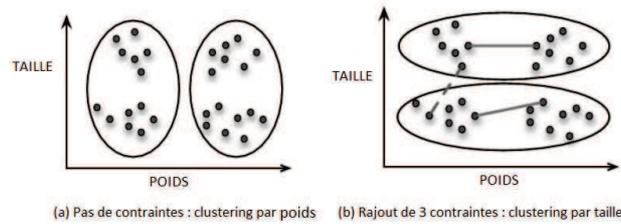


FIGURE 3.14 – Illustration de l’importance des objectifs de l’utilisateur dans le processus de partitionnement : Clustering avec $k = 2$ avec trois contraintes en dur. Les contraintes *must-link* sont représentées par des barres continues et la contrainte *cannot-link* par une barre en pointillée.

Il existe deux manières de regrouper les items présents dans cet ensemble de données : par taille et par poids. Un algorithme entièrement non supervisé va les regrouper selon l’un ou l’autre des deux critères, pour un nombre de cluster fixé à 2. Par l’ajout de deux contraintes *must-link* on décide alors quel critère sera retenu, illustré dans la partie (b) de la figure (ici critère taille). La partition obtenue n’est pas meilleure au niveau de la minimisation de la fonction objective, mais l’est définitivement du point de vue de l’utilisateur. Afin de s’assurer qu’une solution est plus probable d’être trouvée, le MPCK-means associe à chaque contrainte un poids, venant peser comme une pénalité si jamais la contrainte n’est pas satisfaite, mais empêchant ainsi un possible échec de partitionnement. Cette problématique sera également adressée dans l’approche proposée dans ce travail, à travers un relâchement de la satisfaction en dur des contraintes (cf. partie 5).

Les deux techniques que nous venons de mentionner imposent la satisfaction de l’ensemble des contraintes pendant la procédure de clusterisation, mais il existe aussi des techniques permettant de les inclure dans la fonction objective à minimiser ([Demiriz et al., 1999]), similaires en cela aux techniques d’apprentissage de la mesure (cf. section précédente). Une autre approche que nous avons déjà mentionnée (cf. section 3.3.1.3.1) consiste à utiliser un échantillon des données prélabellisé pour générer des *seeds* de cluster, assimilables aux centroïdes de l’algorithme des K-moyennes, pour la génération des contraintes ([Basu et al., 2002]). Toutefois nous n’approfondirons pas cette dernière approche quant à l’utilisation des contraintes ainsi obtenues, du fait qu’elle est inapplicable à notre base de données qui ne s’accompagne pas d’un sous-ensemble prélabellisé.

3.3.1.3.3 Intégrer les deux approches

Comme précédemment mentionné, les algorithmes de clustering semi supervisé se regroupent sous deux catégories générales, l’une utilisant l’ensemble C de contraintes comme d’un guide pour partitionner les données, l’autre s’en servant dans l’apprentissage d’une métrique de distance spécifique. Les deux exemples suivants sont deux méthodes permettant de combiner ces deux approches dans un seul algorithme. Le MPCK-Means ([Bilenko et al., 2004]) est semblable à l’apprentissage de mesures locales en ce sens qu’il permet d’obtenir des mesures customisées pour chaque cluster, autorisant ainsi une variété de forme pour les clusters

d'une même partition. Là s'arrête la ressemblance puisque, contrairement aux techniques d'apprentissage de mesure précédemment mentionnées, cet apprentissage n'est pas désolidarisé de l'algorithme de clustering lui-même. En effet la mesure est modifiée à chaque itération de l'algorithme de partitionnement. La mesure de distance se combine directement à une fonction de coût de violation de contrainte, pour donner la fonction objective suivante à minimiser :

$$\begin{aligned}
 J_{mpckm} = & \sum_{x_i \in X} \left(\|x_i - \mu_{l_i}\|_{A_{l_i}}^2 - \log(\det(A_{l_i})) \right) \\
 & + \sum_{(x_i, x_j) \in M} w_{ij} f_M(x_i, x_j) \mathbb{1}[l_i \neq l_j] \\
 & + \sum_{(x_i, x_j) \in C} \bar{w}_{ij} f_C(x_i, x_j) \mathbb{1}[l_i = l_j]
 \end{aligned}$$

Où le premier terme représente la mesure de distance apprise de manière itérative au cours du processus de clustering avec un recalcul du point μ_{l_i} , représentatif de la gaussienne l_i , après chaque assignement d'un nouvel item à la gaussienne l_i ; où les deux autres termes représentent respectivement les fonctions de coût associées à la violation d'une contrainte *must-link* par la fonction f_M , respectivement *cannot-link* f_C ; où $\mathbb{1}$ est la fonction indicatrice telle que $\mathbb{1} \rightarrow [true] = 1$ et $\rightarrow [false] = 0$.

Le *HRMF-KMEANS* ([Basu et al., 2004a]) basé sur les champs aléatoires de markov cachés combine un guidage de l'algorithme par les contraintes ainsi qu'une mesure de distance spécialement apprise, dans l'esprit d'un algorithme d'espérance maximisation (cf. section 3.2.4.2). Sa méthode est très proche du MPCK-Means mais permet l'intégration de mesures de distances différentes de la classique mesure de distance euclidienne, intégrant la divergence de Bregmann ([Bregman, 1967]). Cet ajout permet ainsi de calculer la norme entre x_i et μ_{l_i} non seulement si x_i et μ_{l_i} sont des distributions de probabilités, mais si x_i et μ_{l_i} sont des vecteurs, ou s'ils représentent un vecteur de longueur unitaire.

Comme mentionné précédemment, l'acquisition des contraintes par paires peut être lente, coûteuse et parfois même difficilement concevable par un expert, et ce plus particulièrement dans les analyses exploratoires comme le clustering, où l'on ne connaît pas vraiment à l'avance ce que l'on va obtenir. Notre algorithme se propose d'opérer une acquisition de contraintes par paires *cannot-link* par comparaison de leurs attributs, comparaison rendue possible par l'intermédiaire d'un ensemble de connaissances logiques issues de l'analyse plus théorique des comportements des apprenants en situation d'acquisition d'une langue étrangère.

A ce jour les auteurs n'ont pas connaissance d'un outil spécifique de découverte de connaissances et de *data mining* pour les bases de données de comportements d'apprentissage des langues secondes, prenant en compte les spécificités du domaine, à savoir les connaissances issues des approches acquisitionnistes, et les différents niveaux linguistique impliqués dans le processus d'acquisition. De fait, nos travaux reposent sur l'inclusion d'une expertise du domaine dans le procédé de découverte des profils d'apprenants, les rendant originaux comparés aux techniques classiques présentées lors de ce chapitre.

Chapitre 4

Modélisation et base de données

Sommaire

4.1 Sources d'influence sur le parcours acquisitionnel dans le projet VILLA	64
4.1.1 La langue maternelle	64
4.1.2 Le type d'enseignement	66
4.1.3 Les caractéristiques de l'input et le type de tâche	69
4.2 Des données à la notion de représentation	72
4.2.1 La base de données	72
4.2.2 Représentation multicritère de données comportementales	75

Comment caractériser les débuts de l'acquisition d'une nouvelle langue par un apprenant adulte ? Une des réponses possibles à cette question est proposée par le projet VILLA. Grâce à l'enseignement contrôlé et la batterie de tests administrés aux participants du projet VILLA, nous pouvons tenter de discerner des parcours acquisitionnels typiques pour les apprenants étudiés. L'objectif à long terme étant bien sûr que ces parcours identifiés soient suffisamment robustes pour correspondre à de nouvelles entrées, c'est-à-dire à des parcours de nouveaux apprenants.

Nous avons déjà évoqué que la comparaison entre deux objets est au cœur de tout algorithme de clustering. Nous souhaitons à présent connaître la proximité de comportement entre deux individus lors de l'acquisition d'une nouvelle langue dans les mêmes conditions. Cependant, avant de pouvoir comparer deux individus, il convient de trouver une manière de représenter un individu. En d'autres termes, il convient de se saisir de l'objet "individu" et de s'atteler à sa description avant de pouvoir le comparer à un homologue. La représentation d'un individu qui tient compte de ses comportements d'apprentissage nous a amené à procéder à plusieurs tentatives. Notre difficulté principale était de faire correspondre une représentation numérique vectorielle aux connaissances du domaine de la RAL sur les interactions entre niveaux linguistiques, et aux propriétés de l'input reçu par les apprenants. Les tests soumis aux apprenants du projet VILLA constituent un accès aux représentations internes que possèdent l'apprenant sur le fonctionnement de la LC. Le type de réponse donné par l'apprenant, erronée ou non, fournit les informations sur son interlangue et permet de comprendre les stratégies, efficaces ou non, qu'il applique.

Cette section a pour objet de présenter les travaux effectués au cours de ce doctorat à des fins de modélisation du parcours acquisitionnel d'un apprenant du projet VILLA. Il existe deux aspects primordiaux dans la réalisation du modèle d'un apprenant d'une L2. En ce qui concerne le premier aspect, il convient de manipuler les données d'acquisition afin d'observer les différentes influences des variables en présence, leurs niveaux d'implication sur le parcours acquisitionnel de l'apprenant, et les interactions potentielles entre ces variables. Ce travail consiste essentiellement à suivre une procédure empirico-formel i.e. consiste en la

formulation d'hypothèses, vérifications statistiques ou qualitatives à partir des données, et interprétation théorique. Par l'analyse successive de divers échantillons de la BDD nous avons pu isoler les effets des différentes variables prises en compte dans l'élaboration du projet VILLA, et constater leur influence relative sur l'acquisition en fonction de leurs modalités d'expressions.

Le deuxième aspect du problème est l'aspect technique. Comment représenter un individu et ses manifestations comportementales d'un point de vue informatique? Les résultats des apprenants sous forme de score purement numérique peuvent-ils être pris tels quels? Nous avons déjà introduit le problème lors de la présentation du domaine de la logique floue (cf. section 3.1). Comment modéliser les relations entre les scores aux différentes tâches et les scores aux différentes propriétés des items utilisés dans les tâches en question? Une jonction réussie de ces deux aspects du problème devrait traduire un score numérique en termes acquisitionnels. C'est la construction de cette jonction que nous proposons de présenter dans ce chapitre.

Pour commencer nous résumons les découvertes issues de l'analyse des différentes tâches du projet VILLA, découvertes basées sur différents échantillons d'apprenants (en terme de LM mais également de méthode d'enseignement reçue) et en fonction des différentes caractéristiques de l'input (cf. section 4.1). Ensuite nous présentons la manière dont on peut représenter formellement un apprenant en prenant en compte toutes les variables présentes dans le corpus de données (cf. section 4.2).

4.1 Sources d'influence sur le parcours acquisitionnel dans le projet VILLA

La base de données (BDD) issue du projet VILLA est conséquente. Bien que le dépouillement et la transcription des données soient encore en cours, de nombreux résultats ont déjà été présentés à partir de sous échantillons de la BDD. Ces données peuvent être analysées sous plusieurs angles. Notre point de départ dans la structuration de ces résultats est présentement de comprendre l'influence des variables telles que la LM des apprenants, le type d'enseignement, les caractéristiques de l'input, le type de tâche sur l'apprentissage d'une nouvelle langue par l'adulte, et de voir à quels niveaux et de quelle manière ces variables impactent l'acquisition. En nous centrant sur l'apprenant, nous tentons ainsi d'expliquer ses capacités de traitement et de production en langue cible, à tous niveaux linguistiques (phonologique, lexical, morphosyntaxique et pragmatique). Ces données existent pour tous les participants du projet VILLA, quelles que soient leurs langues maternelles ou la méthode d'enseignement qu'ils ont suivie (cf. *form-based vs. meaning-based* input, chapitre 2). Les étudiants ont, en effet, tous été soumis aux mêmes tests au cours de la période d'observation (14 heures de cours). Nous nous proposons de présenter ces résultats selon trois grands axes qui correspondent selon nous aux trois grandes sources de variation qui différencient les parcours acquisitionnels des apprenants étudiés : la langue maternelle (4.1.1), la méthode d'enseignement utilisée (4.1.2), et les caractéristiques de l'input et des tests de langue soumis aux apprenants (4.1.3).

4.1.1 La langue maternelle

Le rôle des langues sources, langue maternelle et/ou d'autres langues déjà acquises dans la construction du lecte de l'apprenant est incontestable comme le montrent de nombreux travaux (cf. [Giacobbe, 1992, Odlin, 1989, Moretti, 1989]). Les apprenants adultes font appel aux connaissances linguistiques issues de l'acquisition de la langue maternelle qui constituent « un "cadre" linguistique et conceptuel permettant la déduction des hypothèses qui organisent le matériel linguistique dont dispose l'apprenant » ([Giacobbe, 1992]). En comparant les connaissances issues des langues sources aux moyens linguistiques extraits de l'input de la langue cible, les apprenants procèdent à une « psychotypologie des langues » ([Kellerman, 1995]) qui contribue

à la construction du nouveau système linguistique, le lecte des apprenants (cf. [Perdue et Foundation, 1984]). Cependant, le recours aux langues sources ne permet pas d'expliquer tous les phénomènes attestés dans les productions des apprenants. Par exemple, Klein et Perdue ([Klein et Perdue, 1997]) montrent que les productions des apprenants débutants en milieu naturel se ressemblent indépendamment des couples des langues sources et cibles. En effet, les apprenants font également appel à des principes "language neutral" sous-jacents à la faculté de langage et à la communication (cf. [Klein, 2001]).

Le projet VILLA donne la possibilité d'observer un éventuel impact des connaissances linguistiques préalables dans l'acquisition des langues secondes en observant l'acquisition d'une même langue cible (le polonais) par des apprenants de 5 langues sources (allemand, anglais, italien, français, néerlandais) tout en contrôlant les informations sur d'autres langues maîtrisées par les apprenants étudiés (cf. chapitre 2 section 2.2). Ainsi, la dimension translinguistique rend possible l'identification du poids des spécificités linguistiques des langues sources dans le traitement d'une nouvelle langue. Nous pouvons donc observer non seulement l'existence ou non d'une influence de la LS, mais également, si elle existe, d'en quantifier l'étendue, et d'en saisir les variations en fonction de son éloignement typologique avec la LC.

4.1.1.1 Extraction d'un item

L'une des premières étapes qui permet d'entrer dans le système d'une nouvelle langue correspond à la segmentation du flux sonore afin d'identifier des unités et leur associer un sens. En d'autres termes, il s'agit d'extraire différents items d'un flux continu de parole. Le test de reconnaissance des mots (Word Recognition, WR) proposé dans le projet VILLA vise à analyser la capacité à reconnaître un item donné dans un énoncé.

Chaque étudiant a passé ce test trois fois au cours de la période d'observation : en T0 avant le début des cours, en T1 après 7h30 d'exposition au polonais et en T3 après 13h30 de l'input. Van Bergen, Rast et Shoemaker ([Geertje van Bergen et al., 2014]) qui ont analysé les performances des apprenants néerlandais, français, allemands et italiens, montrent que le score en T0 est fortement dépendant de la langue source des apprenants et de la transparence des items (cf. chapitre 1 section 1.2). Ainsi, les francophones et les italophones réussissent significativement mieux à extraire les mots transparents du flux sonore que les germanophones et les néerlandais. Quant aux mots opaques, cette différence disparaît. L'avantage des francophones et des italophones constaté en T0 s'estompe en T1 et en T2 où tous les apprenants améliorent leurs scores. Cependant la progression des français est là encore différente de celles des autres puisqu'il n'y a aucune amélioration notable entre la deuxième période de tests (7h30) et la troisième (13h30). Ce ralentissement de progression est expliqué par les auteurs comme dû à une insensibilité des français à l'accentuation d'un mot, le français étant une langue à cadence syllabique et non accentuelle, contrairement aux autres LSs en présence, ainsi qu'au polonais.

4.1.1.2 Morphologie nominale : jugements de grammaticalité et questions-réponses à l'oral

L'une des caractéristiques du polonais la plus étudiée dans le projet VILLA est sa richesse en morphologie nominale (deux nombres, trois genres et sept cas). Hinz et al. ([Hinz et al., 2013]) ont analysé cet aspect de l'apprentissage chez les apprenants français et allemands du projet VILLA à la fois en traitement grâce à la tâche de jugement de grammaticalité, et en production grâce à la tâche de question-réponse. Ces deux tâches ciblent l'acquisition de l'opposition instrumental *vs.* nominatif en polonais et ont été proposées deux fois aux apprenants, en T1 après 4h30 et en T2 après 10h30 de cours (cf. chapitre 2). La comparaison entre les apprenants de LM français et allemand est intéressante du fait de l'existence d'une morphologie nominale en allemand, bien que limitée par rapport au polonais, alors qu'elle est absente en français.

Quant à la tâche de jugement de grammaticalité, bien que les allemands réussissent mieux que les français en T1, seuls les français améliorent leurs score entre T1 et T2, comblant ainsi leur "retard" vis-à-vis de leurs homologues allemands. Ainsi, l'exposition à la LC entre deux périodes de test (soit 6h supplémentaire

d'exposition pour cette tâche) pallie le désavantage typologique des français.

Cependant, en ce qui concerne la tâche de question-réponse, les français et les allemands ont montré une amélioration notable entre les deux périodes de tests (après 4h30 et 10h30 d'enseignement respectivement). Également, bien que les allemands semblent avoir un meilleur score toute période de test confondue, cette différence n'est pas significative ([Watorek et al., 2016]). En conclusion, l'avantage de départ dont bénéficie les allemands est observable dans la tâche de traitement (GJ) mais non dans la tâche de production (OQA) où ceux-ci réussissent comme les apprenants français.

4.1.1.3 Morphosyntaxe : compréhension et répétition

La richesse de la morphologie nominale du polonais est accompagnée d'une certaine liberté d'ordre des mots lors de la construction d'une phrase. Ainsi, le statut du constituant dans la phrase, sujet ou objet, est marqué par le système casuel. Afin de comprendre une phrase en polonais, l'apprenant doit faire attention non seulement à l'ordre des mots (caractéristiques syntaxiques) mais également au marquage morphologique du syntagme nominal (caractéristiques de la morphologie nominale). Deux tâches ont été conçues dans le projet VILLA sollicitant le traitement de la morphosyntaxe, en l'occurrence de l'ordre des mots variable (SVO, OVS et OSV) et du marquage du sujet et de l'objet direct par les désinences du nominatif et de l'accusatif. Une de ces tâches permet de voir comment l'apprenant traite ce paradigme en compréhension (tâche de Picture verification, PV) et une autre (Sentence Imitation) vise à vérifier ce traitement lors de la répétition des phrases construites selon ces différents ordres syntaxiques.

Starren et al. ([Marianne Starren et al., 2013]) ont analysé les résultats des apprenants néerlandais et français à la tâche PV. Rappelons que cette tâche consistait à apparier une phrase avec la bonne image représentant une action parmi deux choix (cf. chapitre 2 section 2.5.4). Le néerlandais, contrairement au français, langue dite SVO, est une langue avec le verbe fini en deuxième position. Ainsi, les auteurs ont émis l'hypothèse de l'existence d'une possible différence dans le traitement de la morphosyntaxe en polonais entre les apprenants néerlandais et français. Cependant, aucune différence entre les deux groupes n'est apparue au cours des premières heures d'acquisition de la nouvelle langue. Les phrases de type SVO sont traitées avec un score de réponses correctes remarquablement plus élevé que pour les phrases avec les ordres OVS et OSV, et ce quelle que soit la LM de l'apprenant. De plus les phrases de type OVS n'étaient pas mieux réussies par les néerlandais, qui gardent le schéma d'organisation informationnelle Topique-Focus (cf. chapitre 1) comme schéma de base pour effectuer cette tâche. Les différences entre le groupe des français et des italiens ont également été étudiées ([Saturno, 2015b]). On peut ainsi noter une meilleure performance globale des italiens à cette tâche ainsi que dans la tâche de répétition. En effet, les italiens ont tendance à mieux répéter le marquage nominal des phrases polonaises dans la tâche SI. Ce résultat est expliqué par une plus grande facilité des italiens à apprendre le lexique polonais, ce qui serait dû à l'accentuation variable qui caractérise l'italien ([Valentini et Grassi, 2014]). Cependant, l'observation la plus significative de l'étude de Saturno et collaborateurs reste que les patterns d'acquisition de la morphosyntaxe en polonais par des français et des italiens sont les mêmes : les deux groupes ont attesté d'une meilleure performance dans la tâche de compréhension (PV) que dans la tâche de répétition (SI).

4.1.2 Le type d'enseignement

Un autre facteur envisagé comme source d'influence sur le parcours acquisitionnel des apprenants est le type d'enseignement qu'ils ont reçu. Rappelons que dans chaque pays (sauf l'Allemagne), deux types de cours ont été organisés dont l'un axé sur le sens (*meaning based input*) et l'autre sur la forme (*forme based input*). Même si de façon générale l'approche communicative a été adoptée comme démarche didactique pour les cours du projet VILLA, une variation au niveau de la focalisation sur les règles grammaticales enseignées a été proposée et donne lieu à deux types de cours. La version dite *form-based input* (FB) diffère de celle

dite *meaning-based* input (MB) par une explicitation des formes enseignées. Celles-ci sont mises en relief sur les diapositives du cours (cf. annexe A) et reprises pendant les activités par l'enseignant lors des corrections portant sur ces formes. Cette centration sur la forme vise à entraîner chez les apprenants une réflexion métalinguistique sur le système de la LC et sur les associations forme-fonction. L'impact de la focalisation sur les formes enseignées sur les performances des apprenants du projet VILLA est très variable, comme le montrent les résultats disponibles.

Ainsi, lors de la tâche de discrimination de phonème étudiée par Shoemaker ([Shoemaker, 2014]), il apparaît que le mode d'enseignement n'influe pas sur l'apprentissage phonologique des étudiants. L'auteur explique ce phénomène par le fait que cet apprentissage relève d'un modèle d'apprentissage implicite, donc indépendant du type d'exposition à la langue cible. Par ailleurs, l'explicitation des règles enseignées dans le cours du type FB input ne concerne que le niveau morpho-syntaxique, aucun entraînement de la perception des phonèmes n'a été proposé aux apprenants. De même, dans la tâche de reconnaissance de mots (cf. chapitre 2), Van Bergen et collaborateurs ([Geertje van Bergen et al., 2014]) ne trouvent pas de différences dans les résultats des apprenants qui ont suivi les deux types de cours. Les auteurs expliquent ce résultat par un effet différencié entre les deux types d'enseignement possiblement plus tardif, effet qui est donc non perceptible après seulement 14h de cours.

Seulement, des différences entre les deux groupes existent. En effet, les effets du type d'input reçu semblent entraîner des différences sur le parcours acquisitionnel des apprenants en fonction du type de tâche effectuée en LC. Ainsi, Les apprenants ayant reçu l'enseignement de type FB semblent plus en difficulté que les apprenants MB face à des tâches de jugements grammaticales. Ce résultat est supporté par un écrit de Saturno ([Saturno, 2015a]) qui analyse les résultats des apprenants italiens à une tâche de production sollicitant le système morphosyntaxique de la LC (Sentence Imitation, cf. chapitre 2). Il constate que les apprenants FB ont des résultats plus stables, c'est à dire, que leurs scores ne varient pas en fonction d'autres variables en présence, comme la fréquence d'un item ou la déclinaison sollicitée. Les apprenants MB sont eux plus sensibles aux caractéristiques de l'item cible dans l'input. N'étant soumis qu'à un input oral sans réflexion métalinguistique, il apparaît alors que les apprenants MB sont plus sensibles aux caractéristiques de ce même input (saillance de la déclinaison, nombre d'occurrences de l'item nominal, etc.) tandis que les apprenants FB, pour ces mêmes raisons, sont plus indépendants du contexte, ayant amorcés la création de règles abstraites sur l'association forme-fonction.

Latos ([Latos, 2014]) aborde la question de l'effet différencié de deux types d'input sur les résultats des apprenants à une tâche sollicitant un jugement de la grammaticalité d'une phrase. L'auteure analyse les performances des apprenants francophones ayant reçu les deux types d'input en polonais dans la tâche de jugement de grammaticalité portant sur la morphologie verbale. Ce test a été effectué par les apprenants une seule fois à la fin de la période d'observation donc après les 14 heures d'enseignement. L'hypothèse première de l'auteure était que les apprenants de l'enseignement centré sur la forme réussiraient mieux à cette tâche, du fait de l'explicitation des formes attirant l'attention de l'apprenant. En effet, le score des apprenants issus des enseignements basés sur la forme est légèrement meilleur, d'une différence de 8% du taux de réponses correctes. Cependant, cette différence apparaît faible lorsque l'on sait que les deux groupes ont passé le seuil des 70% de réponses correctes à ce test. L'auteure conclut d'ailleurs que son étude est une preuve directe de l'acquisition des indices flexionnels verbaux par les apprenants après seulement 14 heures d'enseignement, et ce quel que soit l'enseignement fourni. Elle souligne cependant que l'acquisition effective de la morphologie verbale de la langue cible par les étudiants ne peut être attestée qu'à travers une tâche de production. C'est pourquoi, cette même auteure, dans une autre étude ([Latos,] compare les productions issues de la tâche semi-guidée d'indication d'itinéraire et réalisées par les apprenants francophones des deux groupes.

Dans cette étude, Latos analyse les différentes formes produites de l'item nominal "*Ulica*" ("rue"). Cet item est un élément clé dans la réussite de la tâche d'indication d'itinéraire et a donc plus de probabilité d'apparaître dans le discours des apprenants. De plus, cet item est présent dans l'input des deux groupes

d'apprenants sous quatre formes possibles, au nominatif, à l'accusatif, au locatif et à l'instrumental. La variation des formes de l'item dans l'input permet de comparer les effets de la focalisation sur la forme dans l'enseignement sur la production effective de la morphologie nominale du polonais. Le principal constat effectué dans cette étude est que les apprenants ayant reçu l'enseignement focalisé sur le sens ont tendance à produire l'item "*Ulica*" au nominatif. Cette forme de l'item étant la première forme présentée dans le cours, elle semble être devenue la valeur par défaut de l'item aux yeux des apprenants, qui la considèrent donc plus souvent comme non fléchie et invariable. Ce constat n'est pas valable pour le groupe FB dont les apprenants produisent l'item sous différentes formes. De plus, au niveau individuel, cette plus grande variabilité dans la production des formes du même item se constate au sein d'un même énoncé plus souvent si l'apprenant fait partie du groupe FB.

Ainsi, les différences constatées entre les deux groupes d'apprenants semblent plus importantes lors de tâches de production. C'est pourquoi Watorek et collaborateurs ([Watorek et al., 2017]) ont analysé les performances des apprenants francophones aux deux tâches GJ et OQA. L'avantage de ces deux tâches est qu'elles testent le même paradigme linguistique en morphologie nominale, à savoir l'opposition instrumental *vs.* nominatif. La mise en parallèle de ces deux tâches autorise donc une plus grande comparabilité des résultats obtenus. Les analyses issues de cette étude montrent l'absence de différence significative entre les scores des apprenants des deux groupes dans la tâche GJ. En revanche, il existe une différence significative en ce qui concerne le temps de réaction. En effet, les apprenants ayant reçu l'input axé sur la forme (FB input) mettent plus de temps pour juger la grammaticalité d'une phrase que ceux qui ont suivi le cours axé sur le sens. Le traitement d'une phrase dans la tâche GJ serait donc plus long lorsque les apprenants ont accès à une explicitation de la règle enseignée. Selon les auteurs, on peut interpréter ce résultat en termes d'une mise en question plus rapide des hypothèses que les apprenants formulent sur le fonctionnement du système cible. Ils essaient de prendre en compte des indices morphologiques fournis par l'enseignant et de les intégrer dans la structure de leur lecture d'apprenant.

Le temps de réaction plus long indique l'impossibilité à appliquer une règle par l'apprenant du fait qu'elle est devenue « critique » dans le sens de Klein ([Klein, 1989]), et qu'elle est en cours de révision.

La comparaison des résultats de la tâche de traitement GJ avec ceux qui proviennent de la tâche de production ciblée (OQA) montre des différences intéressantes. Ainsi, le type d'enseignement impacte la production de la morphologie nominale (Instrumental *vs.* Nominatif). Les formes de l'instrumental masculin sont produites avec un score similaire par les apprenants des deux groupes. En revanche, les apprenants ayant été exposés à l'input axé sur la forme produisent plus de formes correctes dans la catégorie de l'instrumental féminin. Ce résultat renvoie à la réflexion sur la notion de saillance en acquisition, à laquelle nous reviendrons dans la section 4.1.3. L'instrumental masculin avec sa désinence régulière -em /Em/ pourrait être considérée comme perceptuellement saillante.

Exemple :

Patryk jest lekarzem/ Patryk est médecin (=Instrumental Masc)

La perception et la production de cette forme ne dépendrait donc pas du degré de la focalisation sur la forme (cf. le "principe de la saillance perceptuelle de l'input", [Slobin, 1985]). A l'inverse, l'instrumental féminin semble être moins saillant dans la mesure où il s'agit d'une nasale -ą /ɔ̃/ pouvant subir une dénasalisation à la fin du mot (/ɔ/ ou /ɔm/). La mise en relief d'une telle forme dans l'input (FB) de l'enseignant sur les diapositives (supports des cours) aide les apprenants à la percevoir et à la produire. Cette étude montre donc que l'effet de la focalisation sur la forme peut varier en fonction de l'habileté sollicitée par une tâche linguistique donnée et en fonction des caractéristiques des items utilisés dans la dite tâche. Ce sont les effets de ces deux composantes qui sont présentés dans la section suivante.

4.1.3 Les caractéristiques de l'input et le type de tâche

Les tâches prévues par le protocole du projet VILLA testent les acquis des apprenants dans différents niveaux d'analyse linguistique allant de la perception jusqu'aux premières productions. Ces acquis sont testés à la fois en traitement et en production. L'impact du type de tâche et donc du type de capacité sollicitée est visible dans les performances des apprenants. De façon générale, les apprenants obtiennent des scores plus élevés dans les tâches de traitement que dans celles qui impliquent une production. C'est ce que Saturno et Watorek ([Saturno et Watorek, 2020]) constatent en comparant les résultats des apprenants entre la tâche de compréhension PV et la tâche de répétition SI après 13h30 d'exposition à l'input. Ces deux tâches visent à tester la connaissance du paradigme nominatif *vs.* accusatif pour l'attribution du statut du constituant dans une phrase et comportent donc des phrases construites sur les modèles Sujet-Objet (SO) *vs.* Objet-Sujet (OS). Les auteurs constatent que sur leur échantillon de 89 apprenants, 80 d'entre eux associent correctement une image avec sa description lorsque la description est construite sur le modèle SO. Ce chiffre chute à 35 apprenants lorsqu'il s'agit de répéter des phrases SO en plus de les comprendre. Le même constat est valable lorsque l'on s'intéresse aux phrases de type OS, ordre des mots d'une phrase considérée comme plus difficile à acquérir (cf. chapitre 2). En effet, 42 des apprenants identifient correctement le sujet et l'objet des phrases OS dans la tâche de traitement mais parmi eux seulement 23 réussissent également à produire correctement les désinences appropriées dans la tâche de répétition. Les auteurs concluent que la complexité supérieure de la tâche SI, où non seulement la compréhension d'une phrase, le dessin d'une figure géométrique mais également la répétition de la phrase sont sollicités, fragilise la compréhension du paradigme morphosyntaxique, compréhension par ailleurs attestée dans la tâche PV. Cet argument est également soutenu par Ellis et Sagarra ([Ellis et Sagarra, 2011]) qui démontrent que lorsque la difficulté augmente, même les apprenants ayant démontré leur capacité à s'appuyer sur la morphologie flexionnelle pour traiter la LC ont tendance à s'appuyer uniquement sur le sens des items lexicaux pour comprendre une phrase. Dans les résultats aux tâches PV et SI, cette observation s'étend à la production. En effet, malgré une identification du sujet et de l'objet d'une phrase à partir des déclinaisons des items, cette compréhension n'est pas suffisamment robuste pour persister dans la tâche SI.

De plus, en ce qui concerne les tâches de production du projet VILLA, plus la tâche est ciblée, plus les scores augmentent en terme de production des formes correctes. En revanche, la production des formes correctes dans les tâches semi-libre diminuent de façon significative.

Pour ce qui est de l'impact du type de tâche de production sur les performances des apprenants, Watorek et col. ([Watorek et al., 2016]) mettent en relation les résultats que les apprenants francophones ont obtenus aux trois tâches allant de la reproduction d'une phrase à la production semi-libre. Il s'agit de la tâche de répétition des phrases (SI) où les apprenants doivent reproduire une phrase, de la tâche de production ciblée (OQA) où les apprenants répondent à une question précise avec un seul énoncé et une tâche semi-guidée d'indications (RD) d'itinéraire où les apprenants doivent construire un discours complexe composé de plusieurs énoncés répondant à un but communicatif complexe ([Levelt, 1989]). L'objectif de cette étude est d'évaluer l'effet du degré de liberté laissé à l'apprenant dans sa production sur le marquage morphologique des items nominaux employés.

Watorek et al. ([Watorek et al., 2016]) observent ainsi, en étudiant les résultats des apprenants français, que la nature communicative de la tâche prend le pas sur la structure morpho-syntaxique du discours de l'apprenant, conduisant directement à un taux de formes correctement fléchies très faible comparé à ceux constatés dans les tâches ciblées OQA et SI. Plus la tâche contraint la réponse de l'étudiant, plus celui-ci pourra se concentrer sur la forme de son énoncé, et donc le marquage morphologique. Dans une tâche semi-libre, le locuteur doit atteindre un but communicatif précis en gérant à la fois le niveau phrastique et discursif. Un apprenant débutant focalise ici son attention sur le niveau discursif afin de satisfaire le but communicatif de la tâche, en faisant moins attention à l'appropriation de son discours sur le plan formel. Ainsi, le marquage casuel approprié est plus faible dans cette tâche que dans les deux autres plus ciblées.

Également, les auteures soulignent que le choix des items par les apprenants pour construire leur discours en réponse à la tâche RD ne dépend que faiblement des caractéristiques de l'input de ces items, mais bien

de la structure phrastique, avec des items nominaux constituant des repères sur la carte utilisée en support ainsi que des verbes de mouvement. Ainsi, ces items sont sur-représentés dans les discours des apprenants, indépendamment de leurs caractéristiques dans l'input. À l'inverse, bien que le choix des items ne soit pas laissé à l'apprenant, les caractéristiques des items de SI et OQA influencent de manière importante les résultats des apprenants en terme d'application correcte de la morphologie flexionnelle du polonais.

Tout d'abord, en ce qui concerne la fréquence des items dans l'input, les items de haute fréquence sont considérablement mieux marqués morphologiquement lors de la première passation de la tâche de répétition de phrases en polonais que ceux de basse fréquence. Cet écart n'existe plus lors de la deuxième passation de la tâche, le temps d'exposition supplémentaire (de 4h30 de sessions d'enseignement) annulant la difficulté supplémentaire rencontrée par un apprenant lors de la répétition d'un mot faiblement présent dans l'input. De plus, la transparence d'un item aide également à sa répétition correcte, et ce tout au long de la période d'observation.

Les résultats obtenus pour la tâche OQA en regard des caractéristiques des items sont différents. Dans cette tâche, les items opaques sont plus souvent correctement marqués que les items transparents. L'hypothèse des auteures est qu'un item opaque dans l'input est plus difficilement analysé par l'apprenant. La séparation entre la racine et la désinence de l'item n'est pas effectuée et l'apprenant restitue donc la racine et la désinence comme un tout lors d'une tâche de questions-réponses. Les scores plus faibles des apprenants pour les items transparents dénoteraient d'un début d'analyse du marquage casuel de ces derniers, analyse conduisant à des marquages incorrects. Cette hypothèse est soutenue par le constat d'une baisse des productions des formes correctes des items opaques lorsque ceux-ci sont également fréquents. Le fait que l'apprenant soit confronté plus souvent à cet item et ses différentes formes permet à l'apprenant de commencer son analyse, malgré l'opacité de celui-ci.

Les scores plus faibles des apprenants pour les items transparents dénoteraient d'un début d'analyse du marquage casuel de ces derniers, analyse conduisant à des marquages incorrects.

Ainsi, bien que la fréquence et la transparence d'un item influent sur son analyse et sa réutilisation correcte vis à vis du système dérivationnelle de la LC, le type d'activité exercée par l'apprenant conditionne cette influence. Lors d'une tâche de production où la priorité est donnée à la transmission d'une information, la structure morpho-syntaxique passe au second plan, et l'acquisition de cette dernière par les apprenants, visible dans des tâches ciblées, n'est pas restituée dans leurs discours.

Seulement, une autre explication est envisagée. Les deux tâches SI et OQA ne testent pas l'acquisition du même paradigme du système flexionnelle de la LC (instrumental *vs.* nominatif, et nominatif *vs.* accusatif respectivement). Ainsi, depuis cette publication, les auteures se sont intéressées à une autre hypothèse explicative. Celle-ci stipule que la régularité et l'ambiguïté relative des désinences elles-mêmes, ainsi que leurs propriétés phonotactiques, pourraient être à l'origine de ces résultats.

Dans Rast et collaborateurs ([Rast et al., 2018]) les auteures s'intéressent ainsi non plus aux caractéristiques des items lexicaux permettant leur meilleure rétention et utilisation par l'apprenant, mais aux caractéristiques des structures morphologiques enseignées. À partir d'une analyse plus détaillée de la tâche OQA, les auteures cherchent à identifier les propriétés facilitatrices de l'acquisition d'une flexion par les apprenants du projet. Rappelons que la tâche OQA teste l'acquisition des cas de l'instrumental et du nominatif appliqués à des items nominaux féminins et masculins désignant des professions et nationalités. Ainsi, l'apprenant est confronté à quatre flexions possibles, résumées dans le tableau 4.1.

La question dans cette étude est de déceler parmi ces désinences laquelle est la plus facilement et rapidement identifiée et utilisée.

	Nominatif	Instrumental
Masculin	student-∅	student-em
Féminin	studentk-a	studentk-ą

TABLE 4.1 – Morphologie nominale polonaise présente dans la tâche OQA

Tout d'abord, il convient de décrire les déclinaisons présentes dans la tâche. Comme illustré dans le tableau 4.1, le nominatif masculin est marqué par l'absence d'un suffixe à l'item. Ainsi, le nominatif masculin se caractérise par les différentes consonnes en position finale des items auquel il est appliqué (*Informatyk* : k, *Amerykanin* : n, etc.). Le nombre de variations possibles marquant l'utilisation de ce cas et genre le rend hypothétiquement le moins identifiable des quatre. A l'inverse, le nominatif féminin est régulier puisqu'il n'est marqué que par l'ajout de la désinence -a. Seulement cette désinence est ambiguë, dans le sens où elle sert également à marquer d'autres combinaisons de cas et genre comme le génétif masculin et l'accusatif masculin en polonais. Cette multiplicité d'association de la même forme (-a) à différentes fonctions (génétif-masculin, ...) est potentiellement problématique pour l'apprenant. A contrario, les deux déclinaisons de l'instrumental, féminin et masculin, ne sont associées qu'à une fonction (et donc non utilisées pour d'autres cas), et sont régulières (même forme quel que soit l'item nominal). Ces deux déclinaisons semblent donc être les plus utilisées par les apprenants, en contexte approprié ou non. Le seul bémol à noter concerne l'instrumental féminin -ą qui peut être dénasalisé lors de sa production dans une phrase (se transformant en son /-om/), et peut donc ajouter de la confusion pour son identification.

L'analyse du taux d'items correctement fléchis pour chaque combinaison cas-genre confirme ces suppositions. Les formes de l'instrumental sont toutes les deux mieux produites en accord avec le contexte d'énonciation que les formes du nominatif. De plus, le nominatif féminin est également mieux fléchi que le nominatif masculin. Ces résultats viennent supporter la thèse de la régularité de la forme et de son association à une fonction unique comme deux caractéristiques facilitatrices de leurs acquisitions.

Il est intéressant d'analyser également le cas où l'apprenant ne fléchit pas correctement l'item. Que produit-il à la place? Les auteures notent que dans certains cas les apprenants produisent une flexion de l'item purement idiosyncrasique (c'est à dire ne relevant pas des possibilités en LC mais étant propre à l'interlangue de l'apprenant, cf. chapitre 1) ou ne produisent rien, tandis que dans d'autres cas les apprenants utilisent une des quatre déclinaisons présentes dans la tâches mais de manière inappropriée. Cette sur-généralisation de certaines formes sur d'autres soulignent la rétention importante de ces premières par l'apprenant. Ainsi, on constate une sur-généralisation de l'instrumental dans au moins 20% des phrases requérant un nominatif. Le genre par contre semble respecté (instrumental féminin sur le nominatif féminin et instrumental masculin sur le nominatif masculin). On retrouve ce résultat dans la littérature, le genre ne pose en effet apparemment que peu de problème aux apprenants français du polonais ([Gniadek, 1979]).

La sur-généralisation des formes de l'instrumental sur celles du nominatif peut également être expliquée par la saillance perceptuelle des deux formes de l'instrumental dans l'*input*. La saillance perceptuelle de l'*input* est un principe général de traitement du langage observé par Slobin ([Slobin, 1985]) en L1 et confirmé en acquisition L2 ([Slobin, 1993, Slobin, 2012]). Ce principe énonce qu'une forme perceptuellement saillante et régulière est traitée très précocement. Ce principe est également évoqué dans Watorek et collaborateurs ([Watorek et al., 2017]) pour expliquer les résultats à la tâche OQA.

De plus, si l'on s'intéresse aux caractéristiques des items nominaux dans l'*input*, la production correcte de l'instrumental masculin semble influencée par la transparence de l'item décliné, ce qui n'est pas le cas pour l'instrumental féminin. Ainsi, à la *régularité de la forme*, au *lien unique d'association forme-fonction*, et à la *saillance perceptuelle* doit s'ajouter une autre caractéristique non initialement prévue par les auteures. L'*opposition binaire* claire au féminin, entre l'instrumental et le nominatif, participe de son identification et

donc de son utilisation : - \mathfrak{z} (Instrumental) *vs.* -a (Nominatif).

En conclusion, la LM de l'apprenant, l'enseignement qu'il a reçu, ainsi que les caractéristiques des items lexicaux dans l'input, mais également les caractéristiques des formes grammaticales enseignées sont autant de facteurs qui vont influencer le parcours acquisitionnel de l'apprenant. De plus, ces différentes influences ne s'expriment pas de la même manière en fonction du type d'activité sollicité par la tâche. Au vu de l'objectif de cette thèse, il convient d'essayer de modéliser le processus d'acquisition d'une L2 dans un système complet incluant ces différents facteurs pour ainsi opérer un partitionnement des apprenants en différents profils prenant en compte ces variables et leurs interactions.

4.2 Des données à la notion de représentation

Les études présentées dans la section précédentes ont été effectuées sur des sous-échantillons de la BDD du projet VILLA. Dans cette thèse, nous souhaitons effectuer une analyse du parcours acquisitionnel de l'apprenant en se basant sur toutes les données à disposition. Ainsi, dans cette section, une tentative de représentation unifiée des dimensions caractéristiques des données brutes est proposée. Cette représentation unifiée de la BDD a pour objectif principal la caractérisation multifactorielle d'un apprenant. En effet, le processus de regroupement d'apprenants en profil, objectif de cette thèse, nécessite un moyen de comparaison entre eux. Or, la comparaison de deux objets n'est pas faisable si l'objet lui-même n'est pas caractérisé.

4.2.1 La base de données

Notre point de départ est un tableau des scores obtenus, par chacun des apprenants, et pour chaque combinaison des modalités des variables présentées dans le chapitre 2. Chaque score est une valeur comprise entre 0 et 1. Par exemple, l'étudiant 1102 pourra avoir obtenu un score de 0.7 à la tâche de question-réponse OQA, au temps 1 de test, pour les items fréquents et transparents, dans le cas d'une question sollicitant une réponse marquée au nominatif féminin.

Chaque score stocké dans la base de données est donc défini par rapport à plusieurs dimensions : l'individu ayant obtenu ce score ; le test l'ayant entraîné ; les différentes caractéristiques de l'item auquel il est associé ; la période de passation du test. Ainsi, chacune de ces dimensions peut prendre différentes valeurs. À l'aspect multidimensionnel des données s'ajoute une structure hiérarchique pour certaines de ces dimensions : un test possède un ID, relève d'un paradigme linguistique, et sollicite de la part de l'apprenant un certain type d'activité (traitement *vs.* production). Les dimensions peuvent donc elles-mêmes être abordées selon plusieurs niveaux. La figure 4.1 ([Durand, 2016]) procure, par un diagramme en étoile, une aide à la visualisation de la complexité associée à chaque mesure de la base de données.

