N°d'ordre NNT : 2016LYSE2089

THESE de DOCTORAT DE L'UNIVERSITÉ DE LYON

Opérée au sein de

L'UNIVERSITÉ LUMIÈRE LYON 2

École Doctorale : ED 484
Lettres, Langues, Linguistique et Arts

Discipline : Langues, Littérature anglaise et anglophones

Soutenue publiquement le 28 septembre 2016, par :

Aleksandra LICZNER

Une contribution à l'amélioration des ressources terminographiques :

Étude terminologique fondée sur un corpus de textes de spécialité du domaine du droit de l'Internet

Devant le jury composé de :

John HUMBLEY, Professeur émérite, Université Paris 7, Président
Alain POLGUÈRE, Professeur des universités, Université de Lorraine, Rapporteur
Henri BÉJOINT, Professeur émérite, Université Lumière Lyon 2, Examinateur
Anne CONDAMINES, Directrice de Recherche CNRS, Examinateur
François MANIEZ, Professeur des universités, Université Lumière Lyon 2, Directeur de thèse

UNIVERSITÉ LUMIÈRE LYON 2

École doctorale 484 3LA - Lettres, langues, linguistique et arts

Une contribution à l'amélioration des ressources terminographiques : étude terminologique fondée sur un corpus de textes de spécialité du domaine du droit de l'Internet

Aleksandra LICZNER

Thèse de doctorat en Lexicologie et terminologie multilingues, traduction

dirigée par Monsieur le Professeur François MANIEZ

Présentée et soutenue publiquement le 28 septembre 2016

Membres du jury :

Monsieur Henri BÉJOINT, Professeur émérite, Université Lumière Lyon 2

Madame Anne CONDAMINES, Directrice de Recherche CNRS, Université Toulouse Jean Jaurès

Monsieur John HUMBLEY, Professeur émérite, Université Paris Diderot (rapporteur)

Monsieur François MANIEZ, Professeur des universités, Université Lumière Lyon 2

Monsieur Alain POLGUÈRE, Professeur des universités, Université de Lorraine (rapporteur)

À ma petite Sofia

Remerciements

Je voudrais remercier les nombreuses personnes qui ont contribué de différentes façons à mon travail de recherche et à mon évolution lors de mes études en thèse de doctorat.

Je tiens tout d'abord à exprimer mes plus vifs remerciements à Monsieur François Maniez, mon directeur de thèse à qui je dois d'avoir mené jusqu'au bout ce travail. Je le remercie pour sa patience, pour sa disponibilité toujours bienveillante et surtout pour la confiance qu'il m'a accordée tout au long de ce travail.

J'adresse mes remerciements à Monsieur John Humbley et à Monsieur Alain Polguère pour avoir accepté d'être les rapporteurs de ce travail. Je tiens également à remercier Madame Anne Condamines et Monsieur Henri Béjoint de m'avoir fait l'honneur de participer à ce jury de thèse.

Merci encore une fois à Monsieur Alain Polguère pour ses précieuses remarques qui m'ont permis de mieux comprendre le fonctionnement des FL et m'ont indiqué de nombreuses pistes vers lesquelles orienter le présent travail.

Je remercie également l'équipe du CRTT (Centre de Recherche en Terminologie et Traduction) de l'Université Lyon II, au sein duquel j'ai eu l'occasion de développer mon projet de thèse, pour son accueil chaleureux.

Un grand, grand merci à mes parents pour leur dévouement, tous les sacrifices et l'aide inestimable qui m'ont apportée en s'occupant de Sofia et en m'allégeant de bien des tâches au cours de cette dernière année. Sans eux, ce travail n'aurait jamais pu voir le jour. *Dziękuję Wam, Kochani Rodzice, za nieocenioną pomoc. Bez Was, tej pracy by nie było*.

Merci à mon mari, Aziz, pour sa patience, son soutien et les encouragements constants qu'il m'a apportés jour après jour jusqu'au point final de ce manuscrit. *Mo vi' ĩ no yidumă, a yi'ai ko o wadanma* .

Sommaire

Liste des abréviations

BCT	Base de Connaissances Terminologiques
CRTT	Centre de Recherche en Terminologie et Traduction
DAD	*Dictionnaire analytique de la distribution*
DAFA	*Dictionnaire d'Apprentissage du Français des Affaires*
DAFLES	*Dictionnaire d'Apprentissage du Français Langue Étrangère ou seconde*
DATM	*Dictionnaire analytique de la mondialisation et du travail*
DEC	*Dictionnaire Explicatif et Combinatoire*
FE	Frame Element
FL	Fonction Lexicale
FLNS	Fonction Lexicale non Standard
FLSS	Fonction Lexicale Semi-Standard
LAF	*Lexique Actif du Français*
LEC	Lexicologie Explicative et Combinatoire
MST	Modèle Sens-Texte
OLST	Observatoire de linguistique Sens Texte
RA	Rélation associative
RLF	*Réseau Lexical du Français*
RLS	Rélation lexico-sémantique
RTO	Ressource termino-ontologique
TAL	Traitement Automatique des Langues
TGT	Théorie Générale de la Terminologie
TST	Théorie Sens-Texte

Liste des symboles et notations

⸢S⸣	les sens linguistiques seront présentés entre guillemets simples spéciaux, notation empruntée à la LEC
*S	expression jugée sémantiquement ou syntaxiquement inacceptable, notation empruntée à la LEC
L	Lexie
Fonction Lexicale	fonction lexicale
ART	Article ou autre déterminant dans les régimes des valeurs de FL, notation empruntée à la LEC
~	mot-clé dans les FL ou les régimes de valeurs, notation empruntée à la LEC
{CADRE SÉMANTICO - CONCEPTUELS}	cadre sémantico-conceptuel dans le *DITerm*
CADRE SÉMANTIQUE	cadre sémantique dans le *DITerm*
`terme typique`	terme typique dans le *DITerm*
type circonstant	type de relation circonstancielle dans le *DITerm*
[précisions]	précisions sémantiques dans le *DITerm*

Introduction

Problématique de la recherche

Notre projet tire son origine d'une constatation : les ressources terminographiques conventionnelles (dictionnaires spécialisés, glossaires, thésaurus) destinées aux traducteurs ne répondent pas à tous les besoins de leurs utilisateurs (nous abordons ici l'activité de traduction surtout du point de vue de la production linguistique, c'est-à-dire de la réexpression du sens). En effet, il s'agit la plupart du temps d'outils qui proposent de simples recueils de termes accompagnés de leurs définitions et d'équivalents et ne fournissent pas suffisamment d'informations sur le fonctionnement linguistique des termes dans leur univers discursif, renseignements pourtant nécessaires afin de produire un texte cohérent aussi bien au niveau terminologique que sur le plan stylistique. Nous constatons que les dictionnaires d'encodage qui permettraient aux traducteurs d'acquérir une sorte d'autonomie discursive dans la langue cible font cruellement défaut dans le milieu. Ainsi, pour combler cette absence, les traducteurs sont amenés à faire appel à d'autres types de ressources, telles que des documents parallèles (Williams cité par Bowker 1998), des corpus électroniques (Maniez 2002), ou bien le Web (Maniez 2007), très utile par la possibilité d'accéder à une multitude de contextes illustrant le comportement des unités terminologiques. Cependant, les données provenant de corpus électroniques ou bien directement du Web se révèlent de fiabilité inégale et demandent un certain travail d'analyse. Malheureusement, vu les délais très serrés souvent imposés par les prescripteurs de services, les traducteurs ne sont pas en mesure de consacrer beaucoup de temps à ce type de recherche terminologique. De plus, les ressources en question ne proposent pas d'informations structurées d'ordre encyclopédique. Or, comme le souligne Dancette (2007 : 553),

« […] *le manque d'une connaissance conceptuelle extralinguistique conduit à des imprécisions dans la traduction ou à des hésitations dans le choix des équivalents ou au contraire à des compensations ».*

Ainsi, l'idéal serait d'avoir à sa disposition un outil terminographique qui propose une description complète des unités terminologiques. La question se pose alors de savoir quelles caractéristiques, quelles propriétés du terme il faut prendre en compte pour fournir une telle

description. Pour cela, il est, tout d'abord, nécessaire de se pencher sur la notion même de terme.

Le terme – une unité à dimensions multiples ?

Comme le souligne L'Homme (2005 : 1114), les modèles théoriques de la terminologie, souvent fondés sur des postulats et des objectifs qui leur sont propres, abordent la notion de « terme » sous des angles différents. Ainsi, pour les terminologues « classiques »[1], le terme s'oppose radicalement au mot : « *la distinction entre terme et mot était érigée en principe, et affirmée sur le plan de la signification, celle du mot dépendant en grande partie de l'environnement linguistique alors que celle du terme aurait été liée avant tout à l'environnement pragmatique* ». (Béjoint et Thoiron 2000 : 5). Le terme est donc caractérisé par : « *la monosémie, l'univocité, la précision de sa définition et un sens uniquement référentiel faisant de lui une étiquette apposée sur la chose* » (Béjoint et Thoiron, 2010 : 105). La terminologie traditionnelle s'est ainsi érigée par opposition à la linguistique en plaçant le concept au cœur de la discipline. La fonction du terme en tant qu'élément d'un système linguistique a été réduite à un simple acte de dénomination des concepts. Pourtant, nombreux sont les auteurs qui reprochent à la terminologie wüsterienne de s'être construite sur une description d'idéaux plutôt que sur l'étude des phénomènes réels (Béjoint et Thoiron 2000 : 5) et d'avoir occulté par-là la vrai nature du terme. Humbley (2004, 2007) remarque que la théorie de Wüster a essuyé de sérieuses critiques, surtout de la part de linguistes des pays francophones qui réduisent la portée de son œuvre à des propositions coupées de la réalité des phénomènes linguistiques. Ces remises en question ont même abouti à la formulation de nouvelles approches telles que la socioterminologie (Gaudin 2003, voir aussi la section 2.2) ou la terminologie textuelle (Bourigault et Slodzian 1999, Slodzian 2000, voir aussi le chapitre 4). Pour ces auteurs, la terminologie classique et ses méthodes sont perçues comme une entrave à une approche permettant de situer et d'analyser le terme dans son milieu naturel, à savoir dans le discours. Le fait d'occulter la dimension linguistique des termes a engendré, selon eux, de nombreuses lacunes dans les descriptions terminographiques (L'Homme et Vandaele 2007 : 4). Mais, comme le suggère Humbley (2001 : 728), cette mise

[1] Nous nous référons ici à la Théorie générale de la terminologie (TGT) attribuée à Wüster (Wüster 1976, Felber 1987, voir aussi Humbley 2004, 2007)

entre parenthèses de la linguistique et du sens en terminologie était peut-être le prix à payer pour avancer sur d'autres fronts.

Quoi qu'il en soit, depuis une vingtaine d'années la linguistique et surtout la sémantique, qui a été bannie par la doctrine terminologique classique, reviennent « non pas par la fenêtre, mais par la grande porte », pour reprendre l'expression de Humbley (*ibid.*). En effet, depuis les années 1990, on observe un rapprochement entre la terminologie et la linguistique et on assiste à un réexamen des concepts fondamentaux de la TGT, suscité essentiellement par le changement des conditions d'exercice de l'activité terminologique. Soulignons que l'équipe du Centre de Recherche en Terminologie et Traduction de l'Université de Lyon 2, au sein duquel nous avons l'honneur de mener la présente recherche, a singulièrement nourri le débat sur le statut et la nature du terme, notamment en publiant deux ouvrages collectifs, « Le sens en terminologie », sous la direction d'Henri Béjoint et de Philippe Thoiron (2000) traitant de la question du mode de signification du terme, et « De la mesure dans les termes » dirigé par Henri Béjoint et François Maniez (2005), consacré aux changements méthodologiques liés à l'exploitation des corpus et à l'introduction d'outils informatiques dans la recherche terminologique. Nous voudrions nous référer tout particulièrement au premier recueil, qui rassemble des contributions abordant la problématique qui nous intéresse ici, notamment l'opposition entre terme et mot.

Même si les points de vue sur les rapports entre le terme et le mot diffèrent d'un article à l'autre, les auteurs de ce recueil s'accordent tous sur la nécessité d'intégrer le sens dans la démarche (Humbley 2001 : 728). Comme le remarquent Béjoint et Thoiron (2000 : 16) : « *Les auteurs de ce volume font preuve d'une relative unanimité sur certains points : non, le terme n'est pas qu'une étiquette posée de manière immuable sur une chose qu'il désigne ; oui, le terme est une unité destinée à fonctionner dans un environnement linguistique, la langue, ou plutôt le discours de spécialité et dans un environnement social*. ». Ainsi, d'après Sager (2000 : 41), les termes peuvent être étudiés indépendamment de tout contexte linguistique, en tant qu'instruments de classification destinés à structurer des connaissances (on est ici plus proche de la vision classique de la terminologie), ou en fonction du rôle qu'ils jouent dans la communication, en tant qu'éléments du discours. En situation de communication, « *les termes constituent des ensembles dynamiques en relation avec les mots du discours* » (Sager 2000 : 41). Par conséquent, un discours spécialisé peut être présenté

comme un mélange de termes et de mots où les deux groupes d'unités se présentent sous la même forme et ne différent que par leur fonction (d'où l'approche fonctionnelle du terme). En effet, selon Sager (*ibid.* : 53), la signification des termes est limitée par le système cognitif auquel ils appartiennent (contrairement à la signification des mots, qui n'est limitée que par celle des autres mots avec lesquels ils sont combinés dans le discours). Par conséquent, les termes doivent renvoyer clairement au référent qu'ils désignent et permettre la transmission efficace des connaissances, tandis que les mots sont destinés à l'expression de ce qui est techniquement imprécis (*ibid.* : 54).

Depecker (2000, 2002), quant à lui, considère le terme comme un signe linguistique à part entière, un signe à sens spécialisé, un signe vivant (*ibid.* : 92). L'auteur souligne tout de même que le terme, élément fondamental de la terminologie, est pris entre trois entités : la pensée constituée (le concept), la langue (le signe linguistique), et le réel (l'objet), configuration qui n'a véritablement été mise à profit ni en linguistique ni en terminologie (2002 : 22). En même temps, l'auteur précise que le concept ne se résume pas au signifié. En effet, il démontre que les signifiés d'une langue à l'autre ne décrivent pas les concepts de la même façon ce qui fait que, dans toute opération de traduction, le degré de recouvrement est forcément aléatoire (2002 : 32). C'est pourquoi, selon Depecker (2002 : 33), « *la terminologie s'efforce de dégager des langues les concepts, et de prendre appui sur ces derniers pour reconstituer le matériau linguistique les nommant (désignations) ou les formulant (définitions) : c'est donc que la terminologie tend à faire des définitions de concepts* ». Le terme se définit ainsi comme un signe linguistique qui renvoie à un concept situable en dehors de la langue.

Nous avons l'impression que les propositions théoriques succinctement présentées plus haut, même si elles placent l'étude du terme dans la linguistique, tentent en quelque sorte d'aménager l'héritage wüsterien. Cabré (2000, 2007) quant à elle, propose un modèle fédérateur dans lequel les différentes manières d'aborder la terminologie seraient prises en charge (L'Homme 2005 : 1116). En effet, il s'agit de la Théorie communicative de la terminologie (TCT), une conception théorique qui tente de réconcilier différents points de vue sur le terme et que l'auteur présente comme une théorie linguistique des unités terminologiques à composantes cognitive et communicative (Cabré 2007 : 98-101). Le terme dans la TCT est donc considéré comme une unité polyédrique, c'est-à-dire une unité à dimension multiples. Selon Cabré (2000 :13), les unités terminologiques partagent de

nombreux traits avec d'autres unités de la langue naturelle, et la communication spécialisée n'est pas une forme de communication complètement différente de la communication générale. Cependant, vu son caractère polyédrique, l'accès à l'unité terminologique peut se faire par des portes différentes (d'où la *théorie des portes*) : la linguistique, les sciences cognitives et les sciences de la communication sociale.

Les auteurs évoqués ci-dessus ne prétendent pas que terme et mot soient deux entités radicalement distinctes. Néanmoins, ils ne nient pas l'existence de différences entre les deux types d'unités. Cependant, comme le soulignent L'Homme et Vandaele (2007 : 6), nombre d'auteurs adoptent un point de vue opposé et abordent explicitement ou implicitement le terme comme une unité faisant partie du lexique d'une langue. Pour résumer ce point de vue, les auteures citent les propos de Kocourek (1991b), qui voit le terme comme une unité correspondant à une acceptation spécifique rattachée au mot : « *Les termes étant des unités lexicales définies, ils ne représentent, potentiellement, que certaines acceptions de l'aire sémantique de l'unité lexicale, à savoir celles qui sont définies dans les textes spécialisés, et non par les lexicographes dans les dictionnaires généraux.* « (Kocourek, 1991 : 180, cité dans L'Homme et Vandaele 2007 : 7). Cette approche, que L'Homme (2004 : 32-38) qualifie de *lexico-sémantique*, aborde les termes comme des unités lexicales constituant un sous-ensemble du lexique d'une langue et dont la particularité est d'avoir un sens qu'on peut associer à un domaine de la connaissance humaine. Comme le souligne l'auteur (*ibid.* : 37), les unités terminologiques entretiennent avec d'autres unités lexicales un ensemble complexe de relations sémantiques qu'il est possible de décrire en ayant recours aux modèles empruntés à la lexicologie et plus particulièrement à la sémantique lexicale (Cruse 1986, Mel'čuk *et al.* 1995, Polguère 2008).

En effet, il convient de remarquer que la démarche lexico-sémantique n'a pas fait l'objet d'une théorisation en terminologie ce qui s'explique, d'après L'Homme (2005 : 1112), par le fait que les questions d'ordre théorique ou méthodologique trouvent des réponses dans les modèles sémantiques. Néanmoins, il est nécessaire de souligner que malgré cette absence de théorisation dans la littérature terminologique, de plus en plus de terminographes se tournent vers les modèles de représentation des unités terminologiques basés sur la sémantique lexicale. Comme le souligne L'Homme (*ibid.* : 1122), les projets terminographiques réalisés dans une optique lexico-sémantique mènent à des descriptions qui constituent un reflet des

observations faites sur un ensemble de données linguistiques et des interactions entretenues par les unités. Cette démarche permet donc de retenir des éléments qui échappent à une perspective strictement conceptuelle en enrichissant ainsi la description des termes.

D'innombrables débats entourent la question du statut du terme et comme nous pouvons le constater, il n'y a pas de manière unique d'y répondre. Pour résumer, nous pouvons dire qu'il est difficile de tracer une frontière nette entre l'unité terminologique et le mot de la langue générale. Nous croyons tout de même, à l'instar de Bourigault et Slodzian (1999 : 30) et de L'Homme (2005 : 1113), que la nature du terme ne peut pas appréhendée en faisant abstraction des objectifs d'une application donnée. Ainsi, sachant que la notion de terme est toujours colorée par la perspective du spécialiste qui l'analyse (L'Homme *ibid.*), nous proposons de l'aborder du point de vue d'un linguiste, comme une unité lexicale. Nous admettons par là qu'il est possible de l'analyser et de la décrire en s'appuyant sur des modèles empruntés à la sémantique lexicale. Cependant, tout comme Cabré (2000, 2007), nous considérons qu'il s'agit d'une unité qui a un statut privilégié puisque, sous l'ange cognitif, elle entre dans la structure conceptuelle du domaine. Le terme, tel que nous l'envisageons dans le cadre de cette étude, est donc une unité à dimensions multiples qui demande une description aussi bien sur le plan linguistique que conceptuel.

Objectifs de la thèse

Ainsi, l'objectif principal de notre travail est de proposer un modèle de description complète des unités terminologiques du domaine du droit de l'Internet, qui doit servir de base à la conception d'un dictionnaire spécialisé destiné aux traducteurs dont la langue de travail est le français. Le projet, baptisé *DITerm,* tente avant tout de répondre aux besoins de compréhension et d'autonomie discursive de ces derniers. Il s'agit d'un modèle qui cherche à rendre compte des usages observés en discours spécialisé afin de permettre aux traducteurs de reconnaître et de générer l'ensemble des emplois. En effet, son ambition est de refléter aussi bien la dimension linguistique des termes (leur nature linguistique, le comportement en langue, leurs relations lexico-sémantiques, les combinaisons lexicales typiques dans lesquelles ils se trouvent) que leur dimension cognitive (la place des termes dans la structure conceptuelle). Le but est de fournir pour chaque unité un grand nombre d'informations de nature linguistique et conceptuelle qui permettront aux traducteurs d'insérer les termes

correctement dans les textes spécialisés. Il s'agit donc d'une ressource explicitement dédiée à la mise en discours, une ressource qui avant tout fournit des données nécessaires à l'encodage.

Le *DITerm* doit donc permettre de :

- trouver un répertoire des termes fondamentaux dans le domaine du droit de l'Internet ;
- trouver des descriptions sémantiques fines facilitant la compréhension des termes vedettes ainsi que des unités qui y sont liées;
- trouver des informations de nature conceptuelle permettant de rattacher le terme au schéma notionnel du domaine
- trouver, pour chacun des termes, l'ensemble des autres termes ou unités lexicales partageant avec lui une relation sémantique ou un lien conceptuel, car la mise en relation des termes du même champ permet de rendre compte de la structure conceptuelle et sémantique du domaine et guide le traducteur dans son approche d'un nouveau domaine ;
- trouver, pour chaque terme, l'ensemble des autres termes ou unités lexicales se combinant de façon privilégiée car la mise en lumière des affinités combinatoires (na nature aussi bien conceptuelle que linguistique) permet de refléter la structure lexicale et notionnelle du domaine.

Il convient de souligner que notre projet s'inscrit dans une mouvance plus générale. En effet, l'insuffisance des ressources terminographiques traditionnelles a déjà été dénoncée par de nombreux terminologues et terminographes comme L'Homme (2002, 2004, 2005), Heid et Freibott (1991), Meyer *et al.* (1992), Kocourek (1991a, 1991b), Dancette (2005, 2009), Cohen (1986), Binon *et al.* (2000). Les auteurs s'accordent à considérer que les ressources terminographiques modernes devraient se fixer comme objectif de faire une description globale de la langue de spécialité. Comme le souligne L'Homme (2004), cette volonté d'enrichir le contenu des dictionnaires spécialisés (ou plutôt des dictionnaires de langues de spécialité), se généralise dans le milieu, même si les méthodologies adoptées restent différentes. En proposant notre modèle de description des unités terminologiques du domaine du droit de l'Internet, nous voudrions donc contribuer à l'amélioration des ressources terminographiques existantes. Nous espérons, à travers cette étude, pouvoir répondre, au

moins en partie, à la question suivante : comment on peut rendre les dictionnaires destinés à la traduction spécialisée plus performants et plus utiles aux traducteurs ?

Pour ce faire, nous serons amenée à mettre en œuvre deux stratégies, souvent considérées comme concurrentes (Dancette 2006) ou bien incompatibles (L'Homme 2005), à savoir :

- la description détaillée du fonctionnement linguistique des termes dans leur univers discursif basée sur l'observation des usages dans le corpus ;
- la structuration des connaissances relatives au droit de l'Internet extraites du corpus en établissant des réseaux intentionnels entre certaines séries de termes liés entre eux

Notons que l'analyse des corpus textuels joue un rôle essentiel dans cette démarche.

Bien évidement, nous ne prétendons pas à la création d'une ontologie du domaine du droit de l'Internet. Nous considérons, tout comme L'Homme (2005 : 1123) ou Cabré (2007 : 82), que la tâche de structuration conceptuelle doit être confiée à des spécialistes des domaines et non pas aux linguistes, dont le rôle est la description du comportement des termes en langue. Néanmoins, nous estimons qu'en proposant une description globale des termes, il est impossible de faire abstraction de leurs propriétés cognitives. C'est donc là que nous nous heurtons à un problème fondamental, un problème qui est d'ordre méthodologique. En effet, la question se pose de savoir s'il est possible de trouver une méthode de description qui intègre aussi bien la dimension linguistique que conceptuelle des termes. Comme le souligne L'Homme (2005 : 1121) le terminographe doit faire un choix parmi les modèles théoriques et descriptifs offerts par la terminologie (c'est-à-dire qu'il doit opter soit pour une approche conceptuelle, soit pour une approche lexico-sémantique) et assumer ses conséquences méthodologiques. En voulant proposer une méthode de description terminologique permettant de rendre compte du double statut des termes, nous nous demandons tout de même dans quelle mesure il est possible de rapprocher ces deux démarches.

Organisation de la thèse

La thèse s'articule autour de quatre parties. Les deux premières parties sont consacrées à la présentation des fondements théoriques de notre recherche. Les deux dernières, quant à elles, retracent les différentes étapes de l'élaboration du projet *DITerm*. Les aspects théoriques et pratiques abordés dans la thèse sont répartis en 10 chapitres.

Nous partons de l'hypothèse que le terme est une sorte d'unité lexicale, une unité lexicale à sens spécialisé qui devrait être décrite en tenant compte des relations qu'elle entretient, aussi bien sur l'axe paradigmatique que syntagmatique, avec d'autres unités terminologiques ou lexicales. Nous admettons par là qu'il est tout à fait possible d'appliquer au traitement des termes les principes propres à la lexicologie, et plus particulièrement à la sémantique lexicale. La première partie aborde donc la question de compatibilité des modèles lexicographiques et terminologiques Le premier chapitre a pour objet d'exposer la théorie linguistique qui sous-tend notre travail de description des unités terminologiques du domaine du droit de l'Internet, à savoir la *théorie Sens-Texte* (TST) (Mel'čuk 1997) et plus précisément sa composante lexicale, appelée *Lexicologie Explicative et Combinatoire* (Mel'čuk *et al.* 1995). Après avoir introduit les bases notionnelles de la théorie mel'čukienne, nous passons à la présentation des fonctions lexicales (dorénavant désignées sous le sigle FL), outil formel proposé dans le cadre de la TST et permettant de modéliser l'ensemble des relations sémantiques. Le deuxième chapitre fait un tour d'horizon des projets terminographiques qui s'inspirent de la Lexicologie Explicative et Combinatoire et proposent d'appliquer le modèle des FL à la description terminographique (Frawley 1988, Mortchev-Bouveret 2006, 2007, Binon *et al.* 2000). Nous nous arrêtons plus particulièrement à la présentation de deux dictionnaires en ligne, à savoir le *DiCoInfo* et le *DiCoEnviro,* issus des travaux menés au sein de l'OLST de l'Université de Montréal et dirigés par Marie-Claude L'Homme. Il s'agit de ressources dont nous nous sommes largement inspirée en travaillant sur le modèle *DITerm*. Nous évoquons également deux autres projets, notamment le *DAD* (Dancette et Réthoré 2000 / 2006), le *DAMT* (Dancette 2010) qui proposent, quant à eux, un autre regard sur la modélisation des relations sémantiques entretenues par les termes.

La deuxième partie aborde un autre aspect théorique important pour notre recherche (qui s'inscrit dans une démarche résolument descriptive), notamment le statut du corpus dans

la linguistique moderne en général et en particulier, en terminologie. Dans le troisième chapitre, nous parcourons les différentes étapes de l'histoire de la linguistique de corpus en évoquant les chercheurs qui ont le plus contribué à son développement (Leech (1991, 1992), Halliday (1966, 1985, 1991), Sinclair (1991). En s'appuyant sur la distinction entre les courants *corpus-based* et *corpus-driven*, nous décrivons différentes manières d'aborder la notion de corpus. À l'occasion, nous présentons également les travaux fondateurs du contextualisme britannique. Le quatrième chapitre s'intéresse aux principaux apports des corpus à la terminologie. Nous évoquerons ici la terminologie textuelle (Bourigault et Slodzian 1999, Slodzian 2000), approche qui est le fruit du rapprochement de la linguistique (linguistique de corpus et analyse du discours) et de la terminologie. Nous y passons en revue différentes méthodes et outils d'analyse des données terminologiques basées sur le corpus.

La troisième partie de la thèse présente les différentes étapes de la constitution et du traitement du corpus *DITerm* en vue de l'extraction d'unités terminologiques du domaine du droit de l'Internet. Dans le cinquième chapitre, nous faisons un bilan sur la notion de corpus et présentons le point de vue que nous avons adopté dans le cadre de ce travail en exposant la démarche retenue pour élaborer notre corpus. Nous décrivons les critères qui ont présidé à la constitution de notre corpus. En effet, il s'agit d'un corpus monolingue (français), comptant environ 5 000 000 mots, caractérisé par une variété de genres et de types de textes (arrêts de la cour, directives, ordonnances, articles de presse, rapports). Dans le sixième chapitre, nous présentons les outils utilisés (*NooJ* et *TermoStat*) et les méthodes mises en œuvre (basées sur des indices quantitatifs et linguistiques) pour extraire les candidats-termes. La septième chapitre de la thèse est destiné à l'analyse des données extraites du corpus *DITerm*. Comme nous l'avons remarqué plus haut, l'unité terminologique n'est plus considérée comme une unité que l'on observe à l'état isolé mais comme un élément qui doit être analysé en contexte. Ainsi, après avoir proposé notre classification des candidat-termes extraits du *DITerm*, nous nous concentrons sur l'analyse de l'environnement contextuel des termes retenus. Nous considérons que l'examen de l'univers discursif des termes peut donner accès à un grand nombre d'informations permettant la structuration du domaine du droit de l'Internet aussi bien au niveau lexical que conceptuel.

La quatrième partie représente le cœur de la thèse. En effet, elle est consacrée à notre proposition de modélisation des données terminologiques extraites du corpus *DITerm*. Tout au long de cette partie (qui contient 3 chapitres), nous essayons de systématiser la multitude

de relations que les termes du domaine du droit de l'Internet entretiennent avec d'autres unités aussi bien sur le plan linguistique que conceptuel. Le huitième chapitre est donc destiné à la recherche d'une méthode de description de la dimension linguistique des termes. Nous nous y intéressons au modèle des fonctions lexicales (Mel'čuk *et al.* 1995) et analysons la possibilité d'adaptation de celui-ci à notre projet terminographique. Le neuvième chapitre décrit notre tentative de systématisation des relations conceptuelles entre les termes du domaine du droit de l'Internet. Pour ce faire, nous nous tournons vers la théorie des cadres, entendus ici au sens large (Minsky 1975, Fillmore 1982, 2009). En effet, nous remarquons que les termes (qui sont associés à des structures de nature conceptuelle et cognitive que nous appelons *cadres sémantico-conceptuels* et qui correspondent à un niveau de conceptualisation plus abstrait) se réalisent dans le discours, dans des combinaisons de formes linguistiques qui peuvent être analysées et annotées au moyen de structures schématiques appelées *cadres sémantiques*. Le dernier chapitre présente un modèle de description hybride permettant de rendre compte aussi bien des relations lexico-sémantiques que des liens conceptuels.

PREMIÈRE PARTIE : Lexicographie et terminologie : compatibilité des modèles et des méthodes[2]

Comme nous l'avons annoncé en introduction, l'objectif principal de cette recherche est de trouver une méthode de description complète des unités terminologiques permettant de rendre compte aussi bien de leur dimension linguistique que conceptuelle et ceci par le biais de leur comportement dans l'univers discursif. Ainsi, en vue de proposer notre propre modèle de représentation, nous allons nous intéresser à la fois aux relations contextuelles, conceptuelles et lexicales des termes.

*« Les unités lexicales ne peuvent être appréhendées comme des entités isolées, closes sur elles-mêmes, elles doivent au contraire être définies en termes d'emplois dans le cadre des phrases où elles apparaissent. » (*Mathieu-Colas & Le Pesant 1998 : 6)

Remarquons que d'un côté, de nombreux spécialistes tels que L'Homme, Meyer, Heid et Freibott, Dancette, Cohen, Kocourek, Binon, Selva, Verlinde, (leurs travaux seront évoqués tout au long de notre étude), soulignent que les ressources terminographiques modernes devraient se fixer comme objectif de faire une description globale de la langue de spécialité. Ils constatent qu'il n'est pas suffisant de présenter des règles générales (réduites très souvent à la simple équation : concept et dénomination égale terme). Selon ces auteurs (dont la plupart adhèrent à une optique lexico-sémantique, voir L'Homme 2004 : 32), il faut décrire toutes les propriétés linguistiques de chaque mot, pour être capable de reconnaître et de générer l'ensemble des emplois. Selon Heid & Freibott (1991 :78), le but du traducteur, aussi bien que de l'auteur technique est de fournir un texte non seulement cohérent au niveau terminologique, mais aussi « lisible » et « riche » sur le plan stylistique. Cela est lié à deux

[2] Le titre de cette partie à été emprunté à l'ouvrage l'ouvrage collectif publié sous la direction de Marie Claude L'Homme et Sylvie Vandaele (2007).

phénomènes, notamment à la dérivation sémantique et à la combinatoire lexicale. Pour atteindre le degré nécessaire de lisibilité dans la langue cible, les traducteurs ou bien les auteurs techniques doivent disposer d'un « trésor » collocationnel[3] important (*ibid.*). Comme le souligne Mel'čuk (2003 : 23), c'est la fréquence et la qualité d'usage des unités phraséologiques qui déterminent la différence entre un locuteur natif et un étranger : « *un natif parle en phrasèmes*[4] ». Cette constatation concerne également les langues de spécialité. D'où l'importance de traiter des phénomènes combinatoires ainsi que d'autres phénomènes sémantiques dans les ressources terminographies afin de fournir aux traducteurs (ou auteurs techniques) une variété d'expressions précises et de formulations appropriées à un domaine de spécialité donné.