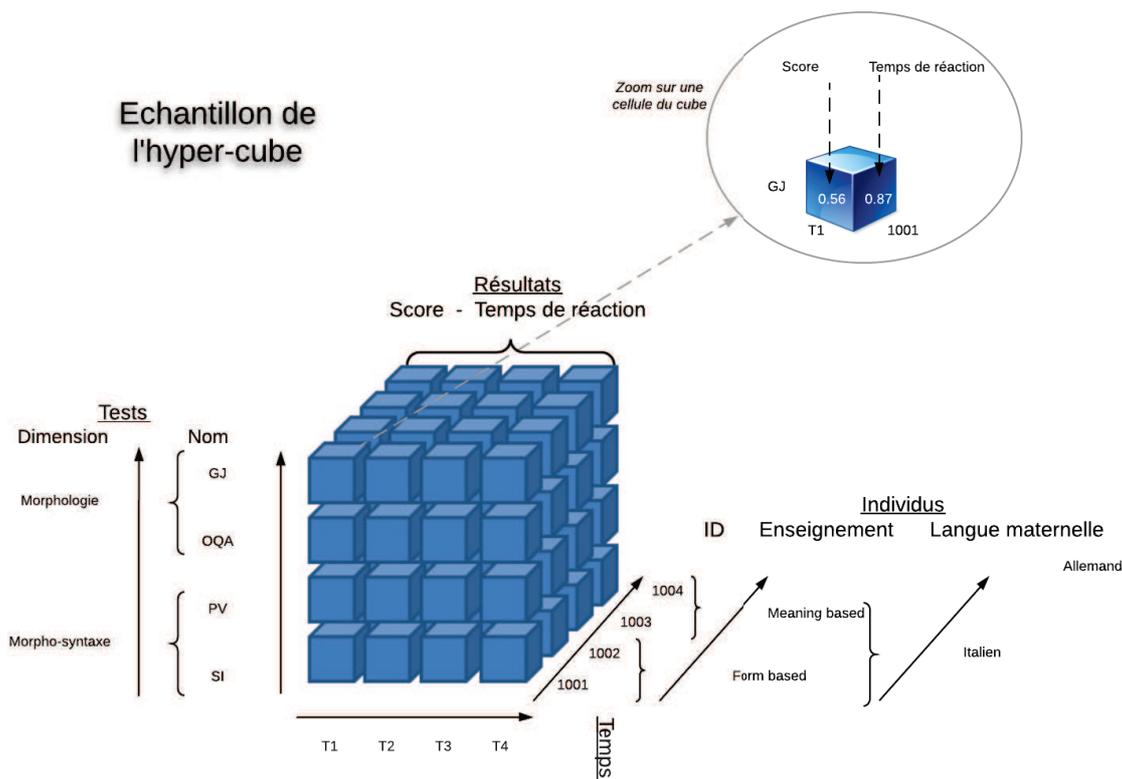


FIGURE 4.2 – Échantillon en 3 dimensions de l’hypercube représentant la BDD du projet VILLA

Il reste important de souligner l’asymétrie des données recueillies au cours du projet VILLA. Par exemple, certains tests linguistiques ont été effectués par les apprenants plusieurs fois (tests à passation multiple) tandis que d’autres une seule. De plus, durant la période d’observation, les tests à passations multiples n’ont pas tous été réalisés le même nombre de fois ni à la même période d’enseignement. Également, nous disposons de la mesure du temps de réaction uniquement pour certains tests programmés sur E-prime, cette donnée n’est donc pas disponible pour tous les tests de traitement. De ces diverses asymétries dans le protocole d’expérimentation découlent directement des données déséquilibrées. C’est un fait qu’il semble important de garder à l’esprit lors de la représentation de la BDD du projet. L’hypercube métaphorique regroupant les données est donc un hypercube amputé. Le fait que les données soient manquantes en raison de la structure même du protocole de collecte des données n’autorise pas l’utilisation des procédures d’analyse robustes aux données manquantes. En effet, ces procédures ne sont utiles que lorsque les données manquantes sont aléatoires et imprévues par le plan expérimental. L’asymétrie du protocole d’expérimentation constitue donc une caractéristique importante qu’il faut prendre en compte lors de l’analyse des données.

L’évaluation d’un apprenant d’une LE s’opère donc par l’analyse de ses compétences sur différents niveaux d’analyse linguistique. Ces différents niveaux sont interdépendants et complémentaires mais non miscibles. La question de la représentation d’un apprenant et de la manipulation de celle-ci à des fins de comparaisons quantitatives se pose.

4.2.2 Représentation multicritère de données comportementales

Nous proposons ici une représentation multicritère d'un apprenant et du formalisme associé pour sa manipulation ([Durand et Truck, 2015]). L'objectif à moyen terme était d'établir un objet numérique formalisé doté d'une mesure de similarité à des fins de comparaison d'un ensemble de ces objets lors du processus de classification.

La problématique principale consiste en l'évaluation d'un apprenant sur ses compétences en LE. Cette évaluation est disponible sous la forme d'expressions linguistiques de jugements émises par un évaluateur sur les résultats d'un apprenant selon différents critères. Les classifications supervisée et non supervisée sont communément utilisées pour des données numériques ou symboliques. Lorsque les données sont des phrases, des opinions, etc., c'est-à-dire lorsqu'elles sont des termes linguistiques, le *computing with words* (calcul à l'aide de mots) propose un ensemble de modèles destiné à leurs analyses. Parmi ces modèles se trouve la notion de 2-tuple. Les 2-tuples linguistiques flous ont été introduits par Herrera & Martínez ([Herrera et Martínez, 2000]) afin d'obtenir le résultat de l'agrégation d'expressions linguistiques exprimées sur une échelle linguistique prédéfinie avec un minimum de perte d'informations sur cette même échelle linguistique plutôt que sur un ensemble de définition propre au résultat de cette agrégation.

Pour un ensemble de termes linguistiques ordonnés $S = \{s_0, s_1, \dots, s_g\}$, où s_0 peut être "très mauvais", s_1 "mauvais" etc., on exprime un jugement linguistique sur cette échelle par l'intermédiaire d'un couple (s_i, α) , où $s_i \in S$ et α est une translation symbolique $\in [-0.5, 0.5]$. La translation symbolique α permet de spécifier le jugement sur l'échelle S sans recourir à un changement d'échelle si celle-ci ne correspond pas à un terme existant dans S .

Graphiquement, l'ensemble de termes linguistiques est composé de SEFs triangulaires distribués uniformément et symétriquement sur l'ensemble $[0, 1]$. Un ensemble d'opérateur est défini sur S :

- l'opérateur de négation $Neg(s_i) = s_j$ tel que $j = g - i$ où $g + 1$ est la cardinalité.
- l'opérateur de maximum : $max(s_i, s_j) = s_i$ si $s_i \geq s_j$.
- l'opérateur de minimum : $min(s_i, s_j) = s_i$ si $s_i \leq s_j$.

Une amélioration de ce modèle a été proposée afin de gérer le cas où les termes linguistiques ne sont pas uniformément distribués sur $[0, 1]$, c'est-à-dire le cas où l'ensemble des termes linguistiques est déséquilibré *i.e* asymétrique (*unbalanced*). Dans ce cas de figure, on définit \mathcal{S} un ensemble de termes linguistiques asymétriques, définit comme l'union de plusieurs ensembles linguistiques symétriques. Chaque ensemble linguistique symétrique dont est issu \mathcal{S} représente un de ses niveaux hiérarchiques. Un terme $s_i \in \mathcal{S}$ peut ainsi être exprimé sur un niveau hiérarchique différent de s_{i+1} . De plus, la partie gauche de s_i peut appartenir à un certain niveau hiérarchique tandis que la partie droite sera issue d'un autre, résultant en une asymétrie du SEF triangulaire de s_i . Pour noter l'appartenance de s_i à la hiérarchie appropriée, on note $s_i^n(t)$ où $n(t)$ est la cardinalité du niveau hiérarchique d'expression de s_i . Par exemple, soit un ensemble $\mathcal{S} = \{F, D, C, B, A\}$ où D est la valeur milieu. L'ensemble est ainsi asymétrique avec plus de termes à droite du terme intermédiaire qu'à sa gauche. Ainsi, après calcul, F est noté s_0^3 , D est noté s_1^3 , exprimés sur la même hiérarchie, et C est noté s_3^5 , B s_7^9 , et A s_8^9 .

Dotée de ces modèles, notre problématique concerne la représentation d'un apprenant, à travers le prisme de plusieurs critères non miscibles, mesurés par des expressions linguistiques de jugement émises par des évaluateurs ou assignées à une valeur numérique β par une fonction Δ telle que :

$$\begin{aligned} \Delta : [0, g] &\rightarrow S \times [-0.5, 0.5] \\ \Delta(\beta) &\mapsto (s_i, \alpha) \quad \text{avec} \begin{cases} s_i, & i = \text{arrondi}(\beta) \\ \alpha = \beta - i, & \alpha \in [-.5, .5) \end{cases} \end{aligned}$$

Nous nous dotons également de sa réciproque Δ^{-1} :

$$\begin{aligned} \Delta^{-1} : S \times [-0.5, 0.5] &\rightarrow [0, g] \\ \Delta^{-1}((s_i, \alpha)) &\mapsto \beta = \alpha + i \end{aligned}$$

Ainsi, un apprenant décrit sur p critères F_1, F_2, \dots, F_p est exprimé comme un vecteur colonne de 2-tuples linguistiques flous :

$$\begin{bmatrix} (s_{i_1}, \alpha_{i_1}) \\ (s_{i_2}, \alpha_{i_2}) \\ \dots \\ (s_{i_p}, \alpha_{i_p}) \end{bmatrix}$$

Pour deux critères F_1 défini sur $\mathcal{S}_1 = \{s_{0_1}, s_{1_1}, s_{2_1}, s_{3_1}, s_{4_1}, s_{5_1}\}$ et F_2 défini sur $\mathcal{S}_2 = \{s_{0_2}, s_{1_2}, s_{2_2}, s_{3_2}\}$, un apprenant P est graphiquement représenté par un tétraèdre avec 5 arêtes dénotées a, b, c, d, e , chacune ayant deux coordonnées $a(x), a(y), b(x)$, etc. En effet, les deux critères étant exprimés sous la forme de 2-tuples, un apprenant devient le produit cartésien de deux 2-tuples linguistiques flous. On constate en figure 4.3 qu'un apprenant P caractérisé sur deux attributs F_1 et F_2 tel que $F_1 = (s_{3_1}, 0.4)$ et $F_2 = (s_{2_2}, -0.25)$ est un 2-tuple linguistique flou 3D.

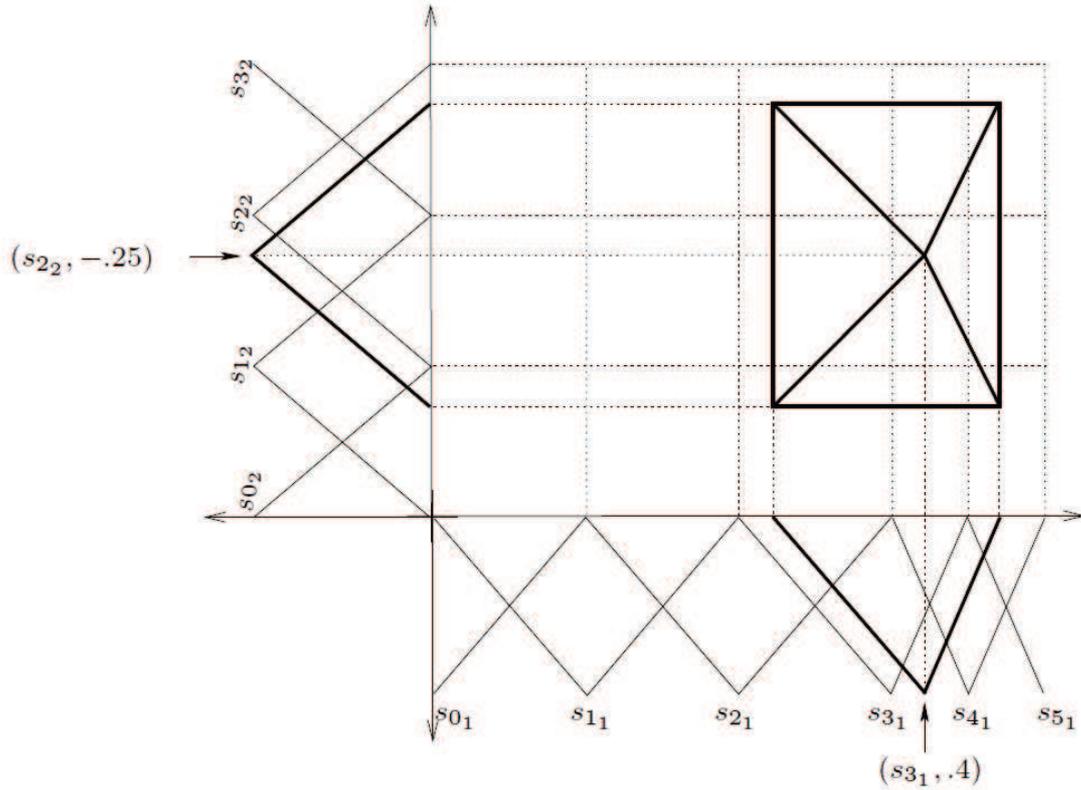


FIGURE 4.3 – Représentation de l'apprenant P en 2-tuple linguistique flou 3D.

L'objectif est maintenant d'établir la similarité entre deux 2-tuples linguistiques flous 3D. Pour ce faire, il faut définir une mesure de distance entre eux. Etant donné que les espaces de définition de chaque attribut définissant un apprenant ne sont pas les mêmes, il convient de calculer une valeur de distance pour chacun des attributs.

La distance est dénotée ς et est composée de plusieurs coordonnées, une par attribut. Pour p critères, ς est exprimée sous la forme d'un vecteur colonne :

$$\varsigma = \begin{bmatrix} \varsigma_1 \\ \varsigma_2 \\ \dots \\ \varsigma_p \end{bmatrix}$$

avec ς_j la $j^{\text{ième}}$ valeur de distance pour la caractéristique j .

Pour $p = 2$ et P et Q deux 2-tuples linguistiques 3D :

$$\varsigma : (S_1 \times [-.5, .5]) \times (S_2 \times [-.5, .5]) \rightarrow [0, 1] \times [0, 1]$$

$$\varsigma(P, Q) = \begin{bmatrix} \varsigma_1(P, Q) \\ \varsigma_2(P, Q) \end{bmatrix}$$

avec

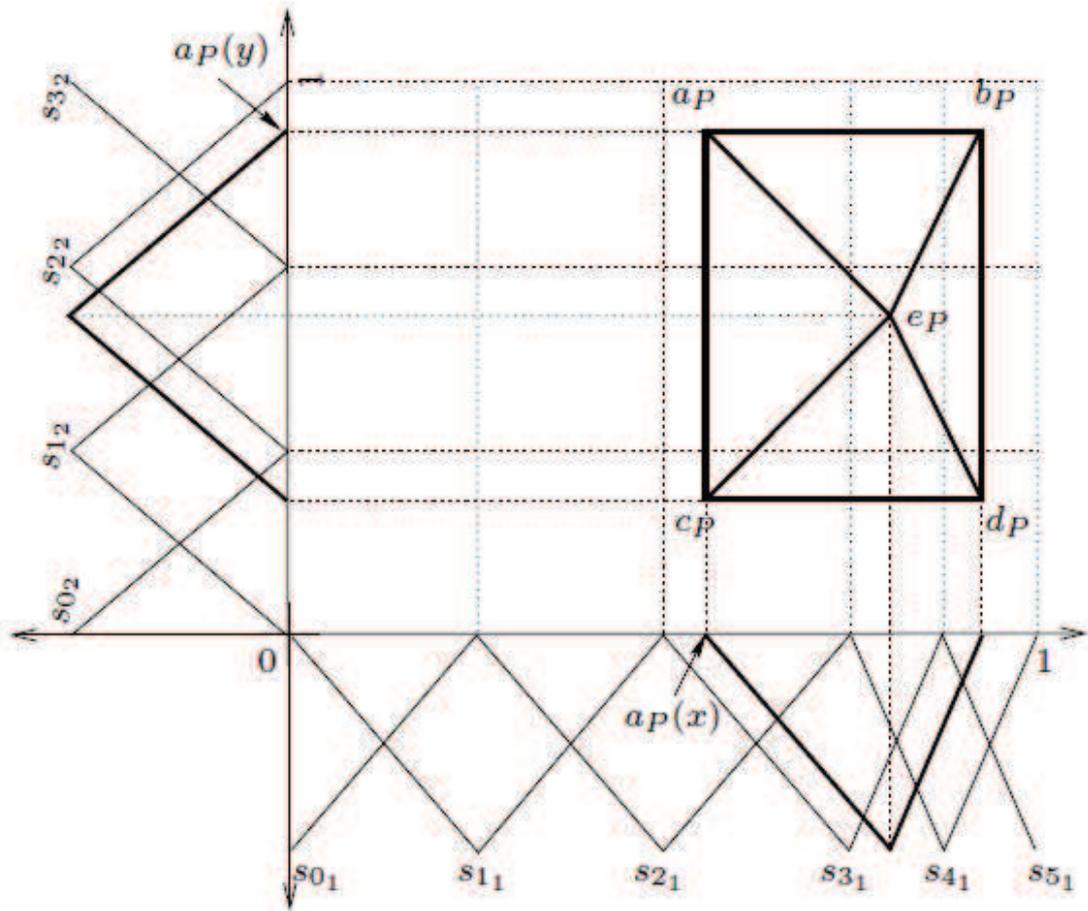
$$\begin{aligned} s_1(P, Q) &= \sqrt{\sum_{\iota=a}^e w_\iota (\iota_P(x) - \iota_Q(x))^2} \\ &= \sqrt{w_a(a_P(x) - a_Q(x))^2 + w_b(b_P(x) - b_Q(x))^2} \\ &\quad + w_c(c_P(x) - c_Q(x))^2 + w_d(d_P(x) - d_Q(x))^2 \\ &\quad + w_e(e_P(x) - e_Q(x))^2} \end{aligned}$$

et respectivement

$$s_2(P, Q) = \sqrt{\sum_{\iota=a}^e w_\iota (\iota_P(y) - \iota_Q(y))^2}$$

où w_ι est le poids assigné à chaque point ι du tétraèdre et $\sum w_\iota = 1$. Étant donné que les points a, b, c et d correspondent à la base du tétraèdre et donc à une valeur d'appartenance de 0 à s_i , leurs poids doivent être considérablement réduits par rapport à e qui correspond à une valeur d'appartenance à s_i de 1.

Par exemple, on assigne à a, b, c et d un poids quatre fois plus petit que celui de e : $w_a = w_b = w_c = w_d = .125$ et $w_e = .5$ avec $\sum w_\iota = 1$. Cela signifie que l'apex est aussi importante que les quatre sommets réunies. Pour le pattern P , avec s_{i_P} dans l'ensemble des termes linguistiques S_1 , $a_P(x) = c_P(x) = \Delta^{-1}((s_{i_P-1}, \alpha_{i_P}))$ (cf. figure 4.4). $b_P(x) = d_P(x) = \Delta^{-1}((s_{i_P+1}, \alpha_{i_P}))$. De même, pour le pattern Q , avec s_{i_Q} dans l'ensemble des termes linguistiques S_1 , $a_Q(x) = c_Q(x) = \Delta^{-1}((s_{i_Q-1}, \alpha_{i_Q}))$. $b_Q(x) = d_Q(x) = \Delta^{-1}((s_{i_Q+1}, \alpha_{i_Q}))$. $e_P(x) = \Delta^{-1}((s_{i_P}, \alpha_{i_P}))$.

FIGURE 4.4 – Sommet et apex du tetrahedron du pattern P .

On procède de même pour obtenir les résultats dans l'ensemble de termes linguistiques S_2 pour les deux patterns P et Q . Par l'intermédiaire de la fonction Δ^{-1} , on calcule ainsi les deux distances entre c_1 et c_2 . La question de l'agrégation des résultats des deux mesures de distance se pose alors, afin d'obtenir une mesure de distance finale. Pour plus de détails, différents opérateurs d'agrégation sont testés sur des données artificielles dans [Durand et Truck, 2015].

La mesure de distance ainsi établie permet d'entrer dans le processus de partitionnement lui-même. Ainsi, si l'on considère l'algorithme des K-moyennes, l'étape suivante sera la comparaison de chaque apprenant avec un nombre prédéfini de clusters également exprimés en 2-tuples. Le calcul de la distance de chaque apprenant avec chacun des clusters permettra de déterminer à quel cluster l'apprenant appartient le plus vraisemblablement.

Cette proposition de représentation d'un apprenant à travers la manipulation de 2-tuples linguistiques flous 3D permet de capturer la nature multicritère de la BDD du projet VILLA. De plus, elle permet la création d'une mesure de similarité entre les apprenants qui respecte l'aspect non miscible des différents attributs correspondant aux différents niveaux d'analyses linguistiques des productions des apprenants. La nature même des 2-tuples autorise également de prendre en compte la possibilité d'une expression imprécise

de jugement des compétences de l'apprenant sur un critère donné, ainsi que les expressions linguistiques n'appartenant potentiellement pas à la même échelle de référence, c'est à dire au même ensemble de termes linguistiques. Cependant, les données effectives disponibles pour le projet VILLA sont des données de scores numériques. La procédure présentée est motivée par une possible application en classe de langue du travail de cette thèse. En effet, en classe de langue l'appréciation du travail de l'apprenant n'est pas exprimée sous forme d'un score testant l'apprenant sur différents niveaux linguistiques mais sur des tests globaux. Ainsi, une appréciation des compétences de l'apprenant pour chaque critère qualifiant son acquisition en LC serait coûteuse en terme de temps pour l'enseignant. Cependant, celui-ci, à partir d'un devoir classique, peut être capable de rapidement formuler un jugement des dites compétences en termes linguistiques. Cette procédure est donc seulement théorique et nécessite d'être évaluée sur de véritables données issues de l'appréciation d'un évaluateur sur les compétences d'un apprenant L2.

Dans ce chapitre nous avons abordé les différents résultats issus de la littérature sur le projet VILLA afin de comprendre les interactions complexes existant entre les facteurs d'influence de l'acquisition d'une LE. La prise en compte de ces différents facteurs nous a amené à les représenter et à réfléchir sur une telle représentation complète et sans perte d'information du parcours acquisitionnel d'un apprenant. L'objectif étant la manipulation de cette représentation, nous avons proposé un modèle permettant d'exprimer les compétences d'un apprenant évaluées linguistiquement par l'intermédiaire des 2-tuples, et nous avons développé une mesure de distance associée, mesure au cœur de tout processus de partitionnement.

Dans le prochain chapitre, nous abordons l'élaboration de l'algorithme de partitionnement que nous souhaitons utiliser pour la mise en exergue des différents profils d'apprenants en présence dans la BDD issue du projet VILLA. A cette fin, nous commencerons par un questionnement sur la notion de comparabilité entre apprenants et sur la capture de l'imprécision dans les données comportementales qui nous permettra de transformer nos données, puis nous énoncerons les différentes étapes de l'algorithme ainsi construit.

Chapitre 5

Algorithme de classification semi supervisée spécifique au contexte

Sommaire

5.1	Pré-traitement des données	81
5.1.1	Vers quelles stratégies ?	82
5.1.2	Transformations en sous-ensembles flous	86
5.1.3	Manipulations d'expressions linguistiques de jugements de performances	91
5.1.4	<i>Cannot-link</i> : qu'est-ce que la comparabilité ?	95
5.1.5	Transformation des données	96
5.2	Les étapes de l'algorithme	100
5.3	Validation de l'algorithme	111

Dans le chapitre précédent, nous avons pu identifier les attributs pertinents pour la caractérisation de nos individus dans leur processus d'acquisition d'une langue étrangère. Un apprenant est maintenant défini par ses scores à différents tests linguistiques et par ses performances dans ces derniers, vis-à-vis des caractéristiques des items qui y sont utilisés. Dans la première partie de ce chapitre nous nous intéressons à l'attribution de sens aux différentes valeurs de ces scores. Un score numérique peut être révélateur de stratégies diverses des apprenants dans la réalisation des tâches auxquels ils ont été soumis. L'objectif est ici de définir quels apprenants ne sont pas comparables entre eux en fonction de ces scores, et donc des stratégies mises en place. Dans la seconde partie de ce chapitre nous présentons les différentes étapes de notre algorithme de classification semi supervisée.

5.1 Pré-traitement des données

L'objectif de ce travail est d'intégrer les connaissances issues des travaux en sciences du langage et des analyses des données du projet VILLA, dans le processus de partitionnement des apprenants. Après avoir modélisé une représentation adéquate de nos apprenants du point de vue acquisitionniste, nous souhaitons attribuer une explication comportementale aux scores des différents attributs pour chaque apprenant. Ces scores permettent l'identification de variations dans les stratégies mises en place par les apprenants dans l'exécution des différentes tâches. Nous serons donc capables de définir si deux individus sont comparables ou non, en fonction de leur stratégie respective. Cette transformation est établie afin de s'affranchir d'une comparaison d'objets purement numériques, opérée par une mesure de distance classique. En ce sens, nous présentons un travail sur le traitement d'expressions linguistiques exprimant un jugement des performances

d'un apprenant. L'objectif de ce pré-traitement des données est de rendre possible la séparation des individus jugés non comparables par la génération de l'ensemble des contraintes *cannot-link* de notre ensemble de données. Un critère de non-comparabilité entre deux apprenants, sur n'importe lequel de leurs attributs, donne lieu à la création d'une contrainte *cannot-link* les reliant. La notion de comparabilité est définie dans le présent chapitre, définition purement **spécifique** au contexte et empirique car issue des travaux en acquisition et des analyses des données issues du projet VILLA. Suit alors dans ce chapitre la présentation d'un algorithme **générique** intégrant ces contraintes spécifiques par l'intermédiaire d'une "mesure" de comparaison, puis de similarité, afin d'obtenir une partition finale des apprenants selon leur profil respectif.

5.1.1 Vers quelles stratégies ?

Notre processus de modélisation a permis de mettre en évidence deux types de sources d'information sur le processus d'acquisition de nos apprenants. Le premier regroupe les scores des apprenants aux tâches du projet VILLA, tâches permettant d'évaluer leurs progressions sur différents niveaux linguistiques pertinents. Le second type d'information accessible concerne le score des apprenants aux différentes tâches en fonction des caractéristiques des items qui y sont utilisés, ainsi qu'en fonction de la période de passation de ces tâches. Ces deux types d'informations sont différents. Pour le premier type, le score d'un apprenant sera révélateur de son avancement dans l'appropriation de la LC. Pour le second type, nous parlons plutôt non pas d'une appropriation par l'apprenant du système de la LC, mais d'une sensibilité à l'input et ses caractéristiques. L'objectif principal étant la création de profils d'apprenants, nous souhaitons observer si des apprenants réagissent différemment à ces caractéristiques d'items. Nous savons déjà par l'analyse partielle et fragmentaire des données VILLA que des effets de la fréquence d'un item, en interaction avec sa transparence, sont observables (cf. chapitre 4). Ces caractéristiques jouent donc un rôle dans l'utilisation appropriée au contexte de tels items. Ce travail propose de visualiser si cet impact est unique ou non pour tous nos apprenants.

Dans ce travail, nous postulons que ces différentes variables agissent différemment sur le parcours acquisitionnel en fonction de l'apprenant considéré. Si les caractéristiques des items impactent différemment l'acquisition des participants au projet, comment qualifier cette différence ? Est-elle suffisante pour être un des facteurs de regroupements de certains de nos étudiants ? Ces questions sont également valables lorsqu'on étudie l'évolution globale des scores de nos apprenants, toutes tâches confondues, entre la première période de passation des tests et la deuxième période. Il est évident que l'instruction en classe de polonais (l'exposition à l'input) va permettre aux apprenants d'avancer dans leur acquisition, surtout qu'au début de l'apprentissage d'une langue étrangère, les participants partent de zéro (cf. chapitre 1). La question est donc la même ici, les effets bénéfiques des séances d'enseignement sont-ils les mêmes pour tous les apprenants de la BDD ou s'expriment-ils différemment ? Existe-t-il une régularité suffisante dans ces différences pour l'identification et la qualification de groupes homogènes d'apprenants, sur la base de leur évolution entre deux périodes de tests ?

Pour une utilisation adéquate des scores des apprenants pour les différents niveaux linguistiques, il convient de les analyser toujours dans le cadre de la tâche de laquelle ils sont issus.

En effet bien que nos tâches testent chacune un paradigme linguistique précis, leur construction, par nature expérimentale, ne permet pas de capturer une compétence générale sur un niveau linguistique mais bien une habileté de l'apprenant à répondre à cette tâche, même si celle-ci fait appel aux représentations qu'a l'apprenant sur le fonctionnement de la LC.

Tout d'abord, les tâches sollicitant un traitement de la part de l'apprenant (et non une production orale originale) comme un jugement de similarité entre deux sons, ou un jugement de grammaticalité d'une phrase, sont construites de manière à ce que l'apprenant n'ait que deux choix de réponses. Cette réponse est généralement donnée par l'intermédiaire du choix entre deux cases à cocher, deux boutons sur lesquels appuyer. Il n'est donc pas impossible que notre apprenant réponde au hasard en choisissant aléatoirement l'un ou l'autre des boutons.

Ensuite, rappelons que chaque tâche examine un paradigme linguistique précis sur un niveau linguistique

donné :

- La tâche *Phoneme Discrimination* relève du niveau phonologique. Elle permet de qualifier la capacité discriminative entre deux phonèmes. Ces phonèmes incluent des sons de la langue polonaise auxquels ne sont pas habitués nos apprenants car ils n'existent pas dans leur langue maternelle respective. La discrimination entre deux paires phonémiques n'est donc possible que si l'apprenant est sensibilisé aux phonèmes de la langue cible.
- La tâche *Word Recognition* relève du niveau lexical. Elle permet de connaître l'étendue du lexique de l'apprenant. L'apprenant est capable d'identifier la présence d'un mot précis dans une phrase si ce mot lui est acquis.

La tâche WR semble donc fortement liée aux caractéristiques des items utilisés lors des séances d'enseignement. L'hypothèse selon laquelle un mot fréquent sera plus facilement retenu peut s'appliquer ici. De même un phonème plus fréquemment entendu sera plus facilement identifiable par l'apprenant. Seulement, la fréquence des phonèmes n'a pas été contrôlé dans l'input du projet VILLA. Cette hypothèse devrait être vérifiable, dans un autre travail, par l'analyse des enregistrements audio du projet. La notion de transparence ne s'applique pas au test *Phoneme Discrimination* mais est pertinente pour la tâche *Word Recognition*. Une des hypothèses à la base de l'utilisation de ce test dans le projet VILLA est de vérifier si les mots transparents seront plus facilement isolés du reste de la phrase de test, et donc identifiés. Le test WR semble donc un bon indicateur de la sensibilité des apprenants aux propriétés de l'input.

- La tâche *Grammaticality Judgement* (GJ) relève du niveau morphologique. Elle est composée de phrases grammaticalement correctes du fait de l'utilisation appropriée de l'instrumental et de phrases grammaticalement incorrectes où le nominatif est utilisé en lieu et place de l'instrumental. Elle permet de tester la sensibilité à la morphologie flexionnelle.
- La tâche *Picture Verification* (PV) relève du niveau morphosyntaxique. Elle est constituée de deux types de phrases, aux propriétés syntaxiques différentes. Le premier type de phrase est construit selon l'ordre d'apparition des mots Sujet puis Objet. Le deuxième type inverse cet ordre. Ainsi l'identification de la fonction de l'item dans la phrase ne repose plus sur son ordre d'apparition mais sur le cas auquel il est décliné, à savoir nominatif s'il est sujet, accusatif s'il est objet. Cette tâche permet d'évaluer la conscience de l'apprenant de l'absence d'un ordre strict des mots en polonais, et de la caractérisation de leur fonction par l'intermédiaire du système morphologique.

Pour réussir ces deux tâches, l'apprenant doit avoir intégré les règles du système casuel en polonais, et avoir réussi l'association correcte des cas avec leur formes. Un tel système n'existe (de manière allégée) que pour une langue maternelle sur les cinq possibles des apprenants, l'allemand.

En ce qui concerne la tâche GJ, pour obtenir un bon score, l'apprenant doit reconnaître dans le contexte phrastique, ce qui implique l'emploi de l'instrumental, que les désinences appropriées sont -em pour les items masculin et -ą pour les items féminin. Les désinences du nominatif masculine ou féminine dans le contexte impliquant l'emploi de l'instrumental doivent être identifiées comme incorrectes, la phrase étant jugée incorrecte.

Sachant qu'il existe 50% de phrases grammaticalement correctes, un score de 50% de bonnes réponses peut indiquer plusieurs états de l'interlangue de l'apprenant. Nous émettons d'abord l'hypothèse de la réponse aléatoire pour expliquer un tel score médian. Cependant, il faut également considérer la possibilité qu'un score de 50% de bonnes réponses soit obtenu, car l'apprenant juge l'ensemble des phrases correctes.

Cet état de fait est possible si l'apprenant n'ayant pas encore assimilé l'existence d'un système morphologique en polonais, juge donc (globalement) toutes les phrases comme grammaticalement correctes, alors que la moitié sont incorrectes. A contrario, l'apprenant étant sensible à la morphologie nominale de la LC, pourra soit obtenir un score tendant vers les 100% de bonnes réponses, si celui-ci a de plus clairement identifié les marques de l'instrumental, soit obtenir un score tendant vers le 0% de bonnes réponses, si son acquisition du paradigme instrumental *vs.* nominatif n'est pas relié aux formes appropriées.

Pour ce qui est de la tâche PV, c'est le paradigme nominatif *vs.* accusatif qui est testé. Ce test sert de support à l'illustration de l'absence d'un ordre strict des mots en polonais, la caractérisation du rôle du constituant dans la phrase se faisant par l'intermédiaire de la morphologie. Pour réussir cette tâche, l'apprenant doit donc être conscient du système casuel, et de son rôle dans l'assignation du statut du constituant dans la phrase (le nominatif exprimant le sujet et l'accusatif l'objet). Sachant qu'un tiers des phrases de ce test sont construites sur le modèle Sujet-Objet et deux tiers sur le modèle Objet-Sujet, nous rééquilibrons le poids de chacun des types de phrases présents dans la tâche afin que l'ordre Sujet-Objet compte autant dans le score final que l'ordre Objet-Sujet. Ainsi, le score global obtenu à cette tâche permet de distinguer les étudiants qui ont amorcé un affranchissement de l'ordre des mots dans la composition de la phrase de ceux qui respectent encore l'ordre Agent-Patient pour l'attribution de la fonction du mot dans la phrase. Dès lors, un score autour de 50% découlera soit d'une stratégie de réponse aléatoire, soit celle de l'application du principe d'analyse Agent-Patient pour l'identification du rôle de chaque constituant de la phrase en LC.

En ce qui concerne les deux tâches de production SI et OQA, les mêmes hypothèses d'association d'un score à une stratégie ne s'appliquent pas. En effet, la production orale de l'apprenant en LC, lorsqu'elle existe, ne peut relever d'une stratégie aléatoire, comme lors d'une activité de traitement, lorsque la réponse consiste à choisir entre deux items. Lorsqu'un apprenant s'exprime effectivement en LC, il est obligé de mobiliser son interlangue. Il convient de s'attendre à de plus faibles score généraux pour une tâche de production. En effet, même pour les apprenants réussissant à traiter la morphologie nominale du polonais n'arriveront peut être pas tous à l'appliquer en production. En effet, les tâches de productions ciblées SI et OQA reposent également (comme pour les tâches PV et GJ) sur une dualité entre deux paradigmes.

- La tâche *Oral Question Answer* (OQA), teste en production la maîtrise du même paradigme que la tâche GJ. En fonction de la question l'apprenant doit utiliser dans sa réponse, soit le marquage casuel au nominatif, soit à l'instrumental (cf. chapitre 2). Cette tâche permet de tester la capacité de l'apprenant à produire des phrases réponses à l'instrumental et au nominatif en fonction du contexte induit par la question. Les résultats à cette tâche peuvent être mis en parallèle avec les résultats de la tâche GJ, puisqu'elles testent toutes deux l'acquisition des mêmes paradigmes linguistiques.
- La tâche *Sentence Imitation* (SI), tout comme la tâche PV, sollicite la répétition de phrases en LC. Rappelons que les phrases stimuli entendues que l'apprenant doit répéter sont des deux types, Sujet-Objet, soit Objet-Sujet. L'hypothèse sous-jacente à cette tâche est que l'apprenant ne peut pas imiter correctement une phrase s'il ne dispose pas d'une représentation de celle-ci. La précision de la répétition reflète sa compétence grammaticale.

Ainsi, pour les tâches de production comme pour les tâches de traitement, il s'agit d'identifier les différences possibles entre le jugement/la production de deux modalités possibles par rapport à une même variable. Cette identification nous permet d'établir des hypothèses sur la stratégie que l'apprenant applique pour répondre à la tâche.

Dans les tâches de traitement, le fait d'identifier correctement une des modalités d'expression de la variable sur les deux suffisaient à réussir le test. Par exemple, dans la tâche GJ, le fait de reconnaître les marques de

l'instrumental suffit pour réussir le test, sans pour autant reconnaître les marques du nominatif. Dans le cas des tâches de production ce constat ne s'applique pas.

Il convient donc de considérer les résultats des apprenants pour chacune des modalités d'une même tâche séparément. Par exemple, pour la tâche OQA, deux scores seront comptabilisés pour un même apprenant, sa production correcte et appropriée dans le contexte impliquant l'emploi de l'instrumental, et sa production correcte et appropriée dans le contexte requérant la forme du nominatif. Ce n'est qu'ainsi que l'on s'assure de la différenciation des apprenants en fonction de leur niveau acquisitionnel, entre les trois états possibles : production des deux modalités, d'une seule ou d'aucune. La production des deux marquages casuels sollicités dans la tâche OQA permet d'affirmer la bonne compréhension de l'apprenant et de sa capacité de mobilisation de ses connaissances lors d'une activité de production. La production de manière correcte d'un seul des marquages casuels dans un contexte approprié peut signifier soit la non-maîtrise de l'autre marquage casuel, soit la surgénéralisation du premier sur le deuxième, etc. Tandis que l'incapacité à produire l'un ou l'autre de manière appropriée ne signifie pas nécessairement l'absence des connaissances nécessaires sur le fonctionnement du système casuel de la LC mais peut dénoter d'une difficulté de l'apprenant à mobiliser ces connaissances lors d'une activité de production. On peut supposer que l'apprenant a intégré une règle si le marquage casuel est productif, c'est-à-dire s'il emploie des désinences différentes en fonction du contexte et que cet emploi est approprié, mais l'inverse n'est pas vrai. Autrement dit, lors de l'analyse des résultats d'un apprenant à une tâche de production, on ne juge pas que la compréhension de l'apprenant mais également sa capacité de restitution orale.

La mise en relation des résultats des apprenants obtenus dans les deux tâches GJ et OQA, ainsi que dans les tâches PV et SI, s'impose dans la mesure où les deux couples de tâches examinent le traitement et la production d'un même paradigme linguistique. Il s'agit respectivement de l'opposition instrumental *vs.* nominative dans le cas de GJ et OQA, et du marquage du statut du constituant de Sujet *vs.* Objet par le marquage casuel nominatif *vs.* accusatif dans le cas de PV et SI. La mise en parallèle d'une tâche de traitement et d'une tâche de production axées sur le même phénomène en LC nous permet d'établir un nouvel attribut pertinent pour la caractérisation du parcours acquisitionnel de l'apprenant. Les capacités à mobiliser les mêmes connaissances dans une activité de traitement *vs.* une activité de production ne sont pas développées de manière simultanée lors de l'apprentissage d'une seconde langue. Un laps de temps est souvent nécessaire à l'apprenant afin de transférer et mobiliser ses connaissances grammaticales, attestées en traitement, pour la production, et ce plus particulièrement lorsque la production est une production libre. En effet le but communicatif de telle tâche semble prendre le pas sur le respect de la forme dans les énoncés des apprenants (cf. chapitre 4). Ainsi, la création d'un attribut "transfert" doit être effectué afin de rendre compte des possibles variations dans le parcours acquisitionnel des apprenants.

Lorsque les apprenants obtiennent des scores élevés en GJ, on admet qu'ils sont sensibles au système casuel de la LC. S'ils arrivent également à utiliser la morphologie nominale de façon productive dans la tâche OQA, ils auront très probablement réussi le transfert du traitement vers la production. A contrario, si ces apprenants ne réussissent pas à produire les désinences casuelles appropriées malgré un score élevé en GJ, cela suggère qu'ils n'ont pas encore effectué le transfert des connaissances traitées en production. Par opposition, les apprenants se heurtant encore à la prise de conscience de l'existence d'un système casuel en polonais n'ont que peu de chance de produire une désinence appropriée au contexte. Il en va de même si l'on considère le système morphosyntaxique de la LC en rapprochant les résultats des apprenants aux tâches PV et SI. Cependant, le faible nombre d'items lexicaux utilisés dans la tâche SI ainsi que la nature même de cette tâche, de répétition plutôt que de production au sens strict, pourra hypothétiquement faciliter le transfert des connaissances de l'apprenant en traitement vers leurs productions.

La structure particulière de chaque test utilisé dans le projet VILLA permet de distinguer aisément à partir d'une valeur numérique l'avancée dans l'appropriation d'un paradigme par l'apprenant. Étant donné que la construction d'un test repose ici sur l'évaluation d'un paradigme linguistique ciblé (cf. chapitre 2), le positionnement d'un individu sur le continuum des valeurs possibles des résultats à ce test est un indicateur

quasi exclusivement de l'appropriation de ce paradigme par l'individu. Dans les données considérées, de traitement comme de production, les variables intervenant dans le phénomène de réponse de l'apprenant sont extrêmement réduites. Des facteurs extérieurs vont naturellement venir perturber cet objectif premier mais dans une moindre mesure comparée à des productions orales spontanées et libres que nous pourrions recueillir. Cet état de fait est une condition préalable de l'application d'une telle analyse et d'un tel découpage sur l'ensemble des scores obtenus.

Une illustration ici concerne la tâche PV. La construction initiale du test permettait de tester trois modèles de constructions de phrases : Sujet Verbe Objet ; Objet Verbe Sujet ; Objet Sujet Verbe. Le test contient donc 1/3 de phrases pour chaque modèle. De cette tâche ont été fusionnés les scores obtenus pour les phrases OVS et OSV nous permettant de nous intéresser exclusivement à la dichotomie SO *vs.* OS. Malgré tout, l'absence première d'une simple étude de l'opposition entre deux modalités d'une même variable d'intérêt empêche une analyse directe de la moyenne globale obtenue par un apprenant à cette tâche. Une moyenne pondérée doit ainsi être utilisée afin de s'assurer que les scores pour les phrases SO et pour les phrases OS représentent chacun la moitié de la moyenne globale utilisée par la suite, à des fins de comparaisons. Cette précaution permet de retrouver la structure utilisée dans les autres tests du projet VILLA, pris en considération dans les présents travaux, et d'illustrer les contraintes pré existantes à l'application de l'analyse effectuée dans ce chapitre. L'utilisation de la moyenne pondérée comme base dans notre analyse des résultats obtenus pour la tâche PV permet ainsi d'interpréter les scores obtenus autour de la valeur 0.5 comme représentatifs d'une stratégie de réponse ne considérant pas l'absence d'un ordre strict des mots dans la phrase.

Pour les tâches sollicitant un traitement, nous pouvons dès lors assumer qu'un score avoisinant les 50% de bonnes réponses est révélateur d'un comportement de réponses aléatoires, de l'absence de la compréhension du paradigme testé, ou d'une sensibilité relativement faible aux caractéristiques des items employés. Pour les tâches sollicitant une production, il est nécessaire de considérer un score pour chaque modalité de la variable testée. Ce score permettra d'attester de la capacité de l'apprenant à produire un marquage casuel correctement, et ce en contexte approprié. Ce n'est que par la suite qu'il sera mis en parallèle avec son équivalent pour la seconde modalité testée dans la tâche dont ils sont tous les deux issus. Cette mise en parallèle (opérationnalisée comme nous allons le voir dans la prochaine sous-section par une procédure de défuzzification) permettra d'identifier le niveau acquisitionnel de l'apprenant sur le paradigme linguistique considéré.

5.1.2 Transformations en sous-ensembles flous

Un score numérique laisse donc transparaître des différences dans les stratégies de réponses aux différents tests, ainsi que différentes sensibilités aux caractéristiques de l'input. Nous avons pu émettre des hypothèses quant à la signification d'un score en termes de stratégie de réponse à une tâche, ainsi qu'en termes de sensibilité aux caractéristiques des items de l'input. Les valeurs autour de 0.5 se distinguent des autres sur l'ensemble d'appartenance des scores aux différentes tâches de traitement. Un score aux alentours de la valeur médiane représente, pour les différentes tâches, soit un comportement aléatoire ; soit marque l'absence d'un début d'appropriation du système morphologique et/ou morphosyntaxique de la LC. Ces hypothèses sont essentielles pour le processus de partitionnement car elles nous permettent de caractériser un apprenant non pas sur un score numérique, mais en terme de stratégie de réponse à une tâche. Elles serviront de base à l'interprétation de la classification des apprenants obtenue par l'algorithme. Ensuite, lors de l'étude de la différence entre utilisation appropriée des items fréquents (respectivement transparents) et utilisation appropriée des items non fréquents (respectivement opaques), il existe une valeur intermédiaire dans l'ensemble d'appartenance de cette différence signifiant un impact faible voir nul de cette caractéristique sur les résultats des apprenants. Ce constat est également valable pour l'attribut capturant l'évolution entre deux périodes de tests, tout au long de l'enseignement en classe de langue.

Les tâches de production comme les tâches de traitement ont pour objectif l'identification de différences possibles entre le jugement/la production de deux modalités d'un même stimulus. Seulement, le champ des erreurs possibles lors d'une production orale est infiniment supérieur à celui engendré par une activité de traitement. Considérer que le score global d'un apprenant à une tâche de production est suffisant pour nous renseigner sur sa maîtrise potentielle de l'une ou l'autre des modalités présentes dans la tâche est potentiellement problématique. Par exemple, pour la tâche OQA, nous divisons le score global en deux : celui du taux de réponse correcte pour les phrases à l'instrumental, et celui pour les phrases au nominatif. Prendre en compte les résultats des productions de l'apprenant pour chacun des paradigmes testés de manière séparée nous permet de nous assurer de la finesse de nos analyses.

Nous souhaitons dès lors mettre en place un moyen de capturer ces différences significatives de score, à l'aide d'un outil approprié. Un premier objectif est de capturer les scores gravitant autour de la valeur intermédiaire 0.5, significatifs d'un comportement différent, ou d'une sensibilité faible aux propriétés de l'input. Le deuxième objectif est d'être capable d'agréger le résultat de la transformation opérée sur les scores afin de pouvoir unifier deux résultats caractérisant le comportement d'un apprenant dans une même tâche lorsqu'il s'agit d'une tâche de production.