De l'autre côté, il ne faut pas négliger la dimension conceptuelle des termes. En effet, comme le souligne Dancette (2003 : 142), la cohérence n'existe pas dans le texte *a priori* mais relève de construits cognitifs. Selon l'auteure, plus les référents cognitifs du traducteur se rapprochent de ceux du scripteur (auteur), plus la cohérence manifestée dans le texte traduit devient compatible avec la cohérence du texte de départ : « *Le dictionnaire basé sur une approche conceptuelle et cognitive offre un réseau de connaissances sur lesquels le traducteur peut échafauder ses référents cognitifs ; en ce sens il peut être un outil précieux en traduction spécialisée.* » (Dancette, *ibid.* : 155). D'où l'importance de la prise en compte des liens conceptuels qu'entretiennent les termes du domaine.

Dans ces deux premiers chapitres, nous nous concentrons sur la dimension linguistique des termes. En effet, nous partons de l'hypothèse que les termes sont des unités lexicales à sens spécialisé. Nous admettons par-là qu'il est tout à fait plausible de s'inspirer des théories et des méthodes développées dans le cadre de la sémantique lexicale. Ainsi, nous proposons tout d'abord, d'analyser différents modèles de description des propriétés lexico-sémantiques des unités en langue générale. Nous nous intéresserons plus particulièrement à la Théorie Sens – Texte de Mel'čuk (1997) et à sa composante lexicale, la Lexicologie Explicative et Combinatoire (Mal'čuk *et al.* 1995) qui constituent le cadre théorique de ce travail.

[3] Nous prenons ici le terme *collocationnel* au sens large, celui des contextualistes britanniques).
[4] Nous définirons le terme dans la suite de notre travail.

Chapitre 1. Traitement des relations sémantiques en lexicologie

1.1 Aperçu des modèles de représentation des unités lexicales en langue générale

Pour pouvoir fournir une description complète des termes qui sont envisagés ici comme des unités lexicales à sens spécialisé, il est tout d'abord nécessaire de s'interroger sur différentes modèles de représentation du lexique en langue générale. En effet, depuis quelques décennies, de nombreux linguistes (Wierzbicka 1996, Lyons 1978, Pottier 1992, Tesnières 1959, Mel'čuk 1997, Cruse 1986) insistent sur la nécessité de placer la sémantique au centre de la description linguistique. Par conséquent, on observe un regain d'intérêt pour le lexique en tant que tel et non uniquement comme élément nécessaire à la description de la syntaxe. Comme le dit Pottier (1992 : 35) :

« *Les deux ouvrages traditionnels dépositaires d'une langue sont le dictionnaire et la grammaire. Mais l'un fait obligatoirement référence à l'autre. Une lexie entraîne un certain nombre de pressions, sémantiques ou syntaxiques, sur son entourage (rections, sélections, affinités)* ».

Murphy (2003, 60-129), dans son ouvrage dédié à la présentation d'une approche pragmatique des relations sémantiques, fait un survol de différents modèles de représentation du lexique aussi bien en linguistique que dans d'autres champs disciplinaires tels que la philosophie, l'anthropologie, la psychologie, l'intelligence artificielle. Ainsi, dans le domaine de la linguistique, l'auteure (*ibid.* : 85) distingue deux approches majeures du lexique : une approche componentielle, appelée également décompositionnelle (*ibid. :* 85-91) et une approche relationnelles (*ibid.* : 104-117). D'après Murphy, la distinction entre ces modèles de représentation du lexique se manifeste à travers deux types de ressources lexicales qui constituent deux propositions de description du lexique complétement différentes à savoir, le dictionnaire (*dictionary metaphor*) et le thésaurus (*thesaurus metaphor*): « *The dictionary defines words, usually by breaking down their meaning into smaller parts, while the thesaurus shows un relations among words*. » (*ibid. :* 85).

En effet, les modèles componentiels proposent de décrire les unités lexicales en décomposant leur sens en unités sémantiques plus petites : « *Such theories break word*

meaning down into semantic primitives *or* semantic features *and their* specification. » (*ibid.*). En présentant cette théorie sémantique, Murphy fait référence aux modèles anglo-saxons tels que l'analyse du sens de Katz et Fodor (1963), basée sur le système de traits sémantiques représenté sous forme d'arborescence propre à la tradition générative, le lexique génératif de Pustejowsky (1995) ou bien les primitifs sémantiques de Wierzbicka (1996). Rappelons que cette dernière a mis place un système de métalangue sémantique naturelle (MSN) composé d'éléments primitifs comme YOU, SOMEONE, THINK, BAD, WANT, BECAUSE, CAN, PEAOPLE qui combinés entre eux, sont destinés à fournir des descriptions sémantiques des unités lexicales. Les primitifs sémantiques se caractérisent par leur aspect primaire et universel ainsi que par l'impossibilité de les définir à l'aide d'éléments plus basiques. Ils possèdent également leur propre combinatoires, elle aussi supposée primitive. À titre d'exemple, nous proposons de reproduire la définition de *unhappy,* citée par Murphy.

Unhappy

X feels something

sometimes a person thinks something like this :

{something bad happened to me; I don't want this;

if I could, I would want to do something because of this}

because of this, this person feels something bad

X feels something like this

Figure 1. Définition de l'unité lexicale *unhappy* au moyen de primitifs sémantiques (Wierzbicka 1996 cité dans Murphy 2003 : 88)

En ce qui concerne la tradition européenne, il convient de rappeler l'approche sémique de Pottier (1992) ainsi que la théorie de sémantique interprétative de Rastier (1987), développées à la suite des travaux de la première génération des structuralistes européens tels que Bréal, Saussure, puis Hjelmslev, Greimas. Selon ces modèles, le signifié se décompose en sèmes qui d'un côté marquent l'appartenance du sémème à une classe sémantique (il s'agit des *sèmes génériques*) et de l'autres côté, distinguent le sémème de tous les autres de sa classe (*sèmes spécifiques*).

En revanche, dans une approche relationnelle, le sens d'un mot est déterminé et représenté par les relations qu'il entretient avec les autres unités du lexique : « *[…], the following approaches treat words as semantically unanalyzable ; Meaning are not « in » lexical entries or concept, but instead are among them. […] In other words, a word's meaning is dependent on all the words with which it enters into relation (and, by extension, all the words with which those words enter into relations, and so on.)* » (Murphy, 2003 : 104). En effet, selon les théories relationnelles que l'on pourrait également qualifier d'associationnistes et de holistes, le sens n'est pas inhérent de l'unité lexicale mais émerge des relations lexicales entretenues par cette dernière. Ces relations forment un réseau lexico-sémantique. Comprendre un mot équivaut à le situer dans le réseau. Ainsi, les modèles relationnels ne proposent pas de décomposition sémantique des unités lexicales mais ils s'intéressent à la description des liens qui les caractérisent.

Un des projets lexicographiques qui reflètent en quelque sorte la théorie associationniste et holiste est la base de données WordNet développée à l'Université de Princeton[5]. C'est une ressource dont l'objectif est la représentation du lexique de la langue anglaise en un système interconnecté correspondant à la façon dont les locuteurs organisent leur lexique mental. Elle s'inspire donc des théories psycholinguistiques sur la mémoire lexicale humaine. En effet, comme nous pouvons le voir à travers l'exemple de *computer* extrait de WordNet, chaque unité lexicale est associée à un synset (*synonym set*), c'est-à-dire à un ensemble de synonymes dénotant une acception donnée. Nous pouvons également constater que même si la synonymie est la relationfondamentale, WordNet propose d'organiser et de relier les différents *synsets* au moyen d'autres types de relations paradigmatiques telles que la hyponymie, la hyperonymie, la méronymie, etc. Remarquons qu'il s'agit ici des relations mixtes aussi bien lexicales que conceptuelles.

Soulignons que les deux types d'approches ne sont pas en concurrence les uns avec les autres. Elles ne s'excluent pas non plus mutuellement et dans certains cas, elles sont considérées comme complémentaires. En effet, comme le souligne Murphy (*ibid.* : 21), les différentes propositions de modélisation du lexique se situent sur un continuum dont les extrêmes correspondent, d'un côté, à des modèles strictement componentiels (matérialisé à travers des dictionnaires traditionnels qui ne fournissent aucune information relationnelle) et

[5] WordNet est accessible à l'adresse suivante : https://wordnet.princeton.edu/.

de l'autre côté, à des modèles purement relationnels de type thésaurus (qui ne contiennent pas de définitions classiques mais proposent de décrire les unités par le biais des relations que ces dernières entretiennent avec les autres unités du réseau lexical).

Ainsi, il existe des approches mixtes : « *approaches to semantic relations that fall between the dictionary models […] and the thesaurus models […]. While these theories acknowledge a need for some sense representation for individual lexical items, they also maintain that explicit representation of lexical semantic relation is necessary.*» (Murphy *ibid.* 91). Selon ces théories, le sens d'une unité lexicale doit être analysé de façon individuelle dans la mesure où il est intrinsèquement associé à cette dernière. En même temps, les approches mixtes reconnaissent la nécessité d'une représentation explicite des relations lexicales. C'est notamment le cas de l'approche contextuelle de Cruse (1986). En effet, selon l'auteur, les relations syntagmatiques et paradigmatiques contribuent au sens des unités lexicales: « *We can picture the meaning of a word as a pattern of affinities and disaffinities with all the other words in the language with which it is capable of contrasting semantic relations in grammatical context.* » (Cruse 1986 : 16)

S: (n) **computer**, computing machine, computing device, data processor, electronic computer, information processing system (a machine for performing calculations automatically)

- ***direct hyponym*** / *full hyponym*
  - S: (n) analog computer, analogue computer (a computer that represents information by variable quantities (e.g., positions or voltages))
  - S: (n) digital computer (a computer that represents information by numerical (binary) digits)
  - S: (n) home computer (a computer intended for use in the home)
  - S: (n) node, client, guest ((computer science) any computer that is hooked up to a computer network)
  - S: (n) number cruncher (a computer capable of performing a large number of mathematical operations per second)
  - S: (n) pari-mutuel machine, totalizer, totaliser, totalizator, totalisator(computer that registers bets and divides the total amount bet among those who won)
  - S: (n) predictor (a computer for controlling antiaircraft fire that computes the position of an aircraft at the instant of a shell's arrival)
  - S: (n) server, host ((computer science) a computer that

provides client stations with access to files and printers as shared resources to a computer network)
        - S: (n) Turing machine (a hypothetical computer with an infinitely long memory tape)
        - S: (n) web site, website, Internet site, site (a computer connected to the Internet that maintains a series of web pages on the World Wide Web)*"the Israeli web site was damaged by hostile hackers"*
    - *part meronym*
    - *domain category*
    - *domain term category*
    - *direct hypernym* / *inherited hypernym* / *sister term*
    - *part holonym*
    - *derivationally related form*
- S: (n) calculator, reckoner, figurer, estimator, **computer** (an expert at calculation (or at operating calculating machines))

Figure 2. Description de l'unité lexicale *computer* dans la base lexicale WordNet (consulté le 30.04.2016)

Cependant, leur mise en évidence n'est pas suffisante pour déterminer pleinement l'identité sémantique d'un mot : « *A particular lexical unit, of course, expresses its semantic identity through such relations, but its essence cannot be exhaustively characterized in terms of any determinate set of such relations.* » (*ibid.* : 49). Rappelons que d'après le modèle de Cruse, le sens de l'unité lexicale se reflète entièrement dans ses relations contextuelles. Comme le souligne Murphy (2003 : 97), selon Cruse : « *Word-senses are semantic traits of other word-senses if they partake in contextual relations with each other, and these semantic traits may differ in their status, that is, the degree of connection between the word-senses* ».

Nous constatons que notre projet, dont l'objectif est de fournir une description complète des unités terminologique, doit chercher des solutions d'ordre théorique et méthodologique dans des modèles mixtes qui prennent en compte aussi bien la nature relationnelle que compositionnelle. Ce constat nous a amenée à nous intéresser plus particulièrement à une théorie linguistique qui propose de décrire les unités lexicales aussi bien en termes de relations que de constituants sémantiques. Il s'agit notamment de la Théorie Sens- Texte de Mel'čuk (1997). En effet, Igor Mel'čuk fait partie des linguistes qui ont beaucoup contribué à la revalorisation de la composante lexicale dans les études linguistiques. L'auteur (Mel'čuk *et al.* 1995 : 16) présente une lexie comme étant une entité trilatérale, qui a :

- un sens (le signifié saussurien)
- une forme phonique/graphique (le signifiant saussurien)
- et un ensemble de traits combinatoire (le syntactique de la théorie Sens-Texte)

Selon Mel'čuk, les lexies ainsi conçues forment la partie primordiale de la langue. Pour lui, la langue est constituée de lexies et de règles servant à manipuler ces dernières

.

« *L'ensemble de ces règles, c'est-à-dire la grammaire de la langue est aux lexies ce que l'ensemble des instructions d'assemblage d'un meuble en pièces détachées est à ces pièces. Les règles en question sont donc obligatoires, mais c'est le lexique d'une langue qui prime logiquement sur sa grammaire* »

(Mel'čuk *et al.* 1995 : 17).

Ainsi, il place l'unité lexicale au cœur de son modèle théorique en développant La Lexicologie Explicative et Combinatoire (LEC), composante lexicale de la théorie Sens-Texte. Il convient de souligner que l'approche mel'čukienne, qui offre une description globale et exhaustive de l'unité lexicale, a déjà séduit un grand nombre de terminologues comme L'Homme, Heid, Dancette, Frawelley. Ils trouvent beaucoup d'avantages dans le modèle définitionnel du LEC qui permet d'éviter toute ambiguïté et de toute forme de vague dans les définitions. De plus, la LEC intègre l'outil des fonctions lexicales permettant de représenter l'ensemble des relations lexicales d'un terme avec d'autres termes d'une langue de spécialité. Nous consacrerons tout un chapitre de notre travail à la présentation des applications de la méthode mel'čukienne en terminologie. Mais, avant cela, il nous paraît important de décrire les fondements théoriques de la linguistique Sens-Texte (qui est à l'origine de la LEC).

1.2 La théorie Sens-Texte

La théorie Sens-Texte est née à Moscou, dans les années 1960 comme réponse aux insuffisances de la grammaire générative et transformationnelle. Un de ses pères fondateurs, Igor Mel'čuk (qui avec Juri Asprejan et Alexander Žolkovskij a formé l'École sémantique de Moscou) est contraint à l'exil. Il poursuit ses recherches au Canada où il publie 4 volumes *Dictionnaire explicatif et combinatoire du français contemporain (DEC) Recherches lexico-*

sémantiques I en 1984, II en 1988, III en 1992 et IV en 1999. En 1997, il présente les résultats de 30 ans de recherches lors de sa leçon inaugurale au Collège de France.

Selon Mel'cuk la langue est considérée comme un mécanisme qui permet au locuteur de faire deux choses : « parler » et « comprendre la parole » :

- *Parler, c'est-à-dire, (être capable de) faire correspondre à un sens, qu'il veut exprimer tous les textes de sa langue qui, d'après lui, peuvent véhiculer ce sens et choisir celui qui passe le mieux dans les circonstances concrètes d'un acte langagier donné*
- *Comprendre la parole, c'est – à dire, (être capable de) faire correspondre à un texte qu'il perçoit tous les sens que, d'après lui, ce texte peut véhiculer et choisir celui qui passe le mieux dans les circonstances concrètes d'un acte langagier donné.* (Mel'čuk 1997 : 1)

La tâche des linguistes consiste donc à construire, pour la langue étudiée L, un système de règles (assimilé par Mel'čuk à un programme informatique) qui définisse les mêmes correspondances entre sens et textes que celles qu'établissent les locuteurs.

1.2.1 Les modèles Sens-Texte

La méthodologie qui se trouve au cœur de cette approche de l'étude des langues consiste en la construction de MODÈLES FONCTIONNELS des langues, appelés modèles Sens-Texte ou MST. Dans la théorie Sens-Texte, le modèle fonctionnel est défini comme suit:

« X est un modèle (fonctionnel) de Y : X est un système d'expressions symboliques créé par le chercheur dans le but de représenter le fonctionnement de l'entité donnée Y qu'il étudie.

(Mel'cuk 1997 : 3).

Afin de bien comprendre le concept de modèles fonctionnels, il est nécessaire de mentionner ces deux particularités. Premièrement, un modèle fonctionnel ne garantit pas la vérité de la description obtenue, mais une simple approximation de la vérité. En second lieu, c'est un modèle fonctionnel permettant de représenter, de modéliser le comportement observable d'un

objet dont la structure interne est inaccessible à l'observation directe, ce que Mel'čuk appelle une « boîte noire ». Mel'čuk évoque 3 postulats pour compléter les fondements de sa théorie :

Le premier postulat concerne la conception générale de ce qu'est la langue qui :

« […] est un système fini de règles qui spécifie une CORRESPONDANCE multi-multivoque entre l'ensemble infini dénombrable de sens et un ensemble infini dénombrable de textes ».
(Mel'čuk 1997 : 4)

Cette correspondance peut être représentée selon la formule suivante :

{RSém$_i$} **langue**;<==>; {RPhon$_j$}

Où :

- **RSém** sont des représentations sémantiques, c'est-à-dire les « objets symboliques formels » représentant les « sens »
- **RPhon** sont des représentations phoniques, c'est-à-dire les « objets symboliques formels » représentant les « sons »

Le deuxième postulat présente les modèles Sens-Texte comme outil de description des langues :

« La CORRESPONDANCE doit être décrite par un DISPOSITIF LOGIQUE, qui constitue un modèle fonctionnel de la langue de type Sens-Texte » (Mel'čuk 1997 : 5)

Ce dispositif doit être organisé à partir du Sens vers le Texte (Sens => Texte), bien que du point de vue formel le passage Sens => Texte et le passage Texte => Sens soient équivalents. Le MST suit donc le parcours onomasiologique : il est élaboré dans le sens de la synthèse, de la production de la parole et modélise l'activité du locuteur (activité demandant uniquement des connaissances linguistiques).

Le troisième postulat introduit, dans la structure du MST, deux niveaux intermédiaires relatifs à la phrase et au mot qui sont considérés comme unités de base de la description linguistique.

« Dans la description de la CORRESPONDANCE[6]*, deux NIVEAUX INTERMÉDIAIRES de représentation des énoncés sont nécessaires pour mettre en lumière les faits linguistiques pertinents : la représentation SYNTAXIQUE [= RSynt], qui correspond aux régularités spécifiques à la PHRASE, et la représentation MORPHOLOGIQUE [= RMorph], qui correspond aux régularités spécifiques au MOT ».* (Mel'čuk 1997 : 6)

C'est dans le cadre de la phrase qu'on peut observer et étudier l'ordre des mots, l'accord et le régime, la structuration communicative, la cooccurrence lexicale restreinte et d'autres phénomènes tandis que le mot permet d'analyser la flexion, la dérivation ainsi que les alternances phonémiques.

En résumant, le MST présente les particularités suivantes :

- c'est un modèle discret et formel, basé sur l'approche onomasiologique,

- à la différence des modèles génératifs et transformationnels, le MST ne génère rien mais il fait correspondre à chaque représentation sémantique (RSém) toutes les représentations phonétiques (RPhon) qui peuvent l'exprimer dans une langue donnée et plus précisément, un MST met en rapport des représentations linguistiques des niveaux adjacents. Il prend une représentation du niveau n et il lui associe toutes les représentations correspondantes du niveau n+1. Les représentations ne subissent pas de modifications. Le MST doit se comporter comme un locuteur, c'est-à-dire, « traduire » un sens donné en un texte qui l'exprime (Mel'čuk 1997 : 6-7). C'est pourquoi ce modèle est qualifié d' « équatif » et de « traductif »
- le MST est fondé sur la paraphrase car à un sens donné correspondent plusieurs texte. C'est le locuteur qui choisit le ou les textes les mieux adaptés à une situation ou à un contexte donné.

- c'est un modèle global et intégral car il intègre toutes les composantes de la langue (lexique ainsi que toutes les parties de la grammaire) qui sont destinées à « collaborer » étroitement au cours du processus de synthèse des textes (Mel'čuk 1997 : 7). La langue est décrite comme un tout indivisible.

[6] C'est moi qui ai mis le terme en majuscules.

- le MST se définit par son caractère stratificationnel. En effet, il est composé de 7 niveaux de représentations linguistiques des énoncés. Tous les niveaux, sauf le niveau sémantique, sont subdivisés en un sous-niveau profond [= -P] et un sous-niveau de surface [= -S]

« Le sous-niveau profond est orienté vers le sens : sa tâche est d'exprimer explicitement toutes les distinctions sémantiques pertinentes à son niveau. Le sous-niveau de surface est orienté vers le texte : sa tâche est d'exprimer explicitement toutes les distinctions formelles pertinentes à son niveau. » (Mel'čuk 1997 : 7)

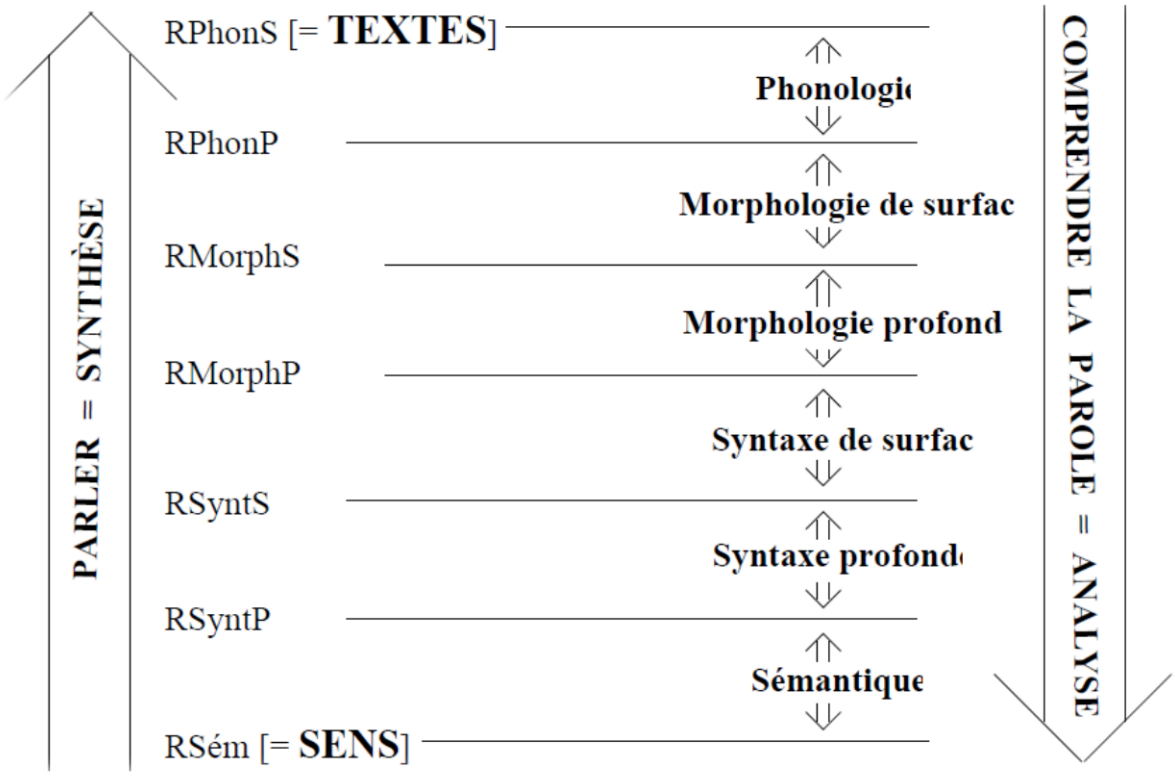


Figure 3. Le MST avec ses composantes et tous les niveaux de représentation (Mel'čuk 1997 : 8)

Le schéma ci-dessus présente la structure détaillée d'un MST avec, à gauche, l'ensemble de 7 représentations linguistiques caractérisées par leur dichotomie « profond *vs* de surface » (présentées à partir du niveau le plus bas ou le plus profond – niveau sémantique, en remontant au niveau le plus haut ou le plus superficiel – niveau phonétique) et à droite, six composantes linguistiques (modules) correspondant aux disciplines de la linguistique moderne (la sémantique, la syntaxe, la morphologie et la phonologie). Une composante du MST (ou un module) est conçue comme :

« (…) un ensemble de règles qui assurent la correspondance entre les représentations de deux niveaux adjacents » (Mel'čuk 1997 : 7)

La flèche de gauche indique le parcours onomasiologique de la synthèse et celle de droite le parcours sémasiologique de l'analyse.

1.2.2 Structures

Chaque représentation est composée d'un certain nombre d'objets formels appelés « structures », notamment d'une structure de base qui constitue un nœud de la représentation ainsi que de structures complémentaires qui superposent ou précisent la dernière. Pour illustrer cette composition, prenons comme exemple, d'après Mel'čuk (Mel'čuk 1997 : 9) (Mel'čuk 1988 : 19) les RSém et RSynP. La structure de base de la représentation sémantique correspond à la structure sémantique qui reflète le sens propositionnel ou objectif d'un énoncé donné. A celle-ci s'ajoutent deux autres structures : la structure sémantico-communicative, qui reflète le sens communicatif (ou subjectif) de l'énoncé et la structure rhétorique, qui reflète « le projet esthétique » du locuteur (l'ironie, le pathétique, les registres de langue, etc.). En ce qui concerne la représentation syntaxique profonde, elle est constituée d'une structure syntaxique profonde (son noyau), d'une structure syntaxique communicative profonde, d'une structure syntaxique anaphorique profonde et d'une structure syntaxique prosodique profonde. Il serait compliqué de présenter en détail toutes structures de toutes les représentations linguistiques. Nous nous limiterons (d'après l'auteur (Mel'čuk 1997 : 10) à décrire uniquement les structures de base.

La structure sémantique (SSém) est un graphe connexe orienté. Les nœuds sont étiquetés par des noms d'unités sémantiques de la langue (sens désambiguïsés des lexies). Ces unités sémantiques sont subdivisées en prédicats et arguments. Les arcs du graphe indiquent des relations prédicat-argument et sont étiquetés par des numéros distinctifs correspondant aux arguments de chaque prédicat. La SSém est écrite dans un langage sémantique, la syntaxe de ce langage étant formalisée par des réseaux où chaque sous-réseau représente une prédication du type *P(x,y)*.

La structure syntaxique profonde SSyntP est un arbre de dépendance dont les nœuds sont étiquetés par des lexies profondes (qui du point de vue de leur nature lexicographique peuvent appartenir à un des trois types suivants : lexème, locution ou fonction lexicale, que nous définirons par la suite) (Milićević 2007 : 40). On constate, à ce niveau, l'absence de pronominalisations, de marqueurs flexionnels, de mots-outils comme les prépositions ou les conjonctions régies, d'auxiliaires. Il est nécessaire de souligner que l'ordonnancement graphique des nœuds n'a aucune pertinence.

La SSynP met en jeu des relations syntaxiques profondes (RSynP), telles que six relations actancielles (I, II, III, …, VI), une relation attributive, une relation appenditive et une relation coordinative représentées par des arcs. Ces relations sont appelées universelles car elles sont suffisantes pour décrire l'organisation syntaxique globale de n'importe quelle phrase dans n'importe quelle langue[7] (Milićević 2007 : 48). Les relations actancielles subordonnent à une lexie donnée un élément de la phrase qui est un Actant Syntaxique Profond (ASynP). La RSyn I correspond aux constructions subjectales, La RSynP II correspond aux constructions avec l'objet principal (grosso modo avec le complément d'objet direct), la RSynP III correspond au complément d'objet indirect et les RSynP de type IV-VI correspondent aux autres compléments de la lexie donnée. La RSynP ATTR (attributive) représente tous les cas de modification, la RSynP APPEND (appenditive) renvoie à toute sorte de constructions plus autonomes syntaxiquement comme les insertions parenthétiques, formes d'adresse, adverbes de phrase. Les trois premiers types de relations sont des relations de subordination. Le dernier type de la RSynP est une RSynP COORD (coordinative) qui représente toute relation de coordination. Mel'čuk fait ici référence à la grammaire de dépendance de Lucien Tesnière (1959 – [1988]).

La structure syntaxique de surface est également un arbre de dépendance non ordonné. Cependant, les nœuds de cet arbre sont étiquetés de tous les lexèmes de la phrase, y compris les mots-outils, éléments pronominaux, etc. En contrepartie, les caractéristiques flexionnelles des lexèmes associés aux nœuds restent incomplètes (ne sont représentées que celles qui portent une charge sémantique, les autres, par exemple celles qui sont liées aux règles d'accord ou de régime apparaissent au niveau suivant) (Mel'čuk 1997 : 13) Les arcs de l'arbre

[7] Claude Hagège dans *La structure des langues* (1982), reconnaît 3 relations syntaxiques universelles : la prédication, la dénomination et la coordination.

SynS ne renvoient plus aux relations universelles. Elles représentent maintenant les constructions syntaxiques particulières d'une langue étudiée.

La structure morphologique profonde est une chaîne (ou un ensemble linéairement ordonné) des représentations morphologiques profondes de tous les mots-formes, c'est-à-dire de tous les lexèmes munis de toutes les valeurs flexionnelles pertinentes. Dans le cas de la langue française (à morphologie assez pauvre), il est difficile d'apercevoir la différence entre la représentation de la structure morphologique profonde et celle de surface (cette dernière constitue également une chaîne de mots-formes de la phrase représentés cette fois-ci comme l'ensemble de morphèmes). À partir de la représentation RMorphS dont la SMorph est la structure de base, la composante morphologique de surface du MST construit toutes les représentations phonémiques RPhonP possibles.

Ainsi, un MST prend, à l'entrée, une Représentation Sémantique (RSém) et produit, à la sortie, les phrases correspondantes en transcription phonétique, en passant par des représentations intermédiaires. C'est une fonction, au sens mathématique du terme, qui fait correspondre à un sens (= son argument) l'ensemble de tous les textes synonymes qui l'expriment (= sa valeur) (Mel'čuk 1997 : 15). Pour obtenir une phrase (Texte), la composante sémantique du MST (ensemble de règles) construit d'abord, en se basant sur la Représentation Sémantique (RSém) dont la Structure Sémantique (SSém) fait partie, une Représentation Syntaxique Profonde (RSynP). Ensuite, la composante syntaxique profonde prend la représentation syntaxique profonde (à laquelle la structure syntaxique profonde SSynP appartient en constituant son noyau), et en dérive toutes les Représentations Syntaxiques de Surface (RSynS) possibles et ainsi de suite. Ainsi, la SSém de départ peut être exprimée par un nombre élevé de phrases. Mel'čuk (1988) démontre que pour un sens suffisamment complexe, on peut construire des (centaines de) milliers de paraphrases plus ou moins synonymes. Il donne comme exemple une phrase tirée d'un magazine français qui a entre deux et trois millions de paraphrases:

« Le style de persécutions policières des gens de lettres en Union Soviétique a évidement connu, depuis un demi-siècle, des changements sérieux ». (Mel'čuk 1988 : 15)

Pour évaluer le nombre total de ces paraphrases, il a divisé la phrase en question ainsi que trois autres phrases basées sur la même représentation sémantique en « tranches »

sémantiques et il a indiqué des variantes d'expression synonymes ou quasi-synonymes pour chaque tranche (Mel'čuk 1988 : 25). En combinant les variantes de chaque tranche, il a obtenu 4 374 000 paraphrases (dont une partie a été éliminée en raison de contraintes de cooccurrence). Cela prouve la richesse et souplesse synonymique des langues naturelles.

« (…) le sens linguistique est l'invariant des paraphrases (synonymes) ; la représentation sémantique est alors un moyen formel de décrire cet invariant – dans le but d'en dériver l'ensemble des paraphrases possibles » (Mel'čuk 1988 : 24)

1.2.3. Composantes - ensembles de règles sémantiques

Comme nous l'avons vu plus haut, ce sont les composantes du MST qui assurent la correspondance entre deux niveaux intermédiaires. Nous avons défini (d'après Mel'čuk), la composante comme un ensemble de règles permettant la construction de la phrase à partir de sa représentation sémantique. Pour illustrer le passage entre les représentations de la phrase et permettre en même temps une meilleure compréhension et lisibilité du MS, il est nécessaire de présenter quelques-unes des règles en question. Nous avons décidé de nous focaliser plutôt sur les règles faisant partie des composantes sémantiques et syntaxiques qui sont plus proche de notre projet.

Ainsi, la composante sémantique établit la correspondance $\{RSém_i\}$ <==>; $\{RSynP_k\}$ au moyen de huit opérations qui sont effectuées par des règles sémantiques (Mel'čuk 1988 : 15) parmi lesquelles on peut citer les règles suivantes :

- R^1 - Règle sémantique lexémique (ou lexico-sémantique) permet d'obtenir, à partir de la SSém, un fragment de l'arbre de la SSynP (nœud avec ses branchements correspondants) pour le lexème en question.