Une valeur numérique représentative d'un comportement humain est par nature imprécise. Aussi la frontière entre deux comportements sur l'axe de l'ensemble des valeurs numériques possibles n'est pas clairement définie. Pourtant, afin d'isoler un comportement spécifique, il faut délimiter une plage de valeurs numériques représentatives d'un tel comportement. Un découpage de l'axe de l'ensemble d'appartenance des valeurs de la base de données en sous-ensembles permet la capture des comportements différents tels que définis dans la section précédente. On opère ainsi une transformation des valeurs numériques de la base de données en valeurs d'appartenance d'un apprenant à un type de stratégie de réponse, et à un niveau de sensibilité aux propriétés de l'input. Pour marquer l'imprécision et nous permettre de manipuler nos ensembles, d'effectuer des opérations de comparaisons entre eux, et d'agréger, tout en gardant l'information fournie par cette imprécision, la frontière entre les différents sous-ensembles doit être progressive. La théorie des sous-ensembles flous introduite par Zadeh (cf. chapitre 3) offre la possibilité d'une caractérisation imprécise et multiple d'appartenance à un ensemble, tout en fournissant les opérateurs essentiels pour la manipulation algébrique de tels ensembles (cf. chapitre 3).

Dans le cas d'une valeur numérique correspondant à une tâche de traitement, il s'agit donc de définir un sous-ensemble flou s'étalant autour de la valeur intermédiaire 0.5 obtenue comme score de réussite. Pour la spécification des niveaux de sensibilité aux caractéristiques des items, les valeurs numériques à regrouper en sous-ensembles correspondent à une différence entre deux valeurs appartenant à l'ensemble $[0; 1]$ et relèvent donc de l'ensemble $[-1; 1]$. Par commodité, l'ensemble d'appartenance du produit de cette différence est redéfini sur l'ensemble $[0; 1]$. Le découpage en sous-ensembles flous de ces valeurs est donc également déterminé par la valeur intermédiaire de 0.5 et ses "environs", que nous souhaitons discriminer du reste de la plage de valeurs, les extrêmes, bornés respectivement par 0 à gauche, de 0.5 et ses environs et 1 à droite. Nous faisons ainsi face à un découpage ternaire.

Sur l'ensemble X de définition de nos valeurs, on identifie trois sous-ensembles flous : S_1 , S_2 et S_3 . Nous choisissons des SEFs trapézoïdaux, afin de rappeler ici la notion de "seuil de significativité" utilisée en statistique, et symétriques, par rapport à la valeur d'intérêt 0.5 (cf. [Dubois et al., 2004]). On définit ainsi $f_{S_i}(x)$, la fonction d'appartenance associant à chaque élément x de X un degré d'appartenance, compris entre 0 et 1, au sous-ensemble S_i .

$$f_{S_i} : X \rightarrow [0; 1]$$

S_2 est le sous-ensemble flou intermédiaire que l'on souhaite isoler. Il doit inclure les valeurs de X autour de 0.5. Ses caractéristiques sont (pour les définitions des caractéristiques d'un sous-ensemble flou voir chapitre 3) :

- $supp(S_2) = [0.5 - \alpha; 0.5 + \alpha]$
- $h(S_2) = 1$
- $noy(S_2) = [0.5 - \alpha'; 0.5 + \alpha']$

où α , une valeur en abscisse, est choisie telle que $f_{S_2}(x) \neq 0$ ssi $0.5 - \alpha < x < 0.5 + \alpha$ et $\alpha \in [0; 0.5]$, et α' , une valeur en abscisse, est choisie telle que $f_{S_2}(x) = 1$ ssi $0.5 - \alpha' < x < 0.5 + \alpha'$ et $\alpha' \in [0; 0.5]$.

A partir de S_2 , on définit S_1 et S_3 de la manière suivante :

- $supp(S_1) = [0; 0.5 - \alpha']$
- $h(S_1) = 1$
- $noy(S_1) = [0; 0.5 - \alpha]$

et

- $supp(S_3) = [0.5 + \alpha'; 1]$
- $h(S_3) = 1$
- $noy(S_3) = [0.5 + \alpha; 1]$

Le découpage de X est illustré dans la figure 5.1. L'ensemble ainsi créé est appelé $F(X)$.

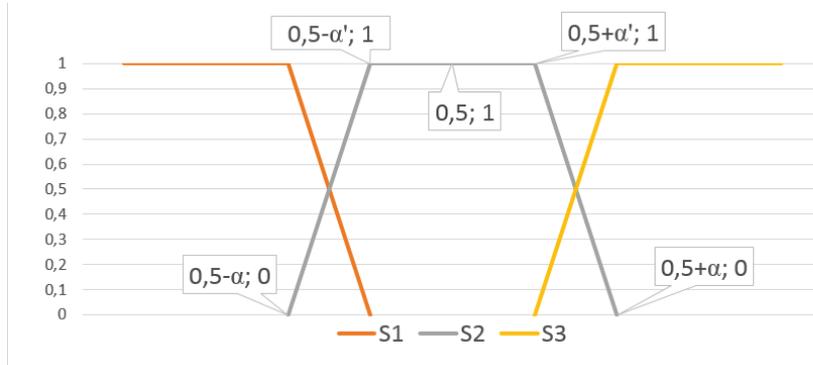


FIGURE 5.1 – Définition des SEFs S_1 , S_2 et $S_3 \in F(X)$ sur l'ensemble de référence X .

Pour les tâches de production, l'approche ternaire de découpage de l'ensemble d'appartenance des scores n'est pas directement identifiable. En effet, un score pour chaque modalité de la variable manipulée dans la tâche est à considérer. Chaque ensemble d'appartenance de cet ensemble de scores (entre 0 et 1) est découpé en deux SEFs. La frontière d'appartenance à l'un ou l'autre de ces deux SEFs se situe autour de la moyenne théorique : 0.5. Un apprenant appartient ainsi à un SEF pour chacune des modalités de la variable testée.

Sur l'ensemble X de définition de nos valeurs on identifie deux sous-ensembles flous : S_1 et S_2 . Nous choisissons des SEFs trapézoïdaux, afin de rappeler ici la notion de "seuil de significativité" utilisée en statistique, et symétriques par rapport à la moyenne théorique 0.5.

La fonction $f_{S_i}(x)$ définie précédemment comme la fonction d'appartenance associant à chaque élément x de X un degré d'appartenance, compris entre 0 et 1, au sous-ensemble S_i , s'applique également.

Les caractéristiques de S_1 et S_2 sont :

- $supp(S_1) = [0; \mu + \epsilon]$
- $h(S_1) = 1$
- $noy(S_1) = [0; \mu - \epsilon]$
- $supp(S_2) = [\mu - \epsilon; 1]$
- $h(S_2) = 1$
- $noy(S_2) = [\mu + \epsilon; 1]$

où μ est la valeur milieu théorique de l'ensemble d'appartenance X et ϵ est une valeur d'erreur possible entre la valeur milieu théorique et la valeur milieu effective. Dans notre cas nous choisissons $\epsilon = 0.5$.

L'ensemble X ainsi découpé est illustré en figure 5.2.

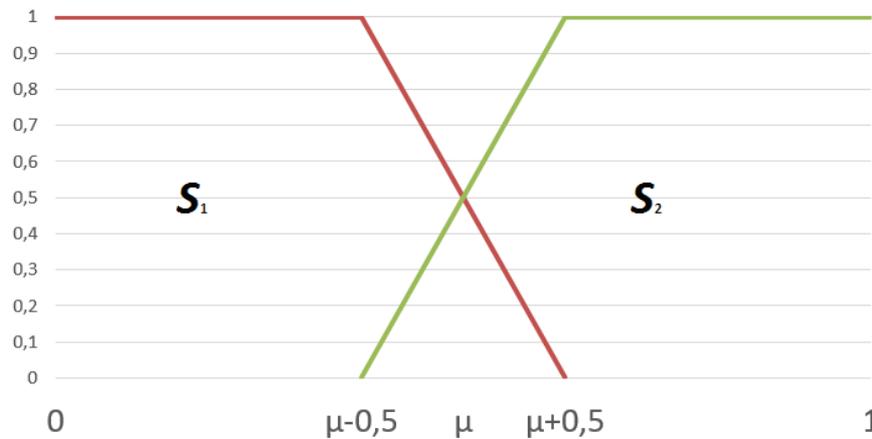


FIGURE 5.2 – Définition des SEFs S_1 et $S_2 \in F(X)$ sur l'ensemble de référence X avec μ la valeur milieu théorique et $\epsilon = 0.5$

Pour une même tâche, il existe autant d'ensembles de référence X_i que de modalités de la variable d'intérêt. Ainsi, le résultat d'un apprenant à la tâche testant la variable V à v modalités est pour l'instant décrit sur v ensembles $F(X_i)$ de SEFs, i allant de 1 à v . Nous souhaitons décrire le résultat de cet apprenant sur un seul ensemble de SEFs. Pour cela nous nous dotons du produit cartésien de SEFs décrit en définition 7.

Définition 7. Soit A et B deux SEFs :

Produit cartésien : $f_{A \times B} = \min(f_A(x), f_B(y))$

Et nous l'étendons pour le calcul du produit cartésien de deux ensembles de SEFs, $F(X_1)$ et $F(X_2)$, décrit en définition 8.

Définition 8. Soit $F(X_1)$ et $F(X_2)$ deux ensembles de SEFs :

Produit cartésien : $f_{F(X_1) \times F(X_2)} = \max(\prod_{i=1}^v \prod_{j=1}^{v'} f_{S_i \times S_j})$

où $S_i \in F(X_1)$ et $S_j \in F(X_2)$

Dans le cadre de ce travail, et dans le cadre du projet VILLA, $v = 2$. L'objectif est de repérer les apprenants dont le SEF d'appartenance n'est pas le même si l'on considère l'une ou l'autre des modalités de la variable manipulée, par exemple l'instrumental ou le nominatif séparément. Ainsi on catégorise un apprenant selon trois états possibles : maîtrise des deux modalités, d'une seule, d'aucune. La valeur d'appartenance la plus élevée obtenue lors du calcul du produit cartésien des ensembles de SEFs détermine à laquelle de ces catégories l'apprenant appartient.

Par exemple, dans la tâche *Sentence Imitation*, un apprenant obtient un score pour les phrases de type SO et un score pour les phrases de type OS. Ses scores x et y définis sur X_1 et X_2 sont transformés pour être décrit sur $F(X_1)$ et $F(X_2)$. Pour chacun de ces ensembles de SEFs, l'apprenant possède deux valeurs d'appartenance : $f_{S_1}(x), f_{S_2}(x) \in F(X_1)$ et $f_{S_1}(y), f_{S_2}(y) \in F(X_2)$. Si :

$$\begin{aligned} f_{S_1}(x) &= 0.8 \\ f_{S_2}(x) &= 0.2 \\ f_{S_1}(y) &= 0 \\ f_{S_2}(y) &= 1 \end{aligned}$$

alors cet apprenant aura pour valeur d'appartenance maximale finale :

$$\begin{aligned} &\max[\min(f_{S_1}(x), f_{S_1}(y)), \min(f_{S_1}(x), f_{S_2}(y)), \min(f_{S_2}(x), f_{S_1}(y)), \min(f_{S_2}(x), f_{S_2}(y))] \\ &= \max[0, 0.8, 0, 0.2] \\ &= 0.8 \\ &= f_{S_1 \times S_2}(x, y) \end{aligned}$$

L'apprenant exemple appartient donc au SEF issu du produit des SEFs $S_1 \in F(X_1)$ et $S_2 \in F(X_2)$. Cela reflète le fait que l'apprenant produit correctement les désinences de l'instrumental en contexte approprié mais pas celle du nominatif. Il maîtrise donc une modalité sur les deux de la variable manipulée dans la tâche. Ce constat aurait été aussi valable si l'apprenant avait appartenu au SEF issu du produit cartésien des SEFs $S_2 \in F(X_1)$ et $S_1 \in F(X_2)$. Ainsi ces deux combinaisons de SEFs reflètent le même pattern dans le lecte de l'apprenant. Les deux autres patterns identifiables sont celui où l'apprenant ne maîtrise aucune des deux modalités, combinaison des SEFs $S_1 \in F(X_1)$ et $S_1 \in F(X_2)$, et celui où l'apprenant maîtrise les deux modalités, combinaison des SEFs $S_2 \in F(X_1)$ et $S_2 \in F(X_2)$.

La fuzzification de l'ensemble d'appartenance de nos données autorise ainsi la discrimination entre plusieurs profils de réponses à un type de tâche. On passe d'une mesure quantitative brute à une mesure floue plus propice à l'analyse de données issues des comportements humains d'acquisition d'une LE.

Par l'utilisation de la théorie des sous-ensembles flous, le processus de regroupement des apprenants s'effectue dès lors par comparaison de leur stratégie respective lors de l'achèvement d'une tâche, ainsi que par leur sensibilité variée aux propriétés de l'input. On s'affranchit de l'utilisation des valeurs numériques directes, les scores, pour privilégier la caractérisation des apprenants en suivant l'approche acquisitionniste. Un autre cas de figure envisageable dans le contexte de l'enseignement des langues serait l'évaluation des performances des apprenants non pas par l'intermédiaire d'une valeur numérique, un score, mais par le jugement direct de l'enseignant ; un jugement exprimé linguistiquement et non pas à travers une note. Pour l'évaluation d'un apprenant, un enseignant est amené à proférer un jugement sur son travail, selon différents critères, à différents moments, jugements facilement exprimables verbalement.

5.1.3 Manipulations d'expressions linguistiques de jugements de performances

Étant donné un objectif de comparaison des élèves à des fins de regroupements/différenciation, comment manipuler des données de performances de ces apprenants si celles-ci sont exprimées linguistiquement par l'évaluateur ? La manipulation de données linguistiques n'est pas un nouveau domaine (cf. chapitre 3). Les chercheurs qui y sont affiliés s'intéressent aux modèles mathématiques par lesquels transformer une expression linguistique en expression abstraite/symbolique aux propriétés algébriques. Pour notre problématique, il apparaît utile de s'intéresser aux différentes formes sous lesquelles les performances des apprenants peuvent être exprimées, puisque qu'elles constituent la base d'une possible comparaison et donc d'un possible partitionnement.

Un professeur de langue évalue les performances d'un élève donc sur plusieurs critères, selon différentes échelles de valeurs, et avec plus ou moins de certitude. Son jugement de la compétence d'un élève peut donc être hésitant. L'hésitation linguistique porte ainsi non pas sur un continuum numérique mais sur le choix entre plusieurs qualificatifs possibles du travail rendu. Ces expressions linguistiques sont traduites dans un certain nombre de travaux comme des 2-tuples linguistiques flous (cf. [Herrera et Martínez, 2000]), permettant ainsi la manipulation, l'agrégation et la comparaison de ces expressions. Dans cette thèse, nous sommes amenés à manipuler des scores numériques compris entre 0 et 1. Malgré tout, il convient de se poser la question de la compatibilité de nos travaux, et notamment de la compatibilité de l'algorithme présenté en section suivante, avec des données d'autres formats. Dans Durand & Truck ([Durand et Truck, 2018]) nous effectuons une proposition de caractérisation d'expressions linguistiques de jugement d'une performance (à un test quelconque), de leur transformation en tuples de valeurs et enfin de leur agrégation par un opérateur adéquat, en l'occurrence la médiane pondérée symbolique ([Truck et Akdag, 2006, Abchir, 2013]). Nous proposons de décrire ici le processus pas à pas.

Une phrase indiquant un jugement peut être analysée et décomposée en différents items remplissant chacun une fonction spécifique. Cette phrase se construit autour d'un item central qualificatif, un adjectif, décrivant l'objet de l'évaluation. Un ensemble de termes linguistiques flous d'hésitation ou HFLTS (*Hesitant Fuzzy Linguistic Term Set*, "ensemble de terme linguistique flou hésitant") ([Rodriguez et al., 2012]), noté H_s , est défini comme un ensemble ordonné et consécutif de qualificatifs. Il peut être caractérisé par sa borne inférieure et supérieure et son niveau de granularité, c'est-à-dire le nombre de qualificatifs intermédiaires le composant. Par exemple un HFLTS de jugement pourra être $= \{t_1, t_2, t_3\}$ où

$$t_i = \begin{cases} \text{"mauvais"} & \text{pour } i = 1 & \text{la borne inférieure} \\ \text{"moyen"} & \text{pour } i = 2 \\ \text{"bon"} & \text{pour } i = 3 & \text{la borne supérieure} \end{cases}$$

Sa granularité est de 3.

Beaucoup d'opérations basiques sur les ensembles ont également été définies pour les HFLTS comme l'intersection, l'union entre deux HFLTS et le complément. Un HFLTS est donc l'ensemble des qualificatifs de référence sur lesquels va se fonder l'expression linguistique du jugement d'un travail. Ces valeurs de références exprimant une graduation sont appelées scalaires. Sapir ([Sapir, 1944]) introduit la notion d'échelle et d'expression de la graduation. Un scalaire représente un terme linguistique sur cette échelle.

Parler d'hésitation revient donc dans ce schéma à hésiter entre plusieurs scalaires d'une même échelle de référence. Un scalaire peut être accompagné d'un adverbe de degré, spécifiquement pour exprimer un doute, une incertitude. Cet adverbe de degré est toujours utilisé en référence à un scalaire, toujours en relation avec une échelle de graduation. Ces adverbes peuvent être divisés en deux familles selon Quirk et collaborateurs ([Quirk, 2010]) : ceux exprimant une mesure (« plus que ») ; et ceux portant une intensité, eux-mêmes divisés en deux catégories, les amplificateurs et les affaiblissants. Les adverbes d'intensité traduisent une modification

de la portée du scalaire qu'ils accompagnent, tandis que ceux de mesures qualifient le scalaire utilisé pour exprimer le doute. Nous nous référons aux adverbes d'intensités comme à des modificateurs (*modifiers*) et aux adverbes de mesures comme à des qualificateurs (*qualifiers*). Ces derniers accompagnent un scalaire (*unary*) ou deux (*binary*). Une expression d'un jugement hésitant pourra être composée à partir de ces éléments en suivant une grammaire non contextuelle permettant de les combiner de différentes manières. A partir d'un ensemble de qualificateur et modificateur, d'un HFLTS \mathcal{L}_M composé de M qualificatifs ordonnés et consécutifs et d'un ensemble de règles de combinaisons R , on obtient un ensemble d'expressions hésitantes de jugement. Cet ensemble regroupe toutes les expressions linguistiques que l'on pourra transformer à des fins de manipulation. La définition 9 donne un exemple d'une grammaire non contextuelle pour l'expression de jugement hésitant.

Définition 9. Soit G_D une grammaire non contextuelle pour l'expression du doute, et $\mathcal{L}_M = \{\tau_0, \dots, \tau_i, \dots, \tau_{M-1}\}$ un ensemble de M éléments ordonnés consécutifs. G_D est défini par un 4-tuples : $G_D = (V, \Sigma, R, S)$ dont la syntaxe est décrite en utilisant la norme étendue de Backus-Naur [Scowen et Grove, 1993] :

$$V = \{\langle \text{valeur de référence} \rangle, \langle S \rangle, \langle \text{relation binaire} \rangle, \langle \text{relation unaire} \rangle, \langle \text{modificateur} \rangle\}$$

$$\Sigma = \{\text{au plus, moins que, au moins, plus que, tout sauf, entre ... et ... , ... ou ... , tendre vers, plutôt, un petit peu, vraiment, } \tau_0, \dots, \tau_i, \dots, \tau_{M-1} \}$$

$$\begin{aligned} R = & \{ S ::= \langle \text{reference value} \rangle | \langle S \rangle \langle \text{reference value} \rangle \\ & S ::= \langle \text{unary relation} \rangle | \langle \text{binary relation} \rangle \langle \text{reference value} \rangle | \langle \text{modifier} \rangle \\ & | \langle \text{modifier} \rangle \langle \text{unary relation} \rangle | \langle \text{modifier} \rangle \langle \text{binary relation} \rangle \langle \text{reference value} \rangle \\ & \langle \text{reference value} \rangle ::= \tau_0 | \dots | \tau_i | \dots | \tau_{M-1} \\ & \langle \text{binary relation} \rangle ::= \text{between ... and} | \dots \text{ ou } \dots \\ & \langle \text{unary relation} \rangle ::= \text{au plus} | \text{moins que} | \text{au moins} | \text{plus que} | \text{tout sauf} \\ & \langle \text{modifier} \rangle ::= \langle \text{weakening} \rangle | \langle \text{reinforcing} \rangle \\ & \langle \text{reinforcing} \rangle ::= \text{vraiment} \\ & \langle \text{weakening} \rangle ::= \text{tendre vers} | \text{plutôt} | \text{un petit peu} \} \end{aligned}$$

La transformation de l'expression linguistique e en valeur numérique s'effectue par l'intermédiaire d'une fonction I permettant l'association de poids w_i aux différents éléments τ_i de \mathcal{L}_m .

Définition 10. Soit $e_{\tau_\alpha}^A$ (respectivement $e_{\tau_\alpha, \tau_\beta}^A$), $\alpha, \beta \in \{0, \dots, M-1\}$, une expression linguistique unaire (respectivement binaire) sur \mathcal{L}_M .

Soit P l'ensemble de modificateurs : $P = \{\text{vraiment, plutôt, a little bit, tend to lean toward}\}$.

Soit Q l'ensemble de qualificateurs : $Q = \{\text{more than, less than, between...and, or, everything except}\}$.

La fonction I attribue des poids pour chaque τ_i , $i = \{0, \dots, M-1\}$:

$$\begin{aligned} I : P \times Q \times \mathcal{L}_M & \rightarrow \mathcal{L}_M \\ e_{\tau_\alpha}^A \text{ (resp. } e_{\tau_\alpha, \tau_\beta}^A) & \mapsto \langle \tau_0^{w_0}, \tau_1^{w_1}, \dots, \tau_{M-1}^{w_{M-1}} \rangle \end{aligned}$$

La répartition des poids dans l'ensemble des qualificatifs s'effectue en accord avec le paramétrage des qualificateurs et des modificateurs. Ce paramétrage dépend des besoins de l'utilisateur. Par exemple, étant donné $L_M = \{\text{néant, faible, moyen, bien, parfait}\}$, et étant donné l'expression hésitante de l'évaluation d'une performance « C'est au moins moyen », on a :

- $\alpha = 3$; $\tau_3 = \text{"moyen"}$
- $A = \{\tau_3, \tau_4, \tau_5\}$, la zone d'influence associée au qualificateur unaire « au moins ».

La répartition des poids se fait sur la zone d'influence (*fuzzy envelope*, [Liu et Rodríguez, 2014]) associée à l'adverbe utilisé. C'est l'établissement de cette zone d'influence, en fonction d'une valeur de référence, et potentiellement soumise à transformation si un modificateur est utilisé, qui est à paramétrer selon les besoins de l'utilisateur et adapté au contexte.

Dans notre exemple, le paramétrage de la répartition des poids pour « au moins » est défini comme suit : $w_\alpha = 0.5, w_{M-1} = 0, w_i = f(i)$ pour $\alpha < i < M - 1$

où $f(x)$

est une fonction linéaire décroissante telle que

$$w_i = (M - 1 - i) / (2 * (M - 1 - \alpha)),$$

puis on normalise la répartition des poids telle que $\sum w_i = 1$. On obtient ainsi pour l'expression « c'est au moins moyen » l'ensemble \mathcal{L}_m pondéré = $\{t_1^0, t_2^0, t_3^{0.5}, t_4^{0.5}, t_5^0\}$.

Ainsi, la transformation elle-même a comme résultat un ensemble de termes pondérés et ordonnés sur une même échelle de graduation. La pondération d'un terme représente le degré de vérité avec lequel ce terme incarne l'expression considérée. C'est à dire, dans notre exemple, si l'on se réfère à la logique multivalente de De Glas (1989, [de Glas, 1986]) :

La performance jugée $J \in_{0.5}$ "moyen" \iff "C'est au moins moyen" est moyen" est 0.5-vrai

TABLE 5.1 – logique multivalente de De Glas

A cette étape, la pondération est répartie sur l'ensemble des qualificatifs \mathcal{L}_M . L'objectif suivant concerne la réduction à un seul terme linguistique de cet ensemble pondéré, terme caractérisant l'expression linguistique hésitante qualifiant l'évaluation des performances d'un apprenant sur un critère donné. Il s'agit ici de lever la marque de l'hésitation et de s'assurer de la pertinence du qualificatif final.

La médiane pondérée symbolique (SWM, *symbolic weighted median*) est un opérateur d'agrégation prenant en entrée un ensemble pondéré de termes linguistiques consécutifs et ordonnés \mathcal{L}_M , pour retourner un terme linguistique unique. La SWM est définie comme suit :

Définition 11. [Truck et Akdag, 2006] Soit $\mathcal{L}_M = \{\tau_0, \tau_1, \dots, \tau_{M-1}\}$ une collection de M éléments ordonnés consécutifs. Lorsque les éléments sont pondérés, la collection est notée $\langle \tau_0^{w_0}, \tau_1^{w_1}, \dots, \tau_{M-1}^{w_{M-1}} \rangle \in \mathcal{B}^{\mathcal{L}_M}$ (ensemble de collections) telle que $\sum w_i = 1, i = \{0, \dots, M - 1\}$. La médiane pondérée symbolique \mathcal{M} est définie comme suit :

$$\begin{aligned} \mathcal{M}: \mathcal{B}^{\mathcal{L}_M} &\rightarrow \mathcal{L}_{M'} \\ \langle \tau_0^{w_0}, \tau_1^{w_1}, \dots, \tau_{M-1}^{w_{M-1}} \rangle &\mapsto \mathcal{M}(\langle \tau_0^{w_0}, \tau_1^{w_1}, \dots, \tau_{M-1}^{w_{M-1}} \rangle) \\ &= \tau_j^{w'_j} \text{ tel que : } \left| \sum_{p=0}^{j-1} w'_p - \sum_{p=j+1}^{M'-1} w'_p \right| < \varepsilon \\ &= m(\tau_i^{w_i}, \mathcal{L}_{M-1}) \text{ avec } w_i = 1 \\ &= m(\tau_i, \mathcal{L}_{M-1}) \end{aligned}$$

où $m(\tau_i, \mathcal{L}_{M-1})$ est un opérateur de modification (ou une composition d'opérateurs de modification) appliqué à un élément de la collection initiale \mathcal{L}_M

et où $\sum_{p=0}^{j-1} w'_p$ (respectivement $\sum_{p=j+1}^{M'-1} w'_p$) est la somme \mathcal{S}_1 (respectivement \mathcal{S}_2) des poids précédant (respectivement suivant) l'élément $\tau_j^{w'_j}$.

Cet opérateur recherche le terme linguistique médian $\tau_j^{w'_i}$, c'est-à-dire dont la somme des poids des termes qui lui sont inférieurs et la somme des poids des termes qui lui sont supérieurs sont égales ou presque (différence négligeable ϵ).

Une famille d'opérateurs de ce type a été défini dans Akdag et al. (2001, [Akdag et al., 2001]), puis repris dans Truck & Akdag (2006, [Truck et Akdag, 2006]) et Truck & Abchir (2014, [Truck et Abchir, 2014]), ce dernier article contient également un algorithme permettant le calcul de la SWM). Les modificateurs symboliques généralisés (*generalized symbolic modifiers*, GSMs) ont été proposés et utilisés pour exprimer le résultat de données agrégées sans perte (ou peu de perte) d'informations liées à l'approximation. L'ensemble \mathcal{L}_M étant par définition un ensemble discret, il est en effet possible que τ_i , le terme linguistique médian recherché, ne soit pas directement trouvé. Un GSM procède alors à un changement de granularité de l'ensemble \mathcal{L}_M , par érosion ou dilatation, jusqu'à trouver l'ensemble contenant le terme linguistique cherché, sans ou avec peu ($< \epsilon$) d'approximation. A chaque changement de granularité, les poids sont re-répartis sur le nouvel ensemble de termes linguistiques, jusqu'à ce que l'ensemble contienne un terme linguistique médian. Les processus d'érosion, de dilatation et de conservation, procédés par lesquels s'opèrent le changement de granularité de l'ensemble des qualificatifs \mathcal{L}_M , sont détaillés dans Truck & Abchir (2014, [Truck et Abchir, 2014]), et un résumé des GSMs est présenté en figure 5.3.

Mode nature	Weakening	Reinforcing
Erosion	$\tau_{i'} = \tau_{\max(0, i - \rho)}$ $\mathcal{L}_{M'} = \mathcal{L}_{\max(1, M - \rho)}$ EW(ρ)	$\tau_{i'} = \tau_i$ $\mathcal{L}_{M'} = \mathcal{L}_{\max(i+1, M - \rho)}$ ER(ρ)
		$\tau_{i'} = \tau_{\min(i + \rho, M - \rho - 1)}$ $\mathcal{L}_{M'} = \mathcal{L}_{\max(1, M - \rho)}$ ER'(ρ)
Dilatation	$\tau_{i'} = \tau_i$ $\mathcal{L}_{M'} = \mathcal{L}_{M + \rho}$ DW(ρ)	$\tau_{i'} = \tau_{i + \rho}$ $\mathcal{L}_{M'} = \mathcal{L}_{M + \rho}$ DR(ρ)
	$\tau_{i'} = \tau_{\max(0, i - \rho)}$ $\mathcal{L}_{M'} = \mathcal{L}_{M + \rho}$ DW'(ρ)	
Conservation	$\tau_{i'} = \tau_{\max(0, i - \rho)}$ $\mathcal{L}_{M'} = \mathcal{L}_M$ CW(ρ)	$\tau_{i'} = \tau_{\min(i + \rho, M - 1)}$ $\mathcal{L}_{M'} = \mathcal{L}_M$ CR(ρ)

FIGURE 5.3 – Résumé des GSMs existants, tiré de [Truck et Abchir, 2014]

Ainsi l'on passe d'un ensemble de termes représentant partiellement une expression, ou plus précisément d'une expression appartenant à un certain degré à différents ensembles (expression $x \in_{\tau_1} A_1, \in_{\tau_2} A_2$ etc., voir table 5.1), à un terme unique qualifiant la performance originalement jugée de manière hésitante, et ce sans recourir à des approximations numériques mais bien par des opérateurs qualificatifs.

Ce travail sur le jugement des performances des apprenants constitue pour l'instant une perspective théorique dans la continuité de cette thèse. En effet, de tels jugements n'existent pas dans les données VILLA à notre disposition. L'application pratique de cette méthode sera sans doute l'objet d'un travail futur.

Dans cette thèse, on l'a vu, la base de données ne comporte que des valeurs numériques, la réussite à un test étant traduite par un score entre 0 et 1. Seulement, là non plus, le simple calcul direct sur ces valeurs numériques n'est guère pertinent. N'ayant pas pour objectif la classification ordonnée des apprenants mais la caractérisation et le rapprochement de leurs processus d'apprentissage, il convient d'identifier au delà de la note, la stratégie mise en place par l'apprenant pour répondre à la sollicitation induite par le test. Dans la section suivante, nous essayons d'identifier les stratégies possibles de réponses aux différentes tâches présentes dans le test VILLA et les différents types de sensibilités possibles aux caractéristiques de l'input, et ce, afin de définir la notion de comparabilité.

5.1.4 *Cannot-link* : qu'est-ce que la comparabilité ?

En classification semi supervisée, l'acquisition des contraintes en contexte classique se fait soit par utilisation d'un sous-ensemble des données déjà labellisées, soit par le recours à un oracle (expert) pour certains des objets à partitionner, choisis pour leur haut potentiel informatif sur les critères du partitionnement souhaité. Par l'expertise effectuée sur les données de cette étude, ainsi que par les connaissances théoriques sur les débuts de l'appropriation d'une langue étrangère, ces critères de partitionnement sont déjà connus. Ils reposent sur la modélisation effectuée de la base de données et sur les dimensions identifiées comme pertinentes pour l'évaluation de l'acquisition de la LC (cf. chapitre 4). Chaque niveau linguistique fait l'objet d'une comparaison entre les deux individus en considération. Ils sont également comparés sur leur sensibilité pour chaque catégorie de caractéristiques des items de l'input. Ainsi une contrainte *cannot-link* entre deux individus est le résultat de la comparaison, un à un, des différents attributs les définissant.

La comparaison ne s'effectue pas sur la valeur numérique associée aux attributs mais sur le sous-ensemble flou d'appartenance de cette valeur numérique. Pour que deux individus soient jugés comme comparables et ne soient donc pas reliés par une relation *cannot-link* ils doivent être comparables sur chacun de leurs attributs. Pour ce faire, nous opérationnalisons les réflexions soumises en 5.1.1.

L'étude et l'identification des stratégies potentiellement utilisées par les apprenants en réponse aux différents tests de langue permet la création de l'ensemble C des contraintes sur lequel va reposer le processus de partitionnement des apprenants. Ainsi, le regroupement des apprenants va être conditionné par leurs comportements stratégiques de réponse. Pour les tâches de traitement, les apprenants adoptant une stratégie de réponse aléatoire, ou n'étant pas entrés dans le système de la LC, c'est-à-dire les apprenants liés au sous-ensemble flou intermédiaire S_2 , ne sont pas comparables aux autres. Ces individus sont jugés comme n'étant pas comparables à leurs homologues appartenant aux deux autres sous-ensembles flous, leur acquisition de la LC n'étant pas encore visible dans leur stratégie de réponse. Par opposition, on jugera comparables des individus ayant amorcé une stratégie de réponse traduisant le début d'un changement dans la représentation du fonctionnement de la LC, et ce, que ce changement soit proche du véritable fonctionnement de la LC ou au contraire éloigné. En effet, l'acquisition d'une structure propre à la LC peut être en cours sans que la manifestation de cette compréhension s'effectue de manière grammaticalement correcte ; la mise en relation forme-fonction pouvant constituer un obstacle supplémentaire. Dans les deux cas, la conscience de l'existence d'un système morphologique riche est présente.

De même, relativement aux sensibilités diverses des apprenants aux différentes caractéristiques de l'input, les apprenants n'ayant pas de grande sensibilité à une catégorie particulière d'items, c'est-à-dire dont la sensibilité à l'input n'est pas significative, sont jugés comparables. On jugera également comparables des individus sensibles aux caractéristiques des items de l'input, que cette sensibilité soit au profit ou au détriment de leurs utilisations appropriées par l'apprenant.

Cependant, pour les tâches de production, les apprenants des SEFs $S_1 \times S'_1$ (produit de deux SEFs, cf. section 5.1.2, avec $S_i \in F(X_1)$ et $S'_j \in F(X_2)$) et $S_2 \times S'_2$ (de même) ne sont pas comparables. Contrairement à une activité de traitement, un score faible obtenu à une activité de production peut indiquer l'absence pure et simple de réponse, ou des réponses complètement inadaptées, tandis que dans une activité de traitement, un tel comportement serait caractérisé par un choix aléatoire entre les deux possibilités de réponses, et correspondrait donc à un score gravitant autour de 0.5.

Les apprenants dont la valeur maximale d'appartenance est issue du produit des SEFs $S_1 \times S'_2$ ou $S_2 \times S'_1$ se caractérisent par la maîtrise d'une des modalités sur deux en présence dans le test. Ils doivent donc être jugés comparables par le procédé de classification.

Cependant, ces apprenants doivent également être différenciés des autres du fait de leur aptitude à produire un marquage casuel approprié pour seulement un contexte sur les deux en présence dans les tâches.

Par exemple, dans la tâche OQA, les apprenants appartenant à $S_1 \times S'_2$ se distinguent des autres par une capacité de production de désinences appropriées dans un contexte instrumental bien supérieur à celui du nominatif. Ceux appartenant à $S_2 \times S'_1$ produisent correctement les désinences du nominatif en lieu et place de celui-ci, mais ne produisent pas celle de l'instrumental lorsque sollicité. Ces deux types de résultats semblent *a priori* comparables. Pour les autres apprenants en revanche, soit ils sont tout simplement incapables de produire des désinences appropriées, voire incapables de produire des désinences tout court, soit ils les produisent correctement. Au vu de nos hypothèses sur l'acquisition de la morphologie et son traitement en LC, on peut supposer que les apprenants appartenant à $S_1 \times S'_2$ formulent également des désinences à l'instrumental lorsqu'un contexte nominatif est demandé. Ce phénomène s'appelle la surgénéralisation et touche tous les apprenants dans une moindre mesure. Cependant, ces apprenants semblent surgénéraliser de manière extrême, et nous souhaitons les séparer des autres. Pour autant, nous ne pouvons considérer comme équivalents les apprenants produisant correctement les désinences de ceux ne les produisant pas correctement, quel que soit le marquage casuel requis.

Ainsi, le résumé des SEFs d'appartenance comparables pour les tâches de traitement et les sensibilités à l'*input* est illustré en tableau 5.2 et pour les tâches de production en tableau 5.3.

SEFs d'appartenance comparables	S_1	S_2	S_3
S_1	X		X
S_2		X	
S_3	X		X

TABLE 5.2 – SEFs comparables pour les tâches de traitement et sensibilités aux caractéristiques de l'*input*

SEFs d'appartenance comparables	$S_1 \times S'_1$	$S_1 \times S'_2$	$S_2 \times S'_1$	$S_2 \times S'_2$
$S_1 \times S'_1$	X			
$S_1 \times S'_2$		X	X	
$S_2 \times S'_1$		X	X	
$S_2 \times S'_2$				X

TABLE 5.3 – SEFs d'appartenance comparables pour les tâches de production. $S_i \in F(X_1)$ et $S'_j \in F(X_2)$

On génère ainsi l'ensemble C des contraintes *cannot-link* associé à l'ensemble de nos données.

5.1.5 Transformation des données

Après avoir présenté les différents procédés appliqués, nous les résumons ici pour obtenir une vision d'ensemble du processus de pré-traitement des données et du résultat final.

Nous sommes, à ce stade, en possession de données recueillies selon la méthodologie du projet VILLA comprenant des scores à une batterie de tests (cf. figure 5.4) et des scores en fonction des caractéristiques de l'*input*.

Pour l'instant la présentation de nos données peut être schématisée par la matrice 5.1.

$$\begin{matrix} & test_1 & test_2 & \dots & input_1 & input_2 & \dots \\ \begin{matrix} \text{étudiant}_1 \\ \text{étudiant}_2 \\ \vdots \\ \text{étudiant}_n \end{matrix} & \begin{pmatrix} v_{1,1} & v_{1,2} & \dots & v_{1,j} & \dots & \dots \\ v_{2,1} & \dots & \dots & v_{2,j} & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ v_{n,1} & v_{n,2} & \dots & v_{n,j} & \dots & \dots \end{pmatrix} & \end{matrix} \quad (5.1)$$

Par notre travail de modélisation et de compréhension des données VILLA (cf. chapitre 4), nous savons qu'une tâche relève d'un niveau linguistique particulier. Nous obtenons la matrice de données en 5.2.

$$\begin{array}{c}
 \text{\textit{étudiant}}_1 \\
 \text{\textit{étudiant}}_2 \\
 \vdots \\
 \text{\textit{étudiant}}_n
 \end{array}
 \begin{pmatrix}
 \text{\textit{niveauLinguistique}}_1 & \text{\textit{niveauLinguistique}}_2 & \dots & \text{\textit{input}}_1 & \text{\textit{input}}_2 & \dots \\
 v_{1,1} & v_{1,2} & \dots & v_{1,j} & \dots & \dots \\
 v_{2,1} & \dots & \dots & v_{2,j} & \dots & \dots \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 v_{n,1} & v_{n,2} & \dots & v_{n,j} & \dots & \dots
 \end{pmatrix}
 \quad (5.2)$$

De plus, il est possible d'identifier les tâches sollicitant une production originale de la part de l'apprenant, des tâches où seul un jugement/traitement est demandé. Pour un même niveau d'analyse linguistique, on établit une capacité de transfert de compétence entre les deux types de sollicitation (du traitement vers la production) pour un même niveau d'analyse linguistique. L'attribut transfert peut ainsi prendre deux valeurs nominales (quel que soit le niveau d'analyse linguistique considéré) : transfert effectué ; pas de transfert. Une troisième valeur nominale est ajoutée à l'ensemble d'appartenance de l'attribut transfert, bien qu'improbable et potentiellement problématique pour son interprétation, au cas où l'apprenant fasse preuve d'une plus grande maîtrise d'un paradigme linguistique en production qu'en traitement.

Les attributs caractérisant nos individus à cette étape du traitement des données sont indiqués dans la matrice 5.3.

$$\begin{array}{c}
 \text{\textit{étudiant}}_1 \\
 \text{\textit{étudiant}}_2 \\
 \vdots \\
 \text{\textit{étudiant}}_n
 \end{array}
 \begin{pmatrix}
 \text{\textit{transfert}}_1 & \dots & \text{\textit{niveauLinguistique}}_1 & \dots & \text{\textit{input}}_1 & \dots \\
 v_{1,1} & \dots & \dots & \dots & v_{1,j} & \dots \\
 v_{2,1} & \dots & \dots & \dots & v_{2,j} & \dots \\
 \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
 v_{n,1} & \dots & \dots & \dots & v_{n,j} & \dots
 \end{pmatrix}
 \quad (5.3)$$

En ce qui concerne les caractéristiques des items employés dans les tâches, il existe une valeur numérique de sensibilité d'un individu à chaque modalité de chaque variable caractérisant ces items. Pour la fréquence et la transparence nous avons deux modalités : fréquent *vs.* non fréquent, transparent *vs.* opaque. Nous définissons ainsi une mesure globale de sensibilité à la fréquence d'un item en faisant la différence entre les scores des deux modalités. Pour la transparence nous ne souhaitons pas obtenir un score unique de sensibilité à cette caractéristique d'item seul, car nous souhaitons observer l'effet de la transparence d'un item en fonction de sa fréquence dans l'input. L'idée ici est d'observer si la transparence d'un item a un effet dès les premières expositions à cet item ou si cet effet ne s'exprime que lorsque l'item est suffisamment fréquent. Il convient donc d'appréhender la sensibilité d'un apprenant à la transparence d'un item toujours à travers le prisme de sa fréquence.

Experiments and elicitation tasks	Time of testing before the first day of language contact (0) or after input sessions (1–9) or wrap-up session (10)										
	0	1	2	3	4	5	6	7	8	9	10
Phoneme Discrimination	X			X					X		
Lexical Decision	X	X			X		X		X		
Word Recognition	X					X					X
Grammaticality Judgement (nominal case)				X					X		
Oral question-answer task (nominal case)				X					X		
Sentence Puzzle (word order)						X			X		
Picture Verification (morpho-syntax; argument roles)							X				X
Sentence Imitation (morpho-syntax; argument roles)							X				X
Grammaticality Judgement (subject-verb agreement)											X
Cloze test (personal pronouns)											X
Elicited production: route direction (discourse)											X
Elicited production: film retelling (discourse)											X

FIGURE 5.4 – Organisation et temps des passations des différents tests lors du projet VILLA

En plus des caractéristiques des items présents dans l’input, est également prise en compte la variable temps d’exposition à l’input. Les modalités de cette variable ne sont pas homogénéisées entre les tests. Certains tests ont été soumis à nos apprenants sur trois intervalles de temps différents (donc trois modalités pour la variable temps), d’autres deux ou une seule fois. Pour calculer une valeur représentant l’évolution d’un individu à travers le temps, nous choisissons de scinder le temps d’exposition à l’input en deux parties. Les scores aux différents tests sont séparés entre ceux obtenus durant les 5 premières sessions d’enseignement (7h30 de temps d’exposition à l’input) et ceux durant les 4 dernières sessions (avec un total de 14h, pour plus de détails se référer au chapitre 2). Ce choix se fonde sur des considérations purement pratiques et sont propres à la configuration de la méthodologie de recueil de données de VILLA. Celle-ci ne permet pas de conserver une granularité de 3 dans le découpage de l’intervalle temps du fait de l’hétérogénéité du nombre et des temps de passation des différentes tâches (cf. figure 5.4, où les temps sont exprimés en nombre de séances d’enseignement reçues). Par différence entre les scores obtenus aux deux périodes de tests retenues, nous calculons un score d’évolution d’un apprenant dans son processus d’apprentissage. La matrice 5.4 résume la configuration théorique des attributs des apprenants et correspond à la matrice de données finale après modélisation de notre problématique de la représentation acquisitionniste d’un apprenant.

$$\begin{pmatrix} \text{étudiant}_1 \\ \text{étudiant}_2 \\ \vdots \\ \text{étudiant}_n \end{pmatrix} \begin{pmatrix} \text{transfert}_1 & \dots & \text{niveauLinguistique}_1 & \dots & \text{fréquence} & \text{transparenceItemsNonFréquents} & \text{transparenceItemsFréquents} & \text{évolution} \\ v_{1,1} & \dots & \dots & \dots & v_{1,j} & \dots & \dots & v_{1,m} \\ v_{2,1} & \dots & \dots & \dots & v_{2,j} & \dots & \dots & v_{2,m} \\ \dots & \dots \\ v_{n,1} & \dots & \dots & \dots & v_{n,j} & \dots & \dots & v_{n,m} \end{pmatrix} \quad (5.4)$$

A chacun des attributs ainsi créés est associée une valeur numérique appartenant à l'ensemble $[0; 1]$ pour les différents niveaux linguistiques, à l'exception des capacités de transfert du traitement vers la production de chacun de ces niveaux linguistiques. Pour la sensibilité des individus aux différentes caractéristiques des items de l'input et pour l'évolution effectuée entre les deux périodes d'enseignement, nous obtenons une

nommons des « *cores* » (noyaux) qui constitueront les groupes de référence pour l'analyse en deuxième étape. Dans cette première étape s'effectue donc une comparaison deux à deux des individus sur chacun de leurs attributs. Le résultat de cette étape se place dans une matrice symétrique binaire $n * n$ nommée **matrice de comparabilité** \mathcal{MC} . Chaque 0 indique l'existence d'une contrainte *cannot-link*, et l'ensemble C des contraintes est ainsi généré. Un core est constitué d'individus non reliés par une contrainte *cannot-link*. Si un core de moins de 5 individus existe, nous le supprimons et rajoutons les individus l'ayant composé à la catégorie des outliers, moins de 5 individus ne pouvant constituer un core à raffiner en second lieu. Un outlier est donc soit un individu comparable uniquement avec lui-même, soit un individu comparable avec moins de cinq autres individus. Nous nous attendons à obtenir énormément de outliers au vu des nombreuses contraintes induites par l'analyse des théories linguistiques dans lesquelles nous plaçons notre étude (cf. section 5.1).

L'identification d'une contrainte *cannot-link* se fait par la comparaison d'individus attribut par attribut. Après transformation de nos données de valeurs numériques en appartenance à un ensemble, une simple règle logique permet de statuer la comparabilité de deux individus sur un même attribut, comparaison effectuée par les fonctions `isEqualTrait()` pour les tâches de traitement et `isEqualProd` pour les tâches de production. Ces deux fonctions se présentent sous la forme de pseudo-contrôleur flou en codes 5.1 et 5.2), et implémentent les tables 5.2 et 5.3 présentées en section 5.1.4. Ces deux codes correspondent au pré-traitement des données.

Code 5.1 – isEqualTrait() et fuzzification

```

FUNCTION_isEqualTrait

VAR_INPUT
  vi,z, vj,z : REEL; // RANGE[0;1] individu i et individu j
  α, α' : REEL; // RANGE]0;0.5[
END_VAR

VAR_OUTPUT
  res : BOOLEEN;
END_VAR

FUZZIFY
  vi,z, vj,z on F(X)
  SEF S1 := (0,1) (0.5-α,1) (0.5-α',0) ;
  SEF S2 := (0.5-α,0) (0.5-α',1) (0.5+α',1) (0.5+α,0) ;
  SEF S3 := (0.5+α',0) (0.5+α,1) (1,1) ;
END_FUZZIFY

f(x) : FUNCTION // fonction de calcul
        //des valeurs d'appartenance de x aux SEFs

RULEBLOCK

  res = FALSE;

  RULE 1:
  IF ((max(f(vi,z)) = S1 OR max(f(vi,z)) = S3)
  AND (max(f(vj,z)) = S1 OR max(f(vj,z)) = S3))
  THEN (res IS TRUE);

  RULE 2:
  IF ((max(f(vi,z)) = S2) AND (max(f(vj,z)) = S2))
  THEN (res IS TRUE);

END_RULEBLOCK
END_FUNCTION_isEqualProd

```

Code 5.2 – isEqualProd() et fuzzification

```

FUNCTION_isEqualProd

VAR_INPUT
  vi,z, vi,z' : REEL; // RANGE[0;1] individu i defined on two variables
  vj,z, vj,z' : REEL; // RANGE[0;1] individu j defined on two variables
  β : REEL; // RANGE]0;0.5[
END_VAR

VAR_OUTPUT
  res : BOOLEEN;
END_VAR

FUZZIFY
  vi,z, vj,z on F(X1)
  SEF S1 := (0,1) (0.5-β,1) (0.5+β,0) ;
  SEF S2 := (0.5-β,0) (0.5+β,1) (1,1) ;

  vi,z', vj,z' on F(X2)
  SEF S'1 := (0,1) (0.5-β,1) (0.5+β,0) ;
  SEF S'2 := (0.5-β,0) (0.5+β,1) (1,1) ;

END_FUZZIFY

fF(X1)×F(X2)(x,y) : FONCTION; //produit cartésien
                                //de deux ensembles de SEFs

RULEBLOCK

res = FALSE;

RULE 1:
IF (fF(X1)×F(X2)(vi,z, vi,z') = fS1×S'1(vi,z, vi,z'))
AND (fF(X1)×F(X2)(vj,z, vj,z') = fS1×S'1(vj,z, vj,z'))
THEN (res = TRUE);

RULE 2:
IF ((fF(X1)×F(X2)(vi,z, vi,z') = fS1×S'2(vi,z, vi,z'))
OR (fF(X1)×F(X2)(vi,z, vi,z') = fS2×S'1(vi,z, vi,z')))
AND ((fF(X1)×F(X2)(vj,z, vj,z') = fS1×S'2(vj,z, vj,z'))
OR (fF(X1)×F(X2)(vj,z, vj,z') = fS2×S'1(vj,z, vj,z')))
THEN (res = TRUE);

RULE 3:
IF (fF(X1)×F(X2)(vi,z, vi,z') = fS2×S'2(vi,z, vi,z'))
AND (fF(X1)×F(X2)(vj,z, vj,z') = fS2×S'2(vj,z, vj,z'))
THEN (res = TRUE);

END_RULEBLOCK
END_FUNCTION_isEqualProd

```

La fonction `isComparable()` (cf. algorithme 5.1) fait appel à la fonction `isEqual()` (appelant elle-même `isEqualTrait()` ou `isEqualProd()`) pour les m dimensions du vecteur représentant un apprenant. Elle effectue donc la comparaison globale de deux individus et, dans le cas où la comparabilité n'est pas vérifiée, elle renvoie le nom de la ou des dimensions sur lesquelles portent la contrainte *cannot-link* identifiée. Ces données sont utilisées afin de créer la matrice de comparabilité indispensable pour le partitionnement des étudiants, ainsi que pour l'analyse future de la partition obtenue. L'identification et le stockage des attributs ayant entraîné l'ajout d'une contrainte *cannot-link* sont nécessaires à la compréhension et l'interprétabilité de la classification obtenue par l'utilisateur. La variable *brokenDim* stocke, lors de la comparaison de deux individus, le ou les attribut(s) de non comparabilité entre deux individus. Cette variable constitue une **trace** du déroulement de la classification. Une démonstration des informations que cette trace permet de recueillir est effectuée dans le chapitre 6 sur deux échantillons de notre base de données.

La répartition des individus en cores est effectuée par le biais de la matrice de comparabilité \mathcal{MC} qui regroupe l'ensemble C des contraintes *cannot-link*.

Algorithme 5.1 Fonction `isComparable(a_1, a_2)` : **tableau SEF** `[m]`, `listDim` : *tableauchaine*`[m]` : **[booléen** `res`, **chaîne** `brokenDim`]

```

/*  $a_1, a_2$  correspondent à deux lignes de  $\mathcal{MD}$  la matrice de données, soient deux objets, ou encore deux
apprenants
/* listDim stocke les  $m$  noms des attributs d'un objet, c'est-à-dire le nom des colonnes de  $\mathcal{MD}$ .
brokenDim ← ""
res ← vrai
for  $i := 1 \rightarrow m$  do
    if !(isEqual( $a_1[i], a_2[i]$ )) then
        brokenDim ← brokenDim + listDim( $i$ )
    end if
end for
if brokenDim!="" then
    res ← faux
end if
→ [res, brokenDim]

```

Algorithme 5.2 Étape 1

```

/* EFFET : Création de la matrice de comparabilité  $\mathcal{MC}$  de dimension  $n * n$ , de la matrice de stockage
des attributs ayant entraîné une contrainte cannot-link et répartition des  $n$  individus en  $c$  cores */
/* État initial :  $\mathcal{MD}$  : tableau SEF  $[n][m]$  la matrice de données avec  $n$  le nombre d'objets et  $m$  le
nombre d'attributs ;  $listDim$  : tableau chaîne  $[m]$  stocke les  $m$  noms des attributs d'un objet de  $\mathcal{MD}$ */
/* État final :  $\mathcal{MC}$  : tableau binaire  $[n][n]$  la matrice de comparabilité ;  $MDim$  : tableau  $[n][n][c]$  chaîne
la matrice stockant pour chaque paire d'individus les attributs les rendant non comparables ;  $listCores$  :
tableau entier  $[c][n]$  la liste du partitionnement des individus (représentés par leur indice de ligne dans
 $\mathcal{MD}$ ) en  $c$  cores */

```

```

/* Création de la matrice de comparabilité */

```

```

for  $i := 1 \rightarrow n - 1$  do
  for  $j := i + 1 \rightarrow n$  do
     $[s, brokenDim] \leftarrow isComparable(\mathcal{MD}(i, :), \mathcal{MD}(j, :), listDim)$ 
    if  $s$  then
       $\mathcal{MC}(i, j) \leftarrow 1$ 
       $\mathcal{MC}(j, i) \leftarrow 1$ 
    else
       $\mathcal{MC}(i, j) \leftarrow 0$ 
       $\mathcal{MC}(j, i) \leftarrow 0$ 
    end if
     $MDim(i, j) \leftarrow brokenDim$ 
     $MDim(j, i) \leftarrow brokenDim$ 
  end for
end for

```

```

/* Création des cores */

```

```

 $c \leftarrow 0$ 
while  $find(\mathcal{MD}(:, 1) \neq -1) \neq \emptyset$  do
  /* Tant que la matrice de données n'est pas entièrement grisée (-1), on continue */
   $c \leftarrow c + 1$ 
   $listCores[i]$ 
   $indiceM \leftarrow find(\mathcal{MD}(:, 1) \neq -1)$ 
   $temp \leftarrow findAll(\mathcal{MC}(indiceM, :) == 1)$ 
  /* Stockage des indices dans le core approprié et grisement de la ligne de données correspondante dans
 $\mathcal{MD}$ */
  for  $i := 1 \rightarrow taille(temp)$  do
     $listCores(c, i) \leftarrow temp(1, i)$ 
     $\mathcal{MD}(temp(1, i), :) \leftarrow -1$ 
  end for ;
end while

```

2. Une fois ces cores obtenus nous effectuons une classification non supervisée sur chacun d'entre eux séparément, afin d'obtenir des clusters d'étudiants **similaires** et non plus seulement **comparables**. Pour cela il convient d'abord de trouver le nombre adéquat de clusters présents dans chacun des cores, plusieurs techniques sont mises en place à cette fin : une visuelle tout d'abord, supportée ensuite par l'utilisation d'indices numériques de cohésion intracluster, et d'hétérogénéité intercluster.

Cette étape utilise la technique de clustering très répandue des K-moyennes. Comme nous l'avons vu dans le chapitre 3, cet outil présente le risque d'une partition non optimale du fait qu'elle repose sur une fonction objective à minimiser, qui est non convexe. Il est possible de minimiser ce risque de deux manières

principales. La première repose sur une initialisation judicieuse des centroïdes utilisés comme base pour le partitionnement. La séparation en cores de notre matrice de données initiale assure une homogénéité intra core. De plus, l'absence d'outliers réduit voire annihile le bruit présent dans les données pouvant perturber le processus de partitionnement. L'algorithme des K-moyennes est l'un des algorithmes de partitionnement les plus simples et dans notre cas son utilisation à l'intérieur de chacun des cores obtenus est stable. Cette stabilité a été vérifiée à travers une procédure de validation croisée, qui constitue la deuxième manière d'éviter le risque d'une convergence en un minimum local.

L'autre difficulté principale de l'algorithme des K-moyennes, qu'il partage avec nombre d'algorithmes de clustering, est le choix optimal du nombre de clusters, choix devant être effectué en amont du processus de partitionnement (cf. chapitre 3). Il existe de nombreux indices permettant de calculer le nombre de clusters présent dans un ensemble. Le package `NbClust()`, package issu du langage R, regroupe une quantité importante de ces indices trouvés dans la littérature en statistiques et les calcule simultanément afin de dégager le choix optimal (le nombre de clusters recommandé par une majorité d'indices) [Charrad et al., 2014]. Cette étape repose donc entièrement sur des outils issus de la littérature scientifique et est complètement automatisée. Pour les besoins de notre étude, une analyse est ajoutée, à des fins de validation de l'algorithme. Cette dernière permet d'identifier les cores non homogènes et, ainsi, de sélectionner les candidats à un raffinement de classification. Bezdek et Hathaway ([Bezdek et Hathaway, 2002]) ont développé une méthode pour juger du nombre de clusters contenus dans un ensemble de données de manière visuelle, la méthode VAT (*Visual Assessment of Tendency*). La méthode se fonde sur une représentation noir et blanc de la matrice de similarité où le contraste fait office de distance entre deux objets. Une version considérablement améliorée esthétiquement mais reposant sur le même principe, est proposée par Havens et Bezdek en 2012 (iVAT, *improved visual assessment of tendency*) [Havens et Bezdek, 2012]. C'est cette version que nous utiliserons lors de l'application de l'algorithme à deux échantillons de notre base de données dans le chapitre 6. L'appréciation visuelle de l'homogénéité des cores par l'algorithme nécessite l'intervention humaine et n'est donc pas automatique, elle n'est ainsi pas décrite dans l'algorithme 5.3 présentant l'étape 2.

Algorithme 5.3 Étape 2

```

/* EFFET : Découvre le nombre de clusters présents dans chaque core et si nécessaire opère le
partitionnement du core en clusters. */
/* État initial : listCores : tableau entier [c][n] Matrice des individus dans chacun des c
cores; NbClust(tableau réel [n][n]) : entier, fonction calculant le nombre optimal de clusters
présents dans une matrice de données, avec un nombre n ≥ 5 d'individus par cluster;
KmeansClust(tableau réel [n][n], entier) : tableau entier [x][2], fonction appliquant l'algorithme des
K-moyennes pour un nombre donné de clusters à une matrice de données et renvoyant une matrice contenant
en première colonne les indices lignes des objets dans MD (IDs) et un numéro de cluster correspondant.
*/
/* État final : listClusters : tableau entier[n-o][2] la matrice de partitionnement temporaire regroupant
les IDs des individus déjà classifiés (au nombre de n apprenants moins o outliers) et leur cluster
d'appartenance*/
numCores ← taille(listCores, 1)
numClustersFinal ← 0
for i := 1 → numCores do
  /*Création de ci à partir de MD et listCores*/
  IDs ← listCores(i, :)
  tailleCore ← taille(IDs)
  for j := 1 → tailleCore do
    ci(j, :) ← MD(IDs(1, j), :)
  end for
  ci(:, $ + 1) ← IDs(1, :)
  numCenters ← NbClust(ci) /* Calcul du nombre de clusters présents dans le core ci*/
  if numCenters == 1 then
    for j := 1 → tailleCore do
      listClusters(IDs(1, j), numClustersFinal + 1)
    end for
  else
    resi ← KmeansClust(ci, numCenters)
    clustTemp ← max(resi(:, 2))
    resi(:, 2) ← resi(:, 2) + numClustersFinal
    numClustersFinal ← numClustersFinal + clustTemp
    listClusters ← addBottom(listClusters, resi)
  end if
end for

```

3. La troisième étape consiste en l'intégration des outliers (cf. algorithme 5.5). Celle-ci s'effectue par le relâchement de l'aspect *crisp* (en dur) des contraintes, le remplaçant par un indice de comparabilité, compris entre 0 et 1, s'apparentant ainsi à une version floue du calcul de cet indice. Lors de la comparaison deux à deux d'individus, chaque contrainte violée viendra peser sur cet indice de comparabilité, et non plus automatiquement générer une contrainte *cannot-link* équivalant à un indice égal à 0. Ces indices sont ensuite fusionnés par cluster. Pour chaque outlier sera ainsi calculé un indice de comparabilité globale par cluster de la partition. Cet indice s'apparente alors à un **indice d'appartenance** à un cluster. Nous utilisons les mêmes contraintes que dans la version *crisp* mais au lieu d'une appartenance exacte à un cluster nous autorisons un étudiant outlier à appartenir plus ou moins (par une valeur comprise entre 0 et 1) à un ou plusieurs clusters. Nous obtenons donc pour chaque outlier un vecteur $1 * c$ où c correspond au nombre de clusters trouvés et où o_i correspond au degré d'appartenance de l'outlier o au cluster i . Les outliers sont directement greffés sur les clusters pour lesquels leur indice d'appartenance est le plus élevé. En cas de degré d'appartenance équivalent, ces individus restent des outliers.

Pour un outlier, nous calculons un score d'appartenance à un cluster en calculant un score de comparabilité entre cet outlier et chacun des individus du cluster par l'utilisation d'une version floue de `isComparable()` : `isComparableFuzzy()` (cf. algorithme 5.4). Est ensuite choisi comme cluster d'appartenance le cluster ayant le score de comparabilité le plus élevé. Les outliers dont le score d'appartenance maximum est attribué à deux clusters différents restent indéterminés. Les outliers dont le score d'appartenance le plus élevé est trop faible, en accord avec un seuil choisi, restent des outliers. La matrice de classification finale peut ainsi être construite.

Algorithme 5.4 Fonction `isComparableFuzzy(a_1, a_2 : tableau SEF [m], $listDim$: tableau chaîne [m])` : [réel $score$, chaîne $brokenDim$]

```

/*  $a_1, a_2$  correspondent à deux lignes de  $\mathcal{MD}$ , soient deux objets, ou encore deux apprenants
/*  $listDim$  stocke les  $m$  noms des attributs d'un objet, soit le nom des colonnes de  $\mathcal{MD}$ .
 $brokenDim \leftarrow ""$ 
 $score \leftarrow 0$ 
for  $i := 1 \rightarrow m$  do
  if  $!(isEqual(a_1[i], a_2[i]))$  then
     $score \leftarrow score + 1$ 
     $brokenDim \leftarrow brokenDim + listDim(i)$ 
  end if
end for
 $score \leftarrow fonctionIntegration(score, m)$ 
 $\rightarrow [score, brokenDim]$ 

```

où `fonctionIntegration()` applique une fonction décroissante f tel que

$$f : \mathbb{N} = \{0, 1, 2, 3, \dots, m\} \rightarrow \text{score d'appartenance} \in \mathbb{R} = [0; 1]$$

Algorithme 5.5 Étape 3

```

/* EFFET : Intégration des outliers à la classification. */
/* État initial :  $MOut$  :  $\mathbf{tab}[o][m]$  la matrice regroupant les outliers avec  $o$  le nombre d'outliers et  $m$  le
nombre d'attributs ;  $listClusters$  :  $\mathbf{tab}[n-o][2]$  la matrice de partitionnement temporaire regroupant les IDs
des individus déjà classifiés et leur cluster d'appartenance ;  $MOutDim$  :  $\mathbf{tab}[o][c][\ ]$  la matrice stockant pour
chaque outlier les attributs ayant diminué leur score d'appartenance à chacun des  $c$  clusters */
/* État final :  $MClassification$  :  $\mathbf{tab}[n][m+1]$  la matrice de partitionnement finale avec les  $n$  individus,
leurs  $m$  attributs et leur cluster d'appartenance en colonne  $m + 1$  */
for  $i := 1 \rightarrow o$  do
  for  $j := 1 \rightarrow c$  do
    for  $k := 1 \rightarrow \text{taille}(listClusters[c])$  do
      [ $s, brokenDim$ ]  $\leftarrow isComparableFuzzy(MOut(i,:), MD(listClusters(c,k,:), listDim)$ 
       $indiceTemp \leftarrow indiceTemp + s$ 
       $MOutDim(i, c, k) \leftarrow brokenDim$ 
    end for
     $indiceTemp \leftarrow indiceTemp / \text{taille}(listClusters[c])$ 
     $MAppartenance(i, c) \leftarrow indiceTemp$ 
  end for
   $\max(MAppartenance(i, :))$ 
end for
 $MClassification(:, :) \leftarrow MD(:, :)$ 
for  $i := 1 \rightarrow c$  do
  for  $k := 1 \rightarrow \text{taille}(listClusters[c])$  do
     $MClassification(listClusters(i, k), \$ + 1) \leftarrow i$ 
  end for
end for
 $indiceOut \leftarrow \text{findAll}(MClassification(:, \$) == \emptyset)$ 
for  $i := 1 \rightarrow \text{taille}(indiceOut)$  do
   $MClassification(indiceOut(i), \$) \leftarrow \max(Mappartenance(i, :))$ 
end for

```

Le résultat consiste en une classification des étudiants selon des critères linguistiques et de sensibilités aux caractéristiques de l'input, pour lesquels ils sont jugés comparables puis similaires. Les différents ensembles d'individus présents dans la partition constitue ainsi ce que nous nommons des profils d'apprentissage. L'ensemble de l'algorithme est résumé en figure 5.5

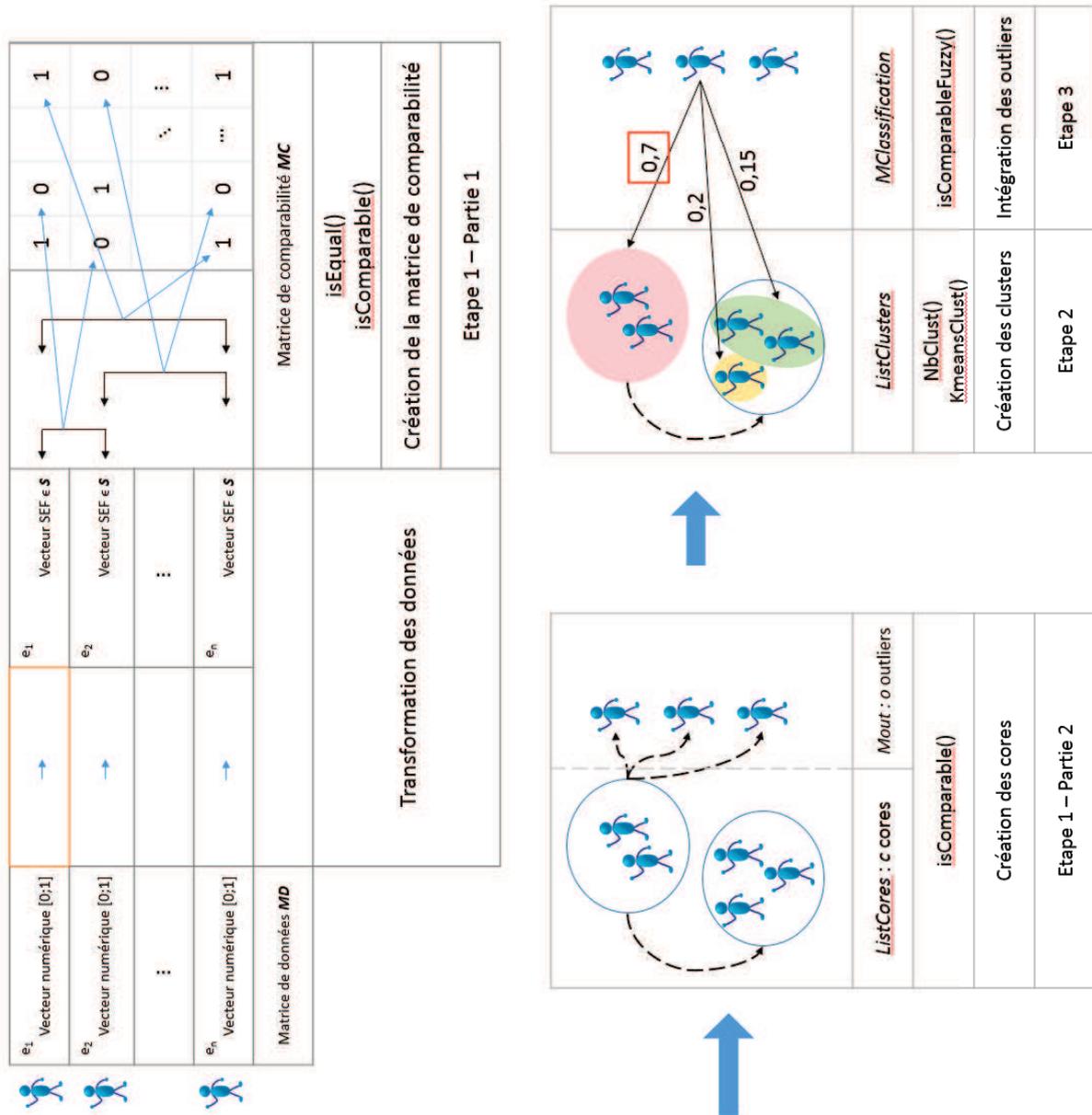


FIGURE 5.5 – Vue d'ensemble des différentes étapes de l'algorithme.

Il convient dès lors de valider la partition obtenue. De par l'objectif initial de ces travaux qui, nous le rappelons, consiste en la découverte de profils d'apprentissage parmi les étudiants considérés, la meilleure validation est l'analyse de la partition obtenue d'un point de vue acquisitionniste. En effet, l'interprétabilité du résultat d'un algorithme de regroupement (c'est-à-dire de la partition des données obtenue) est l'un des enjeux majeurs, sinon l'enjeu unique, de sa validation, comme nous l'avons vu dans le chapitre 3. Cet aspect de la validation constitue l'objectif premier de cette thèse et fera l'objet de sa troisième partie. Malgré tout, la notion d'interprétabilité peut se définir comme la capacité d'un expert du domaine à attribuer du sens à chaque regroupement. Or l'être humain est capable de trouver du sens à quasiment tous les regroupements, suites de nombres, etc. même quand ces ensembles ont été créés aléatoirement. On ne peut se prémunir

de cet état de fait. C'est pourquoi une validation technique est souhaitable. Comme vu dans le chapitre 3, deux types de validation existent : interne et externe. L'interne représente l'interprétabilité et l'utilité de la partition obtenue, que nous verrons en troisième partie ; l'externe fait référence à plusieurs techniques. L'une des principales techniques utilisées dans la littérature consiste à utiliser l'algorithme sur un ensemble de données dont la classification exacte est déjà connue, afin de comparer la partition originale et celle ainsi créée. C'est de l'impossibilité (voire l'inutilité ?) de cette technique de validation du présent algorithme dont nous discutons dans la section suivante. Une autre technique de validation externe concerne la validation croisée (*cross validation*) permettant d'attester de la stabilité de l'algorithme. L'instabilité d'un algorithme de clustering est définie par la distance entre deux partitions C et C' d'un même ensemble de données qu'on aura modifié (perturbé) différemment au préalable. La perturbation d'un ensemble de données peut prendre plusieurs formes. Dans notre cas, nous effectuerons du sous-échantillonnage des données et appliquerons l'algorithme présenté dans cette thèse à ces différents sous-échantillons (cf. le chapitre 6 d'application de l'algorithme).

5.3 Validation de l'algorithme

Nous souhaitons donc valider notre procédé de regroupement par des techniques issues du domaine de la classification non supervisée. Étant donné que notre algorithme est spécifique au contexte, les méthodes à utiliser ne sont pas triviales. Néanmoins nous proposons dans cette sous-partie de réfléchir à ce que nous obtiendrions si nous l'appliquions à des jeux de données publics, classiquement utilisés dans le domaine. Il convient alors de choisir des données de même nature que les nôtres.

L'application d'une transformation de nos données, préalablement au passage de l'algorithme de regroupement, rend ces dernières non plus représentées de manière numérique avec une valeur comprise entre 0 et 1, mais comme une appartenance à un sous-ensemble flou parmi trois préalablement définis. Ces trois sous-ensembles flous viennent partitionner la représentation des données sur un axe $[0; 1]$.

Cette transformation a pour but la création de contraintes *cannot-link* entre individus. Dans ce travail, les contraintes *cannot-link* entre individus sont le résultat de contraintes *cannot-link* sur les attributs de ces individus. Pour un attribut donné, si les valeurs de deux individus ne sont pas comparables, ces individus sont marqués d'une contrainte *cannot-link*. Nous sommes ainsi capables de générer l'ensemble des contraintes *cannot-link* existant dans notre jeu de données.

Pour un jeu de données extérieur, nous n'avons pas accès à cette expertise. Nous ne pouvons donc définir des règles de comparabilités d'individus en nous appuyant sur les valeurs de leurs attributs. Il nous faudra donc utiliser une mesure de distance classique. Plusieurs mesures de distance ont déjà été référencées lors de l'état de l'art (cf. chapitre 3).

La génération de contraintes *cannot-link* entre individus fondée sur une mesure de distance classique ne fait pas sens, la génération de contraintes étant un préalable à la comparaison des individus. Tout du moins l'ensemble des contraintes, s'il n'est pas à la base de la création d'une mesure de distance appropriée, constitue un substitut à cette dernière.

Néanmoins dans ces jeux de données publics, nous avons accès au label de chaque objet, et donc accès à l'ensemble des contraintes *cannot-link*. Il est ainsi possible de passer l'étape de la génération des contraintes telle qu'effectuée dans les présents travaux, et vérifier la validité de l'algorithme présenté en accédant directement à l'ensemble des contraintes du jeu de données. Mais si nous les utilisons toutes, nous obtiendrions directement la partition finale. En effet, la génération de l'ensemble des contraintes par utilisation des labels fournit *de facto*, en plus de l'ensemble des contraintes *cannot-link*, l'ensemble des contraintes *must-link*. L'agrégation de ces deux ensembles représente la partition finale (cf. chapitre 3). Cela ne fait donc pas sens non plus.

La particularité de l'algorithme dédié présenté dans ce travail réside dans le fait que la génération de la totalité des contraintes *cannot-link* n'entraîne pas la connaissance de l'ensemble des contraintes *must-link*.

De par notre définition de la comparabilité, deux objets (individus) comparables, et donc non reliés par une contrainte *cannot-link*, ne sont pas forcément similaires, et ne doivent donc pas nécessairement appartenir au même cluster dans la partition finale (*must-link*). Cette définition toute théorique et spécifique au problème d'application, à l'origine de la création de l'algorithme présenté, rend son application à un ensemble de données connu impropre et inutile.

Dans cette partie nous avons présenté un algorithme semi supervisé spécifique à notre problématique de l'identification de profils d'apprenants en fonction de leurs stratégies de réponses à divers tests en LC et de leurs sensibilités aux caractéristiques de l'input. Pour ce faire, nous avons discriminé les différentes stratégies de réponses possibles pour chaque test, ainsi que les différents degrés de sensibilité possibles aux spécificités des items utilisés dans ces tâches, et nous avons établi leur corolaire en terme de score numérique. Une fois la stratégie d'un apprenant identifié, il s'est vu attribuer un degré d'appartenance à un SEF, reflet de sa stratégie. Nous avons ensuite défini la notion de **comparabilité** entre deux apprenants et décrit les étapes de comparaisons fondées sur cette dernière qui aboutissent à une classification finale. Ces étapes comportent une première phase *crisp* (purement binaire) nous permettant d'obtenir des cores, groupements d'apprenants comparables, et une deuxième phase mettant en œuvre un algorithme classique de clustering, l'algorithme des K-moyennes, permettant le raffinement de ces cores en clusters, lorsque cela est jugé nécessaire sur la base de différents indices d'homogénéité trouvés dans la littérature. Le résultat de cette étape prend la forme d'une partition en clusters provisoire, regroupements d'apprenants non plus comparables mais cette fois similaires. La troisième et dernière partie nous permet d'intégrer les outliers temporaires par le relâchement de la contrainte *crisp* du calcul de l'indice de comparabilité entre deux apprenants et par l'utilisation de cet indice dans le calcul du score d'appartenance d'un outlier aux clusters obtenus en deuxième étape.

Dans la prochaine section nous appliquons cet algorithme sur deux échantillons de la BDD issue du projet VILLA (cf. chapitre 2). L'objectif de cette section est triple. Le premier enjeu consiste en une analyse des données du projet, à des fins de confirmation des différents résultats issus de VILLA. Le deuxième porte sur la découverte de nouvelles façons de procéder à la résolution de tests de langue d'un apprenant, nous renseignant potentiellement sur son lecte et donc son parcours acquisitionnel, et sur la découverte de rapprochements entre apprenants dont les caractéristiques linguistiques et extra linguistiques ne permettent pas *a priori* d'établir une relation entre eux. Le dernier concerne la validation interne et externe de l'algorithme créé. Une validation externe par *cross validation* sur des sous-échantillons est effectuée afin d'attester de la stabilité de la partition finale trouvée dans les données. La plus grande partie de cette validation consiste en une validation interne de l'algorithme, incarnée dans l'analyse et l'interprétation acquisitionniste du déroulement de l'algorithme et des regroupements d'apprenants obtenus à chaque étape.

Chapitre 6

Application

Sommaire

6.1 Premier échantillon : Capacité de traitement en LC	114
6.1.1 Création des cores par l'utilisation de l'ensemble C	115
6.1.2 Raffinement des cores : création des clusters	128
6.2 Deuxième échantillon : Quel transfert du traitement vers la production de la morphologie en LC ?	135
6.2.1 Création des cores par l'utilisation de l'ensemble C pour le deuxième échantillon . .	137
6.2.2 De l'intérêt d'une intégration des outliers	147

Nous avons présenté dans la section précédente l'algorithme général de classification semi supervisé proposé pour l'analyse, la confirmation, et/ou la découverte de patterns dans une base de données.

Nous avons vu que le partitionnement obtenu repose sur la comparaison deux à deux des objets, comparaison qui n'est effectuée qu'après s'être assuré de la *comparabilité* de ces derniers. La notion de comparabilité développée dans le chapitre précédent est spécifique aux cas d'applications de l'algorithme. Cette notion, et sa mesure, s'élaborent premièrement par la connaissance du phénomène étudié, du comportement d'intérêt, ou, d'un point de vue plus technique, de l'objet caractérisé, contenu dans la base de données. De la nécessité de cette connaissance du cadre d'étude découle naturellement la connaissance du cadre de recueil des données, la tâche ou le test ayant permis d'observer et de mesurer le niveau d'acquisition de la LC.

Il convient, à ce stade de la présentation de nos travaux, d'appliquer l'algorithme précédemment décrit à la base de données qui a inspirée sa création (cf. chapitre 2). Notre objectif est double :

- Confirmer et compléter les analyses du projet VILLA déjà publiées ou en cours de publication ;
- Apporter de nouvelles pistes de réflexion par la découverte de patterns encore non soupçonnés dans la littérature.

Ces deux objectifs, s'ils sont atteints, permettent la validation de l'algorithme en affirmant son utilité du point de vue de l'utilisateur, par sa capacité à produire un partitionnement interprétable et comportant de nouveaux patterns.

Dû à différentes contraintes d'accès à la BDD du projet VILLA (contraintes liées à la collecte des données, à leurs analyses parcellaires, ainsi qu'à leur nature qualitative et multidimensionnelle), tous les tests soumis aux apprenants n'ont pas encore été évalués et/ou transcrits. L'intégralité de la BDD n'est pas disponible, nous avons donc élaboré deux extraits de la base de données. En effet, les données des différents tests ne sont pas disponibles pour les apprenants des cinq LMs.

Le premier échantillon comprend les résultats des apprenants à quatre tests recouvrant le niveau phonologique (Phoneme Discrimination, PD), lexical (Word Recognition, WR), morphologique (Grammaticality Judgement, GJ) et morphosyntaxique (Picture Verification, PV) en polonais. Ces quatre tests étant composés d'items variés en terme de fréquence et de transparence, leurs résultats donnent aussi accès aux différentes sensibilités des apprenants aux caractéristiques de l'input. Également, grâce aux tests ayant été administrés plusieurs fois, il est possible d'évaluer l'évolution d'un apprenant au cours des 14 heures d'enseignement. Cet échantillon procure ainsi une représentation relativement exhaustive des compétences en traitement et compréhension du polonais de la part des apprenants des cinq LMs en présence dans le projet.

Pour le deuxième échantillon, l'objectif est de rendre compte des capacités des apprenants en production du polonais, et plus spécifiquement, des éventuels transferts effectués entre connaissances mobilisées en traitement et capacité à les mobiliser en production. Ainsi, le deuxième échantillon regroupe les résultats de quatre tests en LC. Ces quatre tests vont par paire. Deux d'entre eux se centrent autour d'un paradigme d'opposition en morphologie, l'un en traitement (GJ), l'autre en production (Oral-Question Answer, OQA), et deux autres autour d'un paradigme d'opposition du marquage casuel d'ordre morphosyntaxique l'un en traitement (PV), et l'autre en production (Sentence Imitation, SI). Rappelons que le fait que le même paradigme linguistique soit traité dans une tâche sollicitant un traitement et dans une autre, sollicitant une production, nous permet d'observer (ou non) un transfert des capacités en LC des apprenants d'un type d'activité à un autre. La mise en relation des performances des apprenants en traitement et en production permet de confirmer les idées bien connues en acquisition sur la précocité dans la mise en œuvre des moyens linguistiques de la nouvelle langue en traitement par rapport à leur mise en œuvre en production. En effet, le fait que certaines formes ne soient pas produites par un apprenant n'exclut pas la possibilité de sa capacité à les percevoir.

6.1 Premier échantillon : Capacité de traitement en LC

La première étape de l'algorithme conduit à un premier partitionnement des apprenants en groupe d'apprenants **comparables** que nous nommons *core*. Un core constitue donc un regroupement d'apprenants non liés par une contrainte *cannot-link*. La deuxième étape comporte un découpage interne à chaque core, si sa nécessité est avérée, pour finalement obtenir une partition d'apprenants **similaires**, à laquelle nous intégrons les *outliers* quand cela est possible.

Dans cette section nous proposons de présenter les résultats de chaque étape de l'algorithme et d'observer ainsi, pas à pas, la construction du partitionnement final. Celui-ci permet à la fois de discuter les résultats en acquisition et d'apporter de nouvelles perspectives d'analyses de ces derniers, mais également d'effectuer en partie la validation interne de l'algorithme, en évaluant l'interprétabilité de la classification (les profils d'apprenants) obtenue.

L'échantillon utilisé est constitué de 156 apprenants issus des 5 langues maternelles présentes dans le projet VILLA, à savoir le français, l'allemand, l'anglais, le néerlandais et l'italien. Ces apprenants sont caractérisés par une valeur numérique correspondant au score qu'ils ont obtenus aux épreuves en LC sollicitant un traitement sur les niveaux linguistiques suivants :

- la phonologie,
- le lexique,
- la morphologie,
- la morphosyntaxe.

De plus, ces valeurs numériques nous renseignent sur leurs sensibilités aux caractéristiques suivantes de l'input :

- la fréquence d'un item,
- la transparence pour les items fréquents,
- la transparence pour les items non fréquents,
- et la durée d'exposition à l'input.

La sensibilité à une caractéristique de l'input est calculée comme un écart entre les deux scores des deux modalités d'une même variable. Par exemple, la sensibilité à la fréquence s'obtient par l'écart entre le score pour les items fréquents et celui pour les items non fréquents, tous tests confondus. L'écart de réussite entre les deux modalités de la variable fréquence permet de mesurer l'impact positif de cette caractéristique sur le score obtenu par les apprenants. Autrement dit, nous pouvons identifier les apprenants qui réussissent à appliquer des règles grammaticales de la LC plus facilement sur les items fréquents (cf. chapitre 4).

Ces valeurs numériques sont ensuite fuzzifiées et exprimées en degré d'appartenance à un sous ensemble flou (SEF)

, permettant une distinction des stratégies appliquées en réponse à la sollicitation en LC induite par la tâche (cf. chapitre 5). Nous rappelons que pour cet échantillon tous les tests retenus évaluent les capacités de traitement de la LC par un apprenant, et cela pour différents paradigmes d'analyses linguistiques.

6.1.1 Création des cores par l'utilisation de l'ensemble C

6.1.1.1 Ensemble C

Le processus de création de l'ensemble C comprend la comparaison 2 à 2 des individus pour chacune des dimensions les caractérisants (cf. chapitre 5). L'échantillon considéré inclut les résultats de 156 apprenants sur 8 dimensions. Sachant que deux apprenants sont reliés par une contrainte *cannot-link* s'ils sont jugés incomparables sur au moins une de ces dimensions, l'ensemble C peut au maximum inclure 12090 contraintes ($\frac{(n-1)*n}{2}$), si aucun des apprenants n'est jugé comparable avec un autre. Une fois toutes les comparaisons effectuées, l'ensemble C ainsi créé contient 7387 contraintes *cannot-link*.

L'analyse la plus intéressante à effectuer serait de déterminer quelle contrainte a pour origine quelle dimension de la modélisation, et en quelle quantité, afin de démontrer l'importance relative de chaque dimension dans le processus de partitionnement. En d'autres mots, cette analyse nous permettrait d'identifier la dimension créatrice du plus de variabilité parmi nos apprenants, et d'effectuer un classement des dimensions quant à leur propriété d'hétérogénéisation de la population.

Malheureusement, il n'est pas possible de mesurer cette propriété via l'association contrainte-dimension. En effet, il suffit d'une seule dimension pour laquelle deux apprenants sont jugés incomparables pour qu'il y ait création de contrainte. Cependant, il peut également s'avérer que ces deux apprenants soient incomparables par rapport à plusieurs critères. Il n'y a donc pas de correspondance entre le nombre de contraintes de l'ensemble C et le nombre de mesures d'incomparabilité effective. Ainsi, le nombre de contraintes reliant deux à deux les individus de la BDD est plutôt pauvre en signification.

En revanche, l'analyse pour chaque dimension du nombre de fois où un apprenant a été jugé (mesuré) incomparable avec un autre, permet le classement de l'importance relative des facteurs de la modélisation. Par exemple, par le calcul du nombre de fois où deux individus ne possèdent pas la même sensibilité à la fréquence, nous pouvons établir un ordre d'importance de celle-ci par rapport aux autres dimensions, et déterminer ainsi quelle est celle qui crée le plus de variabilité dans le processus d'acquisition d'une L2, pour cet échantillon tout du moins.

Le graphique 6.1 présente le nombre de fois où deux apprenants ont été jugés incomparables, et ce pour chaque dimension. Le chiffre en lui-même n'est pas significatif mais nous permet d'observer une différence de

la taille de l'effet pour chaque dimension dans la création des cores.

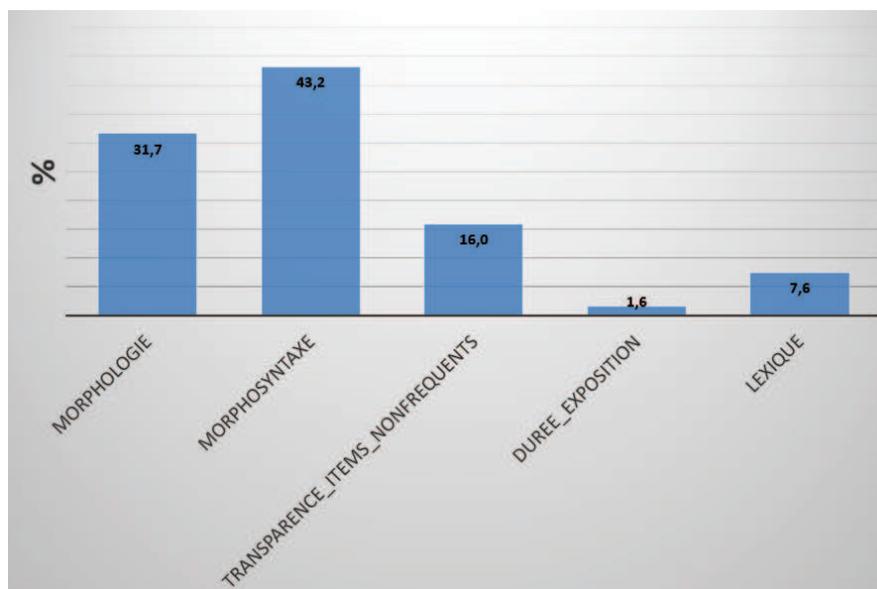


FIGURE 6.1 – Influence relative des dimensions de la modélisation dans le processus de création des cores : échantillon 1

Sur les 8 dimensions retenues pour la caractérisation des apprenants, seules 5 ont été à l'origine de la création de contraintes *cannot-link* :

- le traitement du système morphologique,
- le traitement du système morphosyntaxique,
- la sensibilité à la transparence pour les items non fréquents,

et dans une moindre mesure :

- le traitement du lexique,
- l'effet de la durée d'exposition à l'input.

Les autres dimensions ne semblent pas avoir d'effet discriminant sur la création des cores, n'étant pas responsables de variations dans les comportements des individus.

On observe, toujours sur le graphique 6.1, que lors des comparaisons deux à deux, l'incomparabilité entre les scores de deux apprenants a été constatée 4320 fois. Il semble donc que la morphosyntaxe est la dimension la plus créatrice de variabilité dans le résultat des apprenants. La dimension morphologique est également, mais plus faiblement, facteur de variation dans les résultats acquisitionnel des apprenants. De même, on note que l'impact de la transparence des items non fréquents n'est pas le même pour tous les apprenants, avec une incomparabilité entre deux apprenants constatée à hauteur de 1595 fois. Ce chiffre passe à seulement 755 pour la dimension lexicale, et ne semble donc pas concerner beaucoup d'apprenants. En effet, moins ce nombre est élevé, moins le nombre d'apprenants optant pour une stratégie de réponse différente de celle des autres apprenants est grand. En ce qui concerne la durée d'exposition à l'input, on remarque que les apprenants semblent évoluer globalement de manière homogène au cours des séances d'enseignements.

Les dimensions absentes du graphique 6.1 sont

- la phonologie,
- la fréquence d'un item,
- et la transparence pour les items fréquents

Cette absence signifie que ces dimensions ne sont pas à l'origine de gros écarts dans les résultats des apprenants. Cependant, cela ne signifie pas que la fréquence d'un item n'a pas d'impact sur la rétention et le traitement de celui-ci par l'apprenant. En effet, l'algorithme ne permet pas d'analyser de manière fine les effets des caractéristiques des items sur le processus d'acquisition. La détection n'est possible que si ces effets sont exacerbés. De même, bien que les résultats des apprenants à la tâche phonologique ne sont sûrement pas sensiblement les mêmes, les variations existantes ne permettent pas de justifier d'une stratégie de réponse à la tâche différente. En effet, après observation plus poussée des résultats des apprenants à la tâche PD, on constate que la moyenne de réussite à ce test est de 91% avec un écart type d'environ 0.05. Il apparaît ainsi que non seulement ce test n'est pas à l'origine de variabilités notables dans le processus d'acquisition des apprenants mais que, de plus, ceux-ci réussissent plutôt bien le test de discrimination de phonème. Rappelons tout de même que ce calcul est effectué sur les résultats des apprenants toutes périodes de passation du test confondues.

6.1.1.2 Cores obtenus

Après l'élimination des cores de moins de 5 individus, reclassés dans les *outliers* temporairement, nous obtenons 5 cores à cette étape de l'algorithme, et ainsi 8 *outliers*. Nous pouvons observer le détail de la répartition de la population des apprenants entre les cores sur la figure 6.2.