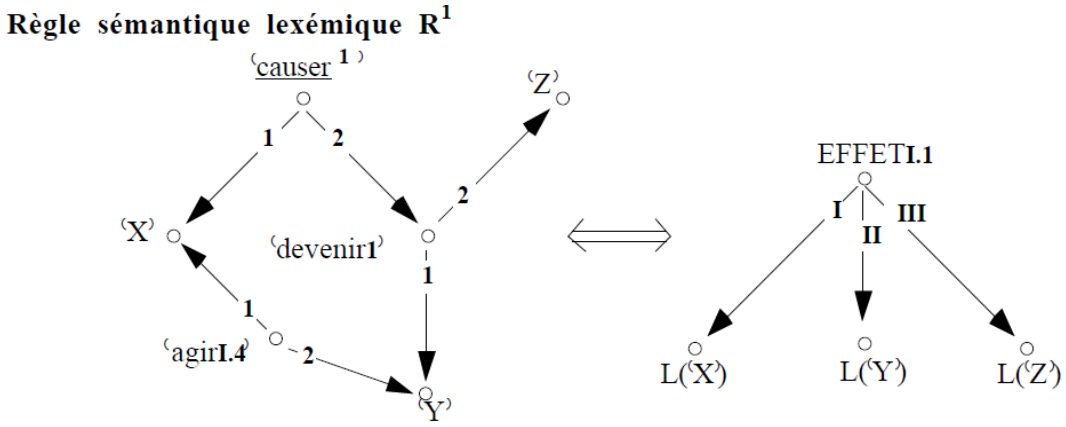


Figure 4. Exemple de règle sémantique lexémique R1 (Mel'čuk *et al.*1999 : 15)

Dans l'exemple présenté ci-dessus (Mel'čuk *et al.* 1999 : 15), la R^1 stipule que le sens ‹W, agissant sur Y, cause que Y devient Z› peut être exprimé par le lexème EFFET : *l'effet Z de X sur Y*. On constate que cette règle constitue la partie centrale de l'article de dictionnaire du lexème en question (on fait bien évidement référence au *Dictionnaire explicatif et combinatoire* (Mel'čuk *et al.* 1984, 1988, 1992, 1999)

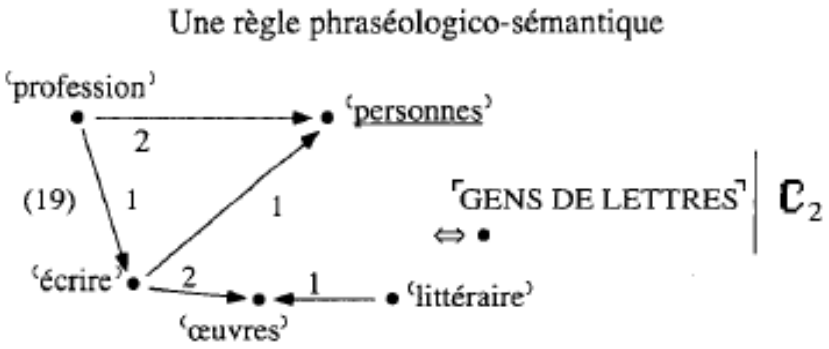


Figure 5. Exemple de règle phraséologico-sémantique R2 (Mel'čuk 1992 : 25)

R^2 – Règle phraséologico-sémantique permet d'obtenir, à partir de la SSém, un fragment de l'arbre de la SSynP (nœud avec ses branchements correspondants) pour, cette fois-ci, le phrasème en question, et elle constitue également un article de dictionnaire – voir l'exemple ci-dessus.

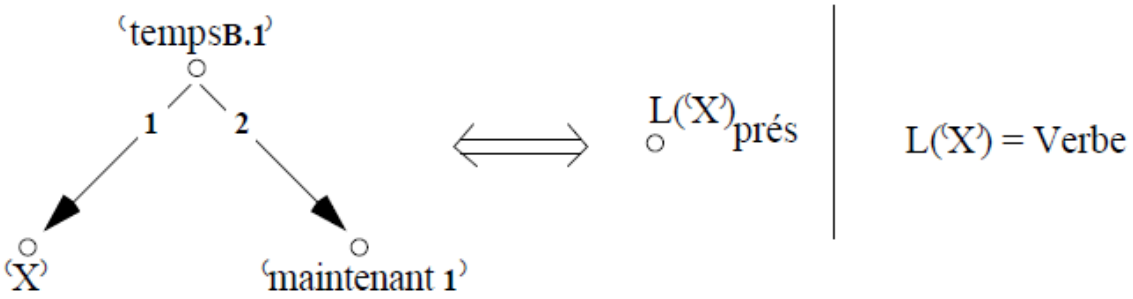


Figure 6. Exemple de règle sémantique flexionnelle R3 (Mel’čuk 1997 : 16)

La figure 6 présente quant à elle une règle sémantique flexionnelle (cet exemple a été présenté par Mel’čuk lors de la Leçon inaugurale au Collège de France). En effet, les règles de ce type décrivent le sémantisme des grammèmes de la langue (= « sémantique morphologique »). Leur nombre, selon Mel’čuk (*ibid.*) (et cela en fonction de la richesse morphologique d’une langue donnée) est estimé à quelques centaines.

En ce qui concerne la composante syntaxique profonde (voir la Figure 1.7), elle établit la correspondance {$RSynP_{k1}$} <==>; {$RSynS_{k2}$} au moyen de cinq opérations qui sont effectuées par les règles syntaxiques profondes.

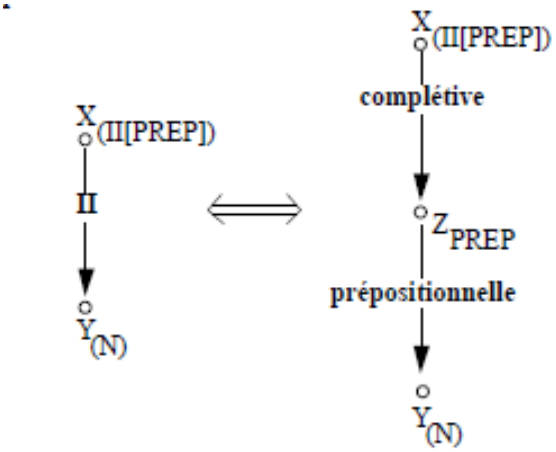


Figure 7. Exemple de règle syntaxique profonde (Mel’čuk 1997 : 17)

Comme nous pouvons le constater sur le schéma ci-dessus (Mel’čuk 1997 : 17), il s’agit d’une règle qui permet d’introduire dans la SSyntS les indications relatives au régime, c’est-à-dire l’ensemble de données sur la combinatoire de la lexie associée au nœud (préposition ou conjonction ou cas demandé par telle ou telle lexie). Ces indications apparaissent dans ce que Mel’čuk appelle le schéma de régime et qui fait partie de la zone de

combinatoire syntaxique du *Dictionnaire explicatif et combinatoire* (il s'agit d'un tableau qui spécifie la forme de l'expression des actants, éléments essentiels qui dépendent de la lexie en question et qui sont prévus par sa définition (Mel'čuk 1997 : 17). D'autres règles permettent de transposer des configurations syntaxiques profondes (qui sont universelles) en constructions syntaxiques de surface (typiques d'une langue donnée). En plus, une partie des règles syntaxiques profondes impliquent des fonctions lexicales en calculant leurs valeurs.

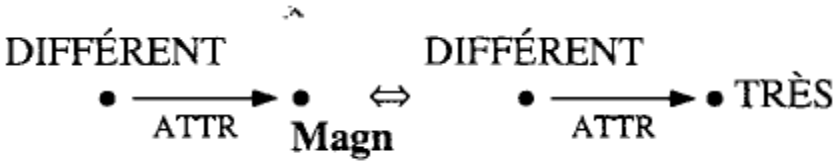


Figure 8. Exemple de règle de syntaxique de surface (Mel'čuk 1992 : 25)

Quant aux règles syntaxiques de surface, elles se divisent en règles locales appliquées au niveau d'une seule branche de l'arbre syntaxique de surface et ainsi qu'en règles globales qui prennent en charge des configurations de nœuds beaucoup plus complexes (Mel'čuk 1997 : 19). Parmi les règles locales, on peut énumérer celle qui établit l'ordre des mots (positionne les prépositions) ou bien celle qui assure l'accord des déterminants en genre et en nombre et définit leur positionnement linéaire par rapport au nom déterminé. Les règles globales rendent également compte des propriétés de l'ordre des mots, mais cette fois-ci au niveau de toute la phrase.

Soulignons que la théorie Sens-Texte conçoit le sens d'une unité lexicale comme un *sens situationnel* (appelé également *propositionnel* ou *déntationnel*). En effet, la définition lexicographique d'une lexie L dans l'approche mel'čukienne correspond à une représentation sémantique [= RSém] qui s'écrit sous la forme d'un réseau dont les nœuds représentent des éléments sémantiques et les arcs identifient les relations « prédicat - arguments » (voir les Figure 5 et Figure 6). Cette structure se reflète ensuite sous forme linéaire, dans une expression de la forme

$$A = (B)$$

où *A* – le *défini* – est la lexie vedette L incluse dans une forme propositionnelle (expression à variable), et (B) – le *définissant* – est une description du sens de L faite dans un métalangage lexicographique (Mel'čuk *et al.* 1995 : 73). La description sémantique d'une unité lexicale consiste donc en une décomposition du sens en éléments sémantiques qui constituent un

réseau sémantique. Il convient également de souligner que selon l'approche Sens-Texte, la plupart des propriétés de comportement d'une lexie sont sous-tendues ou même carrément déterminées par son sens dénotationnel (*ibid.*). La définition lexicographique, telle qu'elle a été définie dans le cadre de la LEC, permet donc d'identifier les relations que la lexie (L) entretient avec d'autres unités lexicales, aussi bien sur l'axe syntagmatique que paradigmatique. En effet, il faut savoir qu'il est impossible de traiter les phénomènes tels que la dérivation sémantique ou la combinatoire restreinte sans avoir formulé le sens des lexies concernées par ces relations. Cela nous amène à aborder le concept clé de la théorie Sens-Texte (et plutôt de sa composante lexicale, la Lexicologie Explicatif et Combinatoire), à savoir la fonction lexicale. Cependant, avant de passer à la présentation de cet outil original d'encodage des liens lexico-sémantiques, nous proposons tout d'abord de décrire plus en détails les deux dimensions des relations lexicales, à savoir la dimension syntagmatique et la dimension paradigmatique.

1.3 Le lexique comme réseau d'unités lexicales connectées les unes aux autres

« Ainsi, dans un état de langue, tout repose sur des rapports ; comment fonctionnent-ils ? »

(Saussure 1972 : 170)

Ferdinand de Saussure révolutionne la linguistique en considérant que la langue est un véritable système, un ensemble cohérent et autonome de dépendances internes. Ainsi, le lexique n'est pas une simple liste de lexies. Comme le souligne Polguère (2008 : 106), chaque lexie prend sa valeur sémantique en langue du fait des liens d'opposition, de similarité, de compatibilité, d'incompatibilité qui les unissent aux autres lexies. Il existe deux types majeurs de liens entre lexies que nous présentons ci-dessus, à savoir l'axe paradigmatique et l'axe syntagmatique.

Comme le rappelle Mel'čuk (1997 : 21), depuis Saussure, l'activité langagière se déroule selon deux axes: l'axe paradigmatique (axe de la sélection ou des choix lexicaux) où le locuteur fait des choix à partir des unités qui s'opposent et s'excluent ; et l'axe syntagmatique (axe de la combinaison ou de la cooccurrence lexicale) où le locuteur combine

des unités déjà sélectionnées. Les relations paradigmatiques sont représentées par une ligne verticale alors que les relations syntagmatiques sont représentées par la ligne horizontale. C'est Saussure qui pose à la base du fonctionnement de la langue deux "formes d'activité mentale" correspondant à "deux ordres de coordination" nettement distincts:

« Les rapports et les différences entre termes linguistiques se déroulent dans deux sphères distinctes dont chacune est génératrice d'un certain ordre de valeurs. (…) Ils correspondent à deux formes de notre activité mentale, toutes deux indispensables à la vie de la langue. »

(Saussure 1972 : 170)

D'une part, il s'agit donc des rapports syntagmatiques, rapports fondés sur le caractère linéaire de la langue et liés aux combinaisons de mots. Saussure décrit un syntagme comme un ensemble d'unités – y compris toute sorte d'unités complexes - consécutives, éléments qui se rangent les uns à la suite des autres sur la chaîne de la parole et n'acquièrent sa valeur que parce qu'ils sont opposés à ce qui précède ou ce qui suit ou à tous les deux. D'autre part, on évoque des rapports associatifs (association de mots qui se forment dans la mémoire) illustrés par des séries associatives basées sur des rapports divers : flexionnelles - *enseignement, enseigner, renseigner* ; sémantique - *apprentissage, éducation, instruction, enseignement.* Les rapports syntagmatiques ont pour support l'étendue et fonctionnent *in praesentia* (*« le rapport syntagmatique repose sur deux ou plusieurs termes également présents dans une série effective »* (*ibid.* : 171). Au contraire, les rapports associatifs (qualifiés de nos jours de *paradigmatiques*) unissent des termes *in absentia* dans *une série mnémonique virtuelle* (*ibid.* : 171).

Cette idée a été reprise et développée par Hjelmslev. Les deux axes fondamentaux du langage sont alors dénommés *système* et *procès*. L'axe du procès est représenté par une ligne horizontale orientée vers la droite, tandis que celui du système est marqué par une ligne verticale coupant la première. Vandendorpe (1990 : 171) qui retrace dans son article l'histoire de la dichotomie paradigme *vs* syntagme, rappelle que selon Hjelmslev seul le procès est directement observable, ce qui lui donne un statut supérieur à celui du système et fonde le texte comme objet privilégié de l'analyse linguistique.

« En regardant un texte imprimé ou écrit, nous voyons qu'il se compose de signes, et que ceux-ci se composent à leur tour d'éléments qui se déroulent dans une direction déterminée

[…]. Les signes forment une chaîne et les éléments de chaque signe forment également une chaîne. » (Hjelmslev 1966 : 56)

Hjelmslev souligne que les signes et les éléments sont reliés entre eux dans la chaîne (syntagme). Le rapport (ou la dépendance) qui existe entre eux à l'intérieur d'une même chaîne est nommé *relation*. En ce qui concerne le système de la langue, le paradigme (« un paradigme d'éléments ») est défini comme "une classe d'éléments qui peuvent être placés à une même place d'une chaîne (Hjelmslev 1966 : 56-57). La fonction existant entre les membres d'un paradigme est appelée *corrélation*. Remarquons qu'il est possible de faire un parallèle avec l'approche systémique de Halliday que nous décrivons plus en détail dans le chapitre 3 de ce travail (voir la section 3.4.1.3). À cette étape de notre étude, soulignons seulement que l'idée de *cline of instanciation* (ou *continuum d'instanciation*) de Halliday qui décrit la relation entre *système* et *texte* (constituant deux pôles du schéma, celui du potentiel et celui d'une occurrence particulière), renouent en quelque sorte avec les notions de *système* et de *procès* de Hjelmslev.

Quant à Jakobson, on peut citer un fragment de l'article « Deux aspects du langage et deux types d'aphasie », où le linguiste se penche sur la pathologie du langage, plus précisément sur sa déperdition : l'aphasie et rend compte de deux types de dysfonctionnement, liés à deux axes mobilisés dans le langage : le paradigme et le syntagme:

« Parler implique la sélection de certaines entités linguistiques et leur combinaison en unités linguistiques d'un plus haut degré de complexité. Cela apparaît tout de suite au niveau lexical : le locuteur choisit les mots (axe paradigmatique) et les combine en phrases (axe syntagmatique) conformément au système syntaxique de la langue qu'il utilise; les phrases à leur tour sont combinées en des énoncés. Mais le locuteur n'est d'aucune manière un agent complètement libre dans le choix des mots : la sélection doit se faire à partir du trésor lexical que lui-même et le destinataire du message possèdent en commun. (...) On peut donc dire que la concurrence d'entités simultanées et la concaténation d'entités successives sont les deux modes selon lesquels nous, sujets parlants, combinons les constituants linguistiques » (Jakobson 1981 : 45-46)

En effet, Jakobson place à la base du langage l'opposition entre métaphore et métonymie. La métaphore est utilisée, comme figure de la similarité, pour rendre compte des phénomènes de

sélection d'entités simultanées et concurrentes, tandis que la métonymie, figure de la contiguïté, correspond aux opérations d'enchaînement et de hiérarchisation :

« *Le développement d'un discours peut se faire le long de deux lignes sémantiques différentes : un thème (*topic*) en amène un autre soit par similarité soit par contiguïté. Le mieux serait sans doute de parler de procès métaphorique dans le premier cas et de procès métonymique dans le second (…).* » (Jakobson 1981 : 61)

Il est nécessaire de souligner que le caractère complexe et irrégulier de la structure sémantique du lexique (aussi bien au niveau des relations lexicales paradigmatique que de la cooccurrence lexicale) constituait, depuis fort longtemps, un élément problématique pour toute description linguistique (surtout dans le cadre des travaux lexicologiques et lexicographiques). L'originalité de la théorie Sens – Texte est de proposer un modèle fonctionnel permettant d'étudier et de modéliser les liens sémantiques qui existent entre les lexies d'une langue. Il s'agit des fonctions lexicales. C'est un outil très efficace qui permet de décrire de façon systématique et régulière d'une part les choix lexicaux sur l'axe paradigmatique et d'autre part la cooccurrence lexicale sur l'axe syntagmatique. Cependant, avant de montrer leur fonctionnement et leurs avantages pratiques, nous souhaiterions décrire plus en détail les phénomènes sémantiques tels que la dérivation sémantique et la combinatoire lexicale.

1.3.1 Liens paradigmatiques – phénomène de dérivation sémantique

Comme nous l'avons évoqué plus haut, le locuteur est amené à faire des choix parmi les lexies qui s'opposent ou s'excluent. Les liens qui existent alors entre les lexies sont qualifiés de paradigmatiques. En effet, il existe un phénomène qui est étroitement lié à la sélection de lexies sur l'axe paradigmatique. Il s'agit de la *dérivation sémantique*. D'après Mel'čuk et Polguère (2007 : 22), une dérivation sémantique est une relation particulière entre deux lexies : une lexie de départ et une lexie « sémantiquement construite » à partir de celle-ci. Pour eux, il s'agit d'un lien lexical orienté : de la base de la dérivation (lexie d'origine) vers le dérivé. Ce type de lien lexical est appelé : lien paradigmatique. Nous proposons de reprendre la définition formelle de la *dérivation sémantique* de Mel'čuk et Polguère (2006 et

2007) présentée dans le cadre de leur projet dictionnairique commun *Lexique actif du français*, présenté à la fin de ce chapitre (mentionnons seulement qu'il s'agit d'un dictionnaire spécialisé qui se focalise sur une description en profondeur de deux phénomènes : dérivations sémantiques et collocations).

Ainsi, une *dérivation sémantique* est une relation entre deux lexies fondée sur une parenté de sens. Une lexie L_2 est un dérivé sémantique d'une lexie L_1 si les trois conditions suivantes sont remplies (Mel'čuk et Polguère 2007 : 18) :

1. *L_2 entretient une relation sémantique avec L_1, c'est-à-dire que L_1 et L_2 possèdent des composantes de sens communes. Dans le cas plus typique, L_2 se définit en termes de L_1.* Par exemple, la lexie *marteau* [= L_2] est définie en termes de *frapper* [= L_1], car le sens ⁽marteau⁾ = ⁽artefact servant à frapper...⁾

2. *La relation sémantique entre L_2 et L_1 est récurrente dans la langue.*

Les relations de type ⁽artefact servant à⁾ comme *couper -> hache* ; *ouvrir* [une porte] -> *clé* ; *fumer -> pipe* sont récurrentes en français.

3. *La relation entre L_1 et L_2 s'exprime éventuellement (mais pas nécessairement), de façon morphologique dans la langue.*

Par exemple, pour la relation ⁽artefact servant à⁾, on trouve les dérivations suivantes : *bouch(-er) -> bouch (-on)* ; *balay(-er) -> balai*

Comme nous le voyons, le dernier point (qui ne constitue pas, à proprement parler, une condition mais plutôt une caractéristique supplémentaire) renvoie à la notion traditionnelle de dérivation, notamment à la *dérivation morphologique* où le lien sémantique est marqué explicitement par un moyen morphologique. La dérivation au sens traditionnel peut être donc considérée comme un cas particulier de *dérivation sémantique* (qui pour sa part renvoie à une notion plus large).

Selon (Mel'čuk et Polguère 2007 : 19), on peut identifier trois grandes familles de dérivations sémantiques. Analysons les trois cas de figure ci-dessous :

1. Les deux lexies possèdent (approximativement) le même sens. Il s'agit ici d'une dérivation sémantique (quasi-)vide, qui correspond aux cas suivants :
- synonymie exacte ou approximative
- conversion exacte ou approximative
- changement de partie du discours (phénomène de *translation* défini par Tesnière)

Par conversion, Mel'čuk et Polguère (*ibid.*) incluent également dans cette première famille les termes génériques (*cours d'eau* pour *rivière, légume* pour *haricot*...).

2. Les deux lexies possèdent des sens opposés. Il s'agit ici de l'antonymie exacte ou approximative.

3. Une des deux lexies désigne un élément de la situation désignée par l'autre. Il peut s'agir d'un participant actant, d'un circonstant ou d'une caractéristique d'un participant ou circonstant En effet, contrairement à la première famille, les dérivés sémantiques regroupés ici partagent une composante de sens non triviale mais ont d'autres composantes sémantiques déférentes qui les distinguent.

Nous souhaiterions revenir rap*ibid.*ent au premier cas de figure et précisément au phénomène du changement de partie du discours qui renvoie à la notion de *dérivation syntaxique.* En effet, la dérivation syntaxique est un phénomène bien connu qui joue un rôle important en permettant le paraphrasage et la reformulation. Charles Bally était un des premiers linguistes français à étudier ce phénomène en l'appelant *transposition*. Tesnière (1988 : 359) l'a qualifié de *translation* et lui a consacré la troisième partie de son ouvrage *Éléments de syntaxe structurale.*

« La translation a pour effet, sinon pour but, de résoudre la difficulté qui surgit pour le sujet parlant lorsqu'il s'est engagé dans une phrase de structure donnée et qu'il se voit obligé, en cours d'élocution, d'employer à l'improviste un mot relevant d'une catégorie qui n'est pas directement connectable avec un des mots de la fraction de phrase déjà énoncée. […]C'est grâce à elle que le sujet parlant ne reste jamais bouche bée, sans pouvoir achever sa phrase. » (Tesnière 1988 : 365).

En effet, c'est une opération qui consiste en un changement de nature syntaxique d'un mot. En subissant une *translation*, un mot assume une fonction qui n'est pas prévue par sa

nature (le changement de fonction constitue une deuxième opération qui résulte de la *translation*). On constate, d'après Tesnière (1988 : 363), que dans certaines expressions comme *le train de Paris, la gare de Sceaux, le livre de Pierre*, les groupes *de Paris, de Sceaux, de Pierre*, étant subordonnés aux substantifs régissants, ont valeur d'adjectif sans être à proprement parler des adjectifs.

«Dans son essence, la translation consiste donc à transférer un mot plein d'une catégorie grammaticale dans une autre catégorie grammaticale c'est-à dire, à transformer une espèce de mot en une autre espèce de mot » (Tesnière, 1988 : 364).

Tesnière introduit une terminologie précise afin de désigner les facteurs essentiels de cette opération,

- le *transférende* est le terme de base, tel qu'il se présente avant d'avoir subi l'opération de la translation
- le *transféré* est le mot qui en résulte
- le *translatif* est un outil grammatical (marquant morphologique) permettant la translation. Il est nécessaire de remarquer qu'il y a des cas de translation sans translatif (où le marquant de la translation est zéro), comme, par exemple *un ruban orange*, *une étoffe citron*

Ainsi, dans la phrase *On n'a pas été payés pendant un an*, le mot *an* (transférende) est un substantif qui est transféré en adverbe grâce au translatif *pendant*. De ce fait, il peut assumer la fonction de circonstant. Ou bien dans les expressions suivantes : *mariage de raison* ou *voyage d'affaires*, *raison* et *affaires* relevant morphologiquement de la catégorie des substantifs, en conservent les caractéristiques, mais deviennent syntaxiquement des adjectifs épithètes.

1.3.2 Liens syntagmatiques - phénomènes phraséologiques en langue générale

Avant de passer à la description des FL syntagmatiques, nous souhaiterions faire un petit tour d'horizon des phénomènes phraséologiques en langue générale. Comme il s'agit d'un sujet complexe, il nous paraît essentiel de consacrer un passage plus long à cette problématique. En effet, depuis les années 80 du siècle, les combinaisons lexicales font l'objet

de multiples études, aussi bien dans le domaine de la lexicologie (Haussmann 1979, Mel'čuk *et al.* 1995, Sinclair 1991, Gross 1996, Gross 1981, 1998, Mathieu-Colas et Le Pesant 1998, Tutin et Grossmann 2002) que de la terminologie (Heid et Freibott 1991, Kocourek 1991a, 1991b, Sager 1991, l'Homme et Meynard 1998, Meyer, Cohen 1986, Maniez 2003). Cependant, pour bien saisir le concept des FL syntagmatiques et ne pas allonger la description, nous avons décidé de nous focaliser, dans cette partie de notre travail, sur la langue générale. Nous reviendrons à la problématique des phénomènes combinatoires dans le chapitre suivant de la thèse, en étudiant le traitement des collocations en langue de spécialité.

Dans son *Traité de syntaxique française* datant de 1909, Charles Bally remarque que certains mots présentent des affinités et tendent à apparaitre ensemble en formant des *séries* ou *des groupements.* Ces groupements peuvent être passagers (« *ils se désagrègent aussitôt après sa formation* »), « *mais à force d'être répétés, ils arrivent à recevoir un caractère usuel et à former même des unités indissolubles. (…). Entre les cas extrêmes (groupements passagers et unités indécomposables), il y a place pour une foule de cas intermédiaires, difficile à classer et à préciser* ». (Bally 1909, [1951] : 66-67)

Comme nous pouvons le constater (d'après Tutin et Grossmann 2002), Bally fait déjà la distinction entre les degrés de figement des syntagmes. Il présente aussi sa typologie des expressions de ce type :

« Les groupes consacrés par l'usage s'appellent locutions phraséologiques *; nous nommerons* série *celles où la cohésion des termes n'est pas relative, et* unités *celles où elle est absolue ».* (Bally *ibid.* : 68)

Ainsi, d'une part, nous avons des expressions complètement figées (opaques et mémorisées) comme *fruits de mer, cordon bleu* ou *la moutarde lui monte au nez*, d'autre part, des associations absolument libres, occasionnelles comme *organisation de séminaires* ou *politicien convaincant*. Finalement, entre les deux, se présente un cas intermédiaire : celui des *collocations*. Pour de nombreux linguistes qui essaient de circonscrire le phénomène, les frontières entre les trois types de combinaisons ne sont pas nettes et leur découpage présente des difficultés sérieuses. De plus, les chercheurs ont recours à diverses dénominations pour désigner ces expressions (ce qui entraîne des divergences définitoires). Ainsi, pour les unités phraséologiques complétement figées (si on utilise la terminologie traditionnelle), on recense

des appellations comme *expressions toutes faites*, *expressions figées* ou *locution* (Rey et Chantereau 1993), *idiotismes*, *unités phraséologique* (Bally *ibid.*), *phrasèmes complets* ou encore *locutions* (Mel'čuk 2003, 2011). Les collocations, à leur tour, bénéficient des dénominations suivantes : *séries phraséologiques* (Bally *ibid.*), *semi-phrasèmes* ou *collocations* (Mel'čuk *ibid.*), encore *collocations* (Haussman 1979) et en langues de spécialité : *cooccurrences lexicales* (Cohen 1986), *phraséologismes* (Pavel et Nolet 2001), *combinaisons lexicales spécialisées* (L'Homme et Meynard 1998).

D'un point de vue quantitatif, les unités phraséologiques abondent dans la langue. Elles sont omniprésentes dans les textes de tous genres, tant en langue générale qu'en langue de spécialité. Dans leur *Dictionnaire des expressions et locutions,* Rey et Chantreau (1989, préface p. X), désignent la phraséologie comme un :

« système de particularités expressives liées aux conditions sociales dans lesquelles la langue est actualisée ; c'est-à-dire à des usages »

Ainsi, c'est un phénomène dynamique qui fait partie du patrimoine de la langue. De plus, comme le souligne Mel'čuk (2011 : 1), les unités phraséologiques (*phrasèmes* dans la terminologie mel'čukienne) représentent un défi pour les linguistes, et en particulier pour les approches purement formalistes car elles ne peuvent pas être « générées » à la façon chomskienne. La seule méthode permettant de les modéliser est de trouver une bonne façon de les décrire. La phraséologie est donc un vaste domaine d'investigations linguistiques. Cependant, malgré la littérature abondante qui lui est consacrée, la communauté linguistique ne s'est toujours pas mise d'accord sur une théorie opératoire rigoureuses (Mel'čuk (2011 : 1) qui soit universellement acceptée et adoptée.

1.3.2.1 Collocation – critères définitoires

Comme le soulignent Mel'čuk & Polguère (2007), il y a une grande variété d'expressions qui recouvrent la notion de collocation, il s'agit de phénomènes omniprésents dans la langue et qui relèvent d'une compétence linguistique difficile à acquérir. En effet, dans la plupart des cas, le locuteur est amené à mémoriser ces associations de lexies, qui s'opposent parfois par des distinctions sémantiques fines et qui ont des comportements syntaxiques très spécifiques (*faire fortune, accumuler une fortune).*

Rappelons rap*ibid.*ent que la notion de *collocation* a été clairement mise en avant sous le nom de *série phraséologique* dans Bally (1909 : 70) :

« il y a série ou groupement usuel lorsque les éléments du groupe conservent leur autonomie, tout en laissant voir une affinité évidente qui les rapproche, de sorte que l'ensemble présente des contours arrêtés et donne l'impression du « déjà vu »

En effet, les phénomènes collocationnels posent beaucoup de problèmes aux linguistes. Tout d'abord, car ils se situent sur un continuum, dans une « zone floue » (Fontenelle 1998) qui s'étend entre la combinaison libre et l'expression figée et dont il est très difficile de délimiter les frontières. Secondement, le terme *collocation* rassemble un groupe de phénomènes de nature différente. Cependant, il est très important d'essayer de les caractériser car leur rôle est primordial. Les linguistes ont adopté un certain nombre des critères pour définir les collocations. Selon les auteurs ou les écoles, on privilégie certains aspects plus que d'autres.

D'après Williams (2001 : 6), nous pouvons distinguer deux grandes tendances :

- la tendance lexicographique qui tend à une formalisation des collocations pour les inclure dans des dictionnaires (Hausmann, Mel' čuk)

- la tendance contextualiste, en ligne directe avec les travaux de Firth, qui considère les collocations comme un phénomène textuel et les définit en fonction de l'apparition de cooccurrences à l'intérieur d'une fenêtre (voir les travaux de Firth, Halliday, Sinclair, auxquels nous consacrons le troisième chapitre de ce travail).

a) Fréquence

La fréquence est considérée par certains linguistes (Dubois 1994, Williams 2001, Sinclair 1991) comme un des éléments fondamentaux pour caractériser et comprendre le concept. Elle figure parmi les premiers critères énoncés pour la reconnaissance des collocations. La notion de cooccurrence *habituelle* (on la retrouve chez Dubois 1994 : 91) renoue avec la tradition contextualiste britannique:

« Collocations of a given word are statements of the habitual or customary places of that word in collocational order ».

(Firth 1957, cité dans Williams 2001 : 2).

En effet, c'est Firth, membre-fondadeur de l'école contextualiste britannique qui utilise le terme *collocation* pour la première fois. Cependant, au début, les contextualises ne cherchent ni à fournir une définition du concept ni à établir une théorie de la combinatoire. Leur objectif est de réhabiliter le rôle du sens des mots dans toute analyse linguistique et de placer l'étude de la langue en contexte. Dans le cadre de cette approche pragmatique, l'analyse du sens est menée en référence au contexte. Le sens du mot découle du contexte mais en même temps, le mot influe simultanément sur ce contexte pour créer l'environnement textuel. Ainsi, l'analyse des collocations se base sur l'observation des relations sémantiques créées par le contexte, à la surface des textes. Il est nécessaire de souligner que la notion de *collocation* dans l'approche contextualiste est beaucoup plus large dans l'approche lexicographique car elle englobe aussi bien des associations lexicales syntagmatiques que des associations lexicales paradigmatiques. Pour Halliday & Hasan (1976 : 284-288), toute paire de mots reliés par une relation lexico-sémantique (aussi bien syntagmatique que paradigmatique) peut être considérée comme une collocation :

« Here we shall simply group together all the various lexical relations (…) – and treat it under the general heading of COLLOCATION, or collocational cohesion, without attempting to classify the various meaning relations that are involved. » (Halliday & Hasan 1976 : 287)

Selon ces linguistes britanniques, les collocations ont un rôle fonctionnel, celui de contribuer à la cohésion du texte :

« […] laugh … joke, blade … sharp, ill … doctor (…) The cohesive effect of such pairs depends not so much on any systematic relationship as on their tendency to share the same lexical environment, to occur in COLLOCATION with one another. In general, any two lexical items having similar patterns of collocation – that is, tending to appear in similar contexts – will generate a cohesive force if they occur in adjacent sentences »

(Halliday & Hasan 1976 : 285-286 cité dans Tutin et Grossmann 2002)

Les recherches de l'école de Birmingham, sous l'égide de Sinclair, ont contribué à l'émergence d'une nouvelle discipline : la linguistique de corpus. L'analyse de grands ensembles de textes a provoqué l'apparition de nouvelles notions : *collocation textuelle, collocation statistique* ou *collocation significative* dont la définition repose sur une dimension purement statistique.

« Collocation is the occurrence of two or more words within a short space of each other in a text. The usual measure of proximity is a maximum of four words intervening. Collocations can be dramatic and interesting because unexpected, or they can be important in the lexical structure of the language because of being frequently repeated. This second kind of collocation, often related to measures of statistical significance, is the one that is usually meant in linguistic discussions."

(Sinclair *et al.* 1991 : 170)

Cette dimension statistique prend généralement la forme d'une mesure, les plus courantes étant le z-score, le t-score et l'information mutuelle, que nous n'analyserons pas ici en détail. Néanmoins, malgré son apport important au TAL, le critère de fréquence comme critère définitoire du concept de *collocation* n'est pas suffisant. Il reste flou. Pour Williams, les mesures statistiques ne fournissent que des « candidats » collocations, etseul le lexicographe, traducteur ou linguiste natif peut décider ce qui constitue une collocation véritable. Nous sommes d'accord avec Williams sur ce point.

b) Aspect arbitraire

Plusieurs linguistes s'accordent à considérer que les collocations sont par nature arbitraires. Hausmann remarque que malgré leur transparence, les collocations restent imprévisibles. A l'instar de Hausmann, Mel'čuk *et al.* (1995 :126) reconnaissent qu'il est impossible de déterminer les collocations par règles, elles sont donc imprévisibles et doivent être apprises. Pour sa part, Benson (1989 : 3), un des auteurs d'un dictionnaire unilingue anglais *The BBI Combinatory Dictionary of English* (dans lequel l'accent est mis sur la combinatoire des lexies), définit une collocation comme une combinaison de mots récurrente mais également arbitraire :

«We can say, that collocations should be defined not just as 'recurrent word combinations', but as arbitrary 'word combinations'. […] The arbitrary (as opposed to free) nature of collocation can be demonstrated when they are juxtaposed with corresponding collocations in other languages. »

Benson (1989 : 3)

L'Homme (1998 : 514) parle du caractère conventionnel des collocations. Les lexèmes (terminologie de Mel'čuk) faisant partie des groupements sont attirés l'un vers l'autre en fonction d'un consensus établi au sein d'un groupe linguistique et non en raison des propriétés linguistiques régulières des unités lexicales qui les composent. Nous retrouvons la même caractéristique chez Tutin et Grossmann (2002 : 3) :

«Si torrentielle *peut en effet apparaître en cooccurrence avec* pluie*, cela paraît nettement moins naturel avec* précipitations torrentielles*. Cela est encore plus manifeste pour les collocations imagées du type* appétit d'ogre *ou* faim de loup *(vs.* *appétit de loup *et* *faim d'ogre*) »*

En effet, selon Williams (2001 : 4), l'aspect arbitraire devient plus significatif dès que l'on aborde le problème des traductions des collocations. D'une langue à l'autre les bases ne sélectionnent pas les mêmes colllocatifs. Prenons l'exemple de Williams (*ibid.* : 1), l'expression anglaise *heavy traffic* est traduite en français par *circulation intense*. De plus, comme le soulignent Heid et Freibott (1991 : 79), l'équivalent d'une collocation ne doit pas forcement être une collocation.

c) Transparence *vs* opacité du sens

Hausmann et Mel'čuk soulignent qu'un grand nombre de collocations peuvent être facilement interprétées par le locuteur (y compris le locuteur non natif). C'est le cas de : *célibataire endurci* (Hausmann 1978 : 191), *heavy drinker* (Cruse 1986, cité dans Williams 2001), *brosser* ou *laver les dents* (Mel'čuk 2003 : 26). Comme nous pouvons le constater, leur sens se déduit, il est transparent. Cependant, il existe des collocations qui ne sont pas immédiatement interprétables : leur sens n'est pas décomposable, l'association de constituants est opaque et arbitraire. Il s'agit ici d'expressions comme *peur bleue* ou *colère noire* (cités

dans Tutin et Grossmann 2002). Compte tenu des exemples évoqués ci-dessus, nous ne pouvons pas considérer le critère de transparence *vs* opacité comme définitoire. Par ailleurs, il est nécessaire de remarquer (d'après Williams 2001) que le critère de transparence *vs* opacité s'applique uniquement aux collocations lexicales (l'auteur met à part les collocations grammaticales qualifiées par Firth de *colligations*).

D'après certains linguistes, les notions de *transparence* et d'*opacité* reposent sur le concept de *compositionnalité* vs *non-compositionnalité* du sens. Comme le souligne Gross (1996 : 10-11), dans le premier cas, le sens d'une séquence est le produit de celui des éléments composants. La séquence (phrase ou syntagme) a une lecture compositionnelle. Dans le second cas, le sens des composants ne permet pas de conclure le sens de l'ensemble. Cependant, selon Mel'čuk (2011), on ne devrait pas confondre ces deux caractéristiques. La transparence vs objectivité a un caractère subjectif, la (non) compositionnalité est un critère complètement objectif et explicite (nous l'expliquons plus loin).

d) Degré de figement

Selon Gross (1996 : 6), le figement devrait être analysé aussi bien du point de vue syntaxique que sémantique car ce sont deux aspects d'un même phénomène qu'il convient de ne pas séparer de façon artificielle. Dans le premier cas, la structure interne d'une suite donnée fait l'objet de restrictions de nature syntaxique comme blocage des propriétés transformationnelles (impossibilité de passivation, pronominalisation, détachement, extraction, relativisation) ou non-actualisation des éléments lexicaux. Quant à Mel'čuk (2003 : 27), le degré de figement est une propriété indépendante du caractère phraséologique d'une expression. En revanche, il doit être pris en compte au moment de la description lexicographique de celle-ci.

En effet, parmi les collocations, on distingue différent degrés de figement. L'expression comme *to pay attention* (cité dans Mel'čuk 2003 : 27) n'accepte aucune transformation (le nom n'admet ni la détermination, ni la modification). En contrepartie, l'expression *to turn one's attention* fonctionne plus librement. Les contraintes particulières des collocations comme *pas d'article, pas de changement de l'ordre des mots, pas de passif* devraient être mentionnées dans les études lexicographiques. Comme nous pouvons le constater, aucun des critères énumérés ci-dessus n'a de caractère définitoire.

Le modèle définitoire dont nous nous inspirerons dans le cadre de ce travail reflète la tendance lexicographique, c'est à dire celle qui tend à une formalisation des collocations pour les inclure dans des dictionnaires. Ce modèle s'inscrit dans une optique d'encodage. La notion de *collocation*, telle que nous la retenons ici, repose sur les critères définitoires élaborés par (Hausmann (1979), développés par (Mel'cuk (2003) et repris par Tutin et Grossmann 2002. Leurs recherches dans le domaine lexicographique ont permis de mettre au jour un certain nombre de caractéristiques linguistiques des collocations. Nous les énumérons ici :

1er critère – aspect polaire (Heid : 79)

Pour Hausmann comme pour Mel'čuk, la collocation se compose de deux éléments. Ainsi, nous distinguons une *base* (dans la terminologie de Hausmann) ou un *mot clé* (dans la terminologie de Mel'čuk) et un *collocateur* (dans la terminologie de Hausmann) ou *collocatif* (dans la terminologie de Mel'čuk). Dans la plupart des cas, les collocations sont constituées de deux mots (lexèmes dans la terminologie de Mel'čuk). Cependant, il arrive que les collocations incluent des constituants plus grands que les lexèmes. Prenons l'exemple de *connaître comme sa poche* (Mel'čuk (2003 : 28) où *comme sa poche* n'est pas un mot simple mais une unité phraséologique (*phrasème* dans la terminologie de Mel'čuk) ou bien de *essuyer un échec cuisant* (Tutin et Grossmann 2002 : 4.) où on a affaire à deux collocations distinctes.

2ème critère – dissymétrie des composants (Tutin et Grossmann (2002 : 4.)

La collocation est une association de deux composants dont le statut n'est pas égal. Selon Hausmann et Mel'čuk, la *base* (ou *mot* clé) de la collocation est un élément autonome qui garde son sens habituel (celui qu'il véhicule généralement). La *base* n'a pas besoin de *collocatif* pour être définie. Quant au *collocateur* (ou *collocatif*), il dépend de la base en acquérant une nouvelle signification au sein de la collocation. Hausmann parle de *combinaison orientée* où l'élément central sélectionne une autre unité lexicale, comme dans :

célibataire ----------→ *endurci*

« La base complète la définition du collocatif, alors que le collocatif se contente d'ajouter une qualité à une base en elle-même suffisamment définie. » (Hausmann 1979 : 192).

Selon Mel'čuk (2003 25 25), le sens ⁽S⁾ d'une *collocation* inclut le sens d'un de ses constituants, celui qui se trouve dans la *position communicativement dominante*. Pour l'autre constituant, son sens peut être ou ne pas être inclus dans le sens de l'expression. Prenons l'exemple de Mel'čuk : *café noir*, collocation qui signifie ⁽café sans produit laitier⁾ où *café* est le mot clé et *noir* son collocatif. Nous voyons que le mot *café* conserve sa signification habituelle et peut fonctionner quoique appauvri, tout seul. En contrepartie, le sens du collocatif *noir* est détourné au profit d'un sens particulier que cet adjectif prend combiné avec *café.*

3ème critère - « combinaison sous contrainte » (Hausman 1979 : 191) ou caractère semi-contraint (Mel'čuk)

Comme nous avons pu le remarquer ci-dessus, le critère concernant la cooccurrence restreinte apparaît comme la continuité du caractère dissymétrique exposé plus haut. En effet, *la base* de la collocation (ou *mots clé* dans la terminologie de Mel'čuk) est sélectionnée par le locuteur de façon régulière (où *de façon régulière* veut dire selon Mel'čuk (2003 : 20): *exclusivement selon un dictionnaire de lexèmes de la langue L et les règles générales de la grammaire de la langue L*) et non contrainte (qui veut dire selon le même auteur : *se permettant l'usage de toute expression synonyme)*. En revanche, le *collocatif* y est sélectionné de façon irrégulière et/ou contrainte et son choix dépend complétement de la *base*. Comme le soulignent Heid et Freibott (1991 : 79) en faisant référence aux travaux de Mel'čuk : « *pour exprimer un contenu donné qui sert à modifier ou déterminer in lexème A, le choix entre différents lexèmes* B_1, B_2, *…,* B_n *est conditionné par le choix de A ».* Ainsi, pour lexicaliser le sens d ⁽intense⁾ en cooccurrence avec *peur*, le locuteur choisira *bleue* (Tutin et Grossmann 2002 : 4). C'est la base qui imposera donc la sélection du *collocatif* et leur cooccurrence ne sera pas libre, mais restreinte.

Après avoir analysé toutes les propriétés linguistiques des collocations, y compris leurs propriétés définitoires telles que : aspect polaire, dissymétrie et caractère contraint, nous proposons de retenir la définition de la *collocation* proposée par Mel'čuk & Polguère (2007 : 20-21), qui nous paraît claire, concise et correspondant bien à notre projet.

*« Une collocation est une combinaison de lexies (lexèmes ou phrasèmes[8]) qui est construite en fonction de contraintes bien particulières : elle est constituées d'une base, que le locuteur choisit librement en fonction de ce qu'il veut exprimer (*argument, méchant, brouillard*…) et d'un collocatif (*massue *pour* argument, comme la gale *pour* méchant, dense *pour* brouillard*…), choisi pour exprimer un sens donné (ici, (intense)) en fonction de la base.»*

1.3.2.2 Typologie des unités phraséologiques selon Mel'čuk

Nous venons de présenter la notion de *collocation* en décrivant assez scrupuleusement ses particularités linguistiques. A présent, il nous paraît important de situer la collocation parmi les autres types de phénomènes phraséologiques pour mettre au relief aussi bien leurs propriétés communes que les différences. Etant donné que notre description est menée dans une optique lexicologique largement influencée par Hausmann et Mel'čuk, nous avons décidé d'analyser la typologie des unités phraséologiques proposées par ce dernier. Rappelons que les deux linguistes s'intéressent aux phénomènes phraséologiques du point du vue de l'encodage. Les cooccurrences sont étudiées dans le cadre de la production.

Avant cela, nous voudrions évoquer les propos de Legallois et Tutin (2013) qui remarquent que la phraséologie, définie comme le domaine qui traite les séquences lexicales perçues comme préconstruites, a considérablement élargi ses objets d'études, ses méthodes et ses approches, En effet, comme le soulignent les auteures dans leur article introducteur à un numéro thématique de *Langage* consacré à la présentation des développements récents dans le domaine, la phraséologie s'émancipe de la lexicologie en intégrant des objets d'études très variés, allant des collocations aux séquences discursives. Les phénomènes phraséologiques commencent à intéresser d'autres disciplines telles que la linguistique du discours, la psycholinguistique, la linguistique informatique. Chaque discipline aborde ces phénomènes sous des angles différents et pas nécessairement par le biais de critères formels syntaxiques et sémantiques. On assiste ainsi à la découverte de nouveaux types de faits phraséologiques tels *segments répétés* (Salem), *patterns* (patrons), *routines conversationnelles* (Coulmas), *speech formulae* (Cowie), *énoncés liés* (Fónagy), *lexical bunles* (« paquets lexicaux » - Biber),

[8] C'est moi qui ai inséré le texte entre parenthèses.

greffes phraséologiques ((voir l'article de Legallois dans le même numéro) *motifs* (voir l'article de Longrée et Mellet 2013 dans le même numéro) qui contribuent à l'élargissement du champ de la phraséologie.

Revenons maintenant à l'approche traditionnelle de l'unité phraséologique basée sur les critères syntaxico-sémantiques et combinatoire du lexique. Ainsi, ces dernières années, Mel'čuk a proposé au moins deux typologies des *phrasèmes*[9] que nous présentons ci-dessous aux Figures 9 et 10. Nous constatons que les schémas évoluent et les définitions de différents types de phénomènes changent. Cela prouve que la phraséologie est en pleine effervescence et que la communauté linguistique est toujours à la recherche d'une théorie satisfaisante. Nous proposons de décrire la dernière typologie mel'čukienne (2011) car elle nous paraît extrêmement rigoureuse et logique.

Le *phrasème* doit être considéré comme un énoncé multilexémique non libre. Le terme *énoncé multilexémique* se définit, quant à lui, comme une configuration de deux ou plus lexèmes syntaxiquement liés (soulignons que dans la version précédente de sa typologie, Mel'čuk (2003) définit le *phrasème* comme *syntagme non libre*). Le caractère libre *vs* contraint doit être déterminé par rapport à l'axe paradigmatique.

« Un énoncé multilexémique est libre si et seulement s'il n'est pas contraint sur l'axe paradigmatique, c'est-à-dire si son sens et chacune de ses composantes lexicales sont sélectionnés par le Locuteur strictement pour ses propriétés linguistiques, c'est-à-dire indépendamment des autres composantes »

(Mel'čuk 2011 : 2).

Ainsi, chaque composante d'un énoncé libre peut être remplacée par n'importe quelle expression synonyme, en préservant la correction linguistique et le sens de cet énoncé. Comme le souligne (Mel'čuk 2011 : 3), un phrasème ne peut pas être librement construit par le locuteur, il doit donc être stocké dans sa mémoire. Dans la typologie de Mel'čuk de 2011, la caractérisation et la classification des phrasèmes se fait selon deux axes : l'axe paradigmatique – « *les contraintes de SÉLECTION de leurs composantes* », et selon l'axe

[9] Phrasème est une unité phraséologique dans la terminologie mel'čukienne.

syntagmatique – « *les contraintes de COMBINAISON de leurs composantes, ou leur compositionnalité* » (Mel'čuk 2011 : 3)

a) **Phrasèmes lexicaux *vs* phrasèmes sémantico-lexicaux**

Selon la nature des contraintes phraséologiques de sélection, Mel'čuk divise les phrasèmes en phrasèmes lexicaux et phrasèmes sémantico-lexicaux. On rappelle que selon la théorie S-T, le locuteur construit des textes en deux étapes, c'est-à-dire qu'il part d'un contenu informationnel C qu'il veut, dans une situation particulière, exprimer par un texte en langue L. Il construit donc pour C une Représentation Sémantique ⁽S⁾, ensuite il construit, pour ⁽S⁾ le texte T. Si la liberté d'un phrasème est enfreinte au moment du choix des lexèmes pour exprimer ⁽S⁾ alors que le locuteur a pu construire librement le sens ⁽S⁾ pour un contenu informationnel donné C, nous avons affaire à un phrasème lexical. Si, en revanche, le sens ⁽S⁾ d'un phrasème n'est pas construit par le Locuteur mais sélectionné comme un tout de façon contrainte, en fonction d'une situation donnée, pour exprimer un contenu informationnel C, nous devons parler d'un phrasème sémantico-lexical. Il est nécessaire de souligner que dans le deuxième cas, les contraintes opèrent à deux étapes en visant le sens et l'expression lexicale. Selon Mel'čuk (2011 :3) *Vous dites ?*, *sauf imprévu*, *Défense de stationner* ou *Ne pas se pencher au dehors* sont des exemples de phrasèmes sémantico-lexicaux.

.

b) **Phrasèmes sémantiquement compositionnels *vs* non compositionnels**

Les phrasèmes sont caractérisés par une propriété importante : leur compositionnalité ou non-compositionnalité sémantique.

Comme le souligne Mel'čuk 2011, un signe linguistique complexe AB est dit compositionnel

$$\text{si } \mathbf{AB = A \oplus B}$$

Où le symbole $\oplus$ représente l'opération d'union linguistique, qui réunit les signes et leurs composantes selon leur nature et leurs propriétés suivant les règles générales de la langue donnée. Ainsi un phrasème comme *porter attention sur* N_Y est compositionnel car son sens peut être distribué entre ses composantes lexicales :

⁽X cause que son attention soit sur Y⁾ = ⁽causer que … soit sur⁾ ←→ *porter* et ⁽attention⁾ ←→ *attention.*

Mel'čuk 2011 souligne que le phénomène de compositionnalité *vs* non compositionnalité ne doit pas être confondu avec celui d'opacité *vs* transparence. Selon lui, la compositionnalité sémantique d'une expression s'analyse en fonction des composantes sémantiques explicites qu'on trouve dans sa définition et dans celles de ses constituants. Mel'čuk donne l'exemple d'une expression sémantiquement non compositionnelle (*prendre le taureau par les cornes*) alors que pour un locuteur, il s'agit plutôt d'une expression transparente. Nous pouvons constater qu'aucun des constituants de cette expression n'apparaît dans sa définition : (traiter la difficulté en question immédiatement et directement). Mel'čuk considère la compositionnalité comme un fait objectif tandis que le caractère de transparence *vs* opacité relève de la sphère subjective.

1.3.4.1 Classification des phrasèmes selon Mel'čuk (2011)

Le croisement de deux dimensions décrites plus haut, notamment la nature des contraintes et la compositionnalité des phrasèmes a permis de mettre au jour trois classes de phrasèmes : locutions, collocations et clichés.

a) Locutions

Les locutions sont des phrasèmes lexicaux (aucun de ses composants lexicaux n'est sélectionné librement) et non compositionnels. La classe des locutions se subdivise en 3 sous-classes : locutions fortes ou complètes, semi-locutions et locutions faibles ou quasi-locutions. Cette division se fait en fonction de l'inclusion du sens des composants A et B dans le sens (S) de la locution AB.

1er cas de figure – locutions fortes ou complètes (il est nécessaire de remarquer que cette nouvelle dénomination correspond aux *phrasèmes complets* présentés dans la typologie de mel'čukienne 2003)

« Une locution forte n'inclut dans son sens aucun des sens de ses composantes :

'AB' ⊅ 'A' et 'AB' ⊅ 'B' » (Mel'čuk 2011 : 4)

C'est notamment le cas des expressions suivantes : *ne pas avoir froid aux yeux¸ au bout du rouleau, mener en bateau, cordon bleu*

Afin de caractériser les deux autres sous-classes des locutions, Mel'čuk introduit la notion de *pivot sémantique* qu'il tient à distinguer explicitement du concept de *composante sémantique communicativement dominante* (qui à son tour correspond au concept de composante générique du ⟨S⟩) :

« Posons qu'un sens ⟨S⟩ peut être divisé en deux parties, ⟨S_1⟩ et ⟨S_2⟩ où ⟨S⟩ = ⟨S_1⟩ ⊕ ⟨S_2⟩.
La partie ⟨S_1⟩ du sens ⟨S⟩ est appelée le pivot sémantique de ⟨S⟩ si l'autre partie ⟨S_2⟩ est un prédicat dont ⟨S_1⟩ est l'argument : 'S' = ⟨S_2⟩ (⟨S_1⟩). » (Mel'čuk 2010 : 5)

Ainsi, pour le sens ⟨réussir un examen⟩, le pivot sémantique est ⟨examen⟩ ou pour le sens ⟨vendre la voiture⟩, le pivot sémantique est ⟨voiture⟩, alors que les composantes sémantiques communicativement dominantes correspondent respectivement à ⟨réussir⟩, ⟨vendre⟩.

2ème cas de figure – semi-locutions (les locutions de ce type faisaient partie des phrasèmes complets dans la typologie mel'čukienne 2003)

« Une semi-locution inclut dans son sens le sens d'une de ses composantes (disons, de A), mais pas en tant que pivot sémantique, et n'inclut pas le sens de l'autre (donc, de B), tout en incluant encore un sens additionnel 'C', qui est son pivot sémantique[10] *:*

'AB' ⊃ 'A', et 'AB' ⊅ 'B', et 'AB' ⊃ 'C' (Mel'čuk 2011 : 5)

C'est le cas des expressions suivantes : *fruits de mer* : ⟨animaux de mer comestibles qui ne sont pas des poissons⟩ où ⟨animaux comestibles⟩ joue le rôle du pivot sémantique, *bain de foule :* ⟨acte qui consiste à se mêler à la foule⟩ où ⟨acte⟩ joue le rôle de pivot sémantique.

3ème cas de figure : locutions faibles ou quasi-locutions (qui correspondent aux *quasi-phrasèmes* dans la typologie mel'čukienne 2003)

[10] Mel'čuk indique les pivots sémantiques par des hachures.

« Une locution faible inclut dans son sens le sens de toutes ses composantes, mais pas en tant que pivot sémantique, en incluant aussi un sens additionnel 'C', qui constitue le pivot :

'AB' ⊃ 'A', et 'AB' ⊃ 'B', et 'AB' ⊃ 'C' (Mel'čuk 2011 : 5)

C'est le cas des expressions suivantes : *donner le sein, rouge à lèvres, point-virgule*.

Analysons, avec Mel'čuk (2011 :6), l'expression *donner le sein* dont la définition (qualifiée de *propositionnelle* dans la terminologie mel'čukienne) est la suivante : (femme X nourrit le bébé Y de son lait en mettant le mamelon d'un de ses seins à la bouche de Y). Même si son sens inclut les sens de deux constituants : (donner) et (seins), c'est (nourrit) qui apparaît dans le rôle du pivot sémantique. Nous avons dont affaire à une quasi-locution (ou locution faible).

c) **Collocation (cette classe de phrasèmes correspond aux *semi-phrasèmes* de la typologie de 2003)**

La deuxième classe de phrasèmes englobe les collocations. Comme dans le cas des locutions, il s'agit des expressions contraintes au niveau des choix lexicaux (ou plutôt semi-contraintes), mais contrairement à ces dernières, les collocations sont compositionnelles.

'AB' ⊃ 'A', et B est sélectionné en fonction de A (Mel'čuk 2010 : 5)

Nous parlons du caractère semi-contraint car une de ses composantes est sélectionnée par le Locuteur librement, juste pour son sens et l'autre est choisie en fonction de la première composante et du sens à exprimer. Nous avons présenté les collocations dans la partie précédente.

Mel'čuk (2011 : 8) divise les collocations en deux sous-classes selon a) leurs caractéristiques quantitatives et b) leur capacité de participer au paraphrasage (phénomène de synonymie de phrases au niveau syntaxique profond). On distingue donc: de très nombreuses *collocations standard* (systématiques et impliquées dans la description du paraphrasage) *et non standard* comme dans *année bissextile, voix flutée, nuit blanche* (dont le lien sémantique entre(A) et (B) n'est pas systématique et les capacités de paraphrasage quasi-nulles) .

d) **Cliché**

Il reste encore une classe des phrasèmes sémantico- lexicaux compositionnels qui est présentée à part dans la typologie de Mel'čuk 2011 à cause de ses propriétés sémantiques. Il s'agit des clichés dont la particularité réside dans le fait qu'ils sont contraints au niveau conceptuel, c'est-à-dire que les contraintes opèrent au moment du passage entre la Représentation Conceptuelle et la Représentation Sémantique. C'est le cas des expressions suivantes : *Défense de stationner, Ne pas se pencher au dehors, Vous dites ? Vous désirez ? Quel âge avez-vous ? Peinture fraîche.* Parmi les expressions citées ci-dessus, on distingue des *pragmatèmes*, une sous-classe des clichés contraints par les conditions pragmatiques de leur emploi, c'est-à-dire par le type de situation dans laquelle le locuteur les utilise. Ainsi, nous rencontrons *Défense de stationner* sur des panneaux officiels et *Ne pas se pencher au dehors* uniquement dans les trains. Selon Mel'čuk, les pragmatèmes ainsi que les conditions de leur utilisation devraient figurer dans les dictionnaires.

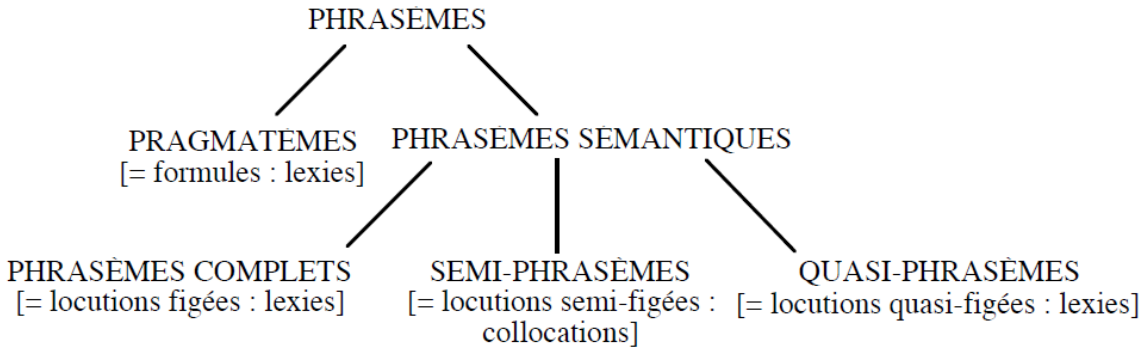


Figure 9. Typologie des phrasèmes (Mel'čuk 2003 : 26)

Si nous comparons les deux typologies de Mel'čuk (Figure .9 et Figure 10), nous constatons que dans la typologie de 2003, les sous-classes des phrasèmes (phrasèmes complets *vs* phrasèmes, semi-phrasèmes *vs* quasi-phrasème) ont été définies par rapport au critère d'inclusion *absolue* (inclus/non inclus) des composantes sémantiques. En revanche, la nouvelle typologie met plus en relief les dimensions de *compositionnalité* vs *non compositionnalité* ainsi que la nature lexicale ou sémantico-lexicale des contraintes. De plus, l'introduction du concept de *pivot sémantique* permet de saisir les différences sémantiques de trois types de locutions.

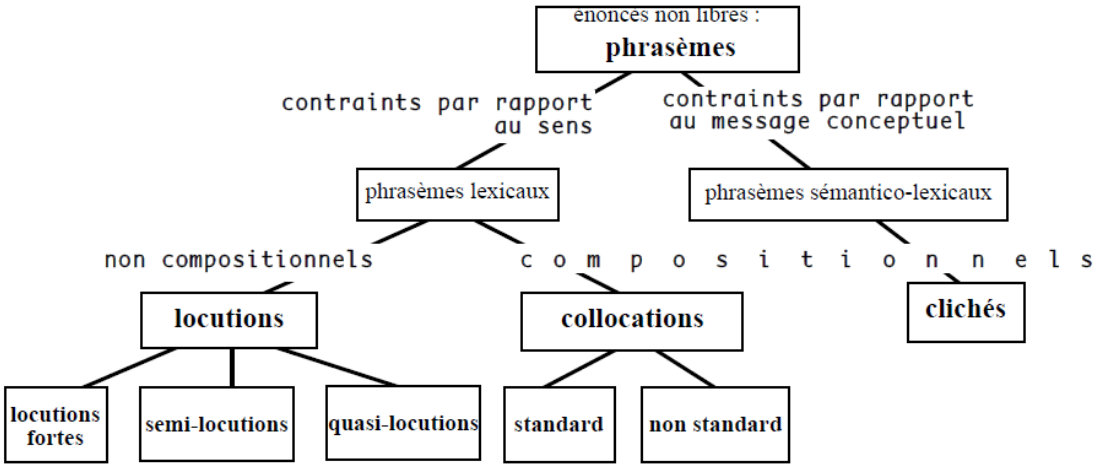


Figure 10. Typologie des phrasèmes (Mel'čuk 2011 : 12)

1.4. Fonctions lexicales comme outil permettant la modélisation des phénomènes sémantiques

Après avoir présenté l'ensemble des postulats et des principes de la théorie Sens-Texte nous aimerions montrer les avantages pratiques qu'offre cette démarche. En effet, l'originalité majeure de l'approche ST réside dans le concept de fonctions lexicales. C'est un modèle fonctionnel qui donne la possibilité d'étudier et de modéliser de façon logique, systématique et rigoureuse les phénomènes tels que la dérivation sémantique et la combinatoire lexicale (décrits en détail plus haut). C'est un outil efficace qui permet de rendre compte des différents types de liens qui peuvent unir les éléments du réseau lexical d'une langue donnée.

1.4.1. Concept de fonction lexicale

Le concept de fonction lexicale repose sur l'hypothèse que les cas de cooccurrence lexicale restreinte (la cooccurrence lexicale restreinte est définie par Mel'čuk (2003 : 25) comme l'ensemble des collocations contrôlées par la lexie L, où la lexie L est la base de collocation) se rencontrent avec un nombre fort réduits de sens spécifiques – très abstraits et

généraux (Mel'čuk 1997 : 23). Il devient donc possible de décrire de façon cohérente ces expressions phraséologiques dont la construction est imprévisible et capricieuse.

Du point de vue formel, une fonction lexicale (=FL) ressemble à une fonction mathématique qui peut être représentée de la manière suivante :

$$f(x) = y,$$

où *x* est l'argument de la fonction (ou son mot-clé) et *y* sa valeur. Ces fonctions sont appelées lexicales car elles n'acceptent en tant qu'argument que des lexies et en tant que valeur, que des ensembles de lexies ((Mel'čuk *et al.* 1995 : 126). Autrement dit, une fonction lexicale est une correspondance f qui associe à une lexie L (argument de f), un ensemble de lexies ou syntagmes figés f(L) – valeur de f. Chaque FL f est associé à un sens ⟨f⟩ très abstrait et très général et, en même temps, à un rôle syntaxique profond. L'argument d'une FL f est la lexie L sur laquelle le sens ⟨f⟩ porte ; et la valeur de la FL f pour un argument donné L est un ensemble de lexies qui peuvent réaliser f (c'est-à-dire, exprimer le sens ⟨f⟩) au lieu de L ou auprès de L. lexies ((Mel'čuk 2003: 30).

Alain Polguère (Polguère 2008 : 160) précise qu'il existe autant de fonctions lexicales qu'il existe de types de liens lexicaux et chaque fonction lexicale est identifiée par un nom particulier. Pour qu'une correspondance lexicale f puisse être considérée comme une fonction lexicale FL normale, elle doit remplir deux conditions (Mel'čuk *et al.* 1995 : 127-128) :

1. Pour toute paire de lexies L_1 et L_2, les lexies $f(L_1)$ et $f(L_2)$ montrent des relations sémantico-syntaxique (presque) identiques à ces lexies :

$$\frac{\langle \mathbf{f}(L_1) \rangle}{\langle L_1 \rangle} = \frac{\langle \mathbf{f}(L_2) \rangle}{\langle L_2 \rangle}$$

2. En règle générale, et au moins pour certains arguments, $f(L_1) \neq f(L_2)$

Il nous paraît indispensable d'apporter quelques explications et illustrations aux formules presque algébriques présentées ci-dessus.