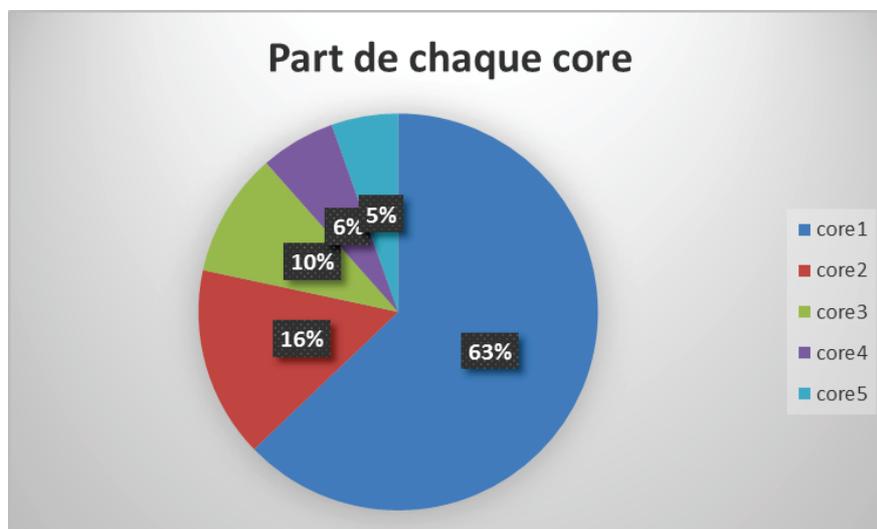


FIGURE 6.2 – Répartition des apprenants entre les différents cores, *outliers* non inclus

Nous observons un core majoritaire comprenant 63% de la population, soit 93 apprenants. Les autres cores se partagent non équitablement les 55 apprenants restants.

La figure 6.3 présente l'appartenance des individus à différents cores identifiés en fonction des SEFs auxquelles ils appartiennent, pour chaque dimension de la modélisation ayant révélé être à l'origine de variations significatives dans les comportements des apprenants. Elle nous permet de comprendre les critères

de répartition de la population dans les différents cores.

Le core 1, majoritaire, regroupe les individus qui semblent avoir commencés à traiter le système morphologique et morphosyntaxique de la LC.

En effet, ses individus appartiennent aux deux SEFs S_1 et S_3 exclusivement, quel que soit le niveau d'analyse linguistique considéré. Leurs appartenances à ces SEFs suggèrent qu'ils ont amorcé une remise en question de leurs représentations initiales du fonctionnement de la LC. Ce changement de stratégie résulte de la mise en question des règles faisant partie du lecte de l'apprenant à un moment donné d'acquisition. Une règle utilisée par l'apprenant suite à l'analyse de l'input et à la formulation des hypothèses sur le fonctionnement de la LC devient critique (dans le sens de [Klein, 1989]) lorsque l'apprenant se rend compte d'un écart qui existe entre ses performances en LC et l'input traité. Ainsi, l'apprenant juge la règle construite dans son interlangue comme contradictoire par rapport au système de la LC. Néanmoins, l'appartenance des apprenants aux deux SEFs S_1 et S_3 indique que cette remise en question n'amène pas obligatoirement ces apprenants à rapprocher leur lecte du fonctionnement réel de la LC. En effet, les apprenants appartenant au SEF S_1 n'ont pas obtenu un score élevé à ces tâches. Ils se démarquent pourtant des apprenants dont le score appartient au SEF S_2 , apprenants qui n'ont pas encore effectué le constat d'une inadéquation entre l'input et leur lecte (pour un rappel sur la construction des trois SEFs, cf. chapitre 5).

Le lecte de l'apprenant est par nature instable (cf. chapitre 1). L'apprenant remet en question ses représentations sur le fonctionnement de la LC lorsque, lors de leur utilisation effective, elles lui apparaissent alors en contradiction avec le système réel de la LC. Les règles de fonctionnement jusqu'alors appliquées sont dédaignées et reléguées en règles critiques. Le changement dans les stratégies de réponses de l'apprenant expriment ce changement interne.

Par exemple, dans la tâche testant le niveau morphosyntaxique (PV), les apprenants optent d'abord pour une stratégie positionnelle qui consiste à attribuer le statut de sujet au premier constituant de la phrase. Cette stratégie est guidée par le principe sémantique Agent-Action-Patient qui correspond à l'ordre SVO (cf. chapitre 1). L'utilisation de cette stratégie dans le traitement des phrases SVO conduit à une ambiguïté entre la structure syntaxique et la structure sémantique. Dès lors que l'apprenant traite correctement les phrases OVS, on peut supposer qu'il applique la stratégie morphosyntaxique. En d'autres termes, lorsqu'il développe une sensibilité au marquage casuel en polonais, il peut identifier le Sujet et l'Objet grâce aux désinences. En effet, au stade dit "positionnel" d'acquisition, si l'on ne considère que les résultats des apprenants concernant les phrases de type SVO, il est difficile de savoir si l'apprenant interprète la phrase à partir des connaissances morphosyntaxiques ou s'il s'appuie sur le schéma sémantique "Agent-Action-Patient". C'est pourquoi il est nécessaire de coupler ces résultats à ceux concernant les phrases de type OVS, ce qui nous fournit l'indice permettant de déduire si l'apprenant se base sur le marquage casuel pour identifier le statut d'un constituant de la phrase. Lorsque l'apprenant appartient au SEF S_2 , on peut supposer que seules les phrases de type SVO (représentant 50% du score global) ont été traitées correctement. Or le core majoritaire ne regroupe pas d'apprenants appartenant au SEF S_2 , quel que soit le paradigme linguistique testé (morphologie, morphosyntaxe, etc.). L'appartenance aux deux SEFs S_1 et S_3 indique toutefois que la correspondance entre le statut du constituant de la phrase avec un marquage casuel approprié n'est pas encore connue pour 31 d'entre eux. Nous reviendrons sur cette réflexion dans l'analyse du core 2.

En morphologie également (test GJ, cf. chapitre 2), les différentes désinences sont discriminées par les apprenants qui différencient donc l'instrumental du nominatif pour l'ensemble du core 1 (cf. chapitre 2), et plus spécifiquement, qui utilisent en contexte approprié l'instrumental féminin et masculin. Seuls deux individus, ne parviennent pas à ce résultat. Cependant, ils n'appliquent pas non plus la stratégie aléatoire (S_2), ce qui suggère qu'ils disposent d'une stratégie cohérente pour eux bien qu'elle ne soit pas en accord avec la LC.

Il apparaît ainsi que les apprenants du core 1 ont tous amorcé une entrée dans le système de la LC, avec encore des difficultés d'association correcte forme-fonction pour certains. Cette différence entre les apprenants du core majoritaire sera prise en compte dans la deuxième étape de l'algorithme, à savoir l'obtention d'une

partition d'apprenants similaires et non plus comparables.

Si l'on s'intéresse aux niveaux linguistiques pour lesquels les apprenants ont atteint la même étape acquisitionnelle, à savoir le lexique et la phonologie, tous les apprenants du core 1 appartiennent au SEF S_3 , correspondant à un score élevé. On note d'ailleurs que pour ces deux dimensions les apprenants des autres cores appartiennent également tous (sauf un) à ce SEF, ce qui explique pourquoi ces deux dimensions n'ont donné lieu à aucune création des contraintes de l'ensemble C . Le peu d'apprenants jugés incomparables quant à leurs résultats au niveau lexical se trouvent dans les *outliers*, apprenants sur lesquels nous reviendrons par la suite.

Les apprenants du core majoritaire possèdent une sensibilité relativement faible aux différentes caractéristiques des items de l'input, ainsi qu'à la durée d'exposition à l'input. Dans la littérature VILLA on constate des effets statistiquement significatifs de la fréquence et de la transparence pour certaines tâches, testant l'acquisition des différents niveaux linguistiques de la LC. Le tableau 6.3 montre que les individus du core majoritaire (core 1) ne présentent pas une sensibilité exacerbée à ces différentes propriétés des items de l'input, et à leur durée d'exposition à ce dernier.

Ce core dépeint un profil très cohérent, d'apprenants étant entrés dans le système de la morphologie flexionnelle de la LC, et possédant une sensibilité faible (car non détectée à ce niveau de granularité, contrairement à certaines publications issues du projet VILLA, [Hinz et al., 2013], [Saturno Jacopo, 2014], [Rast et al., 2018]) et très homogène aux propriétés de l'input.

Pourtant, le système identifie d'autres stratégies de réponses mises en places par les apprenants, entraînant la création d'autres cores. De fait, les autres cores possèdent chacun une caractéristique spécifique. Cette spécificité réside dans les résultats des apprenants pour une dimension spécifique de la modélisation. En effet, tous les apprenants des autres cores sont similaires à ceux du core majoritaire excepté pour un ou deux niveaux linguistiques particuliers, ou encore en regard d'une propriété de l'input.

		Phono	Morphologie	Morpho-syntaxe	Lexique
core1	S ₁	0	2	31	0
	S ₂	0	0	0	0
	S ₃	93	91	62	93
core2	S ₁	0	1	0	0
	S ₂	0	0	23	0
	S ₃	23	22	0	23
core3	S ₁	0	0	8	1
	S ₂	0	15	0	0
	S ₃	15	0	7	14
core4	S ₁	0	0	1	0
	S ₂	0	0	0	0
	S ₃	9	9	8	9
core5	S ₁	0	0	0	0
	S ₂	0	8	8	0
	S ₃	8	0	0	8

		Freq	Transp_nonfreq	Transp_freq	Transp	Evol
core1	-	0	0	0	0	0
	O	93	93	93	93	93
	+	0	0	0	0	0
core2	-	0	0	0	0	0
	O	23	23	23	23	23
	+	0	0	0	0	0
core3	-	0	0	0	0	0
	O	15	15	15	15	15
	+	0	0	0	0	0
core4	-	0	0	0	0	0
	O	9	0	9	9	9
	+	0	9	0	0	0
core5	-	0	0	0	0	0
	O	8	8	8	8	8
	+	0	0	0	0	0

FIGURE 6.3 – Tableau récapitulatif de la répartition des apprenants dans les trois SEFs (S_1, S_2, S_3) de l'ensemble d'appartenance en fonction de l'attribut considéré (partie 1 pour les niveaux linguistiques, partie 2 pour les caractéristiques de l'input) et selon les différents cores. Échantillon 1

Les apprenants du core 2 partagent les mêmes caractéristiques que ceux du core 1, excepté en ce qui concerne leurs représentations morphosyntaxiques en LC. A la morphologie nominale très riche du polonais, s'ajoute une certaine liberté dans l'ordre des mots de la phrase. Étant donné leur appartenance aux SEFs S_1 et S_3 pour leurs résultats en morphologie, ils semblent commencer à traiter celle-ci. Cependant, ils peinent à traiter le niveau morphosyntaxique. Le marquage casuel associé à l'ordre des mots constitue donc un obstacle supplémentaire dans l'acquisition de la LC. Lors de l'élaboration de la mesure de comparabilité (cf. sous section 5.1.1) nous avons admis le fait que les apprenants appartenant au SEF S_2 pour la tâche *Picture verification* n'ont pas encore pris conscience de l'existence de cette liberté dans l'ordre des mots en polonais. Ils appliquent la stratégie positionnelle pour comprendre la phrase. Selon cette stratégie, rappelons-le, l'apprenant attribue le statut de l'agent au SN en position initiale dans la phrase et n'a pas recours au

marquage morphosyntaxique. C'est à dire que ces apprenants traitent toutes les phrases du test comme des phrases construites sur le mode Sujet - Verbe - Objet (SVO), bien que certaines soient construites selon l'ordre Objet - Verbe - Sujet (OVS) ou Objet - Sujet - Verbe (OSV). Cette stratégie renvoie potentiellement au principe sémantique décrit par Klein et Perdue ([Klein et Perdue, 1997]) chez les apprenants débutants en milieu naturel, selon lequel l'apprenant considère le premier constituant nominal d'un énoncé comme l'agent de l'action. A ce stade, il n'est pas possible de prédire avec sûreté que cette opération effectuée par l'apprenant renvoie au niveau syntaxique où le premier constituant serait traité comme le sujet de la phrase. Ces auteurs suggèrent donc par précaution que l'apprenant attribue le statut sémantique de l'agent au premier SN dans l'énoncé. En effet, l'attribution du statut de sujet à un constituant de la phrase renvoie à une analyse se basant sur le marquage casuel de ce constituant, analyse que les apprenants de ce core ne font pas, au vu de leur score sur les phrases de type OVS. En effet, la stratégie des apprenants de ce core est reflétée par un score correct très faible avec les phrases OVS et OSV, et un score correct très élevé pour les phrases SVO, pour un score total approchant la moyenne, et donc appartenant au SEF S_2 .

En d'autres termes, les apprenants du core 2 sont sensibles à la variation des formes du syntagme nominal, comme nous pouvons le constater par leur appartenance aux SEFs S_1 et S_3 pour le test de morphologie (ils discernent l'opposition Instrumental - Nominatif), mais ne l'utilise pas pour la compréhension de la phrase, qui nécessiterait une association entre un marquage casuel et le statut du constituant (Nominatif-Sujet *vs.* Accusatif-Objet).

Le core 5 correspond aux apprenants appartenant à une étape du parcours acquisitionnel encore moins avancée que ceux du core 2. Ces apprenants sont encore réfractaire au système casuel polonais. Ils ne semblent pas sensibles au système casuel de la langue cible. L'identification des désinences comme nécessaires à l'interprétation de phrases grammaticalement correctes constitue une étape complexe dans l'acquisition du polonais, notamment, comme nous le verrons en sous section 6.1.1.3, au vu de leur LM.

Ce core présente des caractéristiques similaires à ce que Klein et Perdue nomment la variété de base (*basic variety*, BV, [Klein et Perdue, 1997], cf. chapitre 1). L'absence totale de morphologie flexionnelle est en effet une des caractéristiques de la BV. Ce stade d'acquisition correspond à l'utilisation d'un noyau "embryon de verbe", précédé de l'agent de l'action, et suivi de l'objet de l'action, les items Agent et Patient étant utilisés sous leur forme invariable et non marquée morphologiquement (en genre, nombre, ou cas). Les résultats de ces apprenants suggèrent également une stratégie positionnelle, et une absence de flexion nominale. Toutefois ce parallèle avec la BV doit être exprimé prudemment. La structure de BV a été décrite et théorisée à partir de l'analyse des données orales provenant de la BDD issue du projet ESF ([Perdue, 1993], cf. chapitre 1). Le projet ESF étudie également grâce à une observation longitudinale les premiers stades d'acquisition. Cependant, la nature de l'input reçu par ces apprenants, ainsi que leurs profils sociobiographique diffèrent entièrement de ceux du projet VILLA. Le projet VILLA concerne les apprenants débutants en milieu guidé où ils sont exposés à des séances de cours de langue axés sur l'enseignement de la morphologie flexionnelle, ce qui introduit un biais différent dans le parcours acquisitionnel. A contrario, les apprenants ESF étaient complètement immergés dans leur pays d'accueil et de ce fait ont reçu un input en milieu naturel. Cette différence fondamentale au niveau de l'input rend difficile la comparaison des parcours acquisitionnel des apprenants des deux projets. De plus, les apprenants ESF étaient des immigrants économique ou politique ayant reçu une instruction limitée dans leurs pays d'origine et pratiquant un métier manuel dans leurs pays d'accueil. On peut ainsi supposer que leurs réflexions métalinguistique autour de leur LM n'est pas aussi développée que les étudiants universitaires ayant participé au projet VILLA. Cependant, leur motivation à apprendre la LC se centre autour de l'intégration sociale et professionnelle, motivation très forte donc en comparaison des apprenants du projet VILLA (dont seulement une partie a reçu une compensation financière).

L'interaction entre l'input et le profil des apprenants permet d'émettre l'hypothèse d'une évolution dans le parcours acquisitionnel sensiblement différente. Il est donc difficile d'effectuer une mise en parallèle des résultats des deux projets. De plus, bien que les apprenants du core 5 n'arrivent pas à traiter le système morphosyntaxique de la LC, de telles données en terme de traitement n'existent pas pour les apprenants du projet ESF. En effet, le recueil des données de ce projet est entièrement basé sur les productions des apprenants et n'incluent pas de tâches évaluant les capacités de traitement de la LC. Or, comme nous le

verrons avec le deuxième échantillon d'analyse (cf. section 6.2), l'absence de production en accord avec le système casuel de la LC n'implique pas nécessairement l'incapacité à le traiter correctement pour un même apprenant.

Le core 3 pose des problèmes d'interprétation si on se place du point de vue d'un continuum acquisitionnel d'une LC au système flexionnel riche comme c'est ici le cas. En effet les apprenants de ce core arrivent à identifier le rôle d'un mot dans la phrase par l'intermédiaire de sa désinence et ont dépassé la stratégie positionnelle au niveau morphosyntaxique au vu de leurs résultats au test PV. Ils sont donc entrés dans le système flexionnel de la LC, et commencent à juger de la grammaticalité d'une phrase en se basant sur le marquage casuel des items la composant. Pourtant, ils n'appliquent pas cette connaissance au niveau morphologique au vu de leurs résultats au test GJ.

ID	score_Instrumental_correct	score_Nominatif_incorrect	stratégie	moyenne	SEF
1105	0,906	0,234	aveugle	0,570	S2
2111	0,719	0,359	aveugle	0,539	S2
2119	0,859	0,094	aveugle	0,477	S2
3105	0,516	0,344	alea	0,430	S2
3113	0,891	0,172	aveugle	0,531	S2
3117	1,000	0,031	aveugle	0,516	S2
3118	0,781	0,406	alea	0,594	S2
3214	0,656	0,266	aveugle	0,461	S2
5205	0,797	0,156	aveugle	0,477	S2
5210	0,688	0,391	aveugle	0,539	S2
5212	0,828	0,344	aveugle	0,586	S2
1210	0,625	0,469	alea	0,547	S2
2201	0,516	0,500	alea	0,508	S2
2206	0,500	0,500	alea	0,500	S2
2214	0,828	0,328	aveugle	0,578	S2

FIGURE 6.4 – Résultats des apprenants du core 3 à la tâche GJ en fonction du genre de l'item et du cas sollicité. Échantillon 1.

Le paradigme morphologique testé (par la tâche GJ, cf. chapitre 2) dans le projet VILLA est celui de l'instrumental masculin et féminin en opposition au nominatif masculin et féminin. Dans ce test, l'apprenant entend une phrase contenant ou non une erreur sur la forme de l'instrumental et doit dire si la phrase est correcte ou non d'après lui. Le test est construit de manière à ce que la phrase est correcte seulement lorsque l'instrumental (féminin ou masculin) est utilisé. Ainsi, si les marques du nominatif sont présentes, l'apprenant doit juger la phrase incorrecte. Après une analyse détaillée des résultats des apprenants du core 3 à cette tâche (cf. figure 6.4), nous observons que les apprenants du core 3 appartiennent à deux catégories. Soit ils jugent quasiment toutes les phrases correctes, alors que seulement la moitié le sont, stratégie dénommée "aveugle" dans la figure 6.4, soit ils répondent de manière aléatoire, obtenant ainsi le même score final autour de 0.5, ce qui conditionne leur appartenance au même SEF S_2 .

Dans les deux cas, nous émettons l'hypothèse que la richesse des désinences et des items en présence dans ce test est à l'origine de leur résultat. La tâche *Gramaticality judgement* comprend 4 types de désinences : instrumental masculin et féminin, et nominatif masculin et féminin. De plus, la désinence du nominatif masculin est la désinence \emptyset ce qui conduit à une multitude de forme en fonction de l'item auquel il s'applique. Chaque item dans ce cas se termine par une consonne. On assiste donc à une variabilité due au fait que différents items au nominatif masculin se terminent par différentes consonnes. Il est donc difficile pour l'apprenant de fixer une règle claire. La richesse des formes soumises à l'apprenant dans ce test est accentuée par le nombre d'items utilisés dans la construction des stimulus qui s'élève à 64 (32 professions et 32

nationalités).

Comme nous l'avons évoqué dans le chapitre 5, Rast et col. ([Rast et al., 2018]) explorent les causes possibles de la saillance. En effet, les auteurs partent de l'hypothèse que la saillance d'un item constitue une aide envers l'apprenant pour sa rétention et son apprentissage, notamment lorsque celui-ci correspond à une nouvelle forme auquel est confronté l'apprenant dans l'input. Leur problématique se centre autour des facteurs de saillance des marqueurs morphologiques du polonais. Leurs principales conclusions se centrent autour de quatre critères. L'un de ces critères clés est la régularité de l'association forme fonction présente dans l'input. La multitude des formes prises par l'instrumental dans la tâche GJ ne permet donc pas à l'instrumental d'être saillant pour la perception et le traitement de cette forme auprès de l'apprenant.

Nous soulignons ici les caractéristiques de cette tâche et sa complexité afin de créer un parallèle avec la tâche *Picture verification*, tâche ayant engendré les résultats en morphosyntaxe. Dans ce dernier test, seuls deux items sont utilisés. La règle morphosyntaxique du polonais concernant le marquage casuel du statut de Sujet et de l'Objet par l'opposition "Nominatif (Sujet) vs. Accusatif (Objet)" n'est donc appliqué qu'à ces deux items nominaux, "brat" (frère) et "siostra" (sœur). Ainsi, la tâche de l'apprenant est ici plus simple que dans le cas du test de GJ. En effet, il doit interpréter seulement 4 formes différentes : nominatif et instrumental féminin du mot "sœur" (siostra vs. siostre,) ainsi que nominatif et instrumental masculin du mot "frère" (brat vs. brata).

La complexité différente de ces deux tâches et notamment la saillance faible de la forme de l'instrumental dans la tâche GJ pourraient être à l'origine du regroupement d'apprenants dans le core 3. Ces apprenants auraient compris l'existence du marquage casuel en LC mais ils ne peuvent le mettre en œuvre que dans une tâche moins complexe, à savoir PV.

Une autre approche pour examiner la logique de création des cores obtenus passe par l'étude des sensibilités des apprenants aux différentes caractéristiques des items utilisés dans les tâches du projet VILLA. Lors de l'étude du core majoritaire, nous avons établi que ses apprenants ne présentaient pas une sensibilité particulière à la fréquence d'un mot ou à sa transparence. En effet, la transparence et/ou la fréquence d'un item n'impliquent pas, dans le cas de ces apprenants, l'augmentation des scores corrects dans les tests. Des effets de ces deux variables et de la combinaison de leurs différentes modalités sont observés par ailleurs dans la littérature du projet VILLA (cf. chapitre 2). Le niveau de granularité de l'analyse présentée ici, ne permet pas de déceler un effet global de ces variables. L'objectif est en effet d'observer des effets différenciés, c'est à dire une certaine variabilité, dans la sensibilité des apprenants à la transparence et la fréquence d'un item. Nous souhaitons ainsi non pas observer un effet de ces caractéristiques des items, mais distinguer une influence exacerbée d'une influence générale, d'où le choix d'une analyse moins fine.

Comme précédemment énoncé, la constitution des autres cores semble tourner autour de ce core majoritaire. Leur existence repose sur la variation spécifique qu'ils présentent avec celui-ci, pour une ou deux des dimensions caractérisant l'état acquisitionnel d'un apprenant dans notre modélisation. Ainsi, les apprenants regroupés dans les cores 2, 3 et 5 ne présentent également aucune sensibilité notable aux caractéristiques des items (transparents et fréquents). En revanche, le core 4 se distingue des autres, non pas par rapport au niveau linguistique testé, mais par une hypersensibilité de ses apprenants à la transparence dans le cas d'items non fréquents. Les apprenants de ce core sont plus aidés par la transparence d'un item, lorsque celui-ci est non fréquent dans l'input reçu pendant les séances d'enseignement. En d'autres termes, ces apprenants sont dotés d'une plus grande facilité d'application du système flexionnel de la LC aux items non fréquents lorsque ceux-ci sont transparents, comparativement au reste de la population des apprenants.

Ainsi, les apprenants du core 4, au nombre de 9, partagent les mêmes caractéristiques que les apprenants du core 1. Les apprenants de ces deux cores ont commencé à mettre en place des stratégies de réponses aux tests linguistiques en LC proches du fonctionnement réel de la LC. Cependant, les apprenants du core 4 surpassent ceux du core 1 lorsque les items utilisés dans les tests linguistiques sont des items transparents qu'ils

n'ont que peu vu ou entendu, c'est à dire non fréquent dans l'input. Le fait que ces apprenants saisissent et traitent mieux un item transparent non fréquent que les autres de l'échantillon, est à l'origine de leur regroupement par l'algorithme. Cette caractéristique est fondatrice de leur profil d'acquisition. La question de l'origine de ce résultat reste indéterminée. On ne peut attribuer cette facilité avec leurs LMs car celles ci sont variées parmi les apprenants du core 4. Une autre possibilité d'éclaircissement de ce résultat peut résider dans les caractéristiques psychométriques individuelles telles que la mémoire phonologique ou le style cognitif d'apprentissage. Ces informations existent et ont été établies en préalable au projet VILLA pour tous les participants. Leur examen est une perspective intéressante de travail futur.

6.1.1.3 Rôle de la langue maternelle dans la création des cores

Du fait de l'importance de la LM dans l'acquisition d'une L2, il est nécessaire de s'interroger sur l'existence d'une influence de la LM des apprenants sur leur répartition dans les différents cores obtenus. La question de l'influence de la LM dans l'acquisition d'une LE est un sujet particulièrement intéressant dans le cadre des études sur les débuts de l'apprentissage d'une L2. En effet, des auteurs en RAL ([Klein et Perdue, 1997]) avancent l'idée que l'influence de la LM dans les productions des apprenants en LC n'est possible qu'à un certain niveau d'acquisition, correspondant à un accès lexical plus riche et des moyens linguistiques (notamment morphologiques) plus variés. L'idée est en effet que, au tout début de l'apprentissage, ce sont des principes d'organisations discursives indépendants de la LM qui sont appliqués, comme observés en BV dans le projet ESF (cf. chapitre 1).

Cependant, les travaux en acquisition des langues secondes s'accordent sur le rôle de la langue source dans la construction du lecte de l'apprenant ([Giacobbe, 1992]). Les apprenants adultes font appel aux connaissances linguistiques issues de l'acquisition de la langue maternelle. En les comparant aux moyens linguistiques extraits de l'input de la langue cible, ils procèdent à une « psychotypologie des langues » ([Kellerman, 1995]) qui contribue à la construction du nouveau système linguistique, c'est-à-dire au lecte des apprenants.

De plus, comme nous l'avons observé dans la section précédente, certains cores regroupent des apprenants qui effectuent une analyse morphologique et/ou morphosyntaxique de la LC. Cette différence avec la BV identifiée chez des apprenants en milieu naturel provient très probablement du guidage de l'acquisition (cf. chapitre 1). Les apprenants de VILLA sont exposés dès le début de leur apprentissage aux formes flexionnelles. Du fait de cet input contrôlé et manipulé pour les besoins de l'expérience, il n'est pas étonnant de constater que les apprenants ont commencé l'intégration du système morphologique et morphosyntaxique de la LC (exception faite du core 5), intégration pouvant être variable en fonction de leur LM. Ainsi, il est doublement important de se poser la question de l'influence de la LM d'un apprenant sur son processus d'acquisition dès les toutes premières heures d'enseignement de celle ci, dans un contexte guidé et orienté sur l'étude de la morphologie flexionnelle.

Un test de relation entre les deux variables nominales LM et Core a été effectué ($\text{Khi}^2 = 29.646$, $p > 0.05$). Nous acceptons donc l'hypothèse nulle selon laquelle les deux variables seraient indépendantes, la force d'association de ces deux variables semblant par ailleurs très faible (V de Cramer 0.2). Cependant les prémisses nécessaires à la réalisation d'un tel test ne sont pas respectées, du fait du faible nombre d'individus et du nombre différent d'apprenant par langue maternelle, spécialement en ce qui concerne les allemands dont les effectifs dans notre base de données sont inférieurs de moitié à toutes les autres langues maternelles (cf. chapitre 2). Cette analyse ne permet donc pas de conclure sur une éventuelle influence de la langue maternelle dans la constitution des cores à cette étape de la classification.

Par rééquilibrage des effectifs pour chaque LM, en constituant des groupes de 20 individus par LM, nombre basé sur l'effectif du groupe des germanophones (le plus petit effectif), tirés aléatoirement dans la

base de données, il est possible d'effectuer une analyse plus juste de l'influence de la LM dans le processus d'acquisition des apprenants. Ce rééquilibrage des effectifs est pré-requis à la comparaison des compositions respectives des cores. De plus ces tirages vont nous permettre de vérifier plusieurs hypothèses :

- i La première concerne la présence pérenne ou non d'un core dont la part relative en nombre d'individus est considérablement supérieure à celle des autres (le core majoritaire).
- ii La seconde concerne la stabilité du nombre de cores obtenus, et donc intrinsèquement du nombre de profils existants parmi la population de nos individus.
- iii Selon la troisième hypothèse nous souhaitons également savoir si nous allons retrouver les mêmes caractéristiques des profils obtenus.
- iv La quatrième hypothèse concerne les outliers qui devraient rester les mêmes quel que soit le tirage, du fait qu'ils ne sont jugés comparables avec aucun autre individu de la base de données, ou avec seulement d'autres outliers.

10 tirages aléatoires sont ainsi effectués. Parmi eux, nous retrouvons chaque fois la présence d'un core majoritaire regroupant en moyenne 66,6% des apprenants. Pour 9 tirages sur 10 nous obtenons 5 cores, le tirage restant comprend 4 cores seulement mais un nombre d'outliers plus élevés. Aussi, les outliers attestés dans 9 des tirages appartiennent tous à la catégorie des outliers dans notre analyse principale sur l'ensemble de l'échantillon. Ces trois constats nous permettent d'attester de la stabilité de notre algorithme de clustering, par rapport à l'identification de profils d'acquisition présents dans l'échantillon.

De plus, les cores obtenus dans chacun des tirages partitionnent les apprenants selon le même schéma que lors du partitionnement de l'échantillon complet. Le core majoritaire regroupe les apprenants utilisant la morphologie flexionnelle pour l'identification du statut du constituant dans la phrase, et la discrimination des paradigmes instrumental et nominatif. Les apprenants de ce core sont également tous relativement peu sensibles aux différentes caractéristiques des items utilisés dans les tâches. Par opposition, on observe l'existence d'un core d'apprenants hypersensibles à la transparence d'un item non fréquent. Cette hypersensibilité se reflète dans leur capacité supérieure à celle des autres apprenants à marquer de manière appropriée un item transparent lorsque celui-ci est non fréquent dans l'input.

Les 3 cores restants se constituent d'apprenants tous peu sensibles aux caractéristiques des items, mais se distinguant par leur entrée ou non dans le système de morphologie flexionnelle de la LC. Parmi les apprenants sensibles aux variations du syntagme nominal dans les différentes tâches, on distingue également ceux utilisant cette variabilité uniquement dans la discrimination des paradigmes nominatif et instrumental, et ceux ne l'utilisant que pour l'attribution du statut du constituant dans la phrase. L'hypothèse iii. est donc également corroborée.

Par ces différents constats, nous confirmons la capacité de notre algorithme à identifier la structure sous jacente présente dans l'échantillon. Cette procédure de validation interne s'apparente aux techniques de validations croisées connues, par division de la BDD en k sous échantillons. Traditionnellement les différents sous échantillons ne contiennent pas les mêmes objets à classifier. Seulement, dans notre cas, le nombre d'apprenants de l'échantillon total ne permet pas l'application d'un tel découpage. Comme nous l'avons précédemment évoqué dans le chapitre 5, la validation de l'algorithme de clustering présenté dans cette thèse repose principalement sur l'interprétabilité de la partition finale d'un point de vue acquisitionniste.

Le rééquilibrage des effectifs par LM dans chacun des tirages permet une meilleure visualisation de l'influence de la LM dans la constitution des cores. L'un des principaux résultats obtenus concerne le regroupement quasi exclusif des allemands dans le core majoritaire. Comme nous le constatons en figure 6.5,

sur 20 allemands, 80% d'entre eux (soit 16) sont systématiquement classés dans le core majoritaire. Bien que ce core soit le plus important obtenu, on remarque que celui-ci regroupe relativement plus d'allemands que d'individus d'autres LMs. Les français, italiens et néerlandais y sont également représentés, de manière approximativement similaire, et enfin les anglais, y sont sous-représentés.

Ces résultats nous apportent une information intéressante sur comment des individus ayant une langue maternelle différente rentrent dans le système grammatical d'une même LC différemment. Nous savons que le core majoritaire représente les individus ayant, entre autre, développés une stratégie de mise en relation forme fonction en morphologie et en morphosyntaxe.

Il semblerait que la présence d'un système de morphologie flexionnelle en L1 permet un premier contact avec le polonais moins abrupt. Les allemands sont les seuls de notre échantillon à posséder un système casuel dans leur LM, bien que moins riche que celui de la LC, et semblent donc plus sensibles aux variations du syntagme nominal dans l'input. Ils ont pu développer, grâce à leur LM, leur capacité d'identification d'une désinence sur un item, et savoir *a priori* que cette désinence est liée au statut de l'item dans la phrase. Cette sensibilité n'est pas réservée qu'aux allemands puisque les autres LMs sont également représentées dans ce core, bien qu'en moindre proportion, spécifiquement pour les anglais. Les 20% restants des allemands se répartissent de façon équilibrés entre les outliers (10%) et le core comprenant les apprenants adoptant une stratégie positionnelle dans l'analyse des constituants d'une phrase, mais étant sensibles à la morphologie flexionnelle (10%).

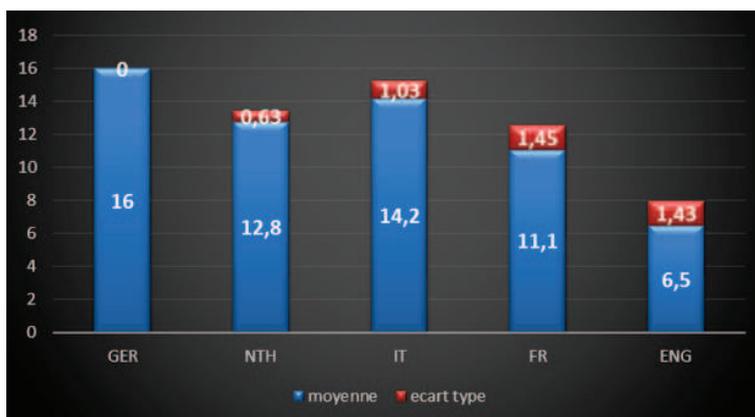


FIGURE 6.5 – Moyenne et écart type des individus appartenant au core majoritaire en fonction de leur LM pour les 10 tirages aléatoires effectués

Du fait de la sous représentation des anglais dans le core majoritaire, ceux-ci sont sur représentés dans les autres cores. Cependant, on constate tout de même que 17% d'entre eux en moyenne appartiennent au core des individus présentant une acquisition de la LC dénuée de morphologie (noté "core double", cf. figure 6.6) contre seulement 5% des italiens, qui constituent pourtant la deuxième LM la plus représentée dans ce core.

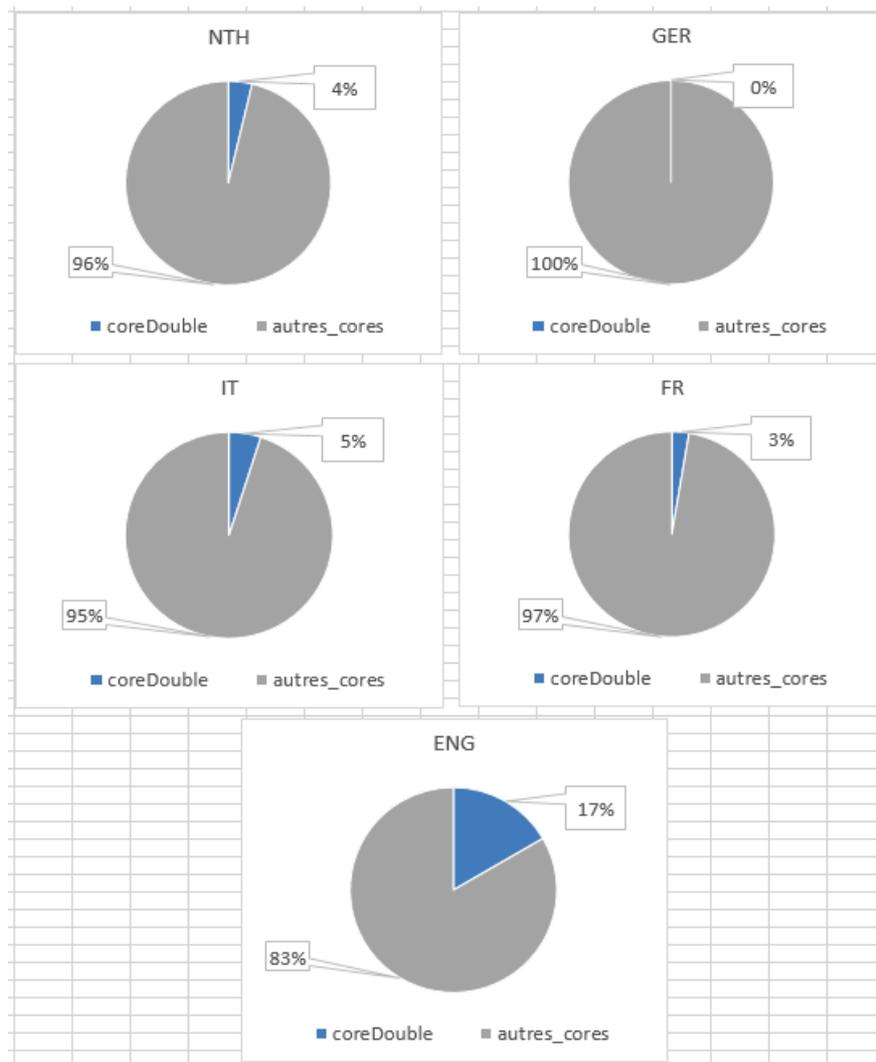


FIGURE 6.6 – Constitution moyenne du core (5) "double" par LM.

Il est intéressant de constater que ces résultats concordent avec des études qualitatives menées par des chercheurs travaillant sur les données VILLA et se basant sur différentes tâches, y compris les tâches de production. Ainsi, Saturno et Watorek ([Saturno et Watorek, 2020]) comparent l'acquisition de la distinction nominatif *vs.* accusatif par les apprenants des 5 LMs. Ces auteurs se basent sur les données provenant des tests *Picture verification* et *Sentence Imitation* et montrent que les apprenants anglophones optent tout au long de la période d'observation (14h) pour la stratégie positionnelle sans recourir aux connaissances sur le marquage casuel dans la répétition et dans la compréhension des phrases. En revanche, les germanophones passent très rapidement vers la stratégie morphosyntaxique. Watorek ([Watorek et al., 2020]) analyse les productions semi-libres (*Route direction*) de ces mêmes apprenants en montrant que les germanophones sont de loin les meilleurs dans l'analyse du système casuel du polonais. Non seulement ils produisent plus d'items avec le marquage casuel approprié mais également ils commettent des erreurs qui prouvent la prise de conscience de l'importance du marquage casuel en LC. En ce qui concerne les apprenants francophones, italo-phones et néerlandophones, ces auteurs observent des stratégies mitigées.

Ces résultats se laissent expliqués par des contrastes typologiques entre les LMs des apprenants VILLA (cf. chapitre 2). Parmi les LMs considérées dans VILLA, le système casuel est seulement présent en allemand (4 cas) contrairement aux autres LMs. De plus, l'anglais est une langue où la morphologie flexionnelle est la plus pauvre et l'ordre des mots SVO le plus rigide.

Cette mise en parallèle de notre analyse des cores intégrant les LMs des apprenants est donc renforcée par la concordance avec les analyses disponibles des données VILLA. Les similarités observées entre les analyses effectuées par les chercheurs participant au projet VILLA et les patterns sous-jacents constatés pour chaque *groupe d'apprenants comparables* tendent à valider cette étape de l'algorithme. De plus, certains de ces groupes amènent à une réflexion sur la construction même des tâches utilisées dans le projet et leur complexité relative (core 3 dit morphologique), et permettent d'observer la présence d'apprenants dont la stratégie d'apprentissage et son succès semblent spécifiquement corrélés aux caractéristiques des items de l'input (core 4).

6.1.2 Raffinement des cores : création des clusters

La répartition des apprenants en cores nous assure d'un regroupement d'individus comparables, c'est-à-dire qui procède de la même stratégie de réponses aux différents tests d'acquisition de la LC, et qui possèdent *vs.* ne possèdent pas une sensibilité exacerbée aux différentes caractéristiques des items utilisés dans les tâches. Désormais l'objectif est d'obtenir des clusters d'apprenants similaires. La notion de similarité est plus restrictive que la notion de comparabilité. Dans ce dernier cas on cherche à séparer les apprenants déployant une stratégie grammaticale pour répondre à une sollicitation en LC ou des apprenants spécialement influencés par l'input, des apprenants adoptant des stratégies de réponses suivant des principes universaux d'organisation discursives, ou répondant aléatoirement, et étant relativement peu influencés par les caractéristiques des items des tâches.

Cependant, comme nous l'avons constaté dans la section précédente, parmi les apprenants sensibles aux variations des syntagmes nominaux, certains ont réussi à associer un marquage casuel approprié au contexte alors que d'autres non. Ces différents apprenants se retrouvent tout de même dans les mêmes cores. L'objectif de cette section est d'analyser l'homogénéité intra core afin de déterminer l'existence de clusters à l'intérieur de chacun d'entre eux et de procéder à leur découpage si nécessaire.

Par définition les individus au sein d'un même core, bien que comparables, ne possèdent donc pas nécessairement le même profil. Nous avons identifié dans l'étape précédente les domaines où leurs spécificités s'expriment. Il est maintenant possible que pour un même domaine ces spécificités s'expriment différemment. La seconde étape dans l'élaboration d'un profil d'apprentissage pour nos apprenants, consiste en une division au sein d'un même core, sous certaines conditions, entièrement basée sur des mesures de distances numériques.

6.1.2.1 Choix du nombre et création de clusters à l'intérieur d'un core

Pour les techniques de clustering se basant sur des fonctions objectives, le plus grand défi consiste d'abord dans le choix du nombre de clusters à élaborer. Étant une technique de classification non supervisée, ce choix doit être spécifié en début d'analyse. L'expérimentateur peut avoir une idée précise, issue de ses connaissances théoriques, du nombre de clusters qu'il souhaite, mais la plus part du temps ce n'est pas le cas, d'où la volonté de voir émerger des données un pattern pouvant nous renseigner sur le nombre de sous-groupes présents dans notre base de données.

Afin de nous familiariser avec la composition de nos cores, et ainsi déterminer si nous devons ou non diviser nos cores, nous souhaitons utiliser une technique de visualisation. Les techniques VAT (*visual assessment of cluster tendency*) et sa version améliorée iVAT (cf. chapitre 5) sont des outils de visualisation largement répandus. Le principe de ces outils est de transformer en nuance de gris la distance numérique entre chaque objet (individu) de notre ensemble de données. Chaque objet devient donc un carré gris plus ou moins clair et est ensuite affichée sur une diagonale. Cette diagonale est réordonnée afin de regrouper les objets les plus

proches numériquement. Ainsi l'apparition de carré gris englobant est la trace de la présence d'un cluster possible.

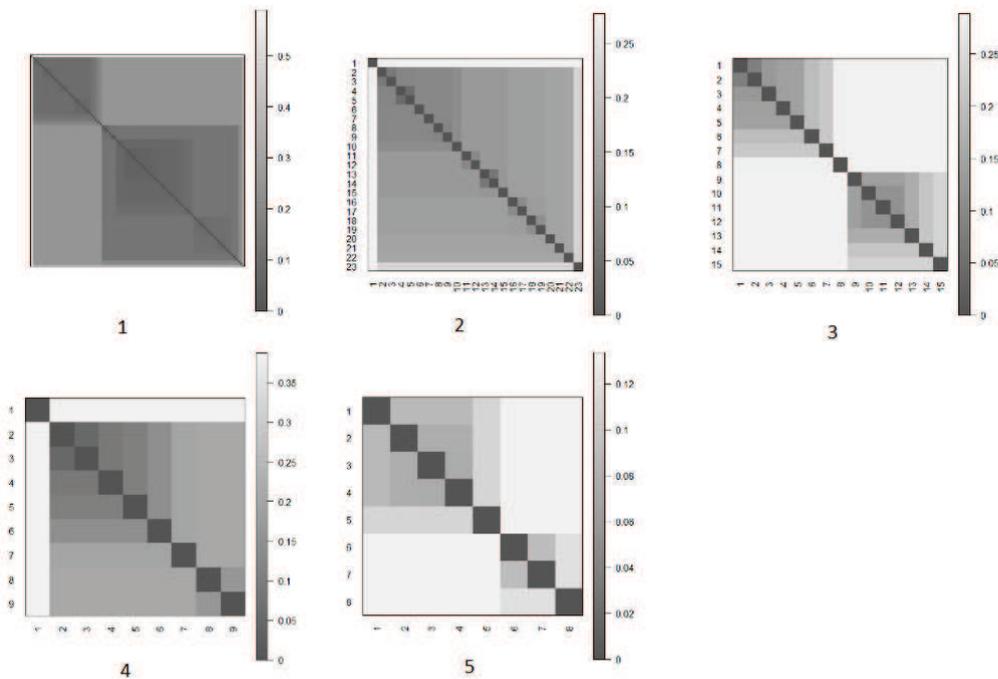


FIGURE 6.7 – Résultats de la méthode iVAT pour les cores 1 à 5 ; la correspondance entre le degré de gris et la distance numérique entre deux objets est légendée sur la droite de chaque core. Échantillon 1

Les images générées par l'outil iVAT (cf. figure 6.7) permettent de mieux appréhender la structure numérique de chaque core. L'interprétation n'est pas toujours évidente, et n'apporte qu'un début de réflexion que nous devons compléter.

Le core n°3 semble receler deux clusters, plus un outlier, bien que l'homogénéité à l'intérieur de chacun de ces clusters semble plus faible que pour le core n°1.

Le core n°1 semble également receler deux clusters distincts. Néanmoins, un des désavantages de l'iVAT s'illustre dans ce core par ce que nous pouvons observer sous la forme de carrés apparaissant à l'intérieur d'un autre. Ici le second cluster visible a l'air de receler lui-même deux clusters. Ainsi la question de la sensibilité de cet outil se pose. A quel niveau de granularité souhaitons-nous découper les données ?