La condition 1 peut être illustrée à l'aide d'un exemple qui a déjà été cité à plusieurs reprises dans les travaux de Mel'čuk (1995, 2003), notamment à l'aide de la FL [Magn] qui exprime le sens de ⟨très⟩, ⟨intense⟩, ⟨à un degré élevé⟩, il est donc un intensificateur. Cette FL

joue le rôle de modificateur adjectival ou adverbial de la lexie-clé L). Voici quelques exemples de son utilisation (tirés de Mel'čuk 1995 : 136 et Polguère 2008 : 169) :

Magn*(pleurer) = amèrement, à chaudes larmes, comme une Madeleine, toutes les larmes de son corps, comme un veau, comme une vache, comme un enfant*
Magn*(pluie) = grosse* I $_{\text{prépos,}}$ *diluvienne, torrentielle, violente*
Magn*(fièvre) = de cheval*
Magn*(funérailles) = imposantes*
Magn*(boire) = comme un trou*

Si l'on applique la formule de la condition 1 à deux premières fonctions, on obtient :

$$\frac{\textit{Comme une Madeleine}}{\text{PLEURER}} = \frac{\textit{grosse}}{\text{PLUIE}} = \ldots$$

Les trois points de suspension signifient que la proportion peut être prolongée à volonté en y ajoutant d'autres paires modélisées par la FL **Magn**). En effet, pour être une FL, une correspondance lexicale doit donner lieu à un grand nombre de proportions de ce type Mel'čuk (2003 : 32).

Comme le remarque Mel'čuk (1995 : 128), tout élément du premier ensemble (en l'occurrence : *comme une Madeleine*) se trouve par rapport à PLEURER (mot clé et, en même temps, base de collocation) dans une relation sémantique et syntaxique qui est identique à la relation qu'entretiennent PLUIE (mot clé et, en même temps, base de collocation) et tout élément du deuxième ensemble (en l'occurrence : *grosse*). Il est évident que l'expression *comme une Madeleine* et l'adjectif *grosse* ne sont ni sémantiquement, ni syntaxiquement égaux. Néanmoins, ils jouent le même rôle par rapport aux lexies qu'ils accompagnent : ils sont tous les deux des modificateurs syntaxiques signifiant, dans ce contexte donné, (intense/intensément).

Si nous analysons les valeurs de la FL **Magn** présentées ci-dessous, nous voyons que la condition 2 est également satisfaite :

comme une Madeleine ≠ grosse ≠ de cheval ≠ imposantes ≠ comme un trou

En effet, l'application de la fonction **Magn** permet d'obtenir un grand nombre de valeurs différentes. Cependant, il existe des correspondances lexicales f qui donnent lieu à un nombre important de proportions mais l'on y trouve le même numérateur pour des dénominateurs différents. C'est le cas du sens (dont le prix est élevé) (dans la langue française) qui appliqué à une série de mots clé présente toujours la même valeur, le lexème *cher*.

$$\frac{\textit{cher}}{\text{VOITURE}} = \frac{\textit{cher}}{\text{VOYAGE}} = \ldots$$

Comme le remarque Mel'čuk (1995 : 128), une telle correspondance lexicale f est triviale et ne présente aucun intérêt car son résultat n'est pas une collocation. Ainsi, on ne peut pas la retenir comme FL.

1.4.2 Typologie des FL

Comme le souligne Jousse (2010 : 73), la notion de FL a un caractère graduel. Ainsi, traditionnellement, on distingue deux sortes de FL : FL standard et FL non standard (voir le *DiCouèbe*). Cependant, selon Jousse (2010 : 85), les FL peuvent s'organiser autour de plusieurs statuts. D'après l'auteur (2007 : 470), plus une relation sémantique est applicable à un large ensemble d'unités lexicale et plus elle est considérée comme standard. En revanche, la relation qui ne concerne qu'un très petit nombre d'unités lexicales, et c'est le cas des langues spécialisées, sera décrite au moyen de FL non standard. Les relations non standard sont encodées selon un métalangage naturel qui met l'accent sur la lisibilité et la clarté de la formulation. On passe ainsi d'une description très synthétique et générale (les FL standard) à une description très précise (les FL non standard). En effet, Jousse (2010 : 85) propose la typologie suivante des FL :

a) FL standard. Construites à partir d'un noyau de 60 FL standard simples (Mel'čuk *et al.* 1995 : 125), elles peuvent être composées d'un seul élément ou combinées entre elles pour :

- former des fonctions complexes. Il s'agit d'un enchaînement de FL simples syntaxiquement liées qui, ayant une valeur globale cumulative, exprime, de façon décomposable, le sens de l'enchaînement entier. Les FL complexes permettent de rendre compte des nuances infimes du sens.

Ex. **S_{res}AntiBonInvolv** (*abeille*) = *piqûre*

- former des configurations de FL. Une configuration de FL est « *une suite de FL simples qui ne sont pas syntaxiquement liées entre elles, mais qui ont le même mot-clé, cette suite ayant une valeur globale cumulative qui exprime de façon indécomposable le sens de la suite entière* ». Mel'čuk *et al.* (1995 : 149)

Ex. **Bon** *(joie)* + **Magn** (*joie*) = *paradisiaque*

- pour former des FL qui ont des valeurs fusionnées, c'est -à dire que le sens de ces fonctions est exprimé par une seule unité lexicale qui comprend le sens du mot clé ainsi qu'une composante indiquant le sens de la fonction (qui habituellement est exprimé par le mot clé + un élément collocatif). Les éléments fusionnés sont indiqués par une double barre inclinée, ce que nous avons déjà pu constater dans 2 exemples cités plus haut. Voici un autre exemple de ce phénomène :

Ex. **AntiBon***(film) = // navet*

Soulignons que les FL standard doivent répondre aux principes d'universalité, de diversité et de cardinalité (voir Polguère 2007 : 48, Jousse 2010 : 73-75)

b) FL standard avec indication de phase :

$Real_2^{I}$ (*avertissement*) = *recevoir* [ART ~]

$Real_2^{II}$ (*avertissement*) = *tenir compte* [de ART ~]

c) FL standard avec pointeurs vers la définition : **$Real_1^{usual}$** (*cigarette*) = *fumer* [ART ~]

d) FL localement standard encodées dans la langue de description selon des formules normalisée : **Essayer_AntiReal$_2$**, (*barricade*) = *attaquer* [ART ~]

e) FL semi-standard constituées d'une FL standard et d'un élément en français venant ajouter une composante de sens non prise en charge par la FL standard : **complètement Real$_2$** (*armoire*) = *remplir* [ART ~]

f) FL non standard écrites intégralement dans la langue de description, en l'occurrence en français: **Chair de ~ utilisé comme aliment** (*mouton*) = *mouton*

1.4.2.1 FL standard simples

Pour qu'une fonction lexicale puisse être qualifiée de standard, elle doit remplir deux conditions (Mel'čuk 2003 : 33) :

1. f doit avoir un grand nombre d'arguments, c'est-à-dire, f doit avoir une vaste cooccurrence sémantique, le sens ⁽f⁾ doit être est suffisamment abstrait et général pour s'appliquer à beaucoup d'autres sens.
2. f doit posséder un grand nombre de valeurs différentes, c'est-à-dire le sens ⁽f⁾ doit assurer un nombre élevé d'expressions possibles (*y*).

Pour parler des conditions déterminant le caractère standard d'une FL, Polguère (2007) propose d'introduire la notion de *standardness*. Selon l'auteur, pour être considéré comme standard, une FL doit répondre aux principes de cardinalité[11] (*Broadness condition*), c'est-à-dire avoir un nombre élevé d'arguments et de diversité (*Diversity condition*), c'est –à-dire retourner des valeurs appartenant à des classes sémantiques différentes. De plus, comme le rappelle Jousse (2010 : 75), une FL standard doit remplir la condition d'universalité, c'est-à-dire, son existence devrait être prouvée dans toutes les langues. C'est notamment le cas de la FL **Magn** dont le sens ⁽très⁾ se combine facilement avec plusieurs lexies et donne lieu (comme dans le cas de PLEURER) à un grand nombre de valeurs.

Soulignons qu'il existe une soixantaine de fonctions lexicales standard (plus exactement, elles sont au nombre de 56). Selon qu'elles expriment des relations paradigmatiques ou des relations syntagmatiques entre les lexies, les fonctions standard sont réparties en deux catégories : FL syntagmatiques et FL paradigmatiques. Chacune de ces FL est identifiée par

[11] Nous nous appuyons sur la traduction proposée par Jousse (2010 : 73).

un nom conventionnel et est traitée comme unité ultime, indécomposable (Mel'čuk 2003 : 35). Nous essayerons, dans les pages qui suivent, d'expliciter leur rôle et utilisation.

1.4.3 FL paradigmatiques

Les FL sont un outil très efficace pour la description de ce que Mel'čuk appelle « les dérivations sémantiques ». Elles permettent de décrire les relations sémantiques entre les lexies, modéliser les liens qui les unissent et leur positionnement dans le réseau lexical. Elles permettent de spécifier l'ensemble de toutes les possibilités d'expression dans le même paradigme sémantique (Mel'čuk *et al.* 1995 : 125). Il s'agit d'abord des FL qui décrivent les relations sémantiques « fondamentales », celles qui (comme le souligne Polguère 2008 : 147), forment la charpente de la structuration sémantique du lexique de chaque langue, notamment : synonymie, conversion, antonymie, hyperonymie (dans la liste des FL de la Théorie ST, on ne trouve pas de FL correspondante à l'hyponymie) – voir les deux premiers cas de figure ci-dessus.

Nous énumérons ainsi les FL suivantes :

- Synonyme $\mathbf{Syn}$, $\mathbf{Syn}_{\supset}$, $\mathbf{Syn}_{\subset}$, $\mathbf{Syn}_{\cap}$

qui distinguent respectivement le cas de synonymie absolu et trois types de synonymes approximatifs : plus spécifique, moins spécifique (inclusion de sens) et à intersection

ex. **Syn** *(voiture) = automobile*

- Conversif **Conv**, FL correspondante au phénomène de conversion, c'est-à-dire d'inversion de deux actants (au moins deux) dans le cas des prédicats dont le sémantisme est identique

ex. **Conv***(effrayer) = craindre* [*La défaite m'effraie = Je crains la défaite*]

- Antonyme **Anti**, FL spécifie les antonymes d'une lexie L en question, c'est-à-dire des lexies dont les sens se distinguent par la négation ou, plus généralement, la mise en opposition d'une de leurs composantes (Polguère 2008 : 152)

Ex. **Anti** (*petit) = (grand)*

- Générique **Gener**, FL qui détermine pour L, un mot générique. En simplifiant légèrement, cette fonction couvre des relations d'hypéronymie.

Ex. **Gener** (*petit) = (grand)*

Comme nous l'avons déjà signalé plus haut, aucune FL ne correspond à la relation d'hyponymie. Cependant, la fonction inverse de **Gener** a été proposée indépendamment par Grimes : il s'agit de la fonction **Spec**:

Ex. **Spec** (*imprimante*) = *imprimante laser, imprimante à jet d'encre*,…

Les trois fonctions suivantes : **Contr, Epit, Figur,** sont considérées comme des types de synonymie et d'antonymie et correspondent respectivement :

- à une expression contrastive consacrée par l'usage, à utiliser dans une figure rhétorique
Ex. **Contr** (*glace*) = *feu,*
- à un épithète pléonastique (adjectif ou adverbe) formant avec le mot clé une expression-cliché
Ex. **Epit**(*océan*) = (*immense*),
- à un métaphore qui forme avec le mot clé un synonyme plus riche de celui-ci

Ex. **Figur**(*fumée*) = *rideau* [de ~]).

Ensuite, nous avons un groupe de FL nominales qui permettent d'exprimer des relations sémantiques plus spécifiques comme : Singulatif **Sing** (quantum régulier) d'un mot clé – ex. **Sing***(riz)* = *grain [de riz]*; Collectif **Mult** (ensemble régulier) d'un mot clé – **Mult***(chien)* = *meute [de chiens]*; **Cap** (chef de) – ex. **Cap***(avion)* = *commandant (de bord)* ; **Equip** (équipe de) - ex. **Equip***(avion)* = *équipage[de* ART *avion]*. Il reste encore quelques autres FL de ce type mais nous ne les présenterons pas toutes ici.

Le troisième groupe de FL correspond à tous les autres cas de dérivation. D'une part, il s'agit des dérivés syntaxiques du mot clé (voir le premier cas de figure – changement de partie de discours) et d'autre part, des dérivés sémantiques nominaux actanciels et circonstanciels et des dérivés sémantiques adjectivaux actanciels, potentiels et virtuels (correspondant à la troisième famille).

Les dérivés syntaxiques sont représentés par quatre fonctions lexicales:
$\mathbf{S_0}$ = nominalisation (ou substantivation), $\mathbf{V_0}$ = verbalisation, $\mathbf{A_0}$ = adjectivisation et $\mathbf{Adv_0}$ = adverbialisation. Ils s'opposent aux dérivés sémantiques par le fait qu'ils présentent le même

contenu sémique que le mot-clé de la fonction, bien que leurs caractéristiques syntaxiques soient différentes.

Ex. **S_0***(tomber)* = *chute* ; **V_0***(chute)* = *tomber ;* **A_0***(ville)* = *urbain* ; **Adv_0**(*honnête)* = *honnêtement*

Dans les exemples ci-dessus, nous voyons que les FL sont réversibles (*chute/tomber*). Il est également nécessaire de souligner que la dérivation syntaxique ne peut pas être assimilée à la dérivation morphologique. Ici, les FL reflètent seulement des relations lexicales pures (Mel'čuk *et al.* 1995 : 133).

Après cette petite parenthèse historique, revenons à la présentation des fonctions lexicales paradigmatiques, relatives, cette fois-ci, aux cas de dérivation sémantique portant sur les actants et les circonstants. Commençons par les FL nominales.

S_1 *(parler) = locuteur*
S_2 *(parler) = paroles, propos, le dit, discours, …*
S_3 *(parler) = allocutaire, destinataire*

} Les FL permettant de dégager les actants ASynP I/II/III [**S_1** parle à **S_3** en lui disant **S_2**]

S_{instr} *(parler) = langue*
S_{loc} *(parler) = parloir*
S_{mod} *(parler) = façon* [de~]

} Les FL permettant de consigner les circonstants comme : instrument, lieu, moyen, manière, résultat

En ce qui concerne les FL adjectivales, nous distinguons trois types de dérivés : dérivés sémantiques adjectivaux actanciels **A_1** **$A_{2,}$** **$A_{3,}$** potentiels **Able**$_{1,}$ **Able**$_{2,}$ **Able**$_3$ et virtuels **Qual**$_1$ **Qua**$_2$ **Qual**$_3$ de la lexie L_2 en tant que ASyntP I/II/III de la lexie L_2

Ex. **A_1** *(mépris) = plein, rempli* [de ~] ou **A_2** *(mépris) = couvert* [de ~]

Ou bien

Able$_1$ *(tromper) = trompeur* ou **Able**$_2$ *(lire) = lisible*

Qual$_1$ *(tromper) = malhonnête* ou **Qual**$_2$ *(tromper) = naïf*

1.4.4 FL syntagmatiques

Les fonctions lexicales FL syntagmatiques modélisent la *combinatoire lexicale restreinte* d'une lexie L donnée. Cette combinatoire est définie par Mel'čuk (2003 : 25) comme l'ensemble des collocations contrôlées par la lexie L. Il s'agit donc de dégager, pour une lexie donnée, des cooccurrents lexicalement contraintes (dont la combinatoire n'est déterminée ni par leur sémantisme ni par leurs propriétés syntaxiques), en permettant ainsi la production de la combinaison adéquate. Les FL représentent un outil efficace pour le recensement des expressions non prévisibles, les expressions qui doivent être apprises car il est impossible de les définir par règles. Rappelons que Bally (1909 : 70-72) a déjà attiré attention sur l'existence des séries usuelles qui représentent des catégories importantes comme séries d'intensité (*chaleur suffocante, accablante, tropicale, torride, sénégalienne*) ou des séries verbales qui périphrasent les verbes (*remporter une victoire, prendre un engagement, avoir l'habitude, avoir coutume)*. Les FL que nous allons présenter rap*ibid.*ent ci-dessus permettent de les systématiser.

Ainsi, parmi les fonctions syntagmatiques, on distingue trois catégories : fonctions adjectivales (parmi lesquelles la fameuse FL **Magn)**, fonctions adverbiales et fonctions verbales (le groupe le plus important). Les fonctions syntagmatiques adjectivales sont des modificateurs adjectivaux ou adverbiaux des mots clé et expriment différents sens comme par exemple: Intensificateur **Magn** pour le sens ⁽très⁾, Comparatif **Plus/Minus** pour les degrés de comparaison, Laudatif **Bon** pour l'approbation subjective du locuteur ou Confirmateur **Ver** pour exprimer le sens ⁽tel qu'il faut⁾.

Ex. **Magn** (*amour*) = *ardent, fou*
Ver (*succès*) = *mérité*
Bon (*conseil*) = *précieux*

La deuxième catégorie concerne les fonctions adverbiales. Les FL $\mathbf{Adv_1}$, $\mathbf{Adv_2}$ $\mathbf{Adv_3}$ correspondent aux dérivés sémantiques adverbiaux actanciels et jouent le rôle des modificateurs adverbiaux typique de la lexie L_2 en tant qu'ASyntPI/II/III de la lexie L_1

Ex. $\mathbf{Adv_1}$ (*joie*) = *avec* [~]

Adv$_2$ (*joie*) = *à* [la ~]

Trois autres fonctions syntagmatiques adverbiales expriment respectivement le moyen (ou l'instrument) Instrumental **Instr**, la localisation Locatif **Locatif** et la cause Conséqutif **Propt**.

Ex. **Instr** (*téléphone*) = *par* [~]
Loc $_{in/ad}$ (*ville*) = *en* [~]
Propt (*jalousie*) = *par* [~]

Le troisième et dernier groupe des FL syntagmatiques est constitué des fonctions lexicales verbales.

Comme le remarquent Mel'čuk & Polguère 2007 : 21), il existe un nombre élevé de collocations où le collocatif ne sert pas à exprimer un sens donné, mais à « supporter » syntaxiquement le nom qui est la base de la collocation. Ce collocatif sert donc à construire une expression fonctionnant comme un équivalent verbal du nom en question. Il s'agit là d'un type fort différent de collocations, les collocations de *verbes supports*. En effet, les FL qui formalisent la notion de *verbe support* sont au nombre de trois **Oper, Func, Labor.** Leurs valeurs sont des verbes sémantiquement vides (ou vidés dans le contexte de leur mot clé) et servant à verbaliser des noms prédicatifs ; leur vocation est simplement syntaxique (Mel'čuk et al ; 1995 : 138)

Ex.
Oper$_1$ *(remarque) = faire* [ART~]
Func$_2$ *(danger) = menacer* [N]
Labor$_{12}$ *(estime) = tenir* [N en ~]

Ensuite, nous distinguons un sous-groupe des FL verbales de réalisation (**Real, Fact, Labreal**) qui expriment la réalisation des « objectifs inhérents de la chose désignée par le mot clé » (Mel'čuk et al ; 1995 : 141). Leurs valeurs sont des verbes sémantiquement pleins.

Ex.
Real$_1$ *(peine) = imposer, infliger* [ART~]
Real$_2$ *(peine) = purger* [ART~]
Fact$_0$ *(film) = être à l'affiche*

Labreal$_{12}$ *(armoire)* = *ranger* [N dans ART ~]

Le troisième sous-groupe des FL verbales comprend **Incep, Fin, Cont**. Elles expriment les trois phases différentes d'un état ou d'un événement (respectivement ⁽début⁾, ⁽fin⁾, ⁽continuation⁾). Leurs valeurs sont des verbes sémantiquement pleins. Ces trois FL ont un aspect aspectuel et doivent donc se combiner avec des verbes.

Ex.

Incep *(dormir) = // s'endormir*

Fin *(dormir) = // se réveiller*

Finalement, nous distinguons un sous-groupe des FL exprimant trois types de causation d'un état ou d'un état **Caus, Liqu, Perm**. Ces FL sont des verbes sémantiquement pleins qui ont respectivement les significations suivantes ; ⁽causer que P⁾ ou (faire en sorte que P a lieu, liquider P⁾ ou ⁽faire en sorte que P non a lieu, ⁽permettre P⁾.

Ex.

CausFunc$_1$ (*forme*) = *donner* [*à* N ART ~]

Mel'čuk définit encore d'autres FL verbales qui permettent d'exprimer d'autres types de verbes (tels que les verbes de manifestation, de préparation, de dégradation, de son typique, etc), mais nous ne les détaillerons pas ici. Nous espérons que la présentation faite ci-dessus sera suffisante pour comprendre les règles de ce mécanisme.

1.4.5 Ressources lexicales fondées sur les principes de la LEC

Nous proposons ici de faire un survol des ressources lexicales fondées sur les principes de la LEC. Nous allons présenter rap*ibid.*ent quatre dictionnaires qui ont recours au formalisme des FL, à savoir : le *DEC*, le *DiCo* (consultable au moyen de l'interface *DiCouèbe*), le *LAF* et le *RLF*. Il s'agit des projets expérimentaux qui contribuent, chacun à sa manière au développement du modèle des fonctions lexicales.

1.4.5.1 Le *DEC*

La Lexicologie Explicative et Combinatoire a trouvé sa première application concrète dans *le Dictionnaire Explicatif et Combinatoire du Français Contemporain* (plus couramment

appelé *DEC*) (Mel'čuk *et al.* 1984, 1988, 1992, 1999), dictionnaire papier, dans lequel chaque lexème (entendu au sens d'une acception de mot) reçoit une analyse aussi complète que possible. Ce modèle dictionnairique a deux particularités, il est *explicatif* et *combinatoire*[12] où l'adjectif *explicatif* signifie que tout élément lexical est accompagné d'une explication sémantique formelle tandis que *combinatoire* indique l'importance particulière portée à la représentation rigoureuse et exhaustive de la combinatoire lexicale. Le *DEC* prend ainsi en compte, de façon systématique, deux axes : l'axe syntagmatique (enchaînement des unités lexicales dans le texte) et l'axe paradigmatique (oppositions sémantiques et sélection sémantique d'unités) (Mel'čuk *et al.* 1995 : 10). Les lexies sont étudiées d'une façon multilatérale en prenant en compte les dimensions sémantique, syntaxique et lexico-combinatoire.

Un article comprend (entre autres) : une définition lexicographique (où le défini est une expression à variable, variables qui représentant les actants), un tableau de régime qui décrit la combinatoire syntaxique de la lexie et fait apparaitre ses actants ainsi que la combinatoire lexicale et les dérivés sémantiques de la lexie modélisés au moyen des fonctions lexicales. Il est important de souligner que la définition joue un rôle primordial car la plupart des propriétés de comportements d'une lexie sont sous-tendues ou carrément déterminées par son sens dénotationnel (Mel'čuk *et al.* 1995 : 73). Le *DEC* fournit donc à l'usager toutes les informations nécessaires pour utiliser une lexie correctement dans n'importe quel contexte.

1.4.5.2 Le *DiCo*[13] (et son interface *DiCouèbe*)

Le *DiCo* est une base de données lexicales développée par Igor Mel'čuk et Alain Polguère à l'Observatoire de Linguistique Sens-Texte de l'Université de Montréal. L'originalité du *DiCo* réside dans son système d'étiquettes sémantiques. En effet, le dictionnaire ne propose pas de définitions classiques. En revanche, il fournit une caractérisation sémantique des unités lexicales au moyen de brèves formules appelées étiquettes sémantiques (Polguère 2003, Polguère 2011).

[12] Ces particularités sont annoncées dans les titres de deux ouvrages de référence : *Dictionnaire explicatif et combinatoire du français contemporain* et sa théorisation *L'introduction à la lexicologie explicative et combinatoire*.

[13] La Base DiCo (ainsi que la documentation concernant le projet) est disponible à l'adresse suivante : http://olst.ling.umontreal.ca/dicouebe/index.php

Comme nous pouvons le lire dans la documentation concernant le projet, du point de vue formel, une étiquette sémantique est un mot ou groupe de mots ayant une signification très générale et pouvant être utilisé comme le terme central dans la définition du mot-vedette. Chaque étiquette correspond donc au genre prochain d'une définition analytique et établit l'appartenance d'une lexie à une classe sémantique. L'ensemble des étiquettes est organisé de façon hiérarchique, des étiquettes les plus générales aux étiquettes les plus pointues. Comme le souligne Polguère (4 :2011), la notion d'étiquette sémantique est purement linguistique, en ce sens qu'il n'est pas nécessaire d'avoir recours à des notions autres que celles de la linguistique pour la construire : « […], les étiquettes sémantiques ne se conçoivent pas relativement au domaine des concepts, mais relativement au système de la langue : signes et règles linguistiques ».

1.4.5.3 Le *LAF*

Le *LAF* (acronyme de *Lexique actif du français* 2007) est un dictionnaire grand public entièrement produit par extraction et formatage des données de la Base *DiCo* (Mel'čuk et Polguère 2006). Conçu comme un dictionnaire d'encodage, il est destiné à être utilisé dans l'apprentissage et dans l'enseignement de la langue française. Comme le soulignent Mel'čuk et Polguère (2007 : 14), le *LAF* se focalise sur une description en profondeur de deux phénomènes lexicaux, à savoir les dérivations sémantiques et les collocations et présente le lexique du français comme un réseau d'unités interconnectées par une multitude de relations lexico-sémantiques. Les auteurs proposent de modéliser les deux phénomènes au moyen d'un même outil descriptif, à savoir des liens lexicaux orientés. En effet, il s'agit de pointeurs qui spécifient les dérivés sémantiques et les collocatifs de chaque lexie décrites dans le dictionnaire (*ibid. :* 20). Les formules décrivant les liens lexicaux sont fondées sur un métalangage compréhensible pour tous – une sorte de français standardisé (*ibid.* : 22).

Ex. ABOIEMENT :
Intense fort < furieux ; féroce
Peu intense *faible // jappement*

Comme nous pouvons le constater les liens lexicaux orientés constituent une version vulgarisée des fonctions lexicales.

1.4.5.4 Le *RLF*

Le *RLF* (acronyme de *Réseau Lexical du Français*) est un projet mené actuellement par l'équipe lexicographique de l'ATILF de l'Université Nancy 2, sous la direction d'Alain Polguère. C'est une ressource lexicale qui propose de modéliser le lexique non pas par un texte dictionnarique, mais sous forme d'un réseau d'unités lexicales. Comme le soulignent Polguère et Sikora (2013 : 3), le *RLF* se situe dans la lignée des travaux menés sur les dictionnaires dits *explicatifs* et *combinatoires* mais se distingue radicalement de ces derniers par sa structure et la méthodologie lexicographique mise en œuvre pour le construire.

En effet, le *RLF* n'est pas un dictionnaire, mais un réseau lexical dont les nœuds sont les lexies de langue et dont les arcs correspondent aux liens paradigmatiques et syntagmatiques connectant ces lexies (*ibid.* : 4). D'après les auteurs, un graphe lexical de type *petit-monde* (Polguère 2014 : 84) présente l'avantage de se rapprocher plus de la structure du lexique mental (en référence à *mental lexicon* d'Aitchison 2003), que « la *plate* page de texte » : « *On doit donc pouvoir effectuer sur une telle structure des opérations de navigation, d'inférence analogique, etc., qui s'approchent de celles effectuées de façon quasi instantanée par le locuteur lorsqu'il accède à ses connaissances lexicales.* » (Polguère et Sikora 2013 : 4). Cependant, il convient de souligner que les systèmes lexicaux du *RLF* sont des modèles non ontologiques, leur ossature est tissée à partir du système des fonctions lexicales.

Après ce rapide tour d'horizon des ressources lexicales basées sur la LEC, nous proposons, dans les pages qui suivent, d'analyser quelques projets terminographiques qui ont intégré le formalisme des FL dans leur description des unités terminologiques.

Chapitre 2. Lexicographie et terminologie : disciplines sœurs ou pratiques distinctes ?[14] – propositions d'application des FL à la terminologie

Comme l'indique son titre (que nous avons emprunté à l'article introducteur à l'ouvrage collectif publié sous la direction de Marie Claude L'Homme et Sylvie Vandaele 2007), ce chapitre sera consacré à la présentation de quelques modèles proposés en terminologie qui s'inspirent largement de l'approche sémantico-lexicale et plus particulièrement de la Lexicologie Explicative et Combinatoire. Cela nous conduira à aborder la question du statut de la terminologie moderne dans une perspective linguistique. Nous mènerons une réflexion sur la compatibilité des méthodes lexicologiques et terminologiques et sur ses conséquences méthodologiques.

En effet, comme nous l'avons déjà annoncé en introduction, depuis un certain temps, on observe un réel renouveau dans les pratiques terminologiques. On considère que les ressources terminographiques modernes devraient se fixer comme objectif de faire une description globale de la langue de spécialité. Comme le soulignent L'Homme et Vandaele (2007), cette volonté d'enrichir le contenu des dictionnaires spécialisés (ou plutôt des dictionnaires de langues spécialisées ou de langues de spécialité), se généralise dans le milieu, même si les méthodologies adoptées restent différentes. Selon les auteurs (*ibid. :* 9), de plus en plus de dictionnaires spécialisés intègrent une composante linguistique et tiennent compte d'éléments descriptifs qui relevaient autrefois du strict domaine lexicographique. On assiste donc à un rapprochement entre la lexicographie et la terminologie, « *deux disciplines qui partagent un objectif commun, à savoir compiler des dictionnaires* » (*ibid.* 3). Cette tendance se manifeste surtout par le fait qu'aussi bien les lexicographes que les terminologues reconnaissent l'importance de rendre compte d'un plus grand nombre de relations syntagmatiques et paradigmatiques entre termes et mots (*ibid. :* 9).

[14] C'est le titre de l'article d'introduction à l'ouvrage collectif dirigé M-C. L'Homme et Sylvie Vandaele (2007) portant sur les points de convergence et de divergence des deux disciplines.

Heid et Freibott (1991) sont parmi les premiers à avoir repensé le modèle classique des bases de données terminologiques. L'originalité de leurs travaux réside avant tout dans la façon de modéliser de l'information collocationnelle. En effet, ils proposent de traiter les collocations comme des entrées à part entière en les distinguant explicitement d'autres phénomènes. Un autre ouvrage de référence cité dans la majorité des travaux qui traitent de combinatoire spécialisée est le Lexique de cooccurrents de la Bourse et de la conjoncture économique de Cohen (1986). L'auteure propose de classer les collocations appelées ici cooccurrences lexicales en fonction des phases du cycle économique. Il s'agit du premier dictionnaire spécialisé qui organise les collocations selon le sens véhiculé par les collocatifs. Un autre projet terminographique, plus récent, où la description des phénomènes collocationnels occupe une place prépondérante est le DAFA (Binon *et al.* 2000). Ses concepteurs ont adopté trois types de critères afin de décrire et de classer les séries collocationnelles, notamment des critères morpho-syntaxiques, sémantiques et pragmatiques. Dans les pages qui suivent, nous proposons de regarder de plus près quelques ressources termonographiques qui proposent une description des relations des unités terminologiques basée sur les principes de la LEC. Nous nous intéresserons plus particulièrement aux projets des terminologues de l'OLST de l'Université de Montréal menés sous la direction de Marie-Claude L'Homme ainis qu'aux travaux de Jeanne Dancette.

2.1 Les travaux précurseurs de William Frawley

L'article de Wiliam Frawley publié dans International Journal of Lexicography en 1988 est considéré par la communauté des terminologues comme un des premiers plaidoyers en faveur d'un changement radical des pratiques terminographiques. En même temps, il s'agit d'une des premières propositions d'application des fonctions lexicales aux langues de spécialités. Dans son article, Frawley démontre qu'il est possible (et même souhaitable) de construire un dictionnaire de spécialité à l'aide des principes de la Lexicologie Explicative et Combinatoire. Selon l'auteur, les dictionnaires spécialisés conventionnels ne répondent pas aux besoins de communication des spécialistes car ils ne fournissent pas suffisamment d'informations sur l'utilisation réelle des termes qu'ils décrivent. S'inspirant largement des

modèles existant pour la langue générale, ils en héritent d'un problème formel fondamental : la définition.

En effet, Frawley (1988 : 192) reproche aux dictionnaires de la langue générale de proposer des définitions vagues :

« […] definition remains vague: i.e., non-specific about the contextualization of the unambiguous interpretation ».

Selon lui, le problème serait lié à la forme de la définition. Les lexicographes concentrent leurs efforts sur la désambiguïsation du sens en négligeant les informations relatives au fonctionnement des unités lexicales en contexte. Le caractère vague correspond donc à l'absence de données concernant ce qu'il nomme *contextualization*, c'est-à-dire la mise en discours des unités lexicales, et non pas à la clarté même de la définition. Ainsi, il remet en cause les pratiques traditionnelles et presque exclusives dans le milieu lexicographe, notamment celle de la définition par genre prochain et différences spécifiques basée sur la paraphrase et celle de la définition fondée sur la synonymie. Même si elles sont profondément ancrées dans les habitudes lexicographiques, elles s'avèrent inadéquates du moment qu'elles ne fournissent pas de renseignements sur les propriétés syntaxiques et sémantiques spécifiques à l'unité lexicale donnée, les renseignements qui déterminent son utilisation en discours. Comme nous l'avons annoncé plus haut, il en va de même pour les outils terminographiques. À titre d'exemple, Frawley cite la définition du terme *endostyle* proposée par Leftwich dans son *A student's dictionary of zoology* (1963 : 82) :

> Endostyle: a glandular ciliated groove running along the floor of the pharynx in Amphioxus, in some Tunicates, and in the larvae of lampreys (Cyclostomata). It produces threads of mucus to which food particles adhere and which are passed dorsalwards round the pharynx and backwards into the gullet by ciliary action. From observations of the development of the lamprey it can be shown that the thyroid gland has evolved from the endostyle.

Figure 11. Définition du terme *endostyle* proposée par Leftwich et citée dans (Frawley 1988: 202)

Malgré la richesse des informations d'ordre encyclopédique et conceptuel permettant d'identifier correctement le référent de *endostyle*, la définition ne contient pas d'éléments renseignant sur la cooccurrence syntaxique, sémantique et lexicale du terme. L'usager n'est pas en mesure d'en déduire la manière d'utiliser le terme correctement dans le discours. Cette définition, malgré sa clarté et son exactitude, reste vague.

« The encyclopedic information given above for endostyle correctly picks out a specific referent – i.e., prodives an EXTENSIONAL CONTEXT – but it is still unclear how the word itself is used in zoological discourse. That is, the INTENSIONAL CONTEXT remains vague. »
(Frawley 1988: 203)

Selon Frawley, le problème du manque d'informations contextuelles peut être résolu en adoptant les règles de la Lexicologie Explicative et Combinatoire. Il propose donc de remanier la définition de *endostyle* de Leftwich en s'appuyant sur les principes de rédaction des articles du Dictionnaire Explicatif et Combinatoire de Mel'čuk (Mel'čuk *et al.* 1984 - 1999), notamment en adoptant une définition formelle sous forme prépositionnelle.

Endostyle:
A Tunicate's, X's, gland for mucus, Y.
Lexical functions
$Func_0$ (endostyle) = run
$Func_2$ (endostyle) = produce
Gener (endostyle) = groove
S_{loc} (endostyle) = pharynx
S_1 (endostyle) = lamprey, Amphioxus, Tunicate
S_2 (endostyle) = mucus
$Qual_0$ (endostyle) = ciliated, glandular

Figure 12. Définition du terme *endostyle* sous forme propositionnelle proposée par Frawley (1988: 202)

La définition spécifie deux participants X et Y où X correspond au propriétaire de l'endostyle (les Tuniciers) et Y au mucus sécrété par l'endostyle. Ce type de définition

contribue à l'élucidation du sens. Les fonctions lexicales qui accompagnent la définition fournissent des renseignements utiles au niveau de la mise en discours. Elles permettent notamment de décrire l'univers lexical du terme, en l'occurrence ses relations actancielles (S_1 et S_1), circonstancielles (S_{loc}), taxonomiques (Gener) et collocationnelles ($Func_0$, $Func_2$, $Qual_0$).Etant donné que Frawley se base uniquement sur la définition proposée par Leftwich, ces informations de nature sémantique restent limitées. Néanmoins, une enquête menée auprès des zoologistes permettrait de compléter la liste des relations typiques que le terme *endose* entretient avec les autres unités lexicales en enrichissant cette description. Il convient de souligner que pour Frawley, il est important de faire appel aux spécialistes afin de recueillir des informations sur le sens spécialisé. Mais, leur rôle d'informateurs devrait s'arrêter là. Le lexicographe est le seul expert en matière de rédaction des dictionnaires car il dispose d'outils formels de représentation du sens et de la combinatoire du terme.

Evoquant le rôle du lexicographe (et non pas du terminographe), Frawley situe les pratiques terminographiques dans la dimension linguistique et préconise l'adoption des principes de la LEC. Pour lui, la possibilité d'une description systématique, rigoureuse et exhaustive qu'offre la méthode mel'čukienne présente un énorme avantage. Elle permet de spécifier de façon succincte et méthodique l'univers syntaxique (les relations grammaticales), sémantique (les informations sémantiques) et lexical (les relations lexicales sur les axes paradigmatique et syntagmatique) du terme en éliminant le caractère vague des définitions traditionnelles.

« (…) the Explanatory Combinatorial Dictionary (ECD), a type of dictionary which addresses directly the problem of vagueness by specifying, formulaically, the things that other dictionaries leave unsaid. » (Frawley 1988: 196)

Un dictionnaire de spécialité conçu à l'instar du *DEC* devient donc un outil de compréhension et d'utilisation des termes très efficace. Même si les travaux de Frawley n'ont jamais été exploités à plus grande échelle, ils ont ouvert la voie à d'autres propositions de modélisation des données terminologiques à l'aide des FL. Dans un ouvrage collectif préparé à l'occasion du 70^{e} anniversaire d'Igor Mel'čuk qui offre une revue transversale des recherches actuelles menées dans le cadre de la TST, Marie Claude L'Homme (2007 : 172) évoque d'autres terminologues qui ont eu recours aux FL. Elle rappelle les travaux d'Yves Gentilhomme (1995) dans le domaine des mathématiques et ceux d'Élisabeth Marcel (2000)

en biologie. On peut également évoquer les auteurs tels que Faber et Sáncher (2001)) qui proposent d'exploiter le système des FL comme une méthode complémentaire de description des unités terminologiques. Pour notre part, nous proposons de présenter succinctement des travaux de Myriam Mortchev-Bouveret pour passer ensuite à une description plus détaillée des projets dictionnairiques developpés d'un côté par l'équipe de l'OLST[15] sous la direction de Marie Claude L'Homme et de l'autre côté par Jeanne Dancette.

2.2 Convertir un dictionnaire spécialisé en un dictionnaire de langue spécialisée – les propositions de Myriam Mortchev-Bouveret

Dans la contribution à l'ouvrage collectif « Lexicographie et terminologie : compatibilité des modèles et des méthodes », dirigé M-C L'Homme et Sylvie Vandaele, Myriam Mortchev-Bouveret (2007 : 293-317) présente les premiers résultats de ses recherches concernant l'encodage des termes du domaine des bio-industries au moyens des fonctions lexicales. Ce travail porte avant tout sur la modélisation des relations paradigmatiques. Il faut rappeller que Bouveret a déjà étudié la possibilité d'adoption des FL à des fins terminologiques dans le cadre d'un large projet dirigé par Marie Claude L'Homme (L'Homme 2002, 20005a, 2005b,),dont nous parlerons plus en détail dans les pages qui suivent. À cette époque, elle a mené des travaux sur la systématisation, au moyen des FL, des séries derivationnelles des termes du domaine d'Internet (Jousse & Mortchev-Bouveret 2003). À présent, il s'agit d'une étude réalisée à partir des données provenant du dictionnaire Biolex, un dictionnaire spécialisé français-anglais-allemand rédigé par l'équipe de terminologie du laboratoire UPRESA 6065 de l'Université de Rouen (actuel laboratoire DYALANG). Il est important de souligner que les travaux sur Biolex s'inscrivent dans l'orientation socioterminologique développée (sous l'impulsion de Louis Guespin) par les terminologues de l'Université de Rouen (Gaudin 2003). Rappelons brièvement que le groupe de Rouen postule la prise en compte de la dimension sociale des termes, c'est-à-dire des influences du fondement culturel, idéologique et religieux des usagers. Pour les socioterminologues, les termes font l'objet de variations en fonction du contexte social dans lequel ils sont utilisés.

[15] Observatoire de linguistique Sens-Texte (OLST) de l'Université de Montréal

Quant au projet de Mortchev-Bouveret, son objectif est de systématiser et d'enrichir la description des termes issus de Biolex en s'appuyant sur la méthode de la LEC. Cependant, soulignons que le dictionnaire original comporte déjà (à côté de 5 rubriques traditionnelles comme *synonyme, hyponyme, hyperonyme, isonyme, antonyme*), des rubriques moins traditionnelles comme *agent typique, action typique, objet typique, application typique* (Mortchev-Bouveret 2007 : 293). Il s'agit des rubriques qui rendent compte des relations prédicatives du terme et permettent *« d'enregistrer des fait syntagmatiques significatifs et non plus seulement des faits paradigmatiques »* (Gaudin et Bouveret, 1996 : 352). Ainsi, l'*action typique* indique l'utilisation la plus fréquente de l'objet décrit, l'*agent typique* – ce qui réalise ou permet la réalisation de l'action typique, l'*objet typique* – ce qui subit l'action typique de l'objet décrit et finalement l'*application typique* renseigne sur les applications, les procédés dans lesquels les actions sont utilisées.

Comme le soulignent Gaudin et Bouveret (1996 : 352), les quatre rubriques sont inspirées des fiches proposées par Pierre Lerat pour les travaux du Centre de terminologie et de néologie (dont il était le fondateur). Afin de consigner des renseignements d'ordre syntagmatique, ce dernier s'est intéressé aux propriétés de prédicat privilégié et d'argument logique contigu dans une relation prédicative de premier ordre. Si la recherche du prédicat privilégié a donné lieu à la création de la rubrique *action typique*, la réflexion sur l'argument logique contigu (objet connexe) a permis de rassembler, dans la rubrique *objet*, toutes les notions qui sont associées le plus spontanément à un terme.

Pour leur part, les auteurs de Biolex ont décidé d'élargir les rubriques à deux autres relations : *agent typique, application typique,* déjà mentionnées ci-dessus. Les quatre rubriques permettent ainsi de situer le terme au sein du champ lexical auquel il appartient et dans le même temps de recenser les cooccurrents les plus fréquents, c'est-à-dire les éléments de discours les plus centraux, les plus courants:

« (…) les relations prédicatives permettent de mettre en lumière le sens le plus fréquemment produit dans une formation discursive.
Pourquoi parler de « formation discursive » ? Pour insister sur le fait qu'il s'agit de rendre compte d'un ensemble de discours tenus au sein d'une sphère d'activité, (…). »

Gaudin et Bouveret (1996 : 352)

C’est ainsi que pour le *filtre bactérien*, nous retrouvons quatre relations prédicatives :

Dictionnaire : Biolex
Entrée: filtre bactérien
Action typique : épuration
Objet typique : effluent urbain
Agent typique : micro - organismes
Application typique : dépollution

Figure 13. Exemple d’entrée dans Biolex, cité par Mortchev-Bouveret (2007 : 295)

Mortchev-Bouveret décide donc de conserver l’originalité d’une double description paradigmatique et syntagmatique du dictionnaire Biolex mais elle propose de systématiser la description de ces deux types de relations lexicales combinatoires (qui pour elle, sont étroitement dépendantes) au moyen des fonctions lexicales. Cette méthode permet de rendre compte du phénomène de la phraséologie considérée ici au sens large, c’est-à-dire englobant tout autant des expressions lexicales que des propriétés sémantiques et syntaxiques de combinatoire entre les unités. Il s’agit donc de décrire les contraintes de sélection aussi bien sur l’axe syntagmatique que paradigmatique en représentant les relations les plus récurrentes et significatives entre les unités terminologiques.

En effet, l’auteure propose de rassembler tous les dérivés sémantiques d’un terme dans des séries dérivationnelles. Si le terme d’entrée est un verbe, l’encodage des relations dérivationnelles sous forme de FL permet de regrouper un ensemble d’actants et de circonstants. La figure ci-dessous (Figure 14) présente trois exemples de telles séries dérivationnelles prédicatives. Nous pouvons constater que ce formalisme se prête très bien à la description linguistique des termes. Cependant, il est nécessaire de mettre l’accent sur quelques questions méthodologiques.

Entrée	**S_0**	**S_1**	**S_2**	**S_{res}**	**S_{instr}**	**S_{med}**	**S_{loc}**
Acidifier	acidification		Lait	Caillé		bactérie lactique	Industrie agroalimentaire

Broyer	Broyage			Broyé	broyeur à billes		Industrie agroalimentaire
Aromatiser	aromatisation1			aromatisation2		arôme	Industrie agroalimentaire

Figure 14. Exemples d'encodage des relations entre les termes du domaine des bioindustries au moyen des FL proposé par Mortchev-Bouveret (2007 : 295)

En effet, la relation d'agent (qui correspond à la rubrique *agent typique*) a été traitée, dans le dictionnaire Biolex, du point de vue référentiel, comme des cas d'origine du procès. En revanche, Mortchev-Bouveret envisage cette relation de point de vue linguistique en conservant à l'actant $\mathbf{S_1}$ son rôle d'actant réservé pour les agents animés du procès. Comme nous pouvons le constater dans les exemples évoqués ci-dessus et comme le souligne Mortchev-Bouveret (2006 :245), on compte peu d'agents animés dans le domaine des bio-industries (la colonne de l'ASyntP1 reste vide). En revanche, ce sont des circonstants tels que le moyen, l'instrument, le mode ou le locatif qui peuvent être amenés à remplir la fonction de sujet syntaxique. On parle alors de medium indirect qui peut être encodé à l'aide des FL $\mathbf{S_{med}}$, $\mathbf{S_{loc}}$, $\mathbf{S_{instr}}$, $\mathbf{S_{mod}}$.

Même si l'auteur situe son travail dans la démarche lexicographique, elle réfléchit à la compatibilité entre approche terminologique et description lexicale selon les principes de la LEC. Elle évoque le statut particulier de la terminologie, ce statut mixte que lui confère l'unité terminologique prise entre propriétés linguistiques et référentielles. Elle rappelle qu'il convient d'être extrement prudent dans ses choix théoriques et méthodologiques à cause des possibles incompatibilités générées par ces choix (Mortchev-Bouveret 2007 : 314). Si l'on considère le terme comme un signe linguistique qui renvoie à un concept situable en dehors de la langue (Depecker 2000 : 92), nous sommes obligés de prendre en compte aussi bien ses caractéristiques linguistiques que taxinomiques. Or la description au moyen des FL est basée essentiellement sur une décomposition sémantique. En effet, le modèle mel'čukien demande une analyse rigoureuse en termes de sèmes. Dans cette approche lexico-sémantique, on s'intéresse aux relations lexicales fondées sur l'héritage de propriétés sémantiques tandis que dans la tradition terminologique, les relations sont établies entre les concepts, le terme étant perçu au sein d'un système. Ainsi, Mortchev-Bouveret propose une représentation médiane

entre relations taxinomiques et composantes linguistiques qui permet de sortir de cette impasse à la fois théorique et méthodologique.

Entrée: bain-marie
Gener (bain-marie) = bain chauffant-refroidissant
Spec (bain-marie) = bain-marie à agitation magnétique
Contr (bain-marie) = bain à lit fluidisé

Entrée : bain marie à agitation magnétique
Gener (bain marie à agitation magnétique) = bain marie
Syn (bain marie à agitation magnétique) = bain-marie agité

Figure 15. Exemples d'encodage des relations paradigmatiques au moyen des FL proposé par Mortchev-Bouveret (2007 : 303)

Comme nous pouvons le constater dans les exemples cités ci-dessus, les fonctions lexicales **Gener** et **Spec** (cette dernière a été proposée par Grimes car il n'existe pas de FL pour l'hyponymie dans la LEC) ont été réservées aux liens hiérarchisant les concepts à l'intérieur du système. Il s'agit des relations d'hyperonymie et d'hyponymie, c'est-à-dire de relations fondamentales en terminologie. Ces fonctions ont été déterminées en termes de référents, c'est-à-dire prenant en compte le système conceptuel auquel appartient le terme en question. En revanche, les fonctions **Contr** et **Syn** (ainsi que les FL **Syn**$_{\cap}$ et **Anti**) ont été choisies pour décrire les différents types de synonymie et d'antonymie et permettent d'affiner les liens à l'intérieur des séries de termes. Contrairement à l'hyperonymie et à l'hyponymie (qui ne répondent pas à l'analyse sémantique), ces relations ont été définies en termes de sèmes demandant une décomposition sémantique stricte et rigoureuse.

D'après Kleiber et Tamba (cités par Mortchev-Bouveret, 2006 : 237), les relations d'hyperonymie et d'hyponymie sont des relations d'inclusion issues de la logique qui caractérisent une relation mixte, à la fois référentielle et linguistique. Mortchev-Bouveret souligne (2007 : 309), qu'il est donc difficile, en terminologie, d'hériter de propriétés strictement sur la base d'une description lexicographique en sèmes (comme le demande la méthode mel'čukienne). Afin de mettre en évidence les liens conceptuels, le terminologue est obligé d'avoir recours à la description en termes de référents. En revanche, l'adoption des

fonctions lexicales permet d'apporter à la rédaction terminologique davantage de précisions linguistiques. Ainsi, l'auteur propose de concilier les approches conceptuelle et lexicographique afin d'enrichir la description des termes dans les dictionnaires, Selon elle, cette représentation intermédiaire, permettant de dégager à la fois des relations hiérarchiques de taxinomies et des relations lexico-sémantiques, paraît tenable pour la rédaction d'un dictionnaire terminologique et constitue une piste théorique et méthodologique.

2.3 Le terme envisagé comme une unité lexicale spécialisée – les travaux de Marie Claude L'Homme et ses collaborateurs de l'OLST

Les auteurs évoqués jusqu'à présent (Frawley 1988, Gentilhomme 1995, Mortchev-Bouveret 2006, 2007), préconisent l'encodage des termes au moyen des FL en vue d'enrichissement et de diversification des descriptions terminographiques sans toutefois l'avoir matérialisé à travers des projets concrets réalisés à grande échelle. Les travaux dont nous allons parler dans cette partie, représentent, au contraire, des entreprises terminographiques à proprement parler. Il s'agit notamment du *DiCoInfo*[16] et du *DiCoEnviro*[17], dictionnaires spécialisés en ligne, élaborés sous la direction de Marie Claude L'Homme au sein de l'Observatoire de linguistique Sens Texte (OLST) de l'Université de Montréal. Le *DiCoInfo* est consacré à la description de la langue de l'informatique et de l'Internet tandis que le *DiCoEnviro* (un projet plus récent) est consacré au domaine de l'environnement. Les données sont disponibles en français, anglais, espagnol et portugais (la version portugaise ne concerne que le *DiCoEnviro*). Nous voudrions attirer l'attention sur le fait que les deux dictionnaires sont toujours en construction et l'état d'avancement de la rédaction varie d'un article à l'autre.

[16] *DiCoInfo*. Dictionnaire fondamentale de l'informatique et de l'Internet, est consultable à l'adresse suivante : http://olst.ling.umontreal.ca/cgi-bin/*DiCoInfo*/search.cgi.

[17] DiCoEnviro. Dictionnaire fondamental de l'environnement, est consultable à l'adresse suivante : http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi.

2.3.1 Le *DiCoInfo* et le *DiCoEnviro* – premières bases de données terminographiques conçues selon les principes de la LEC

D'après la documentation téléchargée sur les sites des *DiCoInfo* et *DiCoEnviro*[18], les dictionnaires s'adressent à tout utilisateur qui souhaite mieux connaître d'un côté la langue de l'informatique et de l'Internet, de l'autre côté, celle de l'environnement. Ils sont donc conçus comme une aide à la rédaction ou à la traduction de textes techniques des domaines donnés. Les deux projets visent à rendre compte des termes fondamentaux, c'est-à-dire des unités lexicales susceptibles de se trouver dans de nombreux textes des deux domaines. Les auteurs écartent ainsi les unités lexicales étroitement attachées à une spécialisation. Il en va de même pour les unités lexicales générales, n'appartenant pas aux domaines donnés, même si elles sont récurrentes dans le corpus. En effet, l'objectif de ces deux dictionnaires est de fournir une information riche sur le fonctionnement des termes dans l'environnement linguistique en donnant un portrait aussi complet que possible de leurs propriétés lexico-sémantiques. Cependant les articles ne proposent pas de longs développements encyclopédiques sur des concepts complexes contrairement aux ressources terminographiques conventionnelles.

Comme nous l'avons annoncé tout au début de cette partie, le *DiCoInfo* et le *DiCoEnviro* ont été conçus selon les principes de la Lexicologie Explicative et Combinatoire (Mel'čuk *et al.* 1995) et exploitent le modèle des fonctions lexicales. Il est nécessaire de souligner qu'à notre conaissance, le *DiCoInfo* constitue une des deux premières bases de données terminographiques qui ont explicitement recouru à ce modèle[19]. Malgré de nombreuses modifications et adaptations méthodologiques que les deux projets de l'OLST (et surtout le *DiCoInfo*) ont subies depuis le début des travaux (qui datent, pour ce dernier de 2002), leur ancrage théorique reste toujours la Théorie Sens-Texte (Mel'čuk 1997). Cependant, comme le souligne Marie Claude L'Homme (2005a : 1122), le choix de la LEC comme cadre a eu d'importantes conséquences méthodologiques et descriptives.

[18] http://olst.ling.umontreal.ca/*DiCoInfo*/manuel-*DiCoInfo*.pdf et http://olst.ling.umontreal.ca/dicoenviro/manuel-DiCoEnviro.pdf consulté le 09.03.2014

[19] La première base de données qui a fait appel aux FL est le *DAD* (Dancette 2006), que nous présenterons en détail dans la partie suivante de ce travail.

En effet, en adoptant la démarche résolument lexico-sémantique, les terminologues de l'OLST décident d'envisager le terme comme une unité lexicale dont le sens peut être associé à un champ disciplinaire préalablement délimité (L'Homme 2005b : 142).

« Les termes sont des unités lexicales particulières. Ils se distinguent des ULG [unités lexicales générales] – les unités lexicales non marquées – ***uniquement*** *par le fait que leur maîtrise linguistique est associée à la maîtrise d'un domaine de connaissance donné (scientifique, technique, etc.). »*

(L'Homme et Polguère 2008)

Ainsi, les auteurs se détachent radicalement de l'optique conceptuelle selon laquelle le terme est considéré comme une expression linguistique de l'organisation des connaissances dans un domaine. Selon L'Homme (2005a : 1123), la description qui présuppose une structuration des connaissances ne peut être réalisée avec succès que par un spécialiste du domaine. L'Homme considère que la démarche terminologique traditionelle se focalisant sur l'importance du système conceptuel et envisageant le terme comme étiquette de concept, reproduit la démarche adoptée par un spécialiste d'un domaine : donner des noms à des concepts (éventuellement les normaliser), délimiter des concepts les uns par rapport aux autres. Or, le terminologue avec sa formation du linguiste ne peut pas prétendre avoir les connaissances d'un expert. Ainsi, il ne devrait pas rivaliser avec ce dernier dans la description du domaine. En revanche, son rôle, le rôle du spécialiste de la langue est de se consacrer à des analyses faisant appel à des connaissances en linguistique et en traitement des corpus.

La perspective adoptée aura d'importantes conséquences sur les descriptions proposées. Optant pour l'approche lexico-sémantique, les terminologues de l'OLST s'intéressent plutôt au sens que revêtent les formes linguistique qu'à la place des concepts (que ces formes linguistiques dénotent) dans un système. Ils appréhendent les termes dans leur fonctionnement linguistique et mènent leurs descriptions en se basant sur des observations faites sur un ensemble des données linguistiques. Comme le remarque L'Homme (2005a : 1122), cela permet de retenir des éléments descriptifs qui échappent à une perspective strictement conceptuelle comme les relations lexicales (y compris les relations collocationnelles) ou les diverses parties du discours (et pas seulement celle du nom, largement représentée dans les ressources terminographiques traditionnelles). Et même, si les descriptions proposées ne peuvent pas prétendre à un modèle structurant des connaissances au

sens de la démarche terminologique traditionnelle, elles permettent de mettre au jour la structure lexicale observable à l'intérieur d'un domaine spécialisé. Cette structure est mise en évidence explicitement par l'énumération et l'explication de la multitude des liens linguistiques existant entre les termes des domaines en question. Et c'est là où réside l'originalité des dictionnaires.

En effet, basés sur l'observation du fonctionnement des termes dans le discours, les articles permettent d'accéder à des renseignements sur la structure actancielle de chaque terme ainsi que sur les liens paradigmatiques (synonymie, antonymie, relations dérivationnelles, etc.), ou sur les liens syntagmatiques (c'est-à-dire les collocatifs), qu'un terme partage avec d'autres termes du domaine. La liste des entrées retenues doit ainsi refléter le plus fidélement possible le réseau lexical du domaine donné. Puisqu'il s'agit des unités lexicales (L'Homme 2005a), les termes représentent chacune des quatres parties du discours majeures (noms, verbes, adjectifs, adverbes) ou bien ils correspondent à des structures plus complexes : syntagmes nominaux, adjectivaux, verbaux et adverbiaux. En effet, les unités entrant dans une relation paradigmatique ou syntagmatique avec un terme préalablement retenu doivent aussi faire l'objet d'une description. Comme l'explique L'Homme (2005b : 143), les adjectifs ou les verbes qui se combinent de façon privilégiée avec les termes nominaux selectionnés feront l'objet d'une entrée au même titre que les termes eux-mêmes. De la même façon seront décrits les actants de termes prédicatifs retenus. De plus, comme le soulignent les auteurs[20], les dictionnaires ne se contentent pas d'énumérer les termes apparentés, ils proposent une explication pour chaque lien repéré. Cela fournit aux utilisateurs une variété d'expressions précises, de combinaisons adéquates et de formulations appropriées au domaine donné.

En pratique, l'application des principes de la LEC se traduit de la manière suivante :

1. Chaque article correspond à une acception spécialisée. Ainsi les unités comme *adresse, page, formater, code, programmer* (dans le *DiCoInfo*), ou *circulation, gaz, stocker, développement* (dans le *DiCoEnviro*), donnent lieu à plus d'une entrée. Par exemple, le vocable *code* fait l'objet de 3 articles différents : $code_1$, (norme d'encodage de caractères), $code_2$ (ensemble d'instructions écrites en langage

[20] http://olst.ling.umontreal.ca/*DiCoInfo*/manuel-*DiCoInfo*.pdf et http://olst.ling.umontreal.ca/dicoenviro/manuel-DiCoEnviro.pdf

constituant un programme)$_1$, code$_3$ (information concernant le processus transmise à l'utilisateur d'un programme)$_2$. Chaque entrée est accompagée d'un numéro d'acception ($_1$, $_2$, $_3$).

2. L'opération de séparation des acceptions, c'est-à-dire le découpage du vocable en lexies se réalise à partir d'un certain nombre de tests contextuels et en fonction des relations sémantiques que les termes entretiennent avec d'autres termes. En effet, les sens des acceptions sont distingués au moyen des critères lexico-sémantiques proposés par la LEC (Mel'čuk *et al.* 1995 : 61-71). Comme le souligne L'Homme (2005a : 1127), l'application des critères tels que la cooccurrence compatible ou différentielle, la substitution par un synonyme, la dérivation différentielle ou morphologique permettent de déterminer et confirmer si l'on a affaire au même sens ou à des sens différents, chacun donnant lieu à une description différente.

3. La structure actancielle joue un rôle central dans la description du sens des unités terminologiques. Comme c'est le cas dans le *DEC* (Mel'čuk *et al.* 1984 - 1999), on montre le nombre et la nature des actants sémantiques et la position respective des actants par rapport au terme. Cependant, comme le souligne L'Homme (L'Homme *et al.* 2011 : 215), le modèle de description de la structure actancielle des termes dans le *DiCoInfo* et le *DiCoEnviro* diffère de celui proposé dans le *DEC*, les actants étant représentés au moyen d'un système d'étiquettes servant à décrire leur rôle par rapport au terme décrit.

4. Afin de rendre compte de la structure lexicale des domaines, l'accent est mis sur la description de l'ensemble des termes avec lesquels le terme en entrée entretient une relation de nature paradigmatique ou de nature syntagmatique. Comme le remarque L'Homme (2005b : 143), la liste des relations sémantiques et leur modèle de formalisation s'inspirent largement des fonctions lexicales du *DEC* (Mel'čuk *et al.* 1984 - 1999). Elle ajoute tout de même que dans le *DiCoInfo* (et le *DiCoEnviro*), l'explication des liens lexicaux est vulgarisée et s'aligne sur les explications données dans la version informatisée du *DEC*, à savoir le *DiCo* (Polguère 2003).

Il est nécessaire de souligner qu'à l'origine, le *DiCoInfo* a été conçu comme un outil de travail adressé à des linguistes, des lexicographes ou à des terminologues qui souhaitent

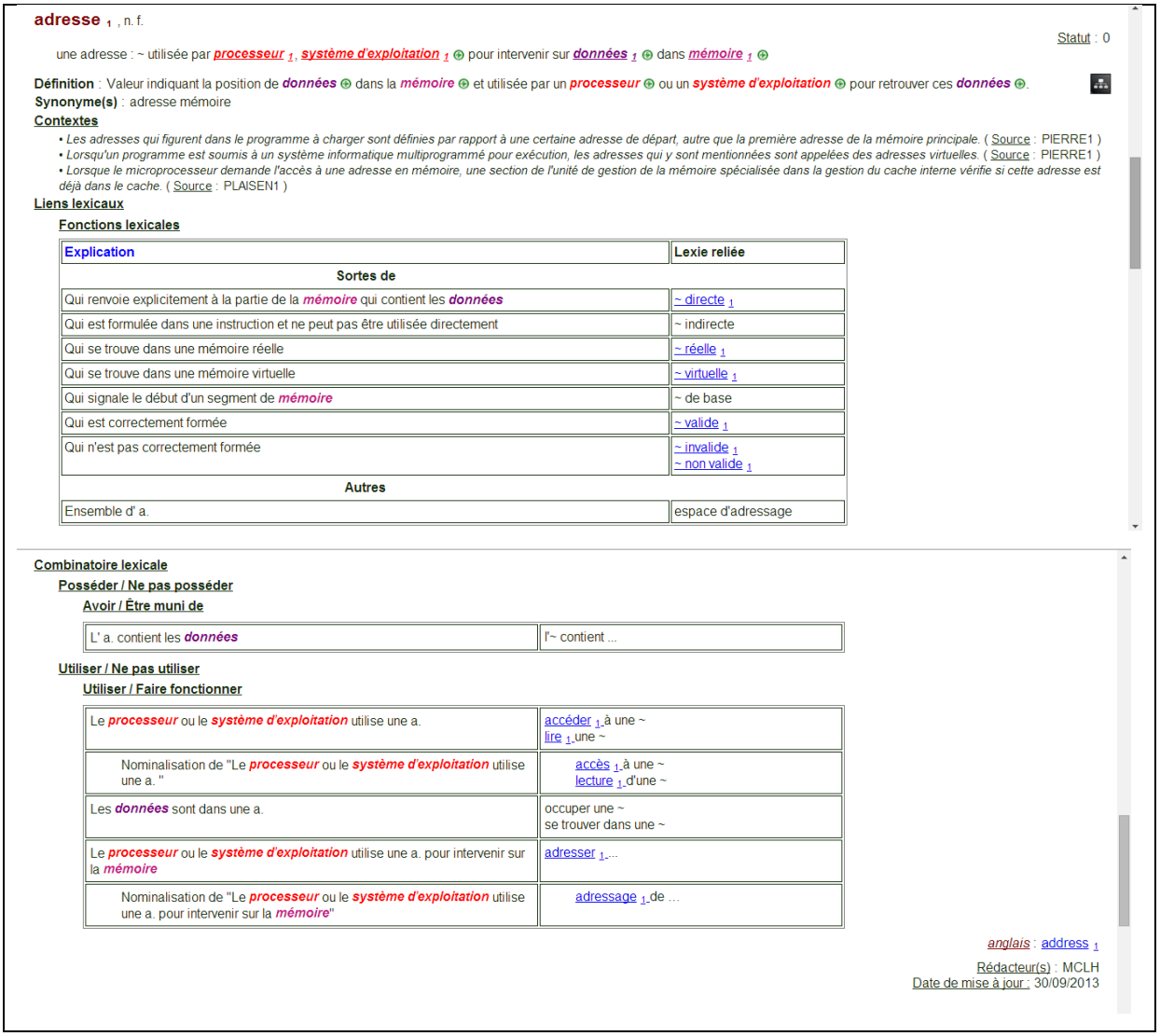
utiliser les données pour d'autres travaux de description des termes. Pourtant, il s'est vite avéré que le projet pouvait s'adapter aux besoins d'autres publics (étudiants, traducteurs, rédacteurs techniques). Par conséquent, une simplification du métalangage sémantique sophistiqué propre à la LEC s'est imposée. Les auteurs ont donc entamé une réflexion sur la modification du modèle abstrait et formel de Mel'čuk en une méthode de description plus accessible aux utilisateurs non familiers avec le cadre de la Théorie Sens-Texte. C'est pour cette raison qu'ils ont proposé deux niveaux d'encodage de certaines catégories de données : le premier, dépouillé du maximum de métalangage technique est plus approprié à une utilisation par un non-linguistite, le deuxième, avec tout son appareillage formel d'encodage propre à la LEC, adressé aux linguistes. Cependant, ce dernier n'est accessible que pour les articles de statut 0, c'est-à-dire dont la rédaction est déjà terminée. Rappellons que les deux dictionnaires sont toujours en construction et la notation des statuts (de 2 à 0), informe l'utilisateur de l'état d'avancement de la rédaction des articles. Lors de notre dernière consultation (le 22.08.2014), le *DiCoEnviro* comportait seulement un article en français de statut 0, celui du terme *réchauffer*$_{1a}$. L'état d'avancement de la rédaction du *DiCoInfo* est supérieur.