Les résultats des autres cores démontrent plus explicitement l'intérêt de l'outil iVAT. Pour le core n°2, peu de structures internes visibles apparaissent, indiquant une grande homogénéité intra core. Le core n°2, regroupant les individus adoptant une stratégie positionnelle dans l'attribution du statut du constituant dans une phrase malgré leur sensibilité à la morphologie de la LC, ne sera donc pas soumis à une seconde étape de division, et est désormais considéré comme un cluster appartenant à notre classification finale.

Les images des core 4 et 5 sont plus complexe. Dans le core 5, on peut déceler un sous-groupe, mais celui-ci semble en lui-même receler encore un sous-groupe. Au vu du faible nombre d'individus contenu dans le core 5, et en l'absence de sous-groupes fortement distincts, nous choisissons également de le rendre indivisible, quitte à nous y intéresser de plus près par la suite, par le biais d'une étude qualitative de ces composants. Le core 4 apparaît homogène excepté un individu, un outlier potentiel dont la présence est évidente sur l'image. Ce core est également rendu indivisible, un seul individu ne pouvant constituer un cluster.

Cette étape est très utile dans le travail d'analyse des données, mais nécessite néanmoins l'intervention

humaine, et n'est donc pas automatisée. Cependant, elle fournit un premier indice sur la composition des cores. Trois des cores ont d'ores et déjà été écartés de la nécessité d'un second passage de division, par le biais de comparaisons purement numériques, pour l'obtention d'une classification finale.

Comme précédemment mentionné, l'outil iVAT a pour fonction première d'observer si un ensemble de données peut et devrait être soumis à une classification de ses composants par visualisation de sa structure interne. Une des conséquences de cette représentation visuelle est de procurer une intuition quant au nombre c de sous-groupes contenus dans les données. Cependant, cette intuition est imprécise du fait de la superposition de clusters. Afin de compléter notre réflexion, un grand nombre d'indices issus de la littérature permettent de déterminer le choix optimal quant à c , le nombre de cluster présents dans un ensemble de données. Ils reposent tous sur le même principe d'évaluation de l'homogénéité intra cluster (*compactness*) et de l'isolation inter cluster (*separation*). Ces différents indices s'appliquant aux méthodes classiques de *clustering* sont rassemblés dans un paquet du logiciel R appelé NbClust (cf. [Charrad et al., 2014]). Au total 30 indices sont inclus, mais tous ne sont pas appliqués, en fonction des données. La présence d'un si grand nombre d'indices n'est pas un avantage car le consensus est rarement atteint, la règle de la majorité est adoptée par défaut pour le choix final du nombre optimal pour c .

Pour le core n°1, parmi les 24 indices utilisés :

- 11 proposent 2 comme le meilleur nombre de clusters
- 4 proposent 3
- 5 proposent 4
- 1 propose 10
- Et enfin 3 proposent 15.

De plus, l'indice D ([lebart et al., 2006]), qui est une méthode graphique, nous indique que $c = 5$ (cf. figure 6.8).

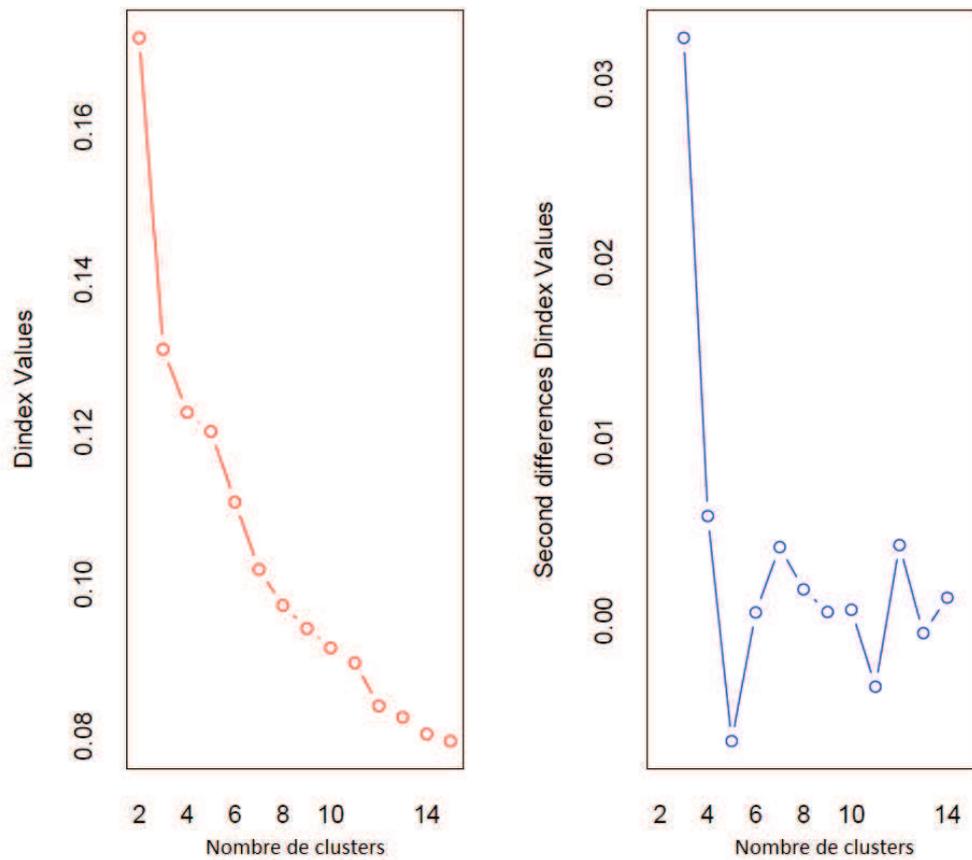


FIGURE 6.8 – Indice D obtenu pour le core 1 ; fonction NbClust() du logiciel R

Supporté par le choix de la majorité des indices, et par l'intuition originale fournie par la visualisation de la structure de clustering fournie par l'outil iVAT, nous choisissons $c = 2$, et procédons au découpage du core 1 en suivant l'algorithme des k-moyennes avec pour mesure de distance une distance euclidienne classique (cf. chapitre 3).

Pour le core 3, la visualisation de la structure du core indique la présence de deux sous groupes et d'un outlier (cf. figure 6.7). Parmi les 24 indices de la méthode NbClust() appliqués au core 3 :

- 8 proposent 2 comme le meilleur nombre de clusters
- 5 proposent 3
- 1 propose 4
- 1 propose 6
- 3 proposent 10
- Et enfin 5 proposent 13.

L'indice D est peu concluant pour ce core (cf. figure 6.9), la courbe des premières différences (*first differences*, en rouge) ne présentant pas de coude.

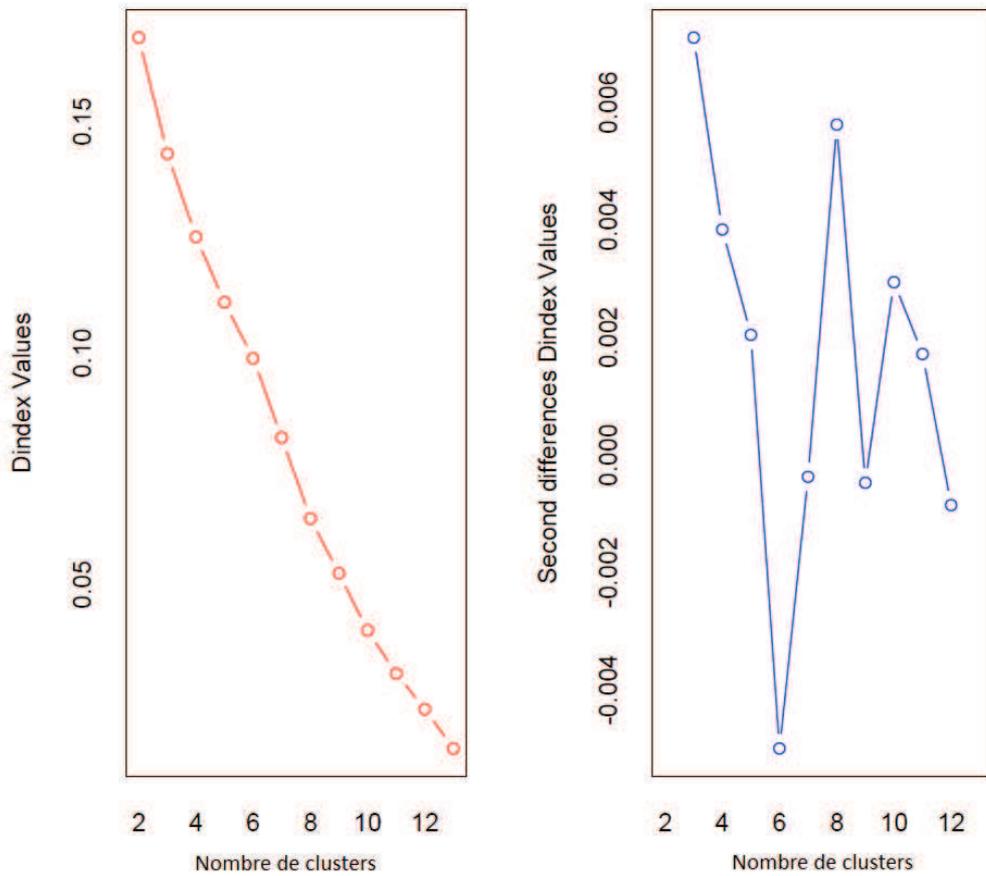


FIGURE 6.9 – Indice D obtenu pour le core 3

Ici la loi de la majorité s'affaiblit en validité puisque seul 8 indices concordent sur $c = 2$. La présence apparente d'un outlier visible sur l'image fournie par l'iVAT tend à indiquer que, pour $c = 3$, un des trois clusters ne sera constitué que d'un individu. Pour tenter de trancher définitivement nous effectuons deux classifications, la première avec $c = 2$, et la deuxième avec $c = 3$. Au vu du faible nombre d'apprenants dans le core 3, nous nous permettons de reporter ces deux classifications dans les figures 6.10 et 6.11.

```
Clustering vector:
1105 2111 2119 3105 3113 3117 3118 3214 5205 5210 5212 1210 2201 2206 2214
  2   2   1   1   2   1   1   1   1   2   2   1   2   2   1
```

FIGURE 6.10 – Partition des apprenants (référéncés avec leur identificateur) du core 3 pour $c = 2$ avec l'algorithme des K-moyennes

```
Clustering vector:
1105 2111 2119 3105 3113 3117 3118 3214 5205 5210 5212 1210 2201 2206 2214
  3   2   1   1   2   1   1   1   1   3   2   1   2   2   1
```

FIGURE 6.11 – Partition des apprenants (référéncés avec leur identificateur) du core 3 pour $c = 3$ avec l'algorithme des K-moyennes

Pour $c = 3$ le troisième cluster est bel et bien constitué uniquement d'un seul individu. Après une analyse détaillée de ces deux partitionnements, en choisissant $c = 3$ au lieu de $c = 2$ la part de variance total expliquée par la partition passe de 59,5% à 66,6%. Ce pourcentage doit être manié avec précaution puisqu'il augmente naturellement avec le nombre de clusters c . L'individu 1105 présente une variation unique dans le panel des différences entre les individus du core 3. Malgré tout, un individu ne pouvant constituer un cluster, nous choisissons de l'intégrer au cluster 2 du core 3. Il est envisageable que cette différence entre l'individu 1105 et les autres individus du sous groupe soit significative, mais la taille de l'échantillon de données ne nous permet pas de la considérer comme facteur de la classification. Avec un plus grand échantillon il serait éventuellement possible d'observer si cette différence se retrouve chez d'autres apprenants, et d'ainsi les isoler des autres dans la classification finale.

Au final, cette étape a permis de distinguer deux groupes dans le core majoritaire. Ce core, regroupant plus de la moitié des apprenants, est caractérisé par les apprenants entrés dans le système de la LC. En effet, ces derniers analysent les désinences des items, et ont donc intégré la richesse de la morphologie flexionnelle dans leurs traitements de la LC. Cependant, certains d'entre eux échouent à utiliser la désinence appropriée au paradigme sollicité. Ces derniers sont donc isolés du reste du groupe, et intègrent le cluster 6 dans la partition, tandis que les autres apprenants du core majoritaire sont regroupés dans le cluster 7.

Le core 3 a également été raffiné, créant ainsi le cluster 1 et le cluster 2. Ces deux clusters regroupent les apprenants sensibles aux variations des syntagmes nominaux et sachant les utiliser pour la caractérisation de leur rôle dans une phrase, mais échouant à la tâche de morphologie GJ, où, malgré leur perception potentielle de la variation des désinences des items en fonction de son contexte, ils ne traitent pas une phrase grammaticalement incorrecte comme telle. Cependant, l'opposition des paradigmes nominatif *vs.* accusatif, permettant de marquer le sujet *vs.* l'objet dans la phrase, est utilisée par ces apprenants. L'utilisation restreinte de deux items pour cette tâche ("frère" et "sœur"), alternant chacun le rôle de sujet et le rôle d'objet, a sûrement procédé à la construction de règles morphosyntaxiques en LC pour ces apprenants. La distinction entre le cluster 1 et le 2 s'effectue sur l'association du rôle du constituant avec le paradigme approprié. En effet, les apprenants du cluster 1 ont inversé l'association forme-fonction, associant le rôle de sujet avec un marquage casuel à l'accusatif, et le rôle d'objet de la phrase avec un marquage casuel au nominatif. Cette étape de l'algorithme permet donc de les distinguer par la création de deux clusters séparés. Cette inversion est également la raison principale du découpage du core 1 en deux clusters.

C'est donc la tâche jugeant de la capacité des apprenants à traiter les flexions liées la morphosyntaxe en LC qui est principalement responsable du raffinement des cores en clusters. Cette étape nous permet de saisir l'importance de ce niveau d'analyse linguistique dans l'introduction de variabilité dans les parcours acquisitionnel des apprenants, importance relativement plus forte que les autres tests traitant d'autres niveaux linguistiques. La partition obtenue par cette étape est représentée dans la figure 6.12.

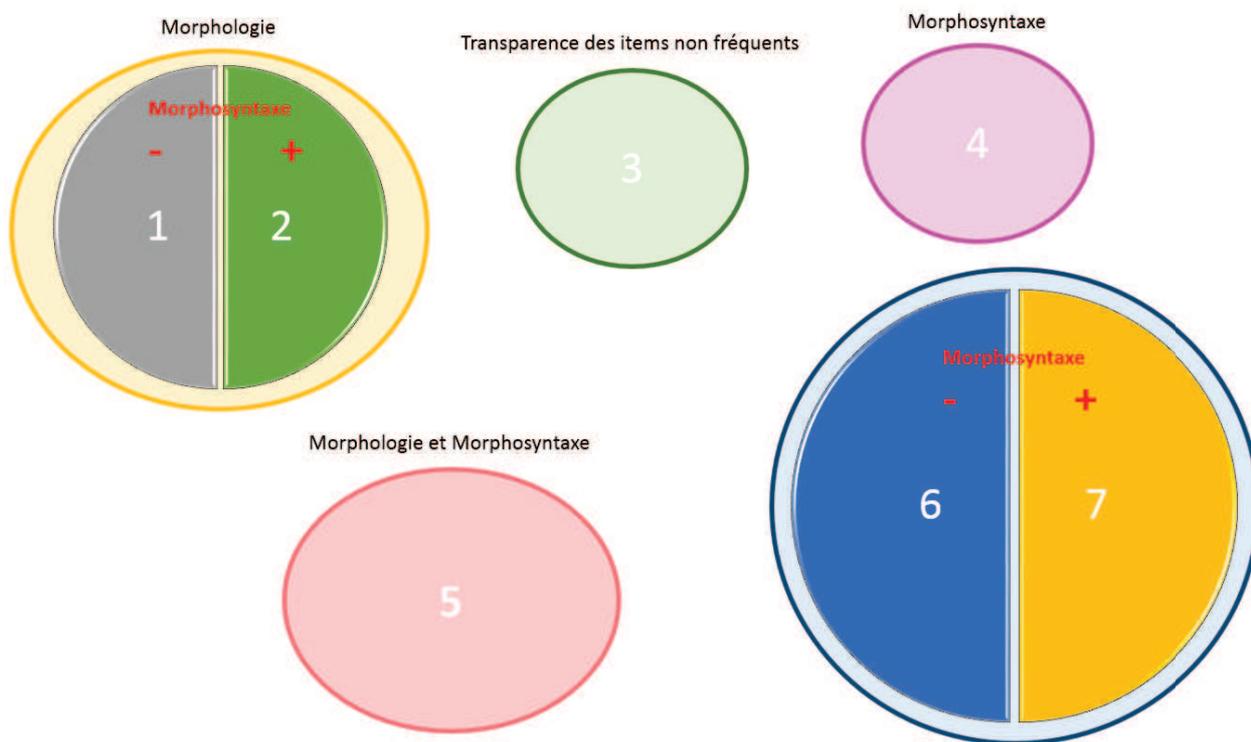


FIGURE 6.12 – Partition obtenue après raffinement des cores en clusters. Échantillon 1

6.1.2.2 Intégration des outliers et partition finale

L'étape finale dans l'établissement des profils des apprenants est l'intégration des outliers aux clusters. Cette étape repose sur le relâchement des contraintes de comparabilité binaire (comparable *vs.* incomparable) et d'appartenance unique (à un seul sous-groupe). La mesure de comparabilité entre deux apprenants, au lieu d'être discrète (0 ou 1) devient continue (entre 0 et 1). Autrement dit nous fuzzifions l'algorithme afin d'intégrer les outliers aux clusters précédemment établis. Pour chaque outliers nous obtenons donc un score d'appartenance à chaque cluster. Ce score d'appartenance correspond à la moyenne des scores de comparabilité entre un outlier et tous les individus du cluster, obtenu par la méthode `isComparableFuzzy()` (cf. chapitre 5).

La partition à ce stade de la classification est composée de 7 clusters, et 8 outliers. L'intégration des outliers nous permettra spécifiquement d'analyser quelles sont les dimensions qui ont empêché leur introduction première dans la partition. A ce stade, nous avons également calculé ce score de comparabilité entre les outliers et les *cores* obtenus après la première étape du processus de partitionnement, à des fins d'observation du comportement de l'algorithme. La figure 6.13 rapporte les scores obtenus.

Pour les individus 1216, 2113 et 2217 le résultat est clair et ils sont intégrés au cluster 5 avec un score de comparabilité de 0.91 (cf. chapitre 5 algorithme 5.4). Pour les autres outliers le résultat est plus délicat. Premièrement, on observe qu'il aurait été plus pertinent d'effectuer l'intégration des outliers lorsque la partition n'était pas encore divisée en clusters mais constituée de *cores*. En effet, la mesure de comparabilité repose sur les mêmes critères que le processus de création des contraintes *cannot-link*, ainsi, pour certains des outliers, l'intégration à un *core* aurait été claire, alors que l'intégration à un cluster est compromise par des degrés d'appartenance équivalents pour plusieurs d'entre eux. Parallèlement à cela, on observe que pour deux des outliers, bien que leur score de comparabilité soit élevé, il existe aussi une égalité entre deux

6.2. Deuxième échantillon : Quel transfert du traitement vers la production de la morphologie en LC ?

outliers	1108	1216	2113	2217	2219	3205	4108	4112
core 1	0,55	0,55	0,55	0,55	0,55	0,91	0,55	0,91
core 2	0,91	0,91	0,91	0,91	0,33	0,55	0,91	0,55
core 3	0,33	0,33	0,33	0,33	0,91	0,55	0,33	0,55
core 4	0,91	0,33	0,33	0,33	0,33	0,55	0,91	0,55
core 5	0,55	0,55	0,55	0,55	0,55	0,33	0,55	0,33
cluster 1	0,33	0,33	0,33	0,33	0,91	0,55	0,33	0,55
cluster 2	0,33	0,33	0,33	0,33	0,91	0,55	0,33	0,55
cluster 3	0,91	0,33	0,33	0,33	0,33	0,55	0,91	0,55
cluster 4	0,55	0,55	0,55	0,55	0,55	0,33	0,55	0,33
cluster 5	0,91	0,91	0,91	0,91	0,33	0,55	0,91	0,55
cluster 6	0,55	0,55	0,55	0,55	0,55	0,91	0,55	0,91
cluster 7	0,55	0,55	0,55	0,55	0,55	0,91	0,55	0,91
gagnant_core	core 2 ou 4	core 2	core 2	core 2	core 3	core 1	core 2 ou 4	core 1
gagnant_cluster	cluster 3 ou 5	cluster 5	cluster 5	cluster 5	cluster 1 ou 2	cluster 6 ou 7	cluster 3 ou 5	cluster 6 ou 7

FIGURE 6.13 – Score de comparabilité des outliers temporaires avec les différents *cores* et clusters.

clusters, mais également entre deux *cores* (*cores* n’ayant pas été divisés à l’étape de clusterisation). Ainsi, bien que le seuil d’acceptabilité pour le score de comparabilité soit largement dépassé, les 5 outliers restant ne sont pas ajoutés à la partition finale.

L’étude du premier échantillon apporte des informations sur les différentes sensibilités des apprenants aux systèmes morphologique et morphosyntaxique de la LC. Nous constatons que les apprenants atteignent des niveaux de traitement différents du système casuel au cours de leur apprentissage. Le troisième versant indissociable de la caractérisation de l’acquisition d’une LE est l’analyse des productions orales des apprenants, c’est ce que nous nous proposons de faire à travers l’étude d’un deuxième échantillon.

6.2 Deuxième échantillon : Quel transfert du traitement vers la production de la morphologie en LC ?

Le projet VILLA comporte des tests allant des productions ciblées aux productions semi-guidées. Nous avons déjà évoqué dans le chapitre 4 l’importance des écarts, à ce niveau d’acquisition, entre les performances des apprenants dans une tâche sollicitant un traitement, une tâche sollicitant une production ciblée et semi libre. Ainsi, il apparaît intéressant d’apporter un nouvel angle d’analyse pour la caractérisation de ces écarts, et d’observer différents comportements d’apprenants dans leur capacité à transférer leurs connaissances mobilisées dans une tâche de compréhension vers une utilisation en production orale.

Dans la batterie de tests à laquelle ont été soumis les participants du projet, il existe quatre tâches nécessitant de la part de l’apprenant une production orale. *Route Direction* (RD) et *Film Retelling* (FR) sont deux tâches de production semi guidée où l’apprenant doit construire un discours sur la base d’un support scénarisé ([Dimroth et al., 2013]). Les tâches *Oral Question-Answer* (OQA) et *Sentence Imitation* (SI) sont des tâches ciblées visant à attester la capacité de production d’un paradigme précis par les apprenants. La tâche OQA requiert de la part des apprenants une production orale à une question sollicitant une réponse soit à l’instrumental soit au nominatif. Les items utilisés pour ces questions peuvent être au féminin ou au masculin. Cette tâche teste les capacités de production d’une morphologie appropriée au contexte de sollicitation. Le paradigme utilisé correspond au même paradigme que celui de la tâche GJ. La tâche SI requiert la distinction des items au nominatif *vs.* accusatif pour la caractérisation du statut du

constituant dans la phrase et relève donc du système morphosyntaxique de la LC. Elle constitue le pendant de la tâche PV testant la compréhension et la mobilisation de ce paradigme dans une activité de traitement. Cependant, la tâche SI est une tâche de répétition de phrase, et ne peut pas être considérée comme une tâche de production au même titre que OQA. Cet état de fait sera pris en compte dans nos analyses.

Cet échantillon regroupe les résultats des apprenants de 4 LMs (français, italiens, anglais et allemands) ayant reçus une instruction *Meaning Based* du Polonais (cf. chapitre 2). Les résultats des apprenants ayant reçus l'instruction *Form Based* n'étant pas encore codés par les différents partenaires du projet VILLA, et l'enregistrement des apprenants néerlandophones pour la tâche SI ayant été corrompu, cet échantillon se constitue au final de 69 apprenants. Contrairement aux données des tâches de traitement, celles de production, avant d'être analysées, doivent être transcrites et codées.

Dans cet échantillon les apprenants sont caractérisés par une valeur numérique correspondant au score qu'ils ont obtenus aux épreuves en LC testant leur capacité de traitement de son système morphologique et morphosyntaxique (tâches GJ et PV), et leur capacité de production en accord avec ces deux systèmes (tâches OQA et SI). Chaque apprenant est donc évalué pour l'instant par 4 valeurs numériques, correspondant à ses résultats pour ces 4 tests :

- Capacité de traitement de la morphologie
- Capacité de production de la morphologie
- Capacité de traitement de la morphosyntaxe
- Capacité de production de la morphosyntaxe

Il convient d'identifier à travers le prisme du score obtenu par l'apprenant dans une tâche de production quelles connaissances du système morphologique et morphosyntaxique de la LC celui-ci mobilise. L'association d'un score numérique à une stratégie de réponse n'est pas la même pour une activité de traitement et pour une activité de production. En effet, la stratégie dite aléatoire potentiellement appliquée par un apprenant lors d'une activité sollicitant le choix entre deux réponses possibles (image A *vs.* image B, phrase 1 *vs.* phrase 2, etc...) n'existe pas dans le cadre d'une tâche sollicitant une production originale. Chaque réponse de l'apprenant, si elle existe, correspond à la mise en œuvre des règles sous-jacentes au lecte d'apprenant à un moment donné d'acquisition. Ainsi, une production orale est moins ambiguë dans son interprétation. Pour observer des différences dans les réponses des apprenants aux tâches OQA et SI une autre lecture du score numérique obtenu par des apprenants à ces tâches est nécessaire, notamment en terme de comparabilité (cf. chapitre 5).

L'objectif de l'analyse de cet échantillon est d'observer des différences dans la capacité des apprenants à transférer leurs compétences en traitement de la LC vers une production en LC. Les deux tâches choisies sont des tâches dites guidées. Les tâches guidées permettent d'isoler la production d'un paradigme précis en ne laissant que peu de marge de manœuvres à l'apprenant pour construire sa production. Ainsi le but communicatif de telles productions est faible voir nul, ce qui permet à l'apprenant de se concentrer plus sur la forme de sa réponse (notamment en terme de morphologie flexionnelle), et moins sur le message véhiculé par sa production.

Le parallèle entre ce type de tâche de production et des tâches de traitement est donc plus facile à opérer (contrairement à des tâches semi libres). Nous ajoutons ainsi deux dimensions caractérisant les apprenants :

- le transfert des capacités de traitement du système morphologique vers sa production
- le transfert des capacités de traitement du système morphosyntaxique vers sa production

La comparaison des performances des apprenants dans les deux types de tâches permet de voir si les connaissances mobilisées par l'apprenant dans une tâche de traitement sous la forme d'une stratégie de réponse donnée, correspondent à celles mobilisées dans une tâche sollicitant une production. Plus précisément, si un apprenant a été caractérisé par une stratégie permettant de constater qu'il a commencé l'analyse du système de morphologie flexionnelle de la LC au vu de ses résultats dans les tâches GJ et PV, est-il capable ou non de procéder de manière similaire lorsqu'il doit produire en LC ?

De nombreux travaux en acquisition et psycholinguistique montrent que le traitement et la compréhension de la langue maternelle chez les enfants et des langues secondes chez les apprenants adultes précèdent la capacité de production. Ainsi, le transfert des connaissances mobilisées en LC en traitement à la production en LC peut être relativement difficile. Dans les tâches de traitement, les apprenants montrent une sensibilité précoce à la morphologie nominale de la LC (cf. [Hinz et al., 2013]) tandis que les productions de ces apprenants se caractérisent par un emploi sporadique de la flexion nominale productive ([Watorek et al., 2017, Watorek et al., 2020]).

Cependant, nous observons, grâce à notre étude, un certain nombre de variations dans la capacité de transfert en fonction du paradigme linguistique testé et des caractéristiques des tâches. En effet, non seulement les deux paradigmes considérés peuvent donner lieu à des difficultés différentes et donc à des réponses différentes, mais également les tâches elles mêmes comportent des difficultés variables.

Le paradigme relevant de la morphologie nominale, instrumental *vs.* nominatif, est testé en production par la tâche de production ciblée (OQA) où l'apprenant produit une réponse à une question précise. Sa réponse est donc contextualisée bien que ce contexte soit minimal. En revanche, le paradigme morphosyntaxique, où l'opposition nominatif *vs.* accusatif est combinée avec l'ordre de mots, est testé par la tâche de répétition de phrase (SI). La répétition d'une phrase décontextualisée ne constitue pas la même activité qu'une tâche de réponse orale à une question (OQA). Nous devons donc tenir compte de cette différence entre les deux tâches de production considérées ici.

L'algorithme de classification semi supervisée est donc appliqué à un échantillon de 69 étudiants caractérisés par 6 dimensions, calculées à partir de 4 tâches différentes testant le niveau de connaissance de l'apprenant sur le système morphologique et morphosyntaxique de la LC, deux sollicitant un traitement de la LC, et deux nécessitant une production orale de l'apprenant.

6.2.1 Création des cores par l'utilisation de l'ensemble C pour le deuxième échantillon

Le processus de création de l'ensemble C de contraintes *cannot-link* permet de créer la matrice de comparabilité des 69 apprenants sur les dimensions retenues. Ainsi, si tous les apprenants sont jugés incomparables entre eux, l'ensemble C contiendra 2346 contraintes.

Pour cet échantillon, la matrice binaire de comparabilité est constitué de 4414 "0", soit un ensemble C constitué de 2207 contraintes *cannot-link*. Dès à présent nous pouvons constater que les dimensions retenues de caractérisation du parcours acquisitionnel des apprenants dans cet échantillon sont bien plus créatrice de variabilité que celles retenues pour le premier échantillon. La figure 6.14 indique que toutes les dimensions retenues ont été responsables de la création de contraintes *cannot-link*. Les dimensions créant le plus d'incomparabilité entre les apprenants sont celles caractérisant leurs capacités de production de la morphologie en LC. En effet, les résultats des apprenants dans des activités de production sont plus de deux fois plus responsables d'un motif d'incomparabilité entre deux apprenants, que leur équivalent dans des activités sollicitant un traitement du système flexionnel du polonais (22,6% de responsabilité en moyenne pour les activités de production contre 10% pour celles de traitement dans la création du total des contraintes

de l'ensemble C).

Même si les apprenants néerlandais ainsi que la moitié des apprenants français, anglais et italiens (ceux ayant suivi la méthode d'enseignement *form based*) sont exclus de cet échantillon, il ne semble pas exister de différence dans la variabilité imputée aux activités de traitement en morphologie et en morphosyntaxe entre les apprenants du premier et ceux du deuxième échantillon. Cet état de fait ne contredit pas, à défaut de garantir, la considération et l'interprétation des analyses de cet échantillon comme un prolongement de celles effectuées sur l'échantillon 1.

La différence entre l'homogénéité relative des résultats en traitement, comparée à l'hétérogénéité des résultats en production, découle directement de la difficulté du transfert des capacités de traitement (perception et compréhension) d'un paradigme linguistique vers sa production appropriée. C'est le constat le plus notable en résultat de cette étape de l'algorithme, constat à l'origine du taux élevé d'incomparabilité. Pour un potentiel maximal de 2346 contraintes, 2207 constituent l'ensemble C de cet échantillon. Cet indice laisse entrevoir un faible nombre d'apprenants intégrés aux cores et un grand nombre d'outliers.

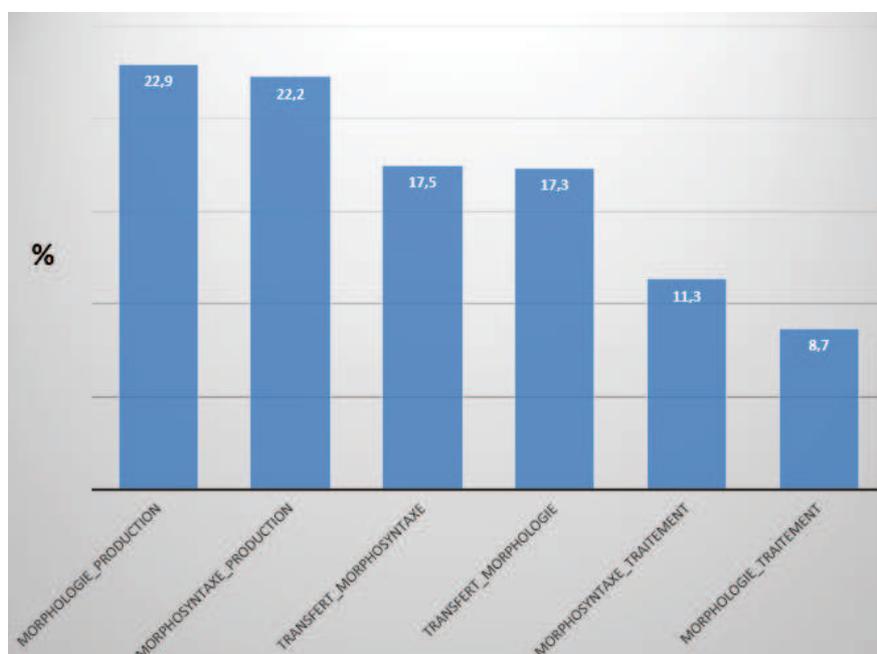


FIGURE 6.14 – Influence relative des différents attributs dans la création de l'ensemble de contraintes C .
Échantillon 2

6.2.1.1 Cores obtenus

Après élimination des cores de moins de 5 individus, individus reclassés en outliers, nous obtenons seulement 2 cores et 48 outliers sur un total de 69 apprenants. La taille relative de chaque core est illustrée en figure 6.15.

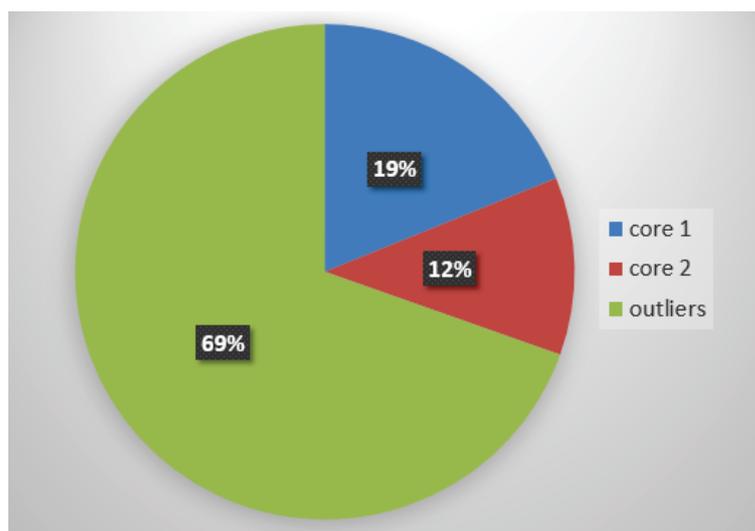


FIGURE 6.15 – Part relative de chaque core et des outliers. Échantillon 2

Parallèlement au nombre impressionnant d'outliers à cette étape du processus de catégorisation, l'absence de core majoritaire reflète l'hétérogénéité de cet échantillon, le plus grand core comprenant 19% des apprenants classifiés, c'est à dire 13 apprenants.

La figure 6.16 est un tableau représentant les appartenances des apprenants de chaque core aux différents SEFs. Ces SEFs, rappelons le, représentent des stratégies de réponses différentes aux 4 tâches, ainsi qu'un transfert différent en morphologie et en morphosyntaxe. Les stratégies de réponses différentes en fonction des tâches et leur SEF correspondant sont détaillés en chapitre 5. Pour ce qui est du transfert, que ce soit d'un paradigme morphologique ou d'un paradigme morphosyntaxique, on identifie trois états possibles, et donc trois SEFs possibles :

1. L'apprenant est meilleur en production qu'en traitement (SEF S_1), cas de figure peu probable au vu de la littérature en RAL.
2. L'apprenant mobilise les mêmes connaissances de la LC en traitement et en production (SEF S_2).
3. L'apprenant a démontré posséder certaines connaissances du système de la LC en traitement mais ne les mobilise pas en production (SEF S_3).

Le nombre de core, 2, en rapport avec le nombre de dimensions, 6, suggère que toutes les dimensions caractéristiques d'un apprenant n'ont pas leur importance dans la création des cores contrairement à ce que suggère la figure 6.14. En analysant simultanément la figure 6.16 et la figure 6.14, on constate que les capacités de traitement de la morphosyntaxe et de la morphologie en LC sont deux dimensions non pertinentes pour l'interprétation des cores, et donc sûrement responsables de la catégorisation temporaire de nombreux individus en outliers. En effet, la figure 6.16 montre que pour les deux cores constitués, les attributs caractérisant les performances des apprenants en traitement appartiennent au même SEF S_3 . Ainsi, tous les apprenants intégrés à un core démontrent une entrée réussie dans l'analyse du système flexionnel de la LC (appartenance au SEF S_3 ou S_1). C'est dans la restitution de cet acquis lors d'une tâche de production que les apprenants vont se différencier.

En ce qui concerne le core 2, tous ses apprenants semblent être à même de d'identifier dans des tâches de traitement la désinence de l'instrumental et celle du nominatif. En revanche, en production, les résultats des

	Morpho_c	Morpho-syntaxe_c		Morpho_p_Inst	Morpho_p_Nom	Morpho-syntaxe_p_SO	Morpho-syntaxe_p_OS		transf_MO	transf_MS
core 1	S1	0	S1	0	0	0	0	S1	0	0
	S2	0	zone floue	0	6	0	0	S2	13	13
	S3	13	S2	13	7	13	13	S3	0	0
core 2	S1	0	S1	0	8	8	6	S1	0	0
	S2	0	zone floue	0	0	0	2	S2	0	0
	S3	8	S2	8	0	0	0	S3	8	8
out	S1	1	S1	10	34	29	19	S1	1	5
	S2	10	zone floue	8	8	4	7	S2	25	20
	S3	37	S2	30	6	15	22	S3	22	23

FIGURE 6.16 – SEF d'appartenance des apprenants de chaque core pour les 6 dimensions du deuxième échantillon (morpho_p_I : morphologie, production de l'instrumental; morpho_p_N : morphologie, production du nominatif; ms_p_SO : morphosyntaxe, production des phrases SVO; ms_p_OS : morphosyntaxe, production des phrases OVS; transf : transfert).

apprenants de ce core montrent qu'ils ne produisent correctement que les phrases à l'instrumental. Pour ce qui est des phrases du test sollicitant un nominatif, la déclinaison utilisée est inappropriée ou inexistante. On observe donc ici le transfert de la capacité à traiter l'instrumental en production de cette forme. Cependant, la caractérisation d'un transfert réussi a été définie comme le succès à produire les déclinaisons des deux cas si les apprenants ont fait preuve d'une capacité à les discriminer en traitement. Ainsi, ces apprenants ont été jugés comme ne réussissant pas leur transfert des compétences en morphologie du traitement vers leur production et appartiennent donc au SEF S_3 pour cet attribut.

Les apprenants du core 2 sont clairement plus faibles lorsqu'il s'agit de mobiliser leurs connaissances de la morphosyntaxe du polonais en production. En effet, pour la tâche SI, ces apprenants échouent à produire des désinences appropriées que ce soit pour des phrases de type SO ou de type OS.

Ainsi, les apprenants du core 2 n'ont pas fait preuve d'une mobilisation de leurs connaissances acquises sur le fonctionnement de la LC, attestées par leurs résultats aux tâches en traitement, dans leurs productions orales, fait conduisant à leur appartenance au SEF S_3 pour leurs attributs de transfert en morphologie et également en morphosyntaxe.

Ces observations sont également intéressantes d'un point de vue méthodologique. En effet, la tâche SI ne semble pas être plus facile pour les apprenants ayant obtenus de bon score à la tâche PV. Pourtant, étant une tâche de répétition d'une phrase décontextualisée, les résultats obtenus à cette tâche devraient être proche de ceux obtenus à la tâche PV. Comme déjà énoncé dans le chapitre 2, une tâche de répétition n'est pas considérée comme une tâche de production classique. Si l'on accepte cette manière d'envisager SI, un core regroupant des apprenants ayant réussi à une tâche de traitement d'un paradigme linguistique mais ne réussissant pas à répéter des phrases en LC testant ce paradigme linguistique ne devrait pas apparaître, ce qui est pourtant le cas ici. L'absence du transfert en répétition de la connaissance du paradigme de l'instrumental *vs.* nominatif est un argument en faveur de la thèse qu'une tâche de répétition présente plus de difficulté pour l'apprenant qu'une tâche de compréhension.

Pour en revenir à l'analyse des apprenants constituant le core 2, leurs faibles performances dans des tâches de production sont à relativiser, du fait de leur réussite à produire de manière appropriée au contexte les formes de l'instrumental. La question qui se pose est de savoir d'où provient l'écart chez les apprenants entre leur traitement réussi des deux désinences (Instrumental et Nominatif) et leur production où seule la désinence de l'instrumental est réalisée en contexte approprié.

Une première explication peut provenir de la construction même de la tâche GJ, utilisée pour attester de l'acquisition du paradigme d'opposition du nominal *vs.* instrumental. Dans cette tâche de jugement de grammaticalité, les phrases tests comprennent soit des items déclinés à l'instrumental soit au nominatif. Seulement, seules les phrases à l'instrumental sont grammaticalement correctes. Ainsi, toutes les phrases construites au nominatifs doivent être jugées incorrectes par l'apprenant. Il est donc possible que la nature même de la tâche GJ ait induit un biais auprès de l'apprenant qui se sera ainsi entraîné à associer l'instrumental à une phrase grammaticalement correcte en LC.

Cependant, cette explication est liée au phénomène de l'entraînement à la tâche et ne prend pas en compte l'input reçu en classe, input comprenant également des phrases grammaticalement correctes à l'instrumental et au nominatif. Ainsi, une autre explication est développée par Rast et collaborateurs ([Rast et al., 2018]). Les auteures ont analysé les résultats des apprenants à la tâche OQA pour les mêmes apprenants que ceux inclus dans notre échantillon. Leur problématique initiale repose sur la description de caractéristiques de la LC permettant à l'apprenant débutant d'acquérir plus facilement sa morphologie flexionnelle. A cette fin, les auteures s'intéressent à la notion de saillance comme caractéristique potentiellement facilitatrice de la rétention et de la restitution de certaines désinences (leurs analyses sont détaillées en sous section 4.1.3). Les auteures expliquent ainsi une meilleure production des désinences de l'instrumental féminin et masculin

de par leur saillance relativement plus élevée que celle des désinences du nominatif féminin et masculin. De plus, les auteures constatent un phénomène de surgénéralisation de ces désinences de l'instrumental en lieu et place des contextes de sollicitation du nominatif. Ainsi on peut supposer que dans le cas d'une sollicitation d'une phrase requérant l'application du nominatif, les apprenants du core 2 produisent une désinence de l'instrumental, ou produisent une désinence idiosyncrasique, ou encore ne produisent pas (dans très peu de cas ils produisent une désinence du nominatif sollicité mais ne s'accordant pas avec le genre de l'item test).

Les apprenants du core 1, dont les performances en traitement de la LC sont équivalentes à celles du core 2, respectent en plus le système casuel de la LC en production orale. En effet, dans les tâches de production ciblées SI et OQA, ceux-ci appartiennent tous au SEF S_3 indiquant une maîtrise du paradigme linguistique testé. Ces apprenants sont donc non seulement sensibles à l'existence d'un système flexionnel en polonais mais ils ont également réussi l'association forme-fonction entre les désinences des items et le cas sollicité dans la phrase, et mobilisent ces connaissances dans leur production orale. Ils constituent ainsi des apprenants ayant réussi leur transfert entre type d'activité (traitement *vs.* production).

Ce résultat est toujours à relativiser vis-à-vis des caractéristiques spécifiques des tâches. Comme nous l'avons évoqué, lors de tâches de production semi-libre, le taux de formes correctement fléchies dans les discours des apprenants chute sensiblement par rapport à celui obtenu lors de tâches de productions ciblées comme celles analysées ici. Watorek et collaborateurs ([Watorek et al., 2016], cf. sous section 4.1.3) soulignent que plus la tâche requiert de l'apprenant la transmission d'un message et offre une liberté dans l'expression de la réponse, moins celui-ci se focalisera sur la forme de son discours pour recentrer son attention sur le but communicatif. Ainsi le constat d'un transfert réussi entre traitement et production pour les apprenants du core 1, présenté sur la figure 6.16 par une appartenance au SEF S_2 pour les attributs `transf_MO` et `transf_MS`, ne peut être attesté que dans le cadre de tâches de production ciblées.

6.2.1.2 Cores non obtenus

L'analyse des cores obtenus est aussi informative sur les parcours acquisitionnels des apprenants que l'analyse des cores non obtenus mais dont on aurait pu théoriser l'existence au vu de nos connaissances sur le projet VILLA. L'absence de certains patterns de regroupement d'apprenants porte en elle-même une signification intéressante pour l'étude de l'acquisition d'une L2.

Par exemple, il est intéressant de constater l'absence d'un core regroupant des apprenants qui auraient mieux réussi à produire en contexte approprié les désinences du nominatif et de l'accusatif dans des phrases d'ordre SO que des phrases d'ordre OS. Une explication possible est que l'ordre des constituants de la phrase ne soit pas un facteur déterminant du taux de formes correctement fléchies en production, comme il l'est en traitement selon l'analyse des résultats des apprenants de l'échantillon 1 à la tâche PV. Saturno ([Saturno, 2016]) étudie cette tendance générale dans les résultats des participants du projet VILLA. Pour ce faire, l'auteur cherche à établir trois scénarios possibles de réponse à la tâche SI :

- les apprenants s'appuient sur une forme d'item lexical non fléchi, sans variation morphologique ;
- les apprenants produisent régulièrement des formes de mots correctement infléchies ;
- les apprenants ont remarqué une certaine variation morphologique dans l'input, mais ils ne peuvent pas encore en rendre compte avec une règle productive. Par conséquent, ces apprenants fourniront des formes d'item de base et fléchies sans régularité apparente.