Dans les deux cas, les articles comportent une dizaine de rubriques. Les rubriques communes aux deux projets, quel que soit leur statut, sont les suivantes : *Entrée*, *Information grammaticale*, *Statut*, *Structure actancielle*, *Contexte* et deux rubriques adminisratives : *Rédacteurs* et *Date de mise à jour*. Comme le soulignent les auteurs[21], les rubriques *Synonyme(s)*, *Liens lexicaux* et *Informations complémentaires* n'apparaissent que si les fiches comportent des données correspondantes. La présence des liens permettant d'accéder aux équivalents espagnol ou anglais (et, dans le cas du *DiCoEnviro*, portugais), ainsi que celle de la rubrique *Définition* dépendent de l'état d'avancement des travaux (*Définition* figure seulement dans les fiches de statut 0). Une partie des rubriques sont affichées par défaut, les autres, sur demande. Comme le souligne L'Homme (2010 : 146), la rubrique consacrée à la déscription des relations syntagmatiques et paradigmatiques, *Liens lexicaux*, est généralement la plus complexe et la plus riche en informations.

De plus, le *DiCoInfo* propose une interface graphique nommé *DiCoInfo Visuel* permettant la visualisation des relations lexicales du terme en question sous la forme de graphes

[21] http://olst.ling.umontreal.ca/*DiCoInfo*/manuel-*DiCoInfo*.pdf et http://olst.ling.umontreal.ca/dicoenviro/manuel-DiCoEnviro.pdf

(réseaux). Le *DiCoInfo* Visuel est également une interface d'interrogation des relations lexicales décrites dans les entrées terminologiques du *DiCoInfo*, mais contrairement à ce dernier, elle décrit uniquement les liens paradigmatiques (taxonomiques, synonymes, contraires, dérivés, voisins et actants). La figure ci-dessous (Figure 16) présente, à titre d'exemple l'article *adresse*$_1$ du *DiCoInfo* dont l'ensembe des rubriques a été déplié.



adresse $_1$, n. f.

Statut : 0

une adresse : ~ utilisée par ***processeur*** $_1$, ***système d'exploitation*** $_1$ pour intervenir sur ***données*** $_1$ dans ***mémoire*** $_1$

Définition : Valeur indiquant la position de ***données*** dans la ***mémoire*** et utilisée par un ***processeur*** ou un ***système d'exploitation*** pour retrouver ces ***données***.
Synonyme(s) : adresse mémoire
Contextes

- *Les adresses qui figurent dans le programme à charger sont définies par rapport à une certaine adresse de départ, autre que la première adresse de la mémoire principale.* (Source : PIERRE1)
- *Lorsqu'un programme est soumis à un système informatique multiprogrammé pour exécution, les adresses qui y sont mentionnées sont appelées des adresses virtuelles.* (Source : PIERRE1)
- *Lorsque le microprocesseur demande l'accès à une adresse en mémoire, une section de l'unité de gestion de la mémoire spécialisée dans la gestion du cache interne vérifie si cette adresse est déjà dans le cache.* (Source : PLAISEN1)

Liens lexicaux

Fonctions lexicales

Explication	Lexie reliée
Sortes de	
Qui renvoie explicitement à la partie de la ***mémoire*** qui contient les ***données***	~ directe $_1$
Qui est formulée dans une instruction et ne peut pas être utilisée directement	~ indirecte
Qui se trouve dans une mémoire réelle	~ réelle $_1$
Qui se trouve dans une mémoire virtuelle	~ virtuelle $_1$
Qui signale le début d'un segment de ***mémoire***	~ de base
Qui est correctement formée	~ valide $_1$
Qui n'est pas correctement formée	~ invalide $_1$ ~ non valide $_1$
Autres	
Ensemble d' a.	espace d'adressage

Combinatoire lexicale

Posséder / Ne pas posséder

Avoir / Être muni de

L' a. contient les ***données***	l'~ contient ...

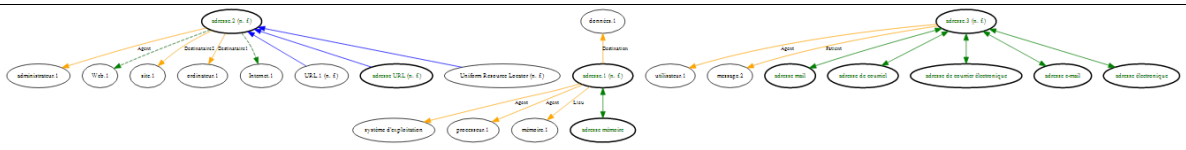
Utiliser / Ne pas utiliser

Utiliser / Faire fonctionner

Le ***processeur*** ou le ***système d'exploitation*** utilise une a.	accéder $_1$ à une ~ lire $_1$ une ~
Nominalisation de "Le ***processeur*** ou le ***système d'exploitation*** utilise une a. "	accès $_1$ à une ~ lecture $_1$ d'une ~
Les ***données*** sont dans une a.	occuper une ~ se trouver dans une ~
Le ***processeur*** ou le ***système d'exploitation*** utilise une a. pour intervenir sur la ***mémoire***	adresser $_1$ ...
Nominalisation de "Le ***processeur*** ou le ***système d'exploitation*** utilise une a. pour intervenir sur la ***mémoire***"	adressage $_1$ de ...

anglais : address $_1$

Rédacteur(s) : MCLH
Date de mise à jour : 30/09/2013



Liens vers les articles des lexies (9 relations)								
taxonimiques	**synonymes**	**quasi-synonymes**	**Agent**	**Destination**	**Lieu**	**Destinataire1**	**Destinataire2**	**Patient**
URL.1.fr Uniform Resource Locator.fr adresse URL.fr adresse.2.fr	adresse de courriel.fr adresse de courrier électronique.fr adresse e-mail.fr adresse mail.fr adresse mémoire.fr adresse électronique.fr adresse.1.fr adresse.3.fr	Internet.1.fr Web.1.fr adresse.2.fr	administrateur.1.fr adresse.1.fr adresse.2.fr adresse.3.fr processeur.1.fr système d'exploitation.fr utilisateur.1.fr	adresse.1.fr données.1.fr	adresse.1.fr mémoire.1.fr	adresse.2.fr ordinateur.1.fr	adresse.2.fr site.1.fr	adresse.3.fr message.2.fr

Figure 16. Exemples d'entrée dans le *DiCoInfo* (consulté le 15.03.2014)

2.3.2 Le *DiCoInfo* et le *DiCoEnviro* – traitement des caractéristiques lexico-sémantiques des unités terminologiques

Comme nous l'avons souligné plus haut, la présentation des unités terminologiques dans le *DiCoInfo* et le *DiCoEnviro* s'appuie, à l'instar du *DEC* (Mel'čuk *et al.* 1984 - 1999), sur la mise en évidence de leur structure actancielle. Il est nécessaire de souligner que depuis au moins une décennie, ce type d'informations a commencé à apparaître dans des descriptions lexicographiques du français, notamment sous forme de définition actancielle ou schémas actanciels. C'est le cas des dictionnaires d'apprentissage comme le *LAF*, le Dictionnaire du Français Usuel, et le DAFLES[22]. En ce qui concerne les dictionnaires spécialisés, nous pouvons citer un seul exemple, le DAFA (Binon *et al.* : 2000). Comme le remarquent Binon *et al.* (2006 : 86), le recours à ces modèles de représentation se révèle un outil très intéressant pour l'élargissement des connaissances lexicales. En effet, la structure actancielle réflète de façon formelle le sens de l'unité lexicale (et dans notre cas, celui de l'unité terminologique). Comme nous l'avons vu dans le chapitre précédent, dans la LEC, on parle du *sens dénotationnel, situationnel* ou *propositionnel : « La plupart des propriétés de comportement*

[22] L'Homme (2010 : 142) fait une comparaison de différents systèmes de représentation des actants proposés dans ces dictionnaires

d'une lexie sont sous-tendues ou même carrément déterminées par son sens dénotationnel. ». Mel'čuk *et al.* (1995 : 73). D'après les auteurs, il est donc impossible de traiter en profondeur les informations syntaxiques ou la cooccurence lexicale des unités avant d'avoir formulé leur sens, c'est-à-dire avant d'avoir identifié la structure actancielle des unités lexicales.

« *Le sens de la grande majorité des lexies ne peut être clairement compris, et donc décrit, sans prendre en compte les* participants *des situations que ces lexies désignent (…). Nous appellerons ces participants* actants sémantiques *(…).* »

(Mel'čuk & Polguère 2007 : 24)

Remarquons que si en sémantique logique on parle des arguments, en lexicologie, on utilise le terme *actant* pour désigner la même chose. En reprenant la citation de Mel'čuk et Polguère, nous voulons rappeler que les actants sont envisagés dans la LEC comme des participants obligatoires contribuant au sens d'unités lexicales de sens prédicatif, contrairement aux circonstants, qui entretiennent un lien avec l'unité lexicale mais ne contribuent pas à son sens. Et même si toutes les lexies d'une langue n'ont pas d'actants sémantiques, la plupart sont considérées comme des unités de nature prédicative (Mel'čuk *et al.* 1995 : 76). Nous reviendrons plus tard sur cette question.

Ainsi, dans le *DiCoInfo* et le *DiCoEnviro* (où les termes sont envisagés comme des unités lexicales à sens prédicatif), les actants jouent un rôle central dans la description terminographique. Tout comme dans le *DEC* (Mel'čuk *et al.* 1984 – 1999), les actants apparaissent sous forme propositionnelle (dans la rubrique *Structure actancielle*). Ensuite, ils sont explicitement indiqués dans les définitions des termes. Comme nous le verrons plus loin, les actants sont également mis en évidence directement dans les contextes extraits de corpus et annotés. Finalement, ils servent à la formalisation des liens lexicaux.

« *Explicit reference to the actants in dictionaries (…) is not only a method to specify the actantial structure of predicative units. It is also a means to account for their central contribution to the meaning of such lexical units."* (L'Homme 2010 : 149)

La structure actancielle peut donc être considérée comme le noyau de chaque article autour duquel s'articule le contenu des autres rubriques. Voici quelques exemples extraits du *DiCoInfo* et du *DiCoEnviro* :

AFFICHER$_1$ **: Agent** (réalisations possibles de l'agent : ***utilisateur$_1$, logiciel$_1$***) **~ Patient** (réalisations possible du patient : ***fichier$_1$, icône$_1$***) à **Destination** (réalisations possibles de la destination : ***écran$_2$***)

UN BALISAGE$_1$ **: ~** de **Patient** (réalisations possible du patient : ***code$_2$***, ***document$_1$***) avec **Instrument** (réalisations possibles de l'instrument : ***balise$_1$***) par **Agent** (réalisations possibles de l'agent : ***concepteur$_1$***)

ACHEMINER$_1$ **: Agent** (réalisations possible de l'agent : ***composant$_1$, routeur$_1$***) **~ Patient** (réalisations possible du patient : ***données$_1$***) de **Source** (réalisations possible de la source : ***composant$_1$, ordinateur$_1$***) vers **Destination** (réalisations possible de la destination : ***composant$_1$, ordinateur$_1$***)

ACCELERER$_{1b}$ **: Cause** (réalisations possible de la cause : ***changement$_{1a}$***) ou **Agent** (réalisations possible de l'agent : ***méthane$_1$***) **~ Patient** (réalisations possible du patient : ***réchauffement$_{1a}$***)

UN ACCUMULATEUR$_1$: ~ fonctionnant avec **Moyen** (réalisations possible du moyen : ***électricité$_1$***) utilisé par **Agent** (réalisations possible de l'agent : ***utilisateur$_1$***) pour intervenir sur **Destination** (réalisations possible de la destination : ***véhicule$_1$***)

Figure 17. Représentation de la structure actancielle des termes dans le *DiCoInfo* et le *DiCoEnviro* (consultés le 12.09.2014)

Comme nous pouvons le constater, les actants n'y sont pas représentés au moyen des variables *X*, *Y*, *Z* comme c'est le cas pour le *DEC* ou le LEC. En effet, les auteurs ont ici recours à un système d'étiquettes dont l'objectif est de typer très généralement le rôle des actants par rapport au terme décrit. Ainsi, le *DiCoInfo* et le *DiCoEnviro*n font respectivement appel à une douzaine d'étiquettes différentes (mis en caractère gras dans notre tableau), ayant chacune une signification spécifique. Comme le soulignent les auteurs (L'Homme 2010 : 148, *DiCoEnviro* Manuel : 12-13[23]), les étiquettes les plus couramment utilisées dans les deux

[23] http://olst.ling.umontreal.ca/dicoenviro/manuel-DiCoEnviro.pdf, téléchargé le 09.03.2014

dictionnaires sont : **Agent** (actant qui renvoie à l'origine de l'action exprimée par le terme ou à l'élément responsable de l'existence ou de l'utilisation d'une entité exprimée par le terme)**,** **Patient** (actant qui renvoie à l'entité subissant l'action exprimée par le terme ou qui désigne l'élément créé ou utilisé par un agent ou sur lequel l'agent intervient) **, Destination** (actant qui renvoie au but visé par une action menée par un agent ou au but visé par la fonction typique d'une entité, **Source** (actant qui renvoie à l'élément à partir duquel une activité est effectuée ou à l'élément à partir duquel la fonction typique d'une entité est réalisée), ou **Lieu** (actant qui renvoie à l'endroit où se déroule l'activité). Les autres étiquettes sont plus spécifiques à chaque dictionnaire et réflètent les relations propres au domaine donné, à savoir l'informatique ou l'environnement. Pour ce dernier, nous pouvons énumérer, entre autres, les rôles sémantiques tels que : **Cause, Moyen, Temps, Ménace, Étendue, Modèle, Partie**. En ce qui concerne le *DiCoInfo*, les étiquettes les plus récurrentes (outre celles signalée plus haut) sont : **Destinataire, Instrument, Support, Point de départ, Substitut** ou comme pour le, *DiCoEnviro*, **Menace**.

De plus, toujours dans un souci de transparence et de lisibilité, les auteurs ont opté pour un système où chaque actant (déjà décrit en termes de rôles sémantiques) est représenté par un terme typique. C'est-à-dire qu'au lieu des variables ou du système de numérotation[24] traditionnellement utilisés pour décrire la structure actancielle, l'usager trouvera l'actant typique qui correspond à l'une des réalisations linguistiques de l'actant (mis en italique dans notre tableau). Il s'agit de l'unité lexicale (ou terminologique) qui apparaît le plus fréquemment dans l'environnement du terme, c'est-à-dire dans les occurrences extraites du corpus et qui peut être considérée comme un terme générique englobant toutes les autres réalisations. Ainsi, la structure actancielle peut être lue de deux manières :

AFFICHER$_1$ **: Agent** affiche **Patient** à **Destination**
AFFICHER$_1$ **:** ***utilisateur$_1$, logiciel$_1$*** affiche ***fichier$_1$, icône$_1$*** à ***écran$_2$***

Sur demande, on peut également accéder à une liste d'autres réalisations linguistiques récurrentes observées dans le corpus. Voici les réalisations linguistiques des actants du terme AFFICHER.

[24] Voir à ce propos L'Homme 2011

Agent	Patient	Destination
logiciel$_1$, programme$_1$, système d'exploitation$_1$, *utilisateur*$_1$	aide$_1$, barre$_1$, boîte de dialogue$_1$, capture d'écran$_{1\ 2}$, caractère$_1$, chaîne de caractères$_1$, commande, courriel, document, données$_1$, dossier$_1$, enregistrement$_2$, fenêtre$_1$, *fichier*$_1$, flèche$_1$, fond d'écran$_1$, formulaire$_1$, *icône*$_1$, image$_1$, information$_2$, ligne$_2$, masque de saisie$_1$, menu$_1$,message$_1$, message$_2$, page$_2$, panneau de configuration$_1$, texte$_2$	écran$_1$, *écran*$_2$

Tableau 1. Liste des réalisations linguistiques des actants du terme *afficher* dans le *DiCoInfo*, consulté le 12.09.2014.

Comme le soulignent les auteurs (L'Homme 2010 : 150, *DiCoEnviro* Manuel : 15), cela contribue à une meilleure compréhension du terme décrit et rend cette description plus accessible et lisible pour les usagers non-linguistes, peu habitués à la présentation formelle des données sémantiques. De plus, la présentation des actants typiques (c'est-à-dire leurs réalisations linguistiques) au niveau de la structure actancielle permet aux usagers d'observer l'emploi des termes dans des contextes concrets et d'accéder à des informations sur la combinatoire. Remarquons que l'on trouve une solution comparable dans le DAFA (Binon *et al.* 2000). Cependant, ce qui est intéressant ici, c'est que le même modèle d'encodage (au moyen des actants typiques) apparaît dans d'autres rubriques de l'entrée, notamment dans la définition construite à partir de la structure actancielle et dans la partie consacrée à l'explication des liens lexicaux. Cela contribue à la clarté et la lisibilité de la description.

La dernière rubrique, *Liens lexicaux*, présente la combinatoire (comprise ici au sens large) des termes. En effet, il s'agit d'une description des relations paradigmatiques et syntagmatiques que le terme apparraissant en entrée entretient avec d'autres termes (dont la plupart sont décrits également dans le dictionnaire). À l'exception des synonymes et des variantes qui, le cas échéant sont énumérés à la suite de la définition, les liens sémantiques sont présentés sous forme d'un tableau. La colonne de droite contient ainsi les termes reliés (assorti d'un lien cliquable et accompagné d'un numéro d'acception permettant de retrouver l'article correspondant, si le terme figure dans le dictionnaire). La colonne de gauche comporte des explications des relations sémantiques. Nous observons que les termes reliés par

un lien paradigmatique apparaissent seuls alors que ceux entrenant une relation de nature syntagmatique avec le terme en entrée sont placés dans un énoncé montrant de quelle manière ils se combinent avec lui. Cependant, les deux types de phénomènes (dérivations sémantiques et collocations) sont présentés ensemble et décrits à l'aide d'un outil commun. Rappelons que selon la LEC, il existe un lien conceptuel profond entre dérivations sémantiques et collocations et que les deux notions ne devraient pas être considérées séparément car elles font appel aux mêmes universaux linguistiques (Polguère 2003 : 118).

Ainsi, les termes reliés sont regroupés et organisés selon l'ordre suivant :

Intitulé	**Type de relation décrite**	**Exemples**
Voisins	Hyperonymie, co-hyponymie ou quasi-synonymie	fichier$_1$ ➔ document$_1$ copie$_{1.1}$ ➔ enregistrement$_{1.1}$, piratage$_2$, installation$_2$, sauvegarde $_{1.1}$, stockage$_{1b}$ poussière$_1$ ➔ particule
Contraires	Antonymie, opposition contrastive, conversivité	accélération$_1$➔ ralentissement$_1$, écologique$_2$ ➔ polluant$_{1,}$ terrestre$_2$➔ marin$_1$, enregistrer$_1$➔ supprimer$_1$, effacer$_1$
Sortes de	Hyponymie ou relation collocationnelle de type : terme + modificateur	copie➔ *mot clé* à la volée copie ➔ *mot clé* illégale, piratage$_2$ anti-logiciel espion$_1$➔ *mot clé* en ligne$_{1a}$ courriel$_1$ ➔ canular$_1$, courriel$_1$ ➔ spam $_1$, pourriel $_1$, *mot clé* non sollicité, *mot clé* indésirable courriel$_1$ ➔ *mot clé* infecté $_1$, *mot clé* viral $_{1b}$, *mot clé* courriel$_1$ ➔ hameçon, *mot clé* frauduleux $_{1a}$, faux *mot clé*, courriel$_1$ ➔ *mot clé* d'hameçonnage, courriel$_1$ ➔ *mot clé* entrant, courriel$_1$ ➔ *mot clé* anonyme $_{1b}$ courriel$_1$ ➔ *mot clé* sortant
Autres parties du discours et dérivés	Dérivation morphologique	chargement$_1$➔ charger$_1$ compilable$_1$➔ compiler$_{1a}$, compilable$_1$➔ compilation$_{1b}$ programmer$_2$➔ programmation$_2$, programmer$_2$➔ reprogrammer$_1$, programmer$_2$➔ programmable$_{1,}$ programmer$_2$➔ reprogrammable$_1$
Autres	Rélations paradigmatiques irregulières	azote$_1$➔ atmosphère$_{1,}$ azote$_1$➔ cycle de l'*mot clé* climat$_1$➔ climatologie$_1$ compostage$_1$➔ usine de de *mot clé*, centre de *mot clé*
Combinatoire lexicale	Collocatifs verbaux et dérivés nominaux ou	fenêtre$_1$➔ la *mot clé* s'affiche, la *mot clé* s'ouvre

(*DiCoInfo*) ou Combinatoire (*DiCoEnviro*)	adjectivaux de ces verbes	fenêtre$_1$➔ affichage$_{1a}$ de la *mot clé*, ouverture$_{1a}$ de la *mot clé* fenêtre$_1$➔ masquer$_1$ une *mot clé*, fenêtre$_1$➔ rafraîchir$_1$ une *mot clé* fenêtre$_1$➔ agrandir$_1$ une *mot clé* fenêtre$_1$➔ maximiser$_1$ une *mot clé* fenêtre$_1$➔ réduire une *mot clé* fenêtre$_1$➔ minimiser une *mot clé*

Tableau 2. Différents types de relations représentées dans le *DiCoInfo* et le *DiCoEnviro*, consultés le 12.09.2014.

Remarquons que parmi les liens lexicaux recensés (surtout dans la rubrique *Autres* qui contient des relations paradigmatiques traditionnellement bien représentées dans les descriptions terminologiques – voir à ce propos Sager 1990 : 28-39), il y a ceux qui pourraient être considérés comme des relations conceptuelles du terme : climat$_1$➔ climatologie$_1$ (science), azote$_1$➔ cycle de l'azote (phases de transformation).

Deuxièmement, il faut noter que chaque lien sémantique (représenté par la flèche➔ dans le Tableau 2) est décrit au moyen d'une paraphrase linguistique qui explicite en détail le sens de la relation en question. En effet, la description des liens paradigmatiques et syntagmatiques dans les *DiCoInfo* et *DiCoEnviro* est basée sur le modèle des fonctions lexicales de la LEC. Cependant, afin de faciliter l'interprétation des relations, les auteurs ont fait appel à un métalangage de vulgarisation des FL en s'inspirant des recherches menées dans le cadre des projets *DiCo* (Polguère 2003) et *LAF* (Mel'čuk & Polguère 2007). Rappelons que Polguère (2003) propose d'envisager la notion de FL en la dégageant du formalisme auquel elle est habituellement associée. Il considère (2003 : 120), que les FL peuvent être comprises et utilisées comme un outil conceptuel indépendamment de leur formalisme d'encodage traditionnel car il s'agit d'universaux linguistiques. D'après l'auteur (2003 :125), une FL **f** données peut être conceptualisée comme une entité linguistique à part, une MÉTALEXIE, existant dans l'esprit du locuteur de façon indépendante des multiples applications possibles **f** (L) de cette fonction. Ainsi, la modélisation offerte par les FL est une pseudo-langue qui à son tour peut être traduite par des paraphrases approximatives des expressions linguistiques qu'elle modélise. Il est donc possible de paraphraser un lien sémantique au moyen d'une « langue contrôlée » (ici français, anglais, espagnol, portugais), fondée sur les universaux sémantiques associés aux FL (Polguère 2003 : 124). En effet, cette paraphrase linguistique est en quelque sorte un travail mental fait par le linguiste qui encode une FL standard pour une lexie donnée. De par son caractère intuitif et explicite, la paraphrase linguistique est une

forme de vulgarisation qui permet aux non-initiés au modélisme formel et abstrait des FL d'accéder à un grande nombre d'informations lexicales. Soulignons que cela ne concerne pas les FL non standard (FLNS) qui sont déjà à l'origine intégralement écrites dans la langue de description.

À titre illustratif, reprenons quelques exemples de liens sémantiques présentés dans le tableau ci-dessus (Tableau 3) :

Terme	Explication	Lexie reliée
courriel$_1$	Qui est non souhaité par le ***destinataire***	spam $_1$, pourriel $_1$, *mot clé* non sollicité, *mot clé* indésirable
courriel$_1$	Qui est envoyé par un ***expéditeur*** qui cherche à obtenir des renseignements personnels du ***destinataire***	hameçon, *mot clé* frauduleux $_{1a}$, faux *mot clé*,
copie$_{1.1}$	Qui est faite sans autorisation	*mot clé* illégale piratage$_2$
azote$_1$	Composant de	atmosphère$_1$
anti-logiciel espion$_1$	Qui est placé sur un ***ordinateur*** distant et qui peut être utilisé lors d'une connexion	*mot clé* en ligne$_{1a}$
compilable$_1$	Nom de sens voisin	compilation$_1$

Tableau 3. Description des relations lexicales proposée dans le *DiCoInfo* (consulté le 12.09.2014).

Ainsi, les deux dictionnaires proposent des descriptions détaillées des propriétés linguistiques des termes en mettant l'accent sur les distinctions sémantiques fines. Il est interéssant de noter que les explications des liens sémantiques sont données en fonction du ou des actants typiques faisant ainsi référence à la structure actancielle du terme en entrée (mis en italiques dans notre tableau). Le fait de garder le même système d'encodage au moyen des actants typiques facilite l'interprétation des données et permet de faire un lien direct entre la structure actancielle et les relations sémantiques du terme.

Cependant, comme nous pouvons le constater dans le tableau ci-dessous (Tableau 4), les auteurs, fidèles aux objectifs initiaux des projets, ont souhaité conserver la notation proposée par la LEC. Les utilisateurs peuvent donc accéder à trois niveaux d'explication : le premier niveau d'explication, affiché par défaut, fait référence aux réalisations linguistiques typiques des actants, le deuxième à la notation des actants en rôles sémantiques et le troisième correspond aux FL. Ainsi, l'originalité des ces deux bases de données terminologiques

consiste à fournir une richesse d'informations lexicales bien structurées et accessibles à tout type de public (linguiste et non-linguiste).

Explication du lien – terme typique	**Explication du lien – rôle actanciel**	**Explication du lien – fonction lexicale**	**Termes liés**
La f. fonctionne		$Fact_0$	la *mot clé* s'affiche$_{1a}$ la *mot clé* s'ouvre
L'***utilisateur*** fait disparaître une f. de l'écran	**Agent** fait disparaître une f. de l'écran	$Liqu_1Func@[@:écran]$	masquer$_1$ une *mot clé*
L'***utilisateur*** met une f. dans un état antérieur	**Agent** met une f. dans un état antérieur	Mettre ~ à jour	rafraîchir$_1$ une *mot clé*
L'***utilisateur*** augmente une f.	**Agent** augmente une f.	$Caus_1PredPlus$	Agrandir$_1$ une *mot clé*
L'***utilisateur*** réduit une f.	**Agent** réduit une f.	$Caus_1PredMinus$	réduire une *mot clé*
L'***utilisateur*** modifie une f.	**Agent** modifie une f.	Changer la taille de ~	Redimensionner$_1$ une *mot clé*
L'***utilisateur*** fait en sorte qu'une f. puisse être utilisée	**Agent** fait en sorte qu'une f. puisse être utilisée	$Caus_1Able_1Real_1$	activer$_1$ une *mot clé* cliquer$_1$ sur une *mot clé*
L'***utilisateur*** prépare une f. en vue d'une utilisation qui lui convient	**Agent** prépare une f. en vue d'une utilisation qui lui convient	$Prepar_1Real_1[BUT:pour_1]$	personnaliser$_1$ une *mot clé*
L'***utilisateur*** utilise une f.	**Agent** utilise une f.	$Real_1$	ouvrir$_1$ une *mot clé* afficher$_{1b}$ une *mot clé*
L'***utilisateur*** met le ***logiciel*** ou le ***fichier*** dans une f.	**Agent** met le **Patient** ou le **Patient** dans une f.	$LabReal_{12}$	ouvrir$_1$ ... dans une *mot clé*
L'***utilisateur*** cesse d'utiliser une f.	**Agent** cesse d'utiliser une f.	$FinRéal_1$	fermer$_1$ une *mot clé*
L'***utilisateur*** fait en sorte qu'une f. ne puisse plus être utilisée	**Agent** fait en sorte qu'une f. ne puisse plus être utilisée	$Liqu_1Able_1Real_1$	*désactiver*$_1$ une *mot clé*

Tableau 4. Collocatifs verbaux du terme *fenêtre*$_1$ (*DiCoInfo* consulté le 12.09.2014).

2.3.2.1. Le *DiCoInfo* et le *DiCoEnviro* – modélisation des relations collocationnelles

L'originalité du *DiCoInfo* et du *DiCoEnviro* réside également dans la façon de modéliser l'information collocationnelle. En effet, il s'agit avant tout des collocatifs verbaux et des dérivés nominaux et adjectivaux de ces verbes qui sont très bien représentés dans les deux projets. Rappellons que hormis quelques rares exceptions (que nous avons mentionnées plus haut), peu de bases de données terminographiques recensent les combinaisons lexicales typiques dans lesquelles se retrouvent les temes. Comme le souligne L'Homme (2008 :79),

les informations de nature linguistique et surtout celles concernant la combinatoire font défaut aux dictionnaires de spécialité. Or, ces renseignements (permettant d'insérer correctement les termes dans les textes), sont particulièrement recherchés par les traducteurs ou les rédacteurs techniques.

« Pour rédiger ou traduire un texte spécialisé, il apparaît nécessaire non seulement de maîtriser des ensembles de termes et les concepts qu'ils désignent, mais de savoir combiner ces termes à des unités lexicales spécifiques. »

(L'Homme 1998 : 513)

Pour parler des phénomènes collocationnels en terminologie, L'Homme (1998 : 513 - 522) introduit la notion de combinaison lexicale spécialisée (desormais CLS). Tout comme les collocations décrites en lexicologie, les CLS ont un caractère conventionnel. En effet, elles font l'objet d'un consensus établi au sein d'un groupe de spécialistes. Comme le remarque l'auteur (1998 : 514) : *« Un non-spécialiste doit apprendre à mobiliser ces usages pour insérer les unités terminologiques dans des environnements idiomatiques. »*. En ce qui concerne la forme de groupements, les CLS sont également composées de deux lexèmes, notamment d'un terme toujours défini comme étant le mot clé ou la base et d'un cooccurrent. Cependant, il est nécessaire de souligner que dans les CLS, les cooccurrents n'acquièrent pas toujours un sens nouveau lorsqu'ils sont combinés à un terme donné. D'après L'Homme (1998 : 516), la question de la compositionalité ne semble pas constituer une caractéristique importante des CLS alors qu'elle est souvent évoquée pour définir les collocations. De plus, il est interéssant de noter que dans les CLS, les cooccurrents spécialisés ont tendance à se combiner avec des termes sémantiquement apparentés, c'est-à-dire des mots qui appartiennent à une même classe conceptuelle. Comme le remarque L'Homme (1998 : 518), cette généralisation des combinaisons semble très productive en ce qui concerne les collocatifs verbaux. Elle souligne qu'au cours du travail de recensement des CLS *verbe + terme* dans le domaine de l'informatique, tous les groupements pouvaient être généralisés à des ensembles de termes appartenant à une même classe sémantique. En effet, le phénomène s'étend à la structure actancielle du verbe et les termes sémantiquement apparentés correspondent aux réalisations linguistiques des actants de ce dernier, comme le montre l'exemple suivant :

lancer$_1$: ***utilisateur$_1$***	lance	***logiciel$_1$***	anti-logicie$_1$ espion$_1$ antivirus$_1$ application$_1$ débogueur$_1$ gestionnaire$_1$ jeu$_1$ navigateur$_1$ programme$_1$ script$_1$ système d'exploitation$_1$

Figure 18. Exemple de classe sémantique dans *DiCoInfo* (consulté le 12.09.2014)

Néanmoins, comme le souligne L'Homme (2002), les classes des termes ainsi dégagés ainsi que leurs collocatifs verbaux doivent être hiérarchisées, mais cette organisation en fonction de la combinatoire n'a rien à voir avec les modèles ontologiques. Il ne s'agit pas ici de traduire l'organisation des connaissances d'un domaine mais plutôt de refléter sa structure lexicale. Ainsi, les efforts de structuration et de hiérarchisation des relations entre les termes vont s'appuyer sur les critères sémantico-lexicaux et non pas sur les aspects conceptuels.