Les deux premiers scénarios sont également considérés dans cette thèse pour les tâches de production. L'auteur explique que le troisième est un scénario où l'apprenant choisit aléatoirement, donc *devine*, quelle désinence appliquée à quel constituant parmi les deux désinences du nominatif `\-a\` et de l'accusatif `\-e\`. Ce scénario de hasard a été écarté de notre travail du fait de l'hypothèse qu'il apparaît improbable de *produire* au hasard, comme cela pourrait être le cas pour l'action de *choisir* entre deux réponses, dans le contexte

d'une tâche de traitement.

Finalement, ce scénario de production aléatoire entre les deux déclinaisons est reformulé par l'auteur. Celui-ci émet l'hypothèse de travail qu'un item décliné au nominatif avec la désinence \-a\ est considérée par l'apprenant comme une forme invariable. L'utilisation de cette forme en contexte approprié ne serait donc que due au hasard et ne dépendrait pas de l'analyse grammaticale de l'apprenant. C'est donc sur la base de la production appropriée au contexte de la forme de l'accusatif \-e\ par les apprenants que l'analyse des résultats à la tâche SI permet de rendre compte de la sensibilité des apprenants au système flexionnel de la LC. En effet, la forme du nominatif \-a\ est la forme non marquée, basique, des items lexicaux en polonais. Elle tend ainsi à apparaître couramment dans les productions des apprenants, de manière appropriée ou non. Cette désinence ne fait apparemment pas l'objet d'une surgénéralisation dans d'autres études (cf. [Rast et al., 2018]), surgénéralisation qui devrait pourtant apparaître si cette forme était effectivement considérée par l'apprenant comme forme de base. On peut en déduire que, dans SI, seul les résultats sur le taux d'items correctement fléchis à l'accusatif rendent compte de l'acquisition du paradigme d'opposition nominatif *vs.* accusatif. Après avoir isolées les réponses des apprenants sur les items sollicités à l'accusatif, l'auteur trouve finalement que les deux facteurs influençant les apprenants dans leur production sont leur LM en premier lieu et ensuite l'ordre des constituants. En conclusion, afin d'observer en production des différences dans les performances morphosyntaxiques des apprenants, et ainsi possiblement voir émerger différents cores basés sur cette différence, il aurait fallu dédaigner toutes les réponses des apprenants sur les items cibles du test devant être conjugués au nominatif. Cette hypothèse nous apparaît plus construite pour expliquer l'absence de variabilité constatée plutôt que l'hypothèse que l'ordre syntaxique de la phrase ne constitue pas une difficulté supplémentaire pour l'apprenant, hypothèse que nous aurions pu déduire au vu des résultats des apprenants des cores 1 et 2 à la tâche SI.

Pour conclure, il faut nous interroger sur la pertinence des informations fournies par le constat de l'absence d'un core. Peut-elle être considérée comme réelle, et donc signifiante et interprétable d'un point de vue psycholinguistique, ou est-elle due à un simple manque de données ? Par exemple, l'absence de core regroupant des apprenants n'ayant pas effectué leur entrée dans le système flexionnel de la LC, ou de core d'apprenants échouant à utiliser la morphologie flexionnelle comme indicateur du statut du constituant dans une phrase, est perturbante. En effet, toujours dans la figure 6.16, on constate que l'échantillon réunit 10 apprenants dans le premier cas (colonne Morpho_c), et 14 dans le deuxième (colonne Morpho-syntaxe_c). Ces individus n'ont pas été classifiés à ce stade du processus de partitionnement et sont donc considérés temporairement comme des outliers.

Ils n'ont *a priori* aucune chance d'inclure dans leurs expressions orales des connaissances sur le système flexionnel de la LC qu'ils n'ont pas su mobiliser dans des tâches de traitement. Il est donc raisonnable d'émettre l'hypothèse que si l'on ne considère que le niveau morphologique (ou que le niveau morphosyntaxique) lors de l'évaluation du parcours acquisitionnel des apprenants de cet échantillon, un core d'apprenants ne faisant pas preuve d'une sensibilité à la morphologie flexionnelle de la LC sera créé par l'algorithme. En réduisant le nombre d'attributs considérés, et donc la complexité de la caractérisation de l'apprenant, de nouveaux patterns, et donc profils, seront obtenus. On peut théoriquement atteindre ce résultat en augmentant le nombre d'apprenants de l'échantillon. En effet, il existe un lien direct entre la capacité de l'algorithme à détecter les profils d'apprenants et le rapport entre le nombre d'attributs et le nombre d'apprenants. En augmentant le nombre d'attributs caractérisant le parcours acquisitionnel de l'apprenant, on augmente sa complexité mais aussi sa validité. Cependant, il faut alors augmenter en conséquence le nombre d'apprenants à partitionner. Ainsi, bien que certaines incohérences semblent exister dans la partition obtenue à ce stade, nous concluons sur le fait que l'absence d'occurrence de certains patterns ne peut prêter à interprétation étant donné le faible nombre d'apprenants considérés.

Avant de passer aux étapes suivantes du processus de partitionnement, il nous apparaît nécessaire de discuter des incohérences obtenues à cette étape de la classification. Les incohérences mentionnées dans le

paragraphe précédent sont surlignées en rouge sur la figure 6.16. Elles concernent le transfert de connaissances sur le système flexionnel de la LC vers leur mobilisation en production. Certains des apprenants appartiennent, pour ces deux attributs de transfert, au SEF S_1 . Pour rappel, l'appartenance au SEF S_1 sur les attributs de transfert signifie que l'apprenant est meilleur en production qu'en traitement. C'est un cas de figure peu probable au vu de la littérature en RAL, mais apparemment existant dans l'échantillon. Sur les 6 apprenants concernés (1 en morphologie, 5 en morphosyntaxe), 5 font preuve d'une stratégie de réponse soit aléatoire soit aveugle dans les activités de traitement. Ils appartiennent donc au SEF S_2 pour les tâches de traitement. Pourtant, ces 5 apprenants produisent les désinences appropriées au contexte, et ce quel que soit le contexte sollicité pour deux d'entre eux. Peu d'hypothèses explicatives sont formulables, et aucune en regard de la littérature en RAL. Cependant, une explication est possible pour quatre d'entre eux. Parmi les six observables, quatre d'entre eux réussissent la tâche SI malgré une apparente insensibilité au système flexionnel de la LC au vu de leur résultat à la tâche PV. Ces apprenants, malgré leur incapacité à traiter les indices morphosyntaxiques d'une phrase, seraient dotés d'une mémoire phonologique et d'une capacité de restitution supérieures, leur permettant de passer outre le traitement et la compréhension de la phrase entendue (hypothèse également avancée dans [Saturno, 2016]). Seulement, sur les quatre apprenants en question, deux d'entre eux ne réussissent la tâche SI que pour un type d'ordre de construction des phrases test (SO pour l'un, OS pour l'autre), contredisant cette hypothèse. La réponse à cette incohérence se décale alors sur un éventuel problème dans leurs performances à la tâche PV, performances qui ne seraient donc pas le reflet de leur niveau d'acquisition et auraient été perturbées soit par les caractéristiques de la tâche, soit par des raisons environnementales invérifiables.

6.2.1.3 Validité des cores obtenus

Par déduction, tous les apprenants (sauf un) intégrés à un core pour cet échantillon sont des apprenants qui appartiennent au cluster 7 ou au cluster 3 (core 4) dans la partition de l'échantillon 1. En effet, les apprenants des cores 1 et 2 ont tous obtenus de bons résultats aux tâches GJ et PV. Où sont passés les apprenants ayant des difficultés en traitement de la LC? Ils ont pour l'instant été qualifiés en tant qu'outliers, c'est-à-dire que ces apprenants ne présentent pas suffisamment de caractéristiques communes pour être considérés comme comparables par l'algorithme, et regroupés dans un même core. Pourtant, dans l'échantillon 1, certains cores ne sont constitués que d'apprenants présentant des difficultés en traitement du système flexionnel la LC.

Ce constat permet de nous interroger sur la faiblesse des techniques de classification quelles qu'elles soient. En effet, la notion de profil d'apprenant d'une L2, bien que dynamique dans le temps, est également dynamique dans l'espace des attributs considérés pour la caractérisation d'un apprenant. Ainsi, bien que la création *bottom-up* d'un ensemble de profils d'apprenants permette l'intégration de l'évolution de l'apprenant au fur et à mesure de son acquisition, le succès d'une telle approche repose sur la représentativité des données utilisées. De plus, ce n'est qu'à attributs équivalents que la comparaison entre deux apprenants est possible. En effet, l'ajout *vs.* la suppression d'attributs dans la caractérisation du parcours acquisitionnel de l'apprenant modifie les patterns et donc les profils repérés par l'algorithme, tout simplement à cause d'une quantité/qualité différente d'informations fournies sur l'apprenant. Ainsi, deux mêmes apprenants pourront être considérés comme similaires dans le premier échantillon et dissimilaires dans le deuxième, ou inversement.

Cependant, le travail développé dans cette thèse a permis l'obtention d'une partition des apprenants et ce malgré les liens évidents entre les résultats aux différents tests de langue. En effet, les différentes tâches utilisées et donc les différents scores obtenus à ces tâches ne sont pas indépendants les uns des autres. Dire le contraire reviendrait à affirmer l'indépendance des différents niveaux d'analyse linguistique. Le processus d'acquisition d'une langue peut s'analyser sur différents niveaux mais reste un phénomène complexe dont les différents niveaux sont en interactions. Les scores des apprenants aux tâches, et en fonction des caractéristiques de l'input, font preuve d'une grande covariance. Cette covariance est le reflet direct des interactions entre les niveaux d'analyse linguistique, puisque chaque tâche est destinée à tester un apprenant

sur un niveau donné. Or les techniques classiques de clustering ne peuvent s'opérer que sur des variables indépendantes. En effet, elles sont inefficaces pour ce type de données de par la colinéarité des variables, et requièrent une stricte indépendance entre attributs. Pourtant, avec la technique développée dans ce travail, le regroupement des données est rendu possible, et chaque variable retenue semble influencer le processus de partitionnement.

6.2.1.4 Rôle de la langue maternelle dans la création des cores

Pour ce qui est des résultats de l'analyse des cores en regard de la LM des apprenants les constituants, tout comme pour le premier échantillon, nous devons prendre des précautions dans leur interprétation, au vu du faible nombre d'apprenants. L'échantillon 2 montre encore plus de disparité que l'échantillon 1. Les trois graphiques inclus dans les figures 6.17, 6.18 et 6.19 représentent la répartition des apprenants dans les deux cores et le groupe des outliers selon leur LM respective.

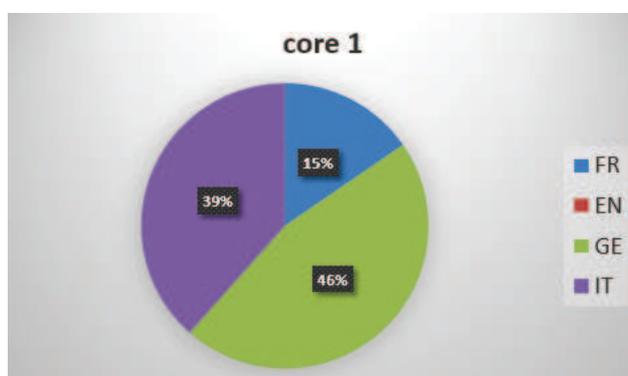


FIGURE 6.17 – LMs des apprenants du core 1. Échantillon 2

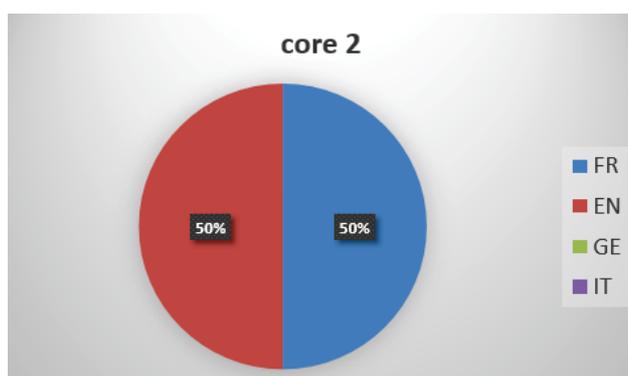


FIGURE 6.18 – LMs des apprenants du core 2. Échantillon 2

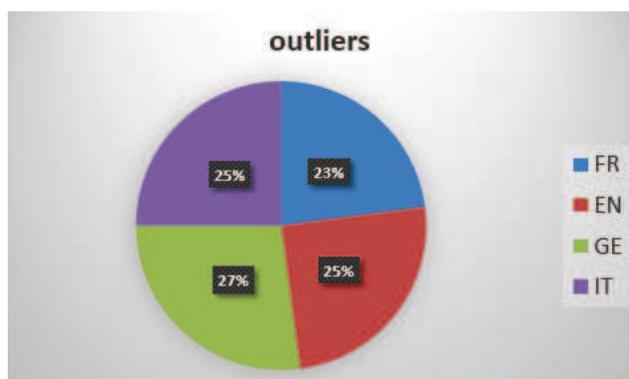


FIGURE 6.19 – LMs des apprenants des outliers. Échantillon 2

Le core 1 regroupe 13 apprenants qui semblent être capables de mettre en œuvre les connaissances sur le système de la LC à la fois en traitement et en production. Malgré le petit nombre d'individus, on peut apercevoir que ces apprenants ont comme LM essentiellement l'allemand et l'italien, et seulement quelques uns sont de LM française. Cet aperçu n'est pas en contradiction avec différentes analyses effectuées par les chercheurs du projet VILLA. D'après les analyses de Saturno [Saturno,], et Saturno et Watorek ([Saturno et Watorek, 2020]) qui mettent en relation PV et SI, il apparaît que les apprenants germanophones et italophones sont meilleurs et évoluent plus rapidement que ceux d'autres LM. En effet, les allemands et les italiens sont les apprenants obtenant les meilleures scores à la fois en compréhension et en répétition du paradigme linguistique nominatif *vs.* accusatif. De plus, Watorek ([Watorek et al., 2020]) qui analyse les productions d'une des tâches semi-guidée du projet (*Route Direction*) montre également que les germanophones commettent des erreurs caractéristiques d'une analyse du nouveau système plus avancée que les apprenants francophones et surtout anglophones. Par exemple, les germanophones tentent de produire une désinence même si elle n'est pas appropriée tandis que les autres produisent davantage des formes nominales invariables.

Le core 2 regroupe huit apprenants qui échouent à mobiliser leurs connaissances du système flexionnel de la LC dans leurs productions. Ce core est composé de quatre apprenants anglophones et quatre francophones. Les travaux du projet VILLA montrent en effet que les apprenants francophones, parfois meilleurs que les anglophones, peinent davantage que les germanophones et italophones pour entrer dans le système casuel et morphosyntaxique du polonais. Les anglophones, non seulement n'arrivent pas à produire des désinences casuelles appropriées mais ils ne commettent pas non plus d'erreurs qui seraient une manifestation d'une analyse du système de la LC ([Watorek et al., 2020]). De plus, les anglophones n'évoluent que très lentement en ce qui concerne l'analyse du système casuel en polonais. Leurs performances ne changent pas beaucoup durant la période d'observation, contrairement aux autres apprenants.

Le faible nombre d'apprenants réunis dans ces cores (1 et 2) nous oblige à être prudents dans l'interprétation de ces données. Cependant, la mise en perspective de notre partition avec les résultats fournis par les travaux du projet VILLA, renforce le résultat issu de notre algorithme.

En effet, il est intéressant de constater que l'observation effectuée dans l'échantillon 1 sur les écarts prononcés de performance entre les apprenants anglais et les apprenants allemands est possiblement transposable à l'analyse du deuxième échantillon. En effet, on remarque que les apprenants anglais et les apprenants allemands ne sont pas présents simultanément dans un même core. Les allemands classifiés appartiennent au core 1, et les anglais classifiés au core 2. Après intégration des outliers, et donc augmentation des effectifs classifiés, l'analyse des LMs de chaque cluster sera plus pertinente.

6.2.2 De l'intérêt d'une intégration des outliers

Cette étape du processus de partitionnement correspond normalement au raffinement des cores en clusters par détection de sous groupes d'apprenants similaires (et non plus comparables) à l'intérieur d'un même core. Seulement, les contraintes spécifiques d'une activité de production orale semblent peser sur la notion même de comparabilité entre apprenants, la rendant probablement plus difficilement atteignable que lors de la comparaison de la réalisation d'une activité de traitement entre apprenants.

En addition de cette hypothèse, la figure 6.16 montre l'homogénéité intracore pour les deux cores obtenus, en terme d'appartenance à un SEF. La méthode iVAT appliquée aux core 1 et 2 en figure 6.20 illustre également ce constat d'homogénéité mais sur une base numérique. En effet, malgré une esquisse de sous groupes potentiels, notamment dans le core 1, ces sous groupes ne se différencient au maximum que par une distance de 0.25 entre deux objets (pour une distance $d \in [0; 1]$), distance trop faible pour justifier une division. Le core 1 devient donc le cluster 1 et le core 2 le cluster 2.

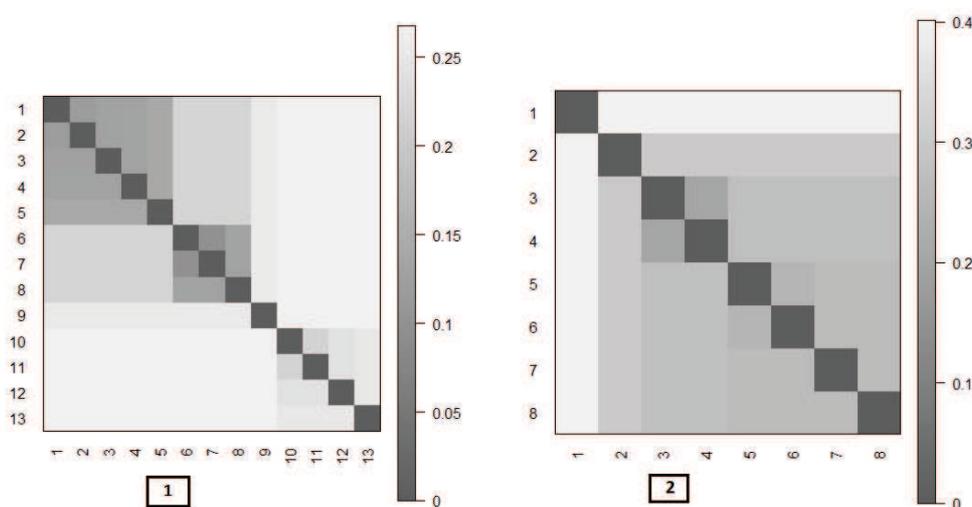


FIGURE 6.20 – Résultat de la méthode iVAT appliquée aux cores 1 à 2. Échantillon 2

La fonction d'intégration des outliers devient ainsi primordiale dans le processus de partitionnement de cet échantillon. Au vu de leur nombre, relativement au nombre d'apprenants des clusters 1 et 2, la classification finale repose sur la justesse de sa conception. La problématique émergente repose sur l'incertitude de la capacité de cette fonction d'intégrer des outliers à un cluster d'apprenants sans observer de différences fondamentales dans le parcours acquisitionnel entre les apprenants déjà classifiés et l'outlier à intégrer.

L'absence de l'émergence de certains patterns en tant que cluster dûe à la petite taille de l'échantillon induit l'hypothèse que certains outliers sont en fait des apprenants au profil potentiellement répandu chez les apprenants débutant du polonais mais sous-représentés dans notre échantillon.

Ainsi, la fonction d'intégration des outliers doit, dans une certaine mesure, être capable de *ne pas* intégrer des apprenants de ce type.

Sur les 48 outliers, 34 obtiennent un degré de similarité avec le cluster 1 ou le cluster 2 supérieur au seuil fixé de 0.55, seuil correspondant à une incomparabilité sur deux attributs au plus. Sur ces 34 outliers avec un degré de similarité suffisant pour être intégré à la classification, 9 possèdent un score de similarité équivalent entre le cluster 1 et le cluster 2. Ainsi, 25 outliers sur 48 sont intégrés à l'un des deux clusters présents dans l'échantillon, 11 au cluster 1 et 14 au cluster 2. La répartition en effectif obtenue après intégration des outliers est illustrée en figure 6.21.

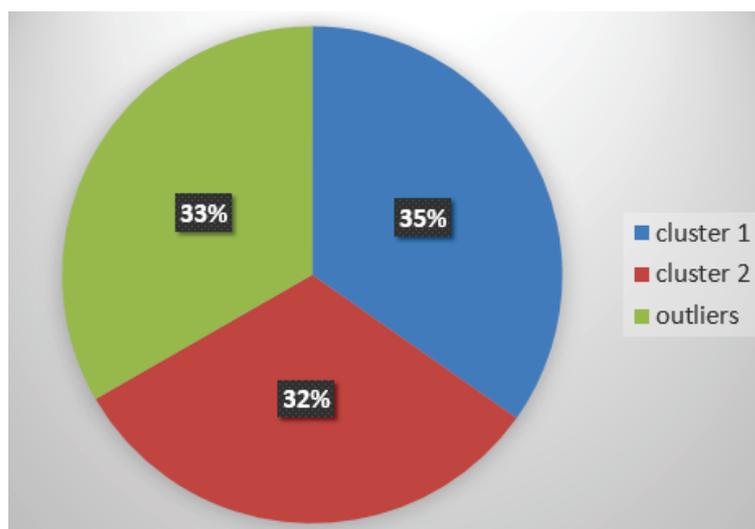


FIGURE 6.21 – Partition finale. Échantillon 2

Les outliers représentent encore un tiers des effectifs totaux de l'échantillon (23/69). Afin d'évaluer la pertinence et l'efficacité de la fonction d'intégration des outliers nous nous proposons d'effectuer une analyse qualitative des différences entre les apprenants originaux des clusters et leur homologues classifiés par cette fonction.

La majorité des apprenants ayant rejoint le cluster 1 par l'intermédiaire de la fonction d'intégration diffèrent de celui-ci de par leurs résultats à la tâche SI. Ces anciens outliers ont en effet toutes les caractéristiques du cluster 1 excepté dans la prise en compte du marquage casuel nominatif *vs.* accusatif pour attribuer le statut du constituant dans la phrase. Ces apprenants tendent à sur-généraliser l'une des deux désinences employées dans le test, en lieu et place du contexte sollicitant l'autre pour un des deux types (SO *vs.* OS) de phrases du test. Ainsi, nous attestons la généralisation de la désinence \-a\ du nominatif féminin dans le contexte où la désinence \-e\ de l'accusatif féminin est requise. Ceci a lieu surtout dans le cas des phrases construites sur le modèle OS. Nous avons identifié également un apprenant qui échoue à marquer le statut du constituant dans les phrases de type SO en se basant sur le marquage casuel, cet apprenant surgénéralise la désinence de l'accusatif et l'emploie dans le contexte qui implique celle du nominatif. Il produit ainsi les deux noms de la phrase avec la même désinence. Les 4 derniers des ex-outliers intégrés au cluster 1 sont des apprenants se différenciant des membres du cluster 1 par leur tendance (à des degrés divers) à surgénéraliser la désinence de l'instrumental dans le contexte du nominatif dans la tâche OQA. Un autre encore des ex-outliers intégré dans le cluster 1 est sujet à la surgénéralisation mais cette fois de la désinence du nominatif dans le contexte de l'instrumental. La question qui se pose est donc : peut-on considérer comme légitime de rassembler des apprenants produisant correctement les désinences de l'instrumental et du nominatif dans le contexte approprié à leurs utilisations, avec des apprenants qui, malgré une sensibilité à la morphologie flexionnelle attestée tant en traitement qu'en production, ne font pas preuve d'une association forme-fonction réussie dans leurs productions ?

La problématique se recentre alors autour de la signification de la connaissance et de l'utilisation d'une forme, et de son suremploi dans d'autres contextes que ceux appropriés. S'agit-il d'une différence foncièrement qualitative, démontrant d'une tendance au *guessing*, comme avancé par Saturno ([Saturno, 2016]) ? En effet, si l'on considère que la désinence \-a\ traduit la forme non marquée, basique, d'un item lexical en polonais, alors sa surgénéralisation dans la tâche SI ne peut pas être attribuée à une bonne utilisation du nominatif. S'agit-il de la saillance d'une forme dans l'input, ce qui rendrait le passage à la production orale plus facile

pour des apprenants débutants ? Ou s'agit-il plus simplement d'une volonté de l'apprenant à marquer les items malgré son manque de connaissance sur le système flexionnel de la LC, le conduisant ainsi au suremploi des désinences connues ? Ces hypothèses ne sont pas aisément réfutables, et ne s'excluent d'ailleurs pas complètement. Nous n'apporterons donc pas de réponses tranchées. Pourtant, ce sont les réponses à ces questions qui nous permettraient de juger de la validité de l'intégration des ex-outliers au cluster 1.

Dans le cas des apprenants incorporés au cluster 2, à cette étape de la classification, trois patterns se distinguent. Le premier regroupe les apprenants ayant le même profil que les individus du cluster 2 au niveau de la morphologie mais différant pour ce qui est de l'analyse morphosyntaxique. Ce premier ensemble d'ex-outliers ne s'appuie pas sur les indices morphologiques pour assigner le statut du constituant dans la phrase mais se repose sur le principe dit "positionnel" attesté dans la variété de base ([Klein et Perdue, 1997]). Ce phénomène est observé dans leurs productions, mais, contrairement aux autres apprenants du cluster 2, également dans leur compréhension de phrases en LC. Le deuxième pattern renvoie à des apprenants qui produisent des désinences pour assigner le statut du constituant en fonction du type de phrase à produire. Ces apprenants marquent les deux constituants d'une même phrase avec des désinences appropriées au contexte mais seulement pour un des deux types de phrases (SO *vs.* OS) rencontrés dans les tests. Ainsi, l'ordre des mots de la phrase influe sur leur capacité à la marquer correctement. Cependant, étant donné qu'un des apprenants marque les deux constituants de manière appropriée seulement pour les phrases SO et que les deux autres ne produisent correctement que pour les phrases de type OS, l'hypothèse d'un effet facilitateur des phrases construites sur le modèle SO pour leur production respectant le système flexionnel de la LC n'est pas applicable à ces trois apprenants.

La troisième et dernière catégorie d'apprenants intégrés au cluster 2 regroupe des apprenants dont seuls les résultats aux tests de compréhension en LC reflètent une ouverture au système flexionnel de la LC. Autrement dit, ces apprenants diffèrent des autres dans leur incapacité à produire les désinences de l'instrumental masculin \-em\et féminin \-a\en contexte approprié tout du moins. Sachant que les apprenants du cluster 2 étaient caractérisés par leur tendance à surgénéraliser les désinences de l'instrumental en contexte impliquant l'emploi du nominatif, l'absence d'un tel phénomène chez ce groupe d'outliers nous interroge sur leur légitimité à être classifiés dans le même cluster que les autres. En effet la surgénéralisation présuppose la compréhension de l'existence d'un système casuel, mais également de l'échec à discriminer les différents cas existants et la mise en relation de ces cas avec leurs marques formelles. Or, l'incapacité de cette catégorie d'apprenants à produire des désinences cibles utilisées dans les tests de production dans le contexte approprié, ne correspond pas au même profil.

Ainsi, la fonction d'intégration des outliers semble perfectible. Le saut qualitatif entre les apprenants originels du cluster 2 et l'un des groupes d'outliers lui étant accroché doit être évité. En effet, la classification en un même ensemble d'apprenants produisant correctement seules certaines désinences et de ceux ne produisant aucun item doté d'un marquage casuel approprié, est difficilement justifiable lorsque l'objectif de la classification se centre autour de l'identification et de la discrimination de profils d'apprenants. Une solution facilement applicable consiste en la révision de la caractérisation des différents niveaux de transfert possible d'un niveau d'analyse linguistique. Par exemple, les apprenants qui surgénéralisent les marques formelles d'un cas en production, alors qu'ils traitent correctement le paradigme linguistique testé, ont été considérés comme ne réussissant pas leur transfert. Cette considération est due à notre définition d'un transfert réussi. Un apprenant qui réussit son transfert est un apprenant faisant preuve de la compréhension de l'opposition entre deux cas et est à même de marquer cette opposition dans ses productions orales. Cependant, le passage du traitement vers la production génère un certain nombre d'obstacles interférant avec les capacités d'analyses grammaticales de l'apprenant et donc de restitution. Cet écart entre le traitement et la production est constaté chez les apprenants du projet VILLA par Watorek et collaborateurs ([Watorek et al., 2016], cf. chapitre 4) qui l'expliquent par la supériorité du besoin communicatif sur le respect des formes grammaticales. Ainsi, l'attention de l'apprenant n'est plus focalisée sur la forme de son énoncé mais sur son sens. Ce constat est d'autant plus vrai que la tâche de

production est libre. Une tâche de production ciblée ne renseigne donc finalement que de manière limitée sur les compétences de production de l'apprenant. Les apprenants obtenant un bon score pour les tâches SI et OQA feraient-ils preuve du même respect du système flexionnel de la LC dans des productions spontanées ?

Si la fonction d'intégration des outliers est pertinente, c'est au niveau du choix et de la création d'attributs que le problème se pose. Bien que l'attribut "transfert" porte en lui-même une information supplémentaire dans la caractérisation du parcours acquisitionnel d'un apprenant, le choix de paramétrage de cet attribut effectué en amont de l'analyse porte en lui-même un enjeu théorique et une hypothèse de fonctionnement du passage du traitement vers la production en LC.

Une fois l'intégration des outliers actée, nous tentons une ré-analyse de la partition en terme de LM des apprenants les constituants. La figure 6.22 illustre cette répartition.

Malgré l'augmentation des effectifs, une analyse statistique n'est pas possible, notamment du fait de l'absence complète du groupe d'apprenant d'une certaine LM dans un des clusters. En effet, les anglais sont absents du cluster 1. Ce cluster a conservé sa composition en terme de LMs avant et après intégration des outliers, avec une majorité d'allemands suivis des italiens et dans une moindre mesure des français. Quant au cluster 2, bien que les anglais et les français constituent toujours une majorité (deux tiers, un chacun) des italiens et des allemands l'ont rejoint. Ainsi, aucun allemand n'est exclu de la partition finale, tandis que les anglais représentent la majorité des outliers (57%).

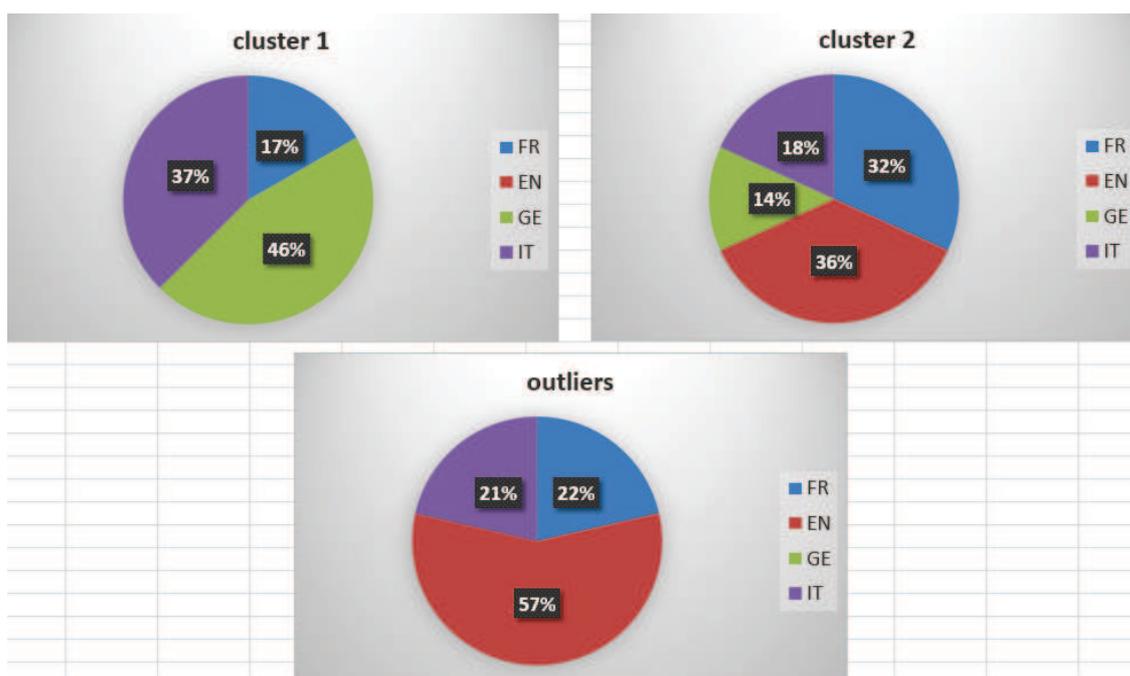


FIGURE 6.22 – Répartition des LMs des apprenants dans la partition finale. Échantillon 2

Bien que toute affirmation soit impossible, l'ensemble de la littérature issue du projet VILLA tend à indiquer que le groupe d'apprenant ayant l'anglais comme LM a rencontré plus de difficulté dans l'acquisition du polonais au cours du projet VILLA, et ce quel que soit le niveau d'analyse linguistique adopté. De plus, le fait que les anglais forment la majorité des outliers résistants pourrait suggérer que ces apprenants ont des niveaux acquisitionnel éloignés des caractéristiques des deux clusters en présence dans la classification. Or les deux clusters illustrent des apprenants éloignés d'une variété de base dénuée de toute flexion, entièrement

en traitement, et partiellement en production. Ces deux faits mis en parallèle, et à manier avec précaution, suggèrent que les anglais pourraient majoritairement incarner un profil d'apprenant que l'algorithme aura échoué à discerner.

Les deux objectifs de l'application de l'algorithme à la BDD sont atteints. Premièrement, les analyses des partitions obtenues corroborent les résultats de la littérature du projet VILLA, que ce soit vis-à-vis des l'influence de la LM sur le parcours acquisitionnel des apprenants ou de phénomènes plus précis, comme par exemple la difficulté à transférer une connaissance du système grammatical attestée en traitement vers sa mobilisation en production. Deuxièmement, certaines pistes apportées par la classification de l'échantillon n'ont pas encore été explorées dans la littérature du projet VILLA. Par exemple, les apprenants issus du core 3 (cluster 1 et 2) correspondent à un profil d'apprenant extrêmement sensible aux caractéristiques de la tâche. Un autre exemple concerne les apprenants de ce même échantillon classés dans le cluster 3, ces apprenants se révèlent dotés d'une grande sensibilité à la transparence des items, sensibilité palliant les difficultés à traiter un item lorsque celui-ci est peu présent dans l'input, et ce malgré les différences de LMs des apprenants de ce profil. Ainsi, les partitions obtenues permettent la découverte de patterns insoupçonnés. De plus, nous avons pu attester de la stabilité des patterns obtenus dans les données par une technique de validation croisée. Cependant, la notion de comparabilité que nous avons établie, en juxtaposition de la notion de similarité, est difficilement applicable lors de l'analyse des résultats issus des tâches de production. La production est en effet une activité suscitant une plus grande variabilité dans les comportements des apprenants débutants. De cette grande variabilité, malgré la fonction d'intégration des outliers, résulte l'impossibilité de classer un tiers des apprenants du deuxième échantillon. Il faut tout de même relativiser ce résultat par la possibilité théorique de son amélioration par l'intermédiaire de l'augmentation du nombre d'apprenants dans la base de données disponible.

Conclusion et perspectives

Les points clés de cette thèse, et donc nos principales contributions, se sont construits autour de trois étapes successives. Il a tout d'abord fallu modéliser un apprenant d'une langue étrangère. Une fois caractérisé, il était alors possible de réfléchir à une procédure de comparaison entre deux apprenants, respectant cette modélisation, afin de l'implémenter dans un algorithme de partitionnement. Enfin, l'application de l'algorithme à deux échantillons de la base de données VILLA a permis d'atteindre un double objectif : corroborer des résultats existants dans la littérature du projet, et apporter de nouvelles pistes de réflexion pour leur interprétation.

Modélisation

La base de données VILLA a donné lieu à de multiples publications. Chacune d'entre elles s'attache à observer l'effet de sources d'influence potentielles sur les performances des apprenants. Les problématiques du projet VILLA se centrent autour de la question de l'interaction entre le lecteur et l'apprenant et l'input. La langue maternelle de l'apprenant fait partie des connaissances internes dont l'apprenant dispose avant toute exposition à la LC. La manipulation du type d'enseignement, basée sur le sens ou basée sur la forme, que l'apprenant va recevoir, permet d'observer, ou non, une acquisition précoce du système flexionnel de la LC. L'application correcte de ce système aux items nominaux est aussi conditionnée par les caractéristiques de ces items dans l'input. La fréquence d'un item dans l'input et sa transparence avec la langue maternelle de l'apprenant, sont deux caractéristiques interagissant et ayant, en fonction de la tâche en LC utilisée, un effet facilitateur de leur restitution correctement déclinée. Il a été particulièrement intéressant de constater des effets différenciés des caractéristiques de l'input en fonction de la nature de la tâche. Les différents résultats évoqués dans la première partie du chapitre 4 permettent de dégager l'idée principale de l'importance de la contextualisation des résultats des apprenants vis-à-vis des caractéristiques de la tâche utilisée pour les obtenir. Le type de sollicitation (traitement, répétition, production) ainsi que le nombre d'items cibles ou encore la saillance des formes testées sont autant de facteurs à prendre en compte lors de l'analyse de la base de données.

La caractérisation d'un apprenant est donc multifactorielle. La complexité de la base de données du projet VILLA a donné lieu à deux représentations schématiques pour une visualisation d'ensemble des informations à notre disposition. La figuration d'un apprenant en un objet géométrique à n dimensions, c'est-à-dire à n caractéristiques pertinentes pour sa caractérisation, est une première piste intéressante, pour sa modélisation en vue d'une manipulation computationnelle. De plus, la perspective d'une évaluation, non plus numérique des performances des apprenants, mais sous la forme de termes linguistiques de jugement comme « mauvais » ou « très bon » a été rendue possible par l'utilisation des 2-tuples linguistiques flous. Cette approche permet en effet d'élargir les possibilités de format des données à notre disposition sans que leur manipulation ne soit compromise.

Comparabilité

Après la représentation d'apprenants se pose la question de leurs comparaisons. Au cœur des algorithmes de clustering réside la notion de distance. Nous sommes allés plus dans le détail en proposant de considérer, avant tout, la comparabilité entre deux apprenants plutôt que de calculer directement une valeur de similarité entre eux. Cette procédure a émergé de la considération de l'objectif principal de l'obtention d'une partition des apprenants : l'interprétabilité de celle-ci d'un point de vue acquisitionniste. L'obtention de profils d'apprenants en fonction de leur score numérique aux différents tests en LC pose en effet des problèmes d'interprétation et la partition finale est inutilisable. Nous avons donc transformé les données pour qu'un apprenant ne soit plus caractérisé par son score mais par sa stratégie de réponse à une tâche. Chaque tâche du projet VILLA teste en effet l'acquisition d'un paradigme linguistique précis sur un niveau d'analyse spécifique.

Pour les tâches sollicitant un traitement de l'apprenant, la construction de la tâche requiert de l'apprenant de choisir pour chaque stimulus entre deux possibilités. Chacune des deux possibilités incarne l'une des modalités du paradigme d'opposition considéré. Ainsi, nous avons traduit le score numérique de l'apprenant en trois stratégies possibles : réponse au hasard ou aveugle ; sensibilité au système flexionnel de la LC ; sensibilité au système flexionnel de la LC et mise en relation forme-fonction réussie.

En ce qui concerne les tâches sollicitant une production, elles testent également un paradigme linguistique précis. Seulement, la même analyse des stratégies possiblement employées par les apprenants n'est pas adéquate. En effet, l'idée de « produire au hasard » est difficilement justifiable. De plus, la production correcte des formes d'un cas en contexte approprié ne garantit pas le même résultat pour les formes de l'autre cas également testé par le paradigme de la tâche.

Algorithme

On considère ainsi deux apprenants comme comparables s'ils présentent tout deux une sensibilité au système flexionnel de la LC, à une caractéristique spécifique de l'input, ou attestent d'une mise en relation forme-fonction réussie dans leurs productions pour le même ou les mêmes cas de déclinaison. On peut ainsi créer l'ensemble des contraintes *cannot-link*. Cet ensemble va nous permettre d'opérer une première partition des apprenants en core, c'est-à-dire en sous-ensembles d'apprenants comparables. Cette première étape nous assure l'intégration de notre expertise sur les données dans le résultat du processus de partitionnement. Il s'en suit le calcul d'une mesure de similarité entre apprenants d'un même core par le biais de leurs scores numérique originaux. Pour finir, l'intégration des outliers dans les clusters obtenus s'opère à travers une réutilisation des contraintes de l'ensemble C pour le calcul d'un indice de comparabilité compris entre 0 et 1.

Applications

L'algorithme a été appliqué à deux échantillons de la base de données du projet dans le but de tester sa validité et d'évaluer son utilité par la découverte de nouvelles pistes de réflexion sur les résultats des apprenants.

L'application de l'algorithme au premier échantillon regroupant les résultats des apprenants à différentes tâches de traitement de la LC a permis de distinguer les apprenants sensibles au système flexionnel de la LC des apprenants ne la traitant pas. Également, la partition isole les apprenants appliquant une stratégie positionnelle pour l'identification du statut du constituant dans la phrase des apprenants considérant la flexion de l'item pour parvenir à cette fin. Les résultats issus de la littérature du projet VILLA tendent à suggérer l'influence de la LM sur le parcours acquisitionnel des apprenants après seulement 14h d'exposition à l'input. Dans l'analyse des cores obtenus, on remarque également cette tendance. Les apprenants allemands semblent appartenir majoritairement au core d'apprenant sensible au système flexionnel de la LC, tandis que les anglais appartiennent au core d'apprenants y étant le plus réfractaire.

Lors de l'analyse de la partition de l'échantillon 2, incluant les performances des apprenants lors de tâches de production en LC, nous avons également pu identifier des phénomènes de surgénéralisation de l'application des désinences d'un cas dans un contexte en sollicitant un autre. L'échantillon 2 autorise l'étude du transfert des connaissances attestées du système flexionnel de la LC en traitement vers leur mobilisation en production. Ce processus ne représente pas le même défi pour tous les apprenants au vu des cores obtenus pour cet échantillon. Cependant, la petite taille de l'ensemble de données en terme de nombre d'apprenants ainsi que l'augmentation de la variabilité dans les résultats des apprenants aux tâches de production comparativement aux tâches de traitement empêche l'émergence de certains profils d'apprenants que notre expertise sur les données nous aurait laissé anticiper.

Les profils obtenus ont permis de retrouver un certain nombre de résultats présentés par les différentes publications autour du projet VILLA. Mais l'utilité d'un tel procédé réside également dans la découverte de profils inattendus. Deux profils particulièrement méritent de retenir notre attention. Le premier concerne un cluster d'apprenants dont les performances en morphosyntaxe surpassent celles attestées dans une tâche de morphologie. Notre hypothèse explicative se centre autour de la nature même de la tâche, rejoignant ainsi un débat plus large sur la nature d'une tâche de répétition et de l'information qu'elle procure quant aux capacités réelles de production. On peut ainsi s'interroger sur l'efficacité peut-être supérieure de leur mémoire phonologique. Le deuxième fait état d'un cluster d'apprenants extrêmement sensibles à une certaine combinaison des caractéristiques des items de l'input. En effet, ces apprenants appliquent le système flexionnel de la LC bien plus facilement que leurs camarades aux items non fréquents dans l'input lorsque ceux-ci sont transparents. Les sensibilités différentes aux caractéristiques de l'input permettent d'envisager une manipulation de l'exposition à l'input rendant celle-ci plus efficace pour l'aide à l'apprenant dans son processus d'acquisition.

Perspectives

"Les parties accessoires de votre thèse deviendront des points de départ pour de nouvelles recherches. Il pourra vous arriver de retourner à votre thèse, même des dizaines d'années plus tard. Elle aura été comme un premier amour : il vous sera difficile de l'oublier. Au fond, c'était la première fois que vous faisiez un travail scientifique sérieux et rigoureux, et c'est une expérience qui compte." [Eco, 2016], p. 223.

Cette citation d'Umberto Eco illustre à mes yeux à quel point un travail de thèse ne représente qu'un commencement, un début de réflexion, et surtout, soulève plus de questions qu'il n'y répond.

Cette thèse ne fait pas exception. L'approche théorique de la définition d'une "comparabilité entre apprenants" paraît légitime, cependant, son opérationnalisation peut être envisagée sous plusieurs formes. L'idée de ne pas comparer des apprenants dont les stratégies de réponses à une tâche sont trop éloignées est un présupposé fort. Pourtant elle semble prometteuse, la comparaison purement numérique des scores des apprenants ne garantissant pas l'interprétabilité et l'utilité d'une catégorisation des apprenants. Il persiste ainsi, à la fin de ce travail, l'envie de reprendre le concept pour l'améliorer. Les pistes possibles de l'amélioration de ce concept sont nombreuses. Par exemple, le nombre de stratégies identifiées pour un ensemble de tâche est trop catégorique. L'évaluation des performances de l'apprenant sur un seul paradigme linguistique ne peut-elle donner lieu qu'à trois types de stratégies de réponse? En production notamment, la réponse semble être négative. En outre, la variabilité des productions des apprenants en comparaison de leurs résultats relativement homogène dans les tâches de traitement n'a pas été gérée par l'algorithme présenté. Une des explications les plus probables semble être la taille de l'échantillon, trop petite pour élever au rang de profil un petit nombre d'apprenant aux comportements similaires. La plus grande perspective de ce travail, vitale pour son amélioration, est l'augmentation de la base de données. Tout d'abord vis-à-vis de la précision des patterns détectés dans l'ensemble de données, mais surtout du point de vue de sa validité en RAL. Cet algorithme explicatif des profils d'apprenants ne peut être considéré que dans le

champ restreint de l'acquisition du polonais par des apprenants adultes *ab initio* de langues maternelles romanes et germaniques. L'incorporation de différentes langues cibles et de différentes langues sources est un incontournable des perspectives de ce travail. A ce sujet, des travaux de réplique du projet VILLA pour l'acquisition de l'arabe et du chinois par des français est en cours. Le design expérimental de ces travaux étant fortement similaires à celui du projet VILLA, la tentation d'y appliquer le travail de cette thèse est intéressante. En ce qui concerne l'incorporation d'apprenants plus avancés dans leur acquisition de la langue, le problème semble plus ardu. Premièrement, le contrôle de l'input et donc la caractérisation de la fréquence d'exposition des apprenants à un item ou à une structure morphologique n'est pas possible d'un point de vue méthodologique. L'attribut de sensibilité des apprenants à cette caractéristique ne sera donc pas prise en compte. Deuxièmement, les tâches utilisées pour évaluer les performances des apprenants ne seront naturellement pas les mêmes. Elles devront en effet correspondre à un niveau d'acquisition plus élevé pour être à même d'observer des différences dans la maîtrise des apprenants des paradigmes testés par ces tâches. De ce constat, quelle comparabilité entre apprenants de niveaux grandement différents? On en revient au cœur de cette thèse sur la notion de comparabilité.

D'un point de vue algorithmique, les possibilités d'améliorations structurelles sont également multiples. Spécifiquement, la création de l'ensemble des contraintes nécessite une expertise en RAL et est ainsi coûteuse en temps et en connaissances. L'objectif de l'intégration d'une expertise pour guider le processus de partitionnement est de l'améliorer mais la question du coût de son implémentation est toujours redoutée. Cependant, l'utilité même d'un tel partitionnement en dépend. Concilier généralisation d'un algorithme à plusieurs types de problèmes et exigences de justesse par des experts du domaine est tout l'enjeu de cette thèse. La réticence des sciences humaines et sociales pour l'utilisation de procédés informatiques standards est due à l'imprécision de ces outils inhérente à leur création en vue d'application à de multiples phénomènes. Cependant, d'autres possibilités de formulation de cette expertise doivent être envisagées afin de conserver l'expertise de la RAL et de lisser les procédures de création de l'ensemble des contraintes.

Bibliographie

- [Abchir, 2013] Abchir, M.-A. (2013). *Vers une sémantique floue : application à la géolocalisation*. Thèse d'université, Université Paris VIII Vincennes-Saint Denis.
- [Akdag et al., 2001] Akdag, H., Truck, I., Borgi, A., et Mellouli, N. (2001). LINGUISTIC MODIFIERS IN A SYMBOLIC FRAMEWORK. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 09(supp01) : 49–61.
- [Allab et al., 2011] Allab, K., Benabdeslem, K., et Aussem, A. (2011). Une approche de co-classification automatique à base des cartes topologiques. *Revue des Nouvelles Technologies de l'Information*, pages 1–24.
- [Ankerst et al., 1999] Ankerst, M., Breunig, M. M., peter Kriegel, H., et Sander, J. (1999). Optics : Ordering points to identify the clustering structure. pages 49–60. ACM Press.
- [Arthur et Vassilvitskii, 2007] Arthur, D. et Vassilvitskii, S. (2007). k-means++ : The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.
- [Augendre, 2008] Augendre, S. (2008). S + V + O : Ordres marqués et non marqués en italien. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (3).
- [Babuška, 2000] Babuška, R. (2000). Fuzzy clustering algorithms with applications to rule extraction. In *Fuzzy Systems in Medicine*, pages 139–173. Springer.
- [Banfield et Raftery, 1993] Banfield, J. D. et Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821.
- [Baraldi et Blonda, 1999a] Baraldi, A. et Blonda, P. (1999a). A survey of fuzzy clustering algorithms for pattern recognition. i. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(6) : 778–785.
- [Baraldi et Blonda, 1999b] Baraldi, A. et Blonda, P. (1999b). A survey of fuzzy clustering algorithms for pattern recognition. ii. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 29(6) : 786–801.
- [Basu et al., 2002] Basu, S., Banerjee, A., et Mooney, R. (2002). Semi-supervised Clustering by Seeding. In *In Proceedings of 19th International Conference on Machine Learning (ICML-2002)*.
- [Basu et al., 2004a] Basu, S., Banerjee, A., et Mooney, R. J. (2004a). Active semi-supervision for pairwise constrained clustering. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 333–344. SIAM.
- [Basu et al., 2004b] Basu, S., Bilenko, M., et Mooney, R. J. (2004b). A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68. ACM.

- [Basu et al., 2009] Basu, S., Davidson, I., et Wagstaff, K. L., éditeurs (2009). *Constrained clustering : advances in algorithms, theory, and applications*. Chapman & Hall/CRC data mining and knowledge discovery series. CRC Press, Boca Raton. OCLC : ocn144226504.
- [Bellet et al., 2013] Bellet, A., Habrard, A., et Sebban, M. (2013). A Survey on Metric Learning for Feature Vectors and Structured Data.
- [Berkhin, 2002] Berkhin, P. (2002). *Survey Of Clustering Data Mining Techniques*. San Jose, CA.
- [Bezdek et al., 1984] Bezdek, J. C., Ehrlich, R., et Full, W. (1984). Fcm : The fuzzy c-means clustering algorithm. *Computers & Geosciences*, 10(2-3) : 191–203.
- [Bezdek et Hathaway, 2002] Bezdek, J. C. et Hathaway, R. J. (2002). VAT : a tool for visual assessment of (cluster) tendency. In *Proceedings of the 2002 International Joint Conference on Neural Networks. IJCNN'02 (Cat. No.02CH37290)*, volume 3, pages 2225–2230 vol.3.
- [Bilenko et al., 2004] Bilenko, M., Basu, S., et Mooney, R. J. (2004). Integrating constraints and metric learning in semi-supervised clustering. In *Proceedings of the twenty-first international conference on Machine learning*, page 11. ACM.
- [Blashfield et Aldenderfer, 1988] Blashfield, R. K. et Aldenderfer, M. S. (1988). The Methods and Problems of Cluster Analysis. In Nesselroade, J. R. et Cattell, R. B., éditeurs, *Handbook of Multivariate Experimental Psychology*, Perspectives on Individual Differences, pages 447–473. Springer US, Boston, MA.
- [Bohné et al., 2018] Bohné, J., Ying, Y., Gentric, S., et Pontil, M. (2018). Learning local metrics from pairwise similarity data. *Pattern Recognition*, 75 : 315 – 326.
- [Bregman, 1967] Bregman, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3) : 200–217.
- [Cai et al., 2016] Cai, L., Yu, T., He, T., Chen, L., et Lin, M. (2016). Active Learning Method for Constraint-Based Clustering Algorithms. In *International Conference on Web-Age Information Management*, pages 319–329. Springer.
- [Carpenter et Grossberg, 1987] Carpenter, G. A. et Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing*, 37(1) : 54–115.
- [Carroll, 1965] Carroll, J. (1965). The General Aptitude Test Battery (GATB). *The sixth mental measurements yearbook*, pages 1027–1029.
- [Carroll, 1981] Carroll, J. B. (1981). Ability and task difficulty in cognitive psychology. *Educational Researcher*, 10(1) : 11–21.
- [Carroll, 1990] Carroll, J. B. (1990). Cognitive abilities in foreign language aptitude : Then and now. *Language aptitude reconsidered*, pages 11–29.
- [Carroll et Sapon, 1955] Carroll, J. B. et Sapon, S. M. (1955). *Modern Language Aptitude Test : Form A*. Psychological Corporation.
- [Carroll, 1999] Carroll, S. E. (1999). Putting ‘input’ in its proper place. *Second Language Research*, 15(4) : 337–388.
- [Carroll, 2001] Carroll, S. E. (2001). *Input and evidence : The raw material of second language acquisition*, volume 25. John Benjamins Publishing.
- [Cha, 2007] Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2) : 1.

-
- [Chapelle et al., 2006] Chapelle, O., Schölkopf, B., et Zien, A., éditeurs (2006). *Semi-supervised learning. Adaptive computation and machine learning*. MIT Press, Cambridge, Mass. OCLC : ocm64898359.
- [Charrad et al., 2014] Charrad, M., Ghazzali, N., Boiteau, V., et Niknafs, A. (2014). NbClust : An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, 61(6) : 1–36.
- [Choi et al., 2010] Choi, S.-S., Cha, S.-H., et Tappert, C. C. (2010). A survey of binary similarity and distance measures. *Journal of Systemics, Cybernetics and Informatics*, 8(1) : 43–48.
- [Chomsky et Lightfoot, 1957] Chomsky, N. et Lightfoot, D. W. (1957). *Syntactic structures*. Walter de Gruyter.
- [Cintrón-Valentín et Ellis, 2016] Cintrón-Valentín, M. C. et Ellis, N. C. (2016). Salience in Second Language Acquisition : Physical Form, Learner Attention, and Instructional Focus. *Frontiers in Psychology*, 7 : 1284.
- [Cohn et al., 2003] Cohn, D., Caruana, R., et McCallum, A. (2003). Semi-supervised clustering with user feedback. *Constrained Clustering : Advances in Algorithms, Theory, and Applications*, 4(1) : 17–32.
- [Collentine, 2004] Collentine, J. (2004). Commentary : Where PI research has been and where it should be going. *Processing instruction : Theory, research, and commentary*, pages 169–181.
- [Connell et Myles-Zitser, 1982] Connell, P. J. et Myles-Zitser, C. (1982). An analysis of elicited imitation as a language evaluation procedure. *Journal of Speech and Hearing Disorders*, 47(4) : 390–396.
- [Corder, 1967] Corder, S. P. (1967). The significance of learner’s errors. *IRAL-International Review of Applied Linguistics in Language Teaching*, 5(1-4) : 161–170.
- [Davidson et al., 2006] Davidson, I., Wagstaff, K. L., et Basu, S. (2006). Measuring constraint-set utility for partitional clustering algorithms. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 115–126. Springer.
- [Davis et al., 2007] Davis, J. V., Kulis, B., Jain, P., Sra, S., et Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM.
- [de Glas, 1986] de Glas, M. (1986). Representation of Łukasiewicz’ many-valued algebras. *Journal of Mathematical Analysis and Applications*, 114(2) : 315–327.
- [Demiriz et al., 1999] Demiriz, A., Bennett, K. P., et Embrechts, M. J. (1999). Semi-supervised clustering using genetic algorithms. *Artificial neural networks in engineering (ANNIE-99)*, pages 809–814.
- [Dempster et al., 1977] Dempster, A. P., Laird, N. M., et Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.
- [Desjardins et al., 2007] Desjardins, M., MacGlashan, J., et Ferraioli, J. (2007). Interactive visual clustering. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 361–364. ACM.
- [Diday, 1971] Diday, E. (1971). Une nouvelle méthode en classification automatique et reconnaissance des formes : la méthode des nuées dynamiques. *Revue de statistique appliquée*, 19(2) : 19–33.
- [Dimroth, 2006] Dimroth, C. (2006). The finite story. *Max Planck Institut for Psycholinguistics* : http://corpus1.mpi.nl/ds/imdi_browser.
- [Dimroth et al., 2013] Dimroth, C., Rast, R., Marianne, S., et Watorek, M. (2013). Methods for studying the acquisition of a new language under controlled input conditions : The VILLA project. *EUROSLA Yearbook*, 13 : 109–138.

- [Doughty, 1991] Doughty, C. (1991). Second Language Instruction Does Make a Difference : Evidence from an Empirical Study of SL Relativization. *Studies in Second Language Acquisition*, 13(4) : 431–469.
- [Doughty et Williams, 1998] Doughty, C. et Williams, J. (1998). *Focus on Form in Classroom Second Language Acquisition*. Cambridge University Press. Google-Books-ID : dUc7sLSt1DIC.
- [Doughty et al., 2010] Doughty, C. J., Campbell, S. G., Mislevy, M. A., Bunting, M. F., Bowles, A. R., et Koeth, J. T. (2010). Predicting near-native ability : The factor structure and reliability of Hi-LAB. In *Selected proceedings of the 2008 Second Language Research Forum*, pages 10–31. Cascadilla Proceedings Project Somerville, MA.
- [Dubois et al., 2004] Dubois, D., Foulloy, L., Mauris, G., et Prade, H. (2004). Probability-Possibility Transformations, Triangular Fuzzy Sets, and Probabilistic Inequalities. *Reliable Computing*, 10(4) : 273–297.
- [Dunn, 1973] Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.
- [Durand, 2016] Durand, M. (2016). Using the CLIP approach to study second language acquisition.
- [Durand et Truck, 2015] Durand, M. E. et Truck, I. (2015). A first attempt towards a fuzzy c-means for the linguistic 2-tuple model. In *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2015)*, Istanbul, Turkey.
- [Durand et Truck, 2018] Durand, M. E. et Truck, I. (2018). A new proposal to deal with hesitant linguistic expressions on preference assessments. *Information Fusion*, 41 : 176 – 181.
- [E. Carroll, 2005] E. Carroll, S. (2005). Input and SLA : Adults’ sensitivity to different sorts of cues to French gender. *Language Learning*, 55(S1) : 79–138.
- [Eco, 2016] Eco, U. (2016). *Comment écrire sa thèse*. Flammarion. Google-Books-ID : HzDRDAAAQBAJ.
- [Ellis, 1994] Ellis, N. C. (1994). Implicit and explicit processes in language acquisition : An introduction. Academic Press.
- [Ellis, 2009] Ellis, N. C. (2009). Optimizing the Input : Frequency and Sampling in Usage-Based and Form-Focused Learning. In *The Handbook of Language Teaching*, pages 139–158. John Wiley & Sons, Ltd.
- [Ellis et Sagarra, 2011] Ellis, N. C. et Sagarra, N. (2011). LEARNED ATTENTION IN ADULT LANGUAGE ACQUISITION : A Replication and Generalization Study and Meta-Analysis. *Studies in Second Language Acquisition*, 33(4) : 589–624.
- [Ellis, 2016] Ellis, R. (2016). Focus on form : A critical review. *Language Teaching Research*, 20(3) : 405–428.
- [Ellis et al., 2002] Ellis, R., Basturkmen, H., et Loewen, S. (2002). Doing focus-on-form. *System*, 30(4) : 419–432.
- [Ester et al., 1996] Ester, M., Kriegel, H.-P., Sander, J., et Xu, X. (1996). A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD’96*, pages 226–231. AAAI Press.
- [Flege, 2009] Flege, J. E. (2009). Give input a chance. *Input matters in SLA*, pages 175–190.
- [Fraley et Raftery, 1998] Fraley, C. et Raftery, A. E. (1998). How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal*, 41(8) : 578–588.
- [Fraser et al., 1963] Fraser, C., Bellugi, U., et Brown, R. (1963). Control of grammar in imitation, comprehension, and production. *Journal of verbal learning and verbal behavior*, 2(2) : 121–135.

-
- [Gass, 1997] Gass, S. M. (1997). *Input, interaction, and the second language learner*. Routledge.
- [Geertje van Bergen et al., 2014] Geertje van Bergen, Rebekah Rast, et Ellenor Shoemaker (2014). Recognizing lexical forms in the speech stream at first exposure.
- [Genesee, 1987] Genesee, F. (1987). *Learning through two languages : Studies of immersion and bilingual education*. Newbury house publishers.
- [Giacobbe, 1992] Giacobbe, J. (1992). *Acquisition d'une langue étrangère : cognition et interaction : études sur le développement du langage chez l'adulte*. CNRS éd.
- [Gniadek, 1979] Gniadek, S. (1979). *Grammaire contrastive franco-polonaise*. Państwowe Wydawnictwo Naukowe.
- [Goldschneider et DeKeyser, 2001] Goldschneider, J. M. et DeKeyser, R. M. (2001). Explaining the “Natural Order of L2 Morpheme Acquisition” in English : A Meta-analysis of Multiple Determinants. *Language Learning*, 51(1) : 1–50.
- [Grigornko et al., 2000] Grigornko, E. L., Sternberg, R. J., et Ehrman, M. E. (2000). A theory-based approach to the measurement of foreign language learning ability : The Canal-F theory and test. *The Modern Language Journal*, 84(3) : 390–405.
- [Guillaumin et al., 2009] Guillaumin, M., Verbeek, J., et Schmid, C. (2009). Is that you? Metric learning approaches for face identification. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 498–505. IEEE.
- [Gullberg et al., 2010] Gullberg, M., Roberts, L., Dimroth, C., Veroude, K., et Indefrey, P. (2010). Adult language learning after minimal exposure to an unknown natural language. *Language Learning*, 60 : 5–24.
- [Hatch, 1983] Hatch, E. M. (1983). *Psycholinguistics : a second language perspective*. Newbury House. Google-Books-ID : tGh5AAAAIAAJ.
- [Havens et Bezdek, 2012] Havens, T. et Bezdek, J. (2012). *An Efficient Formulation of the Improved Visual Assessment of Cluster Tendency (iVAT) Algorithm*, volume 24.
- [Hendriks et Prodeau, 1999] Hendriks, H. et Prodeau, M. (1999). Isn't Dutch a Mixture of English and German? : On the influence of a second language on the acquisition of a third one. In *Ninth Annual European Second Language Association Conference, Lund*.
- [Herrera et Martínez, 2000] Herrera, F. et Martínez, L. (2000). A 2-tuple fuzzy linguistic representation model for computing with words. *Fuzzy Systems, IEEE Transactions on*, 8(6) : 746–752.
- [Hinz et al., 2013] Hinz, J., Krause, C., Rast, R., Shoemaker, E. M., et Watorek, M. (2013). Initial processing of morphological marking in nonnative language acquisition : Evidence from French and German learners of Polish. *EUROSLA Yearbook*, 13 : 139–175.
- [Hulstijn, 2015] Hulstijn, J. H. (2015). Discussion : How Different Can Perspectives on L2 Development Be? : Perspectives on L2 Development. *Language Learning*, 65(1) : 210–232.
- [Jain, 2010] Jain, A. K. (2010). Data clustering : 50 years beyond K-means. *Pattern Recognition Letters*, 31(8) : 651–666.
- [Jain et Dubes, 1988] Jain, A. K. et Dubes, R. C. (1988). *Algorithms for Clustering Data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
- [Jain et al., 1999a] Jain, A. K., Murty, M. N., et Flynn, P. J. (1999a). Data Clustering : A Review. *ACM Comput. Surv.*, 31(3) : 264–323.

- [Jain et al., 1999b] Jain, A. K., Murty, M. N., et Flynn, P. J. (1999b). Data Clustering : A Review. *ACM Comput. Surv.*, 31(3) : 264–323.
- [Ji He et al., 2004] Ji He, Man Lan, Chew-Lim Tan, Sam-Yuan Sung, et Hwee-Boon Low (2004). Initialization of cluster refinement algorithms : a review and comparative study. In *2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541)*, volume 1, pages 297–302.
- [Jolliffe, 1986] Jolliffe, I. (1986). Principal component analysis. *Springer Series in Statistics, Berlin : Springer, 1986*.
- [Kachinske et al., 2015] Kachinske, I., Osthus, P., Solovyeva, K., et Long, M. (2015). Implicit learning of a L2 morphosyntactic rule, and its relevance for language teaching. *Implicit and explicit learning of languages*, pages 387–417.
- [Kellerman, 1995] Kellerman, E. (1995). Crosslinguistic Influence : Transfer to Nowhere?*. *Annual Review of Applied Linguistics*, 15 : 125–150.
- [Khan et Ahmad, 2004] Khan, S. S. et Ahmad, A. (2004). Cluster center initialization algorithm for K-means clustering. *Pattern Recognition Letters*, 25(11) : 1293–1302.
- [Klein et al., 2002] Klein, D., Kamvar, S. D., et Manning, C. D. (2002). From instance-level constraints to space-level constraints : Making the most of prior knowledge in data clustering.
- [Klein, 1989] Klein, W. (1989). *L'acquisition de langue étrangère*. Armand Colin.
- [Klein, 2001] Klein, W. (2001). Elementary forms of linguistic organisation. In Trabant, J. et Ward, S., éditeurs, *New Essays on the Origin of Language*. DE GRUYTER MOUTON, Berlin, New York.
- [Klein et Dittmar, 1979] Klein, W. et Dittmar, N. (1979). Developing grammars.
- [Klein et Perdue, 1992] Klein, W. et Perdue, C. (1992). *Utterance Structure : Developing grammars again*, volume 5 de *Studies in Bilingualism*. John Benjamins Publishing Company, Amsterdam.
- [Klein et Perdue, 1997] Klein, W. et Perdue, C. (1997). The Basic Variety (or : Couldn't natural languages be much simpler?). *Second language research*, 13(4) : 301–347.
- [Klein et Von Stutterheim, 1987] Klein, W. et Von Stutterheim, C. (1987). Quaestio und referentielle Bewegung in Erzählungen. *Linguistische Berichte*, 109 : 163–183.
- [Klir et Yuan, 1995] Klir, G. et Yuan, B. (1995). *Fuzzy sets and fuzzy logic*, volume 4. Prentice hall New Jersey.
- [Kohonen, 1990] Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9) : 1464–1480.
- [Krashen, 1992] Krashen, S. (1992). The Input Hypothesis : An Update.?. *Linguistics and language pedagogy : The state of the art*, pages 409–431.
- [Krashen, 1980] Krashen, S. D. (1980). *The input hypothesis : Issues and implications*. Addison-Wesley Longman Ltd.
- [Kulis, 2013] Kulis, B. (2013). Metric Learning : A Survey. *Foundations and Trends® in Machine Learning*, 5(4) : 287–364.
- [Lahousse, 2003] Lahousse, K. (2003). La complexité de la notion de topique et l'inversion du sujet nominal. *Travaux de linguistique*, (2) : 111–136.
- [Lance et Williams, 1967] Lance, G. N. et Williams, W. T. (1967). A General Theory of Classificatory Sorting Strategies : 1. Hierarchical Systems. *The Computer Journal*, 9(4) : 373–380.
- [Latos,] Latos, A. From input to output. *INALCO*.

-
- [Latos, 2014] Latos, A. (2014). The Effects of Meaning-Based and Form-Based Input on the Initial L2 Acquisition of Polish Verbal Morphology. In *Studi italiani di linguistica slava : strutture, uso e acquisizione*. Firenze University Press, Firenze.
- [lebart et al., 2006] lebart, piron, et morineau (2006). *Statistique exploratoire multidimensionnelle*. Sciences sup, Dunod.
- [Levelt, 1989] Levelt, W. (1989). *Speaking : From intention to articulation*. ACL-MIT Press series in natural-language processing.
- [Li et al., 2008] Li, Z., Liu, J., et Tang, X. (2008). Pairwise constraint propagation by semidefinite programming for semi-supervised classification. In *Proceedings of the 25th international conference on Machine learning*, pages 576–583. ACM.
- [Lisker, 2001] Lisker, L. (2001). Hearing the Polish sibilants [s š ś] : Phonetic and auditory judgements. *To Honour Eli Fischer-Jørgensen (Travaux du Cercle Linguistique de Copenhague XXXI)*, pages 226–238.
- [Liu et Rodríguez, 2014] Liu, H. et Rodríguez, R. M. (2014). A fuzzy envelope for hesitant fuzzy linguistic term set and its application to multicriteria decision making. *Information Sciences*, 258 : 220–238.
- [Loewen, 2005] Loewen, S. (2005). INCIDENTAL FOCUS ON FORM AND SECOND LANGUAGE LEARNING. *Studies in Second Language Acquisition*, 27(03).
- [Long, 1983] Long, M. H. (1983). Does Second Language Instruction Make a Difference? A Review of Research. *TESOL Quarterly*, 17(3) : 359–382.
- [Long, 1991] Long, M. H. (1991). Focus on form : A design feature in language teaching methodology. *Foreign language research in cross-cultural perspective*, 2(1) : 39–52.
- [Long, 1998] Long, M. H. (1998). Focus on form Theory, research, and practice. In *Focus on form in classroom second language acquisition*, pages 15–41.
- [Long et Doughty, 2011] Long, M. H. et Doughty, C. J. (2011). *The handbook of language teaching*, volume 63. John Wiley & Sons.
- [MacWhinney, 2009] MacWhinney, B. (2009). 1. Une histoire de paradigmes. In Kail, M., Fayol, M., et Hickmann, M., éditeurs, *Apprentissage des langues*, pages 29–46. CNRS Éditions.
- [Macwhinney, 2010] Macwhinney, B. (2010). Computational models of child language learning : an introduction. *Journal of Child Language*, 37(03) : 477.
- [Mallapragada et al., 2008] Mallapragada, P. K., Jin, R., et Jain, A. K. (2008). Active query selection for semi-supervised clustering. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE.
- [Marianne Starren et al., 2013] Marianne Starren, Marzena Watorek, Agnieszka Latos, Rebekah Rast, et Heather Hilton (2013). Processing morpho-syntax at first exposure : The role of source language, input and learner variability.
- [McDade et al., 1982] McDade, H. L., Simpson, M. A., et Lamb, D. E. (1982). The use of elicited imitation as a measure of expressive grammar : A question of validity. *Journal of Speech and Hearing disorders*, 47(1) : 19–24.
- [McGuire, 2007] McGuire, G. (2007). English Listeners’ Perception of Polish Alveopalatal and Retroflex Voiceless Sibilants : A Pilot Study. *UC Berkeley PhonLab Annual Report*, 3(3).
- [Meisel et al., 1981] Meisel, J. M., Clahsen, H., et Pienemann, M. (1981). On determining developmental stages in natural second language acquisition. *Studies in second language acquisition*, 3(2) : 109–135.

- [Moore, 1988] Moore, B. (1988). Art 1 and pattern clustering. In *Proceedings of the 1988 connectionist models summer school*, pages 174–185. Morgan Kaufmann, San Mateo, CA.
- [Moretti, 1989] Moretti, B. (1989). H. Ringbom : The role of the first language in foreign language learning, Multilingual Matters, Clevedon - Philadelphia 1987. *Studi italiani di linguistica teorica e applicata*, 18 : 517–521.
- [Murtagh et Contreras, 2017] Murtagh, F. et Contreras, P. (2017). Algorithms for hierarchical clustering : an overview, II : Algorithms for hierarchical clustering. *Wiley Interdisciplinary Reviews : Data Mining and Knowledge Discovery*, 7(6) : e1219.
- [Myles, 2012] Myles, F. (2012). Complexity, accuracy and fluency The role played by formulaic sequences in early. *Dimensions of L2 performance and proficiency : Complexity, accuracy and fluency in SLA*, 32 : 71.
- [Naiman, 1974] Naiman, N. (1974). *The Use of Elicited Imitation in Second Language Acquisition Research. Working Papers on Bilingualism, No. 2.*
- [Norris et Ortega, 2000] Norris, J. M. et Ortega, L. (2000). Effectiveness of L2 Instruction : A Research Synthesis and Quantitative Meta-analysis. *Language Learning*, 50(3) : 417–528.
- [Odlin, 1989] Odlin, T. (1989). *Language Transfer : Cross-Linguistic Influence in Language Learning.* Cambridge Applied Linguistics. Cambridge University Press.
- [Oyama et Tanaka, 2008] Oyama, S. et Tanaka, K. (2008). Distance metric learning from cannot-be-linked example pairs, with application to name disambiguation. In *Constrained Clustering : Advances in Algorithms, Theory, and Applications*, pages 357–374.
- [Perdue, 1993] Perdue, C. (1993). *Adult Language Acquisition : Cross-linguistic Perspectives.* Cambridge University Press.
- [Perdue, 1996] Perdue, C. (1996). Pre-basic varieties : The first stages of second language acquisition. *Toegepaste taalwetenschap in artikelen*, 55(1) : 135–149.
- [Perdue et Foundation, 1984] Perdue, C. et Foundation, E. S. (1984). *Second language acquisition by adult immigrants : a field manual.* Newbury House Publishers. Google-Books-ID : mVxiAAAAMAAJ.
- [Piske et Young-Scholten, 2008] Piske, T. et Young-Scholten, M. (2008). *Input matters in SLA.* Multilingual Matters.
- [Quirk, 2010] Quirk, R. (2010). *A Comprehensive Grammar of the English Language.* PE.
- [Rast, 2008] Rast, R. (2008). *Foreign language input : Initial processing*, volume 28. Multilingual Matters.
- [Rast, 2010] Rast, R. (2010). First exposure : Converting target language input to intake. In *Converging Evidence in Language and Communication Research (CELCR)*, pages 99–115.
- [Rast, 2017] Rast, R. (2017). *Foreign Language Learning and Teaching : From first exposure to first productions.* Thèse d’université, Université américaine de paris.
- [Rast et Dommergues, 2003] Rast, R. et Dommergues, J.-Y. (2003). Towards a characterisation of saliency on first exposure to a second language. *EUROSLA Yearbook*, 3 : 131–156.
- [Rast et al., 2014] Rast, R., Watorek, M., Hilton, H., et Shoemaker, E. (2014). Initial processing and use of inflectional markers : evidence from French adult learners of Polish. In Han, Z. et Rast, R., éditeurs, *First Exposure to a Second Language : Learners’ Initial Input Processing*, pages 64–106. Cambridge University Press.
- [Rast et al., 2018] Rast, R., Watorek, M., Starosciak, K., et Durand, M. E. (2018). Saliency revisited : What helps absolute beginners learn L2/L3 inflectional morphology ? In *EuroSLA*, Münster, Germany.

-
- [Robinson, 2005] Robinson, P. (2005). Aptitude and second language learning. *Annual Review of Applied Linguistics*, 25(1) : 46–73.
- [Robinson, 2007] Robinson, P. (2007). Aptitudes, abilities, contexts, and practice. *Practice in a second language : Perspectives from applied linguistics and cognitive psychology*, pages 256–286.
- [Robinson, 2012] Robinson, P. (2012). Individual differences, aptitude complexes, SLA processes, and aptitude test development. In *New perspectives on individual differences in language learning and teaching*, pages 57–75. Springer.
- [Robinson et Ellis, 2008] Robinson, P. J. et Ellis, N. C. (2008). *Handbook of cognitive linguistics and second language acquisition*. Routledge, New York.
- [Rodriguez et al., 2012] Rodriguez, R. M., Martinez, L., et Herrera, F. (2012). Hesitant Fuzzy Linguistic Term Sets for Decision Making. *IEEE Transactions on Fuzzy Systems*, 20(1) : 109–119.
- [Saffran et al., 1996] Saffran, J. R., Aslin, R. N., et Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294) : 1926–1928.
- [Sapir, 1944] Sapir, E. (1944). Grading, A Study in Semantics. *Philosophy of Science*, 11(2) : 93–116.
- [Saturno,] Saturno, J. PERCEPTUAL PROMINENCE AND MORPHOLOGICAL PROCESSING IN INITIAL SECOND LANGUAGE ACQUISITION. page 20.
- [Saturno, 2015a] Saturno, J. (2015a). Effects of input condition on case ending processing in initial Polish L2. In *Within Language, Beyond Theories (Volume II) : Studies in Applied Linguistics*. Cambridge Scholars Publishing.
- [Saturno, 2015b] Saturno, J. (2015b). Perceptual prominence and morphological processing in initial second language acquisition. *DISUCOM PRESS*.
- [Saturno, 2016] Saturno, J. (2016). *utterance structure in initial L2 acquisition : Acquiring Polish morphology in light of semantics, pragmatics, and the input*. Thèse d’université, Bergamo, Italie.
- [Saturno et Watorek, 2020] Saturno, J. et Watorek, M. (2020). The emergence of functional case marking in initial varieties of Polish L2. *Langage, Interaction & Acquisition*.
- [Saturno Jacopo, 2014] Saturno Jacopo (2014). Case-Ending Processing in Initial Polish L2 The Role of Frequency, Word Order and Lexical Transparency. pages 341–353, Frankfurt.
- [Schmidt, 2010] Schmidt, R. (2010). ATTENTION, AWARENESS, AND INDIVIDUAL DIFFERENCES IN LANGUAGE LEARNING. page 21.
- [Schmidt, 1990] Schmidt, R. W. (1990). The Role of Consciousness in Second Language Learning1. *Applied Linguistics*, 11(2) : 129–158.
- [Scowen et Grove, 1993] Scowen, R. S. et Grove, B. (1993). Extended BNF — A generic base standard. page 10.
- [Selinker, 1972] Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching*, 10(1-4) : 209–232.
- [Shah et al., 2012] Shah, G. H., Bhensdadia, C., et Ganatra, A. P. (2012). An empirical evaluation of density-based clustering techniques. *International Journal of Soft Computing and Engineering (IJSCE) ISSN, 22312307* : 216–223.
- [Sharwood-Smith, 1986] Sharwood-Smith, M. (1986). Comprehension versus acquisition : Two ways of processing input. *Applied linguistics*, 7 : 239.

- [Shoemaker, 2014] Shoemaker (2014). The Development of Perceptual Sensitivity to Polish Sibilants at First Exposure.
- [Shoemaker et Rast, 2013] Shoemaker, E. et Rast, R. (2013). Extracting words from the speech stream at first exposure. *Second Language Research*, 29(2) : 165–183.
- [Skehan, 2002] Skehan, P. (2002). Theorising and updating aptitude. *Individual differences and instructed language learning*, 2 : 69–94.
- [Skehan, 2013] Skehan, P. (2013). Language aptitude. In *The Routledge handbook of second language acquisition*, pages 399–413. Routledge.
- [Skehan, 2016] Skehan, P. (2016). Foreign language aptitude, acquisitional sequences, and psycholinguistic processes. *Cognitive individual differences in second language processing and acquisition*, pages 17–40.
- [Slobin, 1993] Slobin, D. (1993). *Adult language acquisition : A view from child language study*.
- [Slobin, 2012] Slobin, D. (2012). Child language study and adult language acquisition : twenty years later. In *Comparative Perspectives on Language Acquisition : A Tribute to Clive Perdue*, pages 245–262. Multilingual Matters. Google-Books-ID : KgrPBQAAQBAJ.
- [Slobin, 1985] Slobin, D. I. (1985). Crosslinguistic evidence for the language-making capacity. In *The crosslinguistic study of language acquisition, Vol. 1 : The data ; Vol. 2 : Theoretical issues.*, pages 1157–1256. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.
- [Sparks et Ganschow, 2001] Sparks, R. et Ganschow, L. (2001). Aptitude for learning a foreign language. *Annual Review of Applied Linguistics*, 21 : 90–111.
- [Sublemontier, 2012] Sublemontier, J.-H. (2012). *Classification non supervisée : de la multiplicité des données à la multiplicité des analyses*. PhD Thesis, Université d’Orléans.
- [Swain, 1985] Swain, M. (1985). Large-scale communicative language testing : A case study. *New directions in language testing*, pages 35–46.
- [Tan, 2006] Tan, P.-N. (2006). *Introduction to data mining*. Pearson Education India.
- [Tian, 2011] Tian, Z. (2011). The Influence of Linguistics Theories on Foreign Language. *Journal of Qiqihar Junior Teachers’ College*, (6) : 29.
- [Truck et Abchir, 2014] Truck, I. et Abchir, M.-A. (2014). Toward a Classification of Hesitant Operators in the 2-Tuple Linguistic Model : HESITANT OPERATORS IN THE 2-TUPLE LINGUISTIC MODEL. *International Journal of Intelligent Systems*, 29(6) : 560–578.
- [Truck et Akdag, 2006] Truck, I. et Akdag, H. (2006). Manipulation of qualitative degrees to handle uncertainty : formal models and applications. *Knowledge and Information Systems*, 9(4) : 385–411.
- [Tung et al., 2008] Tung, A. K., Han, J., Lakshmanan, L. V., et Ng, R. T. (2008). Privacy-Preserving Data Publishing : A Constraint-Based Clustering Approach. *Constrained Clustering : Advances in Algorithms, Theory, and Applications*, page 375.
- [Valentini et Grassi, 2014] Valentini, A. et Grassi, R. (2014). The role of input properties on lexical development in foreign language acquisition : Transparency and frequency.
- [Van Patten, 2004] Van Patten, B. (2004). *Input and Output in establishing form meaning connections*. In B. Van Patten, J. Williams, S. Rott, & M. Overstreet (eds.), *Form-meaning connections in second language acquisition (pp. 29-47)*. Mahwah, NJ : Lawrence Erlbaum.
- [VanPatten, 1996] VanPatten, B. (1996). *Input processing and grammar instruction in second language acquisition*. Greenwood Publishing Group.

-
- [VanPatten, 2004] VanPatten, B. (2004). *Processing Instruction : Theory, Research, and Commentary*. Routledge.
- [Verley,] Verley, J. L. Espaces métriques. *Encyclopaedia Universalis [en ligne]*.
- [Vinther, 2002] Vinther, T. (2002). Elicited imitation :a brief overview. *International Journal of Applied Linguistics*, 12(1) : 54–73.
- [Vu et al., 2010] Vu, V.-V., Labroche, N., et Bouchon-Meunier, B. (2010). Boosting Clustering by Active Constraint Selection. In *ECAI*, volume 10, pages 297–302.
- [Véronique, 1994] Véronique, D. (1994). Quel profil d’apprenant ? Réflexions méthodologiques. *Acquisition et interaction en langue étrangère*, (4) : 109–129.
- [Wagstaff et Cardie, 2000] Wagstaff, K. et Cardie, C. (2000). Clustering with Instance-level Constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning, ICML ’00*, pages 1103–1110, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- [Wagstaff et al., 2001] Wagstaff, K., Cardie, C., Rogers, S., et Schrödl, S. (2001). Constrained k-means clustering with background knowledge. In *ICML*, volume 1, pages 577–584.
- [Watorek et al., 2020] Watorek, Rast, et arslangul (2020). Texte de clôture : Interface acquisition/didactique –vers une recherche-action. *INALCO*.
- [Watorek, 1996] Watorek, M. (1996). Le traitement prototypique : définition et implications. *Toegepaste Taalwetenschap in Artikelen*, 55(1) : 187–200.
- [Watorek et al., 2016] Watorek, M., Durand, M., et Starosciak, K. (2016). L’impact de l’input et du type de tâche sur la production de la morphologie nominale en polonais par des apprenants francophones débutants. *Discours. Revue de linguistique, psycholinguistique et informatique. A journal of linguistics, psycholinguistics and computational linguistics*, (18).
- [Watorek et al., 2017] Watorek, M., Rast, R., Durand, M., Dimroth, C., et Starren, M. (2017). L’influence du type d’enseignement sur l’appropriation de la morphologie au début de l’apprentissage d’une langue étrangère. *Le Français dans le monde. Recherches et applications*.
- [Weinberger et al., 2006] Weinberger, K. Q., Blitzer, J., et Saul, L. K. (2006). Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480.
- [Weinberger et Saul, 2009] Weinberger, K. Q. et Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb) : 207–244.
- [Wen et al., 2017] Wen, Z. E., Biedroń, A., et Skehan, P. (2017). Foreign language aptitude theory : Yesterday, today and tomorrow. *Language Teaching*, 50(01) : 1–31.
- [White et White, 2003] White, L. et White, L. (2003). *Second language acquisition and universal grammar*. Cambridge University Press.
- [Williams, 2001] Williams, J. (2001). The effectiveness of spontaneous attention to form. *System*, 29(3) : 325–340.
- [Wong, 2004] Wong, W. (2004). The Nature of Processing instruction. In *Processing Instruction : Theory, Research, and Commentary*, pages 33–65. Routledge.
- [Wowczko, 2013] Wowczko, I. A. (2013). Density Based Clustering with DBSCAN and OPTICS - Literature Review.

- [Xing et al., 2003] Xing, E. P., Jordan, M. I., Russell, S. J., et Ng, A. Y. (2003). Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, pages 521–528.
- [Xiong et De la Torre, 2013] Xiong, X. et De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 532–539. IEEE.
- [Yan et al., 2008] Yan, R., Zhang, J., Yang, J., et Hauptmann, A. G. (2008). Learning with Pairwise Constraints for Video Object Classification. *Constrained Clustering : Advances in Algorithms, Theory, and Applications*, page 397.
- [Ying et Li, 2012] Ying, Y. et Li, P. (2012). Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research*, 13(Jan) : 1–26.
- [Zadeh, 1965a] Zadeh, L. A. (1965a). Fuzzy sets. *Information and control*, 8(3) : 338–353.
- [Zadeh, 1965b] Zadeh, L. A. (1965b). Information and control. *Fuzzy sets*, 8(3) : 338–353.
- [Zadeh, 1999] Zadeh, L. A. (1999). Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 100 : 9–34.

Annexe A

Meaning based et form based illustration

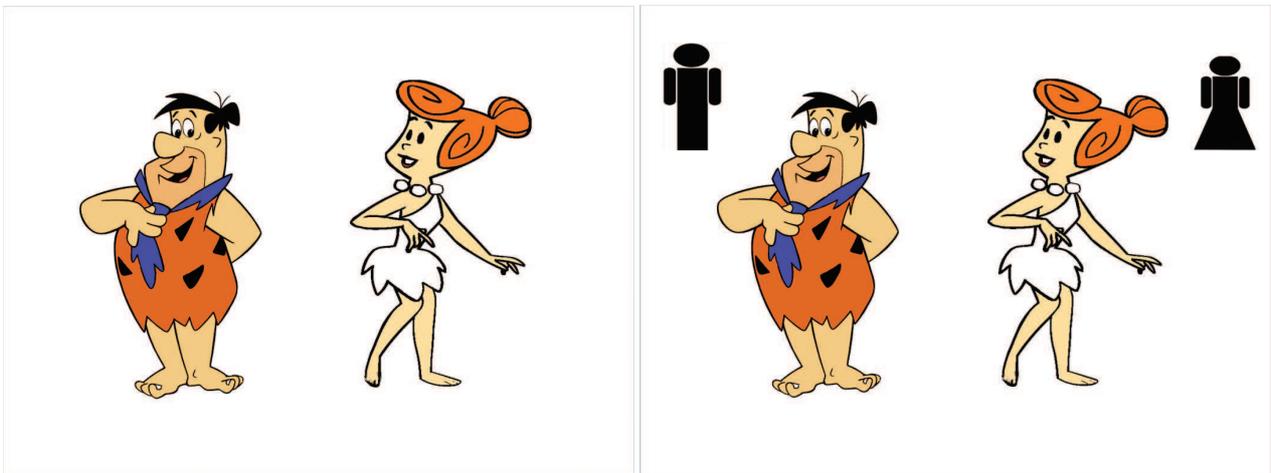


FIGURE A.1 – Différence entre les supports de cours *meaning based* (première image) et *form based* (deuxième image). Le genre des entités est ici souligné dans le cours form based.

Résumé

L'objectif de cette thèse est d'élaborer une méthodologie efficace permettant de décrire le profil de l'apprenant d'une L2 à partir de données d'acquisition (perception, compréhension et production). La méthodologie proposée appartient au domaine de l'intelligence artificielle, plus spécifiquement aux techniques de clustering [Jain et al., 1999b].

Le clustering est une méthode de classification non supervisée qui vise à créer des sous-groupes d'objets appelés clusters. L'objectif est de détecter des régularités dans les comportements acquisitionnels de sous-groupes d'apprenants, en tenant compte de l'aspect multidimensionnel du processus d'apprentissage L2.

Notre algorithme a été appliqué à la base de données du projet VILLA [Dimroth et al., 2013], qui inclut les données d'acquisition d'apprenants de 5 langues sources différentes (français, italien, néerlandais, allemand et anglais) avec le polonais comme langue cible. Chaque apprenant a été testé avec une variété de tests en polonais pendant 14h de session d'enseignement, à partir de l'exposition initiale. Par conséquent, la base de données stocke l'évolution du processus d'acquisition de 156 adultes.

Ainsi, nous avons appliqué l'algorithme aux performances des apprenants sur les niveaux d'analyse linguistique que sont la phonologie, la morphologie, la morphosyntaxe et le lexique. La base de données inclut également la sensibilité des apprenants aux caractéristiques de l'input, telles que la fréquence et la transparence des éléments lexicaux utilisés dans les tâches linguistiques.

Notre algorithme revisite la mesure de similarité utilisée dans les techniques classiques de clustering, afin d'évaluer la distance entre deux apprenants d'un point de vue acquisitionniste. La mesure que nous avons utilisée, appelée " mesure de comparabilité ", repose sur l'identification de la stratégie de réponse de l'apprenant à une structure de test linguistique spécifique. Nous montrons que cette mesure permet de détecter la présence ou l'absence dans les réponses de l'apprenant d'une stratégie proche du système flexionnel de la LC. Ce procédé fournit une classification des apprenants cohérente avec la recherche sur l'acquisition de la langue seconde [Klein et Perdue, 1997, Rast, 2008] et apporte de nouvelles pistes de réflexions sur les parcours acquisitionnels des apprenants *ab initio* .

Mots-clés: acquisition d'une langue seconde; petit ensemble de données; profil de l'apprenant L2; classification semi-supervisée

Abstract

The purpose of this study is to elaborate an efficient methodology allowing the description of learner's profile of an L2 based on acquisition data (perception, comprehension and production). The proposed methodology belongs to Artificial Intelligence field, more specifically to clustering techniques [Jain et al., 1999b].

Clustering is a non-supervised classification methodology which aims at creating subgroups of objects called clusters. The purpose is to detect subgroup regularities with respect to acquisition among the learners (objects), by taking into account the multidimensional aspect of the L2 learning process.

Our algorithm has been applied to the data base of the VILLA project [Dimroth et al., 2013], which includes the performance of learners from 5 different source languages (French, Italian, Dutch, German and English) with Polish as the target language. Each learner was tested with a variety of Polish tests during 14h of teaching session, starting from initial exposure. Consequently, the data base stores the evolution of the acquisitional process of 160 adults.

Thus, we have applied our algorithm on the learners' performances at phonological, morphological, morphosyntactic and lexical levels. These performances also include results with regards to input characteristics, such as frequency and transparency of lexical items used in language tasks that may have influenced the learner outcomes.

Our algorithm revisits the similarity measure used in classical clustering techniques, in order to evaluate the distance between two learners from a SLA point of view. The measure we used, called "comparability measure", relies on the identification of a learner's response strategy to a specific language test structure. We show that this measure allows the detection of the presence or the absence of a coherent strategy in learner's answers, and so enables our algorithm to provide a resulting classification consistent with second language acquisition research [Klein et Perdue, 1997, Rast, 2008]. As a result, we claim that our algorithm might be relevant in the empirical establishment of learners' profiles and the discovery of new opportunities for reflection or analysis.

Keywords: second language acquisition data analysis ; small dataset ; L2 learner's profile ; semi-supervised clustering