Nous proposons de revenir aux exemples du Tableau 4 où nous avons énuméré les collocatifs verbaux du terme *fenêtre*. Remarquons que la liste (quoique non exhaustive), est très longue et présente différentes relations. Il faut savoir que certains articles peuvent contenir jusqu'à 100 relations collocationnelles (comme c'est le cas pour le terme *fichier$_1$*). Ainsi, afin de faciliter l'accès à l'information collocationnelle, si riche et diversifiée soit-elle, les auteurs ont décidé d'organiser les données en prenant en compte trois propriétés linguistiques des collocations, conformément aux principes de la Lexicologie Explicative et Combinatoire. Comme le remarque L'Homme (2009 :240), les rares projets terminographiques comportant des informations sur les collocations proposent différentes formes de description et d'organisation des données : selon les parties du discours (Termium Plus, Meynard 2000, Cohen 1998, DAFA 2000), selon la position syntaxique occupée par rapport au terme (Meynard 2000, DAFA 2000) ou selon le sens (Cohen 1998, DAFA 2000). Comme nous venons de le mentionner, les concepteurs du *DiCoInfo* et du *DiCoEnviro* ont décidé de combiner trois critères, notamment : la combinatoire syntaxique, la structure actancielle, le sens du collocatif.

a) La combinatoire syntaxique (également appelée le régime), désigne les propriétés de combinatoire grammaticale du terme ou autrement dit, le contrôle qu'exercent le terme

et ses cooccurrents sur leur structure de complémentation. Nous reprenons ici l'exemple donné par l'Homme et al (2012 : 218) qui distinguent deux cooccurrents du terme *mot de passe*, notamment : utiliser$_1$ un *mot de passe* et accéder$_2$ à … avec un *mot de passe*. Dans la premère combinaison, le terme *mot de passe* est un complément d'objet direct alors que dans la deuxième combinaison, il occupe la place de complément d'objet indirect. Les deux collocatifs verbaux seront donc présentés séparément.

b) On remarque une certaine analogie entre la combinatoire syntaxique et la structure actancielle du terme (*mot de passe* remis par ***utilisateur$_1$*** à ***fournisseur$_1$***, ***ordinateur$_1$*** pour intervenir sur ***compte$_1$, site$_1$***). Comme nous pouvons le constater, le sens de l'énoncé : utiliser$_1$ un *mot de passe* implique le premier actant du terme *mot de passe* (***utilisateur$_1$***) alors que le sens de : accéder$_2$ à … avec un *mot de passe* en implique deux (***utilisateur$_1$***, ***compte$_1$, site$_1$***). La prise en compte de la structure actancielle dans l'explication des liens colllocatifs permet de faire des distinctions sémantiques fines.

c) Le dernier critère pris en compte est le sens des collocatifs verbaux et de leurs dérivés. En effet, comme le remarquent L'Homme *et al.* (2012 :218), les termes se combinent avec des cooccurrents verbaux qui véhiculent différents sens. Ainsi, les verbes collocatifs utiliser$_1$ et accéder$_1$ correspondent à des réalisations typiques du terme *mot de passe*. En revanche, les verbes : créer$_1$, définir$_1$, activer$_1$, taper$_1$, refuser$_1$ expriment différentes phases dans l'utilisation du *mot de passe*. Ces propriétés sémantiques sont déjà décrites au moyen des fonctions lexicales : **Real$_1$**(mot de passe) = utiliser$_1$, **Labreal$_{12}$**(mot de passe) = accéder$_1$ à…avec ~, **Caus$_1$Func$_0$**(mot de passe) = créer$_1$, définir$_1$. Cependant, afin de faciliter l'accès aux données collocationnelles et d'améliorer la lisibilité des entrées, les collocatifs verbaux et leurs dérivés sont organisés selon l'ordre dans lequel les activités liées au *mot clé* (c'est-à-dire au term) sont normalement effectuées. Voici, comment on été classés les collocatifs verbaux du terme *fenêtre* presenté plus haut :

Fonctionner ou faire ce qu'on attend de	
Fonctionner	La *mot clé* s'affiche La *mot clé* s'ouvre
Mettre quelque part	
Faire disparaître / Ne pas mettre	masquer$_1$ *Le mot clé*

Transformer / Augmenter / Réduire	
Mettre ~ à jour	rafraîchir$_1$ *mot clé*
AugmenterDiminuer / RéduireChanger la taille de ~	agrandir$_1$ *mot clé* réduire une *mot clé* minimiser$_1$ *mot clé* redimensionner$_1$ *mot clé*
Utiliser / Ne pas utiliser	
Permettre l'utilisation / Activer	activer$_1$ *mot clé* cliquer$_1$ sur une fenêtre
Préparer l'utilisation / Le fonctionnement	personnaliser$_1$ *mot clé*
Utiliser / Faire fonctionner	ouvrir$_1$ une *mot clé* afficher$_{1b}$ une *mot clé*
Cesser d'utiliser / De faire fonctionner	fermer$_1$ une *mot clé*
Empêcher l'utilisation / Désactiver	désactiver$_1$ une *mot clé*

Tableau 5. Organisation des collocatifs verbaux du terme *fenêtre*$_1$ (*DiCoInfo* consulté le 12.09.2014).

Nous pouvons constater que la présentation des termes proposée dans le *DiCoInfo* et le *DiCoEnviro* se détache radicalement de l'optique conceptuelle : les termes y sont envisagés comme des unités lexicales et appréhendés dans leur fonctionnement linguistique. Les descriptions constituent un reflet des observations faites sur un ensemble de données linguistiques et des interactions entretenues par les unités. Pour L'Homme (2005 : 1123), cette optique est plus compatible avec les méthodes actuelles de confection de dictionnaires (généraux et spécialisés) qui s'appuient principalement sur des corpus et qui adoptent une démarche sémasiologique. En revanche, cette démarche ne peut pas prétendre à une structuration des connaissances au sens de l'approche conceptuelle. Selon L'Homme (2005 : 1130), le terminologue doit faire un choix entre deux conceptions centrales : 1) le terme comme étiquette de concept ; 2) le terme comme véhicule d'un sens spécialisé. D'après elle, chacune des options a des conséquences méthodologiques et descriptives importantes qui mènent à des descriptions incompatibles.

2.4 Formaliser les relations sémantiques[25] afin de refléter la structure conceptuelle d'un domaine de spécialité – les travaux de Jeanne Dancette

Dans cette partie, nous proposons de présenter deux projets dictionnairiques qui exploitent les relations sémantiques en s'inspirant du modèle mel'čukien sans toutefois s'appuyer sur la Lexicologie Explicative et Combinatoire. En effet, il s'agit du *DAD*[26] et du *DAMT*[27], deux dictionnaires multilingues en ligne, spécialisés (l'un dans le domaine du commerce de détail et l'autre dans le domaine de la mondialisation économique), conçus comme des ouvrages de référence dans leur domaine (Dancette 2011b : 288). Tous deux sont destinés à un public de traducteurs, professionnels, étudiants et rédacteurs et doivent servir à la fois d'ouvrage encyclopédique et de dictionnaire multilingue en incluant aussi bien les connaissances contextuelles (le schéma d'emploi du terme donné en discours) que conceptuelles et encyclopédiques. Comme le souligne Dancette (2011a : 162), l'originalité de ces projets consiste à mettre en valeur un système développé de relations sémantiques et associatives afin de de répondre aux besoins de documentation (informer sur les concepts) et de recherche lexicale (préciser le sens et l'emploi des termes et des collocations spécialisées, dans plusieurs langues) des utilisateurs.

2.4.1 L'intégration de la dimension cognitive dans les modèles descriptifs des langues de spécialité

La recherche documentaire étant un volet important du travail des traducteurs (comme celui des étudiants ou des rédacteurs dans le domaine de spécialité visé) ; l'un des objectifs

---

[25] Nous avons repris ici l'expression « relations sémantiques » telle qu'elle a été utilisée par Dancette (2007, 2011b), soulignant, d'après l'auteur qu'elle couvre l'ensemble des relations entre les termes, même si théoriquement une distinction doit être faite entre les relations d'ordre sémantique et celles d'ordre purement conceptuel.

[26] DANCETTE (Jeanne) et RÉTHORÉ (Christophe), 2000 et 2006 pour la version électronique, *Dictionnaire analytique de la distribution. Analytical Dictionary of Retailing*. Montréal, Les Presses de l'Université de Montréal. Version électronique sur le site de l'OLST (2006). Consultée le 04 avril 2014, http://falbala.ling.umontreal.ca/*DAD*/.

[27] DANCETTE (Jeanne), 2014, *DAMT – Dictionnaire analytique de la mondialisation du travail/Analytical Dictionary of Globalization and Labour/Diccionario analítico de la globalización del trabajo*, Consultée le 4 avril 2014, http://zedamt.herokuapp.com/more.

majeurs de ces deux projets dictionnairiques est la structuration des connaissances qui a été réalisée ici par l'étiquetage des relations sémantiques. Il est évident que la traduction spécialisée implique non seulement un traitement des unités linguistiques, mais aussi des connaissances. Comme le souligne Dancette (2005 :549), c'est une activité qui nécessite une expertise. C'est pour cela qu'elle insiste sur le rôle des construits cognitifs (connaissances encyclopédiques) dans le processus de traduction qui pour elle, contribuent à l'élaboration de la cohérence du texte traduit:

« (…) plus les référents cognitifs du traducteur se rapprochent de ceux du scripteur (auteur), plus la cohérence manifestée dans le texte traduit devient compatible avec la cohérence du texte de départ (cohérence qui, à son tour, pourra contribuer, parmi d'autres éléments, à la compréhension du texte par le nouveau lecteur (destinataire de la traduction). »

Dancette (2003 : 142)

Selon Dancette et Halimi (2005 : 553), le manque d'une connaissance conceptuelle extralinguistique conduit à des imprécisions dans la traduction ou à des hésitations dans le choix des équivalents. Il est donc indispensable que le traducteur puisse accéder à l'ensemble des référents cognitifs que partagent les spécialistes du domaine donné dans les deux communautés linguistiques mises en contact par la traduction. D'où l'importance de l'organisation et de la représentation des connaissances en terminologie. En effet, comme le soulignent les auteurs: *selon l'approche cognitive, le processus de traduction se définit essentiellement par le traitement de l'information contenue dans le texte à traduire et par sa mise en relation avec les connaissances antérieures* (Dancette et Halimi 2005 : 548). Les auteurs font ainsi référence à l'approche constructiviste de l'apprentissage et à ce qu'on désigne, en sciences de l'éducation sous le terme de « scaffolding » (Dancette 2011 : 286), c'est-à-dire, le processus par lequel l'apprenant construit son savoir en situant les nouvelles notions par rapport aux schèmes de pensée préexistants. Dans le cas de la traduction, ce sont les concepts qui constituent des éléments de construction de connaissances.

« (…) le traducteur, pour travailler avec intelligence et assurance sur un texte spécialisé, doit se sentir autonome dans sa recherche d'équivalents lorsqu'il passe de la langue de départ à celle d'arrivée. Cette autonomie dépend de deux facteurs interreliés: la compréhension de la structure globale du champ conceptuel où il évolue et la maîtrise des notions essentielles. »

(Dancette et Réthoré 1997: 230)

Dans le même temps, Dancette met l'accent sur l'efficacité des relations logiques comme moyen d'organisation des connaissances. (Dancette et Halimi 2005 : 553, Dancette 2011 : 286, Dancette 2006 : 144). D'après l'auteur, l'identification des liens entre les concepts aide à structurer les connaissances. En s'appuyant sur les recherches en éducation et en sciences cognitives, elle avance l'hypothèse selon laquelle les cartes conceptuelles (ou tout autre modèle de représentation de relations conceptuelles sous forme procédurale) faciliterait le processus de conceptualisation et, de là, la production langagière, et plus précisément le processus de traduction. Ces modes de représentation servent notamment de soutien au travail de décodage des textes spécialisés, facilitent l'acquisition du vocabulaire spécialisé et permettent de résoudre des problèmes concernant la recherche des équivalents. La figure ci-dessous (Figure 19) représente le réseau conceptuel du terme VENTE AUX ENCHÈRES [28] au moyen d'une « carte conceptuelle ». Il s'agit d'un schéma sous la forme d'un graphe où les différents concepts sont rassemblés par champs et reliés entre eux. Comme nous pouvons le constater, les différentes relations que ces concepts entretiennent les uns avec les autres sont identifiées et classées par type de relations.

Comme le souligne Dancette (2011b : 287), la carte conceptuelle donne une idée de la nature de l'information que l'on cherche à « encadrer ». Le cas échéant, elle permet de visualiser l'univers du terme « VENTE AUX ENCHÈRES», c'est-à-dire, toutes les connaissances que l'on possède de ce concept à un moment donné. Une telle représentation graphique fournit une image plus « parlante » et permet de mettre en lien un grand nombre d'informations. Nous y retrouvons donc des termes qui sont associés à « VENTE AUX ENCHÈRES » par différentes relations telles que des relations hiérarchiques (Spec), des relations de synonymie (Syn, $Syn_{\cap}$), des relations d'agent et d'objet (Ag, AgSpec, Obj), de lieu typique (Loc) ou bien des relations dérivationnelles (V). Comme nous pouvons le constater, ce schéma met en évidence une relation de deuxième degré (*fol enchérisseur* est un spécifique de l'agent *enchérisseur*) en relevant une structure hiérarchique complexe.

[28] Comme le souligne l'auteur, le réseau conceptuel a été établi sur la base de l'article « auction» tiré du corpus *DAD* (Dancette 2000&2006)

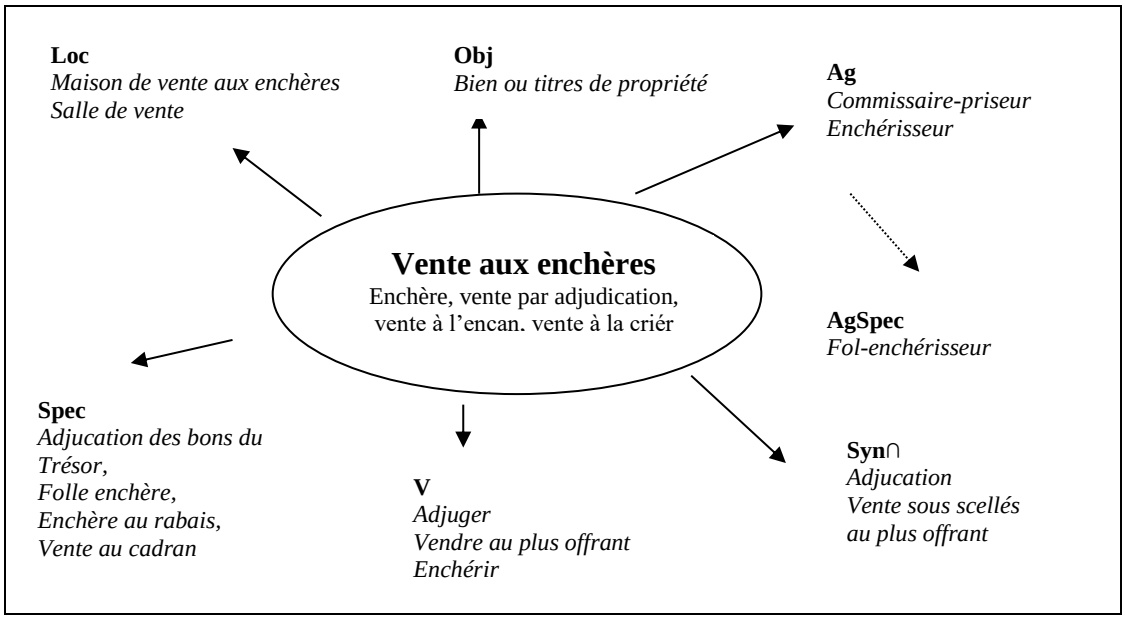


Figure 19. Réseau conceptuel du terme *vente aux enchères* proposé par Dancette (2006 :150)

Dancette fait ainsi référence à la théorie des cadres (ou *frames*), développée parallèlement par les représentants de deux écoles : d'un côté, Fillmore et la Sémantique des cadres (*Frame Semantics*), de l'autre, Minsky et son système de représentation des connaissances à l'aide des cadres (*Frame System Theory*). Même si (comme nous allons le voir dans la partie suivante de notre étude), le premier courant est nettement linguistique et le deuxième se situent plutôt dans le domaine des sciences cognitives et de l'intelligence artificielle, ils partagent la même origine. En effet, ils découlent de l'idée selon laquelle notre lexique mental est organisé en un immense réseau sémantique et qu'il est possible de le représenter à l'aide de différents formalismes.

Ainsi, les outils favorisant l'organisation et le transfert des connaissances doivent être privilégiés en terminologie. Selon Dancette (2003 :155), le dictionnaire basé sur une approche conceptuelle et cognitive offre un réseau de connaissances sur lesquelles le traducteur peut échafauder ses référents cognitifs. Ses deux projets dictionnairiques (à savoir le *DAD* et le *DAMT*) ont donc été conçus comme des bases de connaissances terminologiques en référence à la notion de BCT (TKBs en anglais) introduite, comme nous allons le voir plus en détail à la section 4.3.1, par Ingrid Meyer. Cependant, son idée va plus loin. L'objectif de Dancette est de proposer un modèle dictionnairique combinant les traits d'une encyclopédie, d'une

ontologie et d'une ressource linguistique, un modèle où l'utilisateur devient lui-même le gestionnaire de ses connaissances grâce aux différents modes d'accès aux informations (Dancette 2011a : 162). Cela fait penser aux articles dynamiques (*dynamic articles*) de Trap (2009 : 57). Ainsi, l'enjeu principal est de trouver un formalisme qui donne accès à toute la richesse de l'information, en la systématisant afin de permettre à l'utilisateur de l'exploiter au maximum.

2.4.2 Les représentations sémantiques comme moyen de structuration des connaissances dans les domaines spécialisés

Comme nous l'avons déjà mentionné ci-dessus, l'originalité de ce modèle consiste à mettre en valeur un système développé de relations sémantiques (RS) et associatives (RA) reflétant l'univers des domaines de spécialité. En effet, ce modèle permet de relier chaque terme à un ensemble d'autres termes du même champ en reproduisant ainsi la structure conceptuelle du domaine en question. D'après Dancette (2011b), le balisage de l'information relative aux concepts au moyen des RS et RA facilite l'accès aux données contenues dans les dictionnaires et optimise l'acquisition des connaissances chez l'utilisateur. Même si l'idée de l'exploitation des relations sémantiques dans la description terminologique n'est pas nouvelle (voir entre autres Otman 1996), ce qui distingue ce modèle des autres est le nombre important des liens décrits et leur étiquetage par classes de relation. En effet, les liens que l'on y cherche à formaliser sont de nature différente. À côté des relations taxonomiques et partitives (comme l'hyperonymie, l'hyponymie, la co-hyponymie, la méronymie), traditionnellement exploitées dans les modèles ontologiques, on y retrouve d'autres liens : conceptuels (cause, instrument, but, agent, lieu typique, etc.), paradigmatiques (synonymes, antonymes), ou syntagmatiques (verbes support). Étant donné l'hétérogénéité des types de liens, Dancette a donc décidé de faire une distinction[29] nette entre les relations d'ordre sémantique (RS) et celles d'ordre purement conceptuel (RA) en essayant ainsi de concilier les approches lexicographique et terminologique. Afin de comprendre la démarche de Dancette, il nous paraît important de présenter rap*ibid.*ent les étapes successives de ses recherches qui ont abouti à la rédaction de ses deux dictionnaires, à savoir le *DAD* (2006) et le *DAMT* (2010).

[29] Cela concerne surtout son dernier projet, le *DAMT* (2010)

2.4.2.1 Le *DAD*: entre dictionnaire de langue et encyclopédie[30]

Le point de départ est la version papier du Dictionnaire analytique de la distribution (*DAD* 2000). C'est un dictionnaire de nature encyclopédique, très spécialisé, qui contient 350 articles représentant les concepts clés de la distribution. A l'intérieur des articles, les concepts sont décrits au moyen de neuf rubriques : vedettes anglaises, équivalents français, définition minimale, précisions sémantiques, relations internotionnelles, compléments d'information encyclopédique, informations linguistiques, contextes, exemples. Les deux premières rubriques sont consacrées aux données interlinguistiques. Les caractéristiques sémantiques ainsi que les informations d'ordre conceptuel sont consignées dans les quatre rubriques suivantes ce qui permet d'identifier un ensemble d'autres termes relevant du domaine de la distribution et reliés à la vedette.

La figure ci-dessous (Figure 20), reprend l'article du terme *auction* (vente aux enchères), souvent cité comme exemple (Dancette etL'Homme 2004, Dancette 2006). Nous pouvons constater que la microstructure est très riche et englobe toutes sortes d'informations : description détaillée des notions, liens conceptuels entre termes, possibilités combinatoires, fonctionnement discursif, aspects culturels et interculturels. Ainsi, on y retrouve des relations fondamentales en terminologie comme l'hyperonymie et l'hyponymie *(vente, adjudication des bons du Trésor, folle enchère, enchère au rabais, vente au cadran*), synonymie (*adjudication*), ainsi que d'autres sortes de relations comme par exemple celle d'agent (*commissaire-priseur, enchérisseur, fol enchérisseur*), d'objet (*biens meubles, biens immeubles*), de lieu (*salle de vente*), de moyen (*offre*), d'action typique associée au terme *auction* (*vendre au plus offrant, offrir un prix, enchérir*), qui reflètent la structure du domaine en question. Cependant, vu que les données ont été présentées sous forme de textes descriptifs, les types de liens entre les termes n'ont pas été déterminés de manière formelle. L'utilisateur du dictionnaire est donc amené à les induire des phrases contenues dans les articles :

« La nature des relations sémantiques y est explicitée aussi précisément que possible (...). Toutefois, cette mise en évidence des relations utilise toutes les ressources du langage

[30] Le présent sous-titre reprend le titre de l'article de Dancette et Réthoré (1997)

naturel, avec ses nuances, voire ses ambiguïtés, dans une multitude de formulations idiosyncratiques, contextuelles. »

(Dancette etL'Homme 2001 : 390)

En effet, le format traditionnel du dictionnaire papier (l'ordonnancement linéaire des pages et l'ordre alphabétique) ne permet pas de mettre en valeur toutes les données. Ainsi, afin de faciliter le traitement des informations, les auteurs se sont tournés vers la version électronique. Les travaux sur le projet d'informatisation de la version papier du *DAD* (Dancette et Rhétoré 2000), ont amené Dancette et sa collaboratrice Marie Claude L'Homme à chercher un formalisme qui donne accès à toute la richesse de l'information du dictionnaire en permettant à l'utilisateur de l'exploiter au maximum. Leur objectif était de systématiser l'ensemble des données en proposant un modèle uniforme pour la modélisation de toutes les relations sémantiques et conceptuelles extraites du corpus.

Le modèle classique des traits sémantiques a constitué une première tentative de formalisation des relations entre les termes (Dancette & L'Homme 2001).

TERME	TRAITS SÉMANTIQUES	TERMES APPARENTÉS
VENTE AUX ENCHÈRES	*Prix non fixe*	*Vente au plus offrant*
	Catalogue	*Vente par catalogue*
	Lieu physique	*Vente en magasin*
	Marchandise usagée	*Vente de charité*
	Non-intervention du vendeur	*Vente impersonnelle*
	Support électronique	*Vente électronique*

Tableau 6. Modèle des traits sémantiques dans le projet *DAD* (Dancette 2006 : 149)

AUCTION

auction sale, sale by auction
vente aux enchères, enchère,
vente par adjudication, vente à l'encan,
vente à la criée (Fr.)

DÉFINITION

Vente publique au plus offrant de biens ou de titres de propriété.

PRÉCISIONS SÉMANTIQUES

Le **commissaire-priseur** (*auctioneer*) procède aux **enchères** qui se font de vive voix dans une **salle de vente** (*auction room*). Il présente l'article et demande une première **offre** (*bid*) ou annonce lui-même un prix initial minimal. Les acheteurs potentiels font des offres pour **enchérir** (*to bid, to make a bid*), chaque offre étant supérieure à la précédente. Le commissaire-priseur doit **adjuger** (*to knock down, to strike off*) l'article au dernier **enchérisseur** (*bidder*). Si celui-ci n'est pas capable de payer la somme offerte, s'il a fait une **folle enchère** (*false bidding*), l'article sera remis en vente. Si le prix obtenu la deuxième fois est inférieur à celui de la folle enchère, le **fol enchérisseur** (*false bidder*) devra payer la différence.

Les objets vendus aux **enchères** sont de nature et d'origine diverses. Il s'agit tant de biens meubles (bijoux, objets d'art, animaux) que d'immeubles. La **vente aux enchères** peut être volontaire ou forcée, par exemple, à la suite d'une faillite ou d'une saisie.

Le gouvernement a aussi recours à la **vente aux enchères** pour se défaire d'objets volés ou trouvés non réclamés, de biens saisis à la douane ou de biens en surplus.

La **vente aux enchères** est parfois utilisée sur le marché des **bons du Trésor** (*treasury bills*), sous le nom d'**adjudication des bons du Trésor** (*treasury bill auction, bill auction* [É.-U.]).

RELATIONS INTERNOTIONNELLES

Il existe quelques variantes de la **vente aux enchères** décrite plus haut.

Lors d'une **enchère au rabais** (*Dutch auction, Chinese auction*), ou **vente aux sous-enchères**, le prix initial est fixé plus haut que le prix que l'on estime obtenir, pour ensuite être baissé jusqu'au moment où quelqu'un accepte le prix. Ce type d'**enchère** est utilisé, entre autres, dans la vente d'**animaux** (*livestock*) et dans la vente de poissons et de produits agricoles dans le domaine du commerce de gros.

Dans sa forme la plus moderne, la **vente au cadran** (*clock auction, clock auction sale*), ou **marché au cadran**, ce processus est automatisé. Les acheteurs sont assis dans un amphithéâtre muni d'un grand cadran où les prix sont affichés. Les prix baissent jusqu'à ce qu'un acheteur arrête le mécanisme au moyen d'un bouton posé devant lui et fasse ainsi une offre. Si la marchandise est vendue en lots, le premier offrant peut choisir la quantité qu'il veut, et le reste sera vendu aux acheteurs subséquents.

Dans la **vente sous scellés au plus offrant** (*sealed bid auction*), la vente n'est pas publique, et les offres sont faites par écrit. L'objet sera adjugé au plus offrant.

Adjudication 1 (*adjudication*) est le terme juridique pour désigner la mise en vente aux enchères d'un bien. Le terme français désigne aussi la déclaration par laquelle le commissaire-priseur vend le bien au plus offrant. Il ne faut pas le confondre avec son homonyme **adjudication 2** (*tender*), qui désigne une soumission dans le cadre d'un appel d'offres. Le terme anglais *tender* peut aussi renvoyer à une offre d'achat, comme dans les expressions **vente par adjudication** (*sale by tender*), ou **vente par soumission**.

On appelle aussi **vente à la criée** (*hawking*), ou **criée**, la méthode de vente dans les marchés publics où les exposants interpellent les consommateurs.

COMPLÉMENTS D'INFORMATION

Hérodote rapporte la pratique de la **vente aux enchères** à Babylone dès le VI[e] siècle av. J.-C. Les Romains recouraient aux **enchères** dans le commerce régulier, mais aussi dans des cas particuliers (empereurs vendant du mobilier royal pour payer leurs dettes, soldats vendant leur butin de guerre).

L'Hôtel Drouot de Paris est la plus ancienne **maison de vente aux enchères** (*auction house*) publique du monde (objets d'art et mobiliers anciens). Notons aussi Sotheby's (É.-U.) et Christie's (G.-B.), fondées à Londres en 1733 et en 1766 respectivement, spécialisées dans les articles de luxe (bijoux, tableaux, etc.).

vendre au plus offrant : *to sell to the highest bidder*
offrir un prix, enchérir : *to bid on*

CONTEXTES

***Auctions** [...] have traditionally provided a rapid and effective means of disposing of goods, especially perishable products. **Auctions** are also frequently used to sell products directly to the consumers, especially if the value cannot readily be precisely determined, as is the case of works of art or antiques.* (*Britannica Micropædia* 1991)

*Aucune formalité spéciale n'est prescrite dans les **enchères** de meubles. Mais, dans les adjudications d'immeubles, pour laisser aux intéressés le temps de réfléchir, le Code de procédure civile prescrit l'emploi de bougies pouvant rester allumées une minute environ. L'adjudication ne peut être prononcée qu'après l'extinction successive de trois bougies.* (*Grand dictionnaire encyclopédique Larousse* 1987)

Figure 20. Exemple d'article dans le *DAD* version papier (Dancette et Réthoré 2000 : 9-10)

L'extraction des informations contenues dans l'article *auction* de la version papier du *DAD* a permis de dégager 6 traits sémantiques caractérisant le concept. Ces traits sémantiques permettent de rapprocher des termes appartenant à une même classe, en l'occurrence à la classe de « VENTE ». Comme le souligne Dancette (Dancette 2006 : 149), ce modèle fait apparaître la parenté sémantique (traits partagés et traits qui s'opposent), la hiérarchie (l'hyponyme a un trait spécifique que n'a pas l'hyperonyme) ou bien les relations de synonymie et d'équivalence interlinguistique (traits identiques), Cependant, il présente deux limites : a) il ne se prête pas à la génération car il y a presque autant de familles de traits que de notions ; b) il rend impossible de saisir tous les types de relations comme les relations syntagmatiques.

Dancette et L'Homme se tournent alors vers un modèle plus puissant, celui des fonctions lexicales de Mel'čuk. Comme le précisent les auteurs (Dancette & L'Homme 2002 : 599, Dancette & L'Homme 2004 : 118, Dancette 2006 : 204), ce formalisme rend compte de nombreux types de relations sémantiques (syntagmatiques et paradigmatiques) et permet une très grande généralisation grâce au nombre fini de FL. Nous présentons ci-dessous un exemple d'application de la méthode mel'čukienne au domaine de la distribution. Les relations entre le terme *vente aux enchères* et d'autres unités terminologiques et lexicales identifiées dans les exemples précédents sont formalisées ici à l'aide des FL.

vente aux enchères

Fonctions lexicales :

Syn (vente aux enchères) = enchère 1

$\mathbf{S_1}$ (vente aux enchères) = commissaire-priseur

$\mathbf{S_{loc}}$ (vente aux enchères) = salle de vente

$\mathbf{Oper_1}$ (vente aux enchères) = procéder à ~

Figure 21. Exemples d'encodage des liens sémantiques du domaine de la distribution au moyen des FL proposé par Dancette et L'Homme (2004 : 118)

Il est nécessaire de préciser que dans ce travail de formalisation des liens sémantiques au moyen des FL, les auteurs ont porté une attention particulière à la conversion des relations de type taxonomique et méronymique. Selon les auteurs (Dancette et L'Homme 2002 : 599),

ce type de relations est considéré comme crucial pour la description terminologique car il donne accès à l'organisation hiérarchique des connaissances. Cependant, les travaux sur la modélisation ces relations ont relevé quelques incompatibilités entre la LEC et le projet terminologique. Certaines adaptations ont donc été apportées.

Ainsi, quatre fonctions ont servi à la modélisation des relations taxonomiques, notamment les FL **Gener** et **Spec**[31] pour rendre compte des relations d'hypéronymie et d'hyponymie et **Syn**$_{\cap}$ et **Anti** pour décrire les relations de co-hyponymie. Quant à la méronymie, elle a été formalisée à l'aide de deux FL classiques, introduites par Mel'čuk : **Mult** (ensemble de) et **Sing** (unité minimale) ainsi que trois nouvelles fonctions : **Part**[32] (partie fonctionnelle de), **Part$_{occ}$**, **Phase** (phases du processus). A titre d'illustration, voici quelques exemples proposés par Dancette et L'Homme (2002) :

Gener (debit card) = magnetic card
Spec(good) = convenience good, unsought good, white good, brown good, durable good, non durable good
Anti (durable good) = non durable good
Syn$_{\cap}$ (convenience good) = unsought good
Mult (customer) = // clientele
Sing (factory outlet center) = // factory outlet store
Part (label) = bar code
Phase (prospection) = canvassing

Figure 22. Exemples d'encodage des relations hiérarchiques du domaine de la distribution proposé par Dancette et L'Homme (2002)

Comme nous pouvons le constater ci-dessus, malgré quelques modifications, le modèle mel'čukien s'adapte bien à la description des relations terminologiques traditionnelles. Cependant, il s'avère plus difficile de l'utiliser pour décrire d'autres types de relations, notamment les relations ontologiques.

[31] Rappelons que la FL Spec ne fait pas partie de la liste des 56 FL proposées dans le cadre de la LEC ; elle a été introduite par Grimes (1990).

[32] Cette nouvelle fonction lexicale a été proposée par Fontenelle (1997) afin de modéliser la relation de type *partie fonctionnelle de*.

D'après Dancette (2011b : 285), le monde des connaissances n'est pas structuré de manière parfaitement logique, notamment dans les sciences humaines. D'où le problème du flou des relations, dès qu'on s'éloigne des liens hiérarchiques (générique, partitif, synonymie, antonymie). Comme le précise Dancette (2003 : 146), les articles du *DAD* version papier contiennent beaucoup de relations indirectes ou de relations de parenté sémantique pauvre qu'il est impossible de saisir au moyen des fonctions lexicales classiques. En effet, la Lexicologie Explicative et Combinatoire propose un modèle basé sur une description lexicographique en sèmes alors que dans le modèle terminologique, on ne traite pas des mots mais des concepts.

« If we see terminology as reflecting the conceptual system, then terms are the labels of conceptual frames, i.e., constructs that encode entities and event types that are basic to human experience, in a given field. » (Dancette 2007: 207)

Ainsi, il s'agit là d'une tentative de formalisation des liens conceptuels et non proprement sémantiques, d'où le problème d'héritage des propriétés sémantiques des termes associés rendant difficile l'utilisation des fonctions lexicales.

De plus, l'application des fonctions lexicales requiert une définition linguistique rigoureuse entraînant l'identification de tous les actants. Or, les définitions du *DAD* version papier, qui comportent de nombreuses informations d'ordre encyclopédiques, n'ont pas été construites selon les principes de la Lexicologie Explicative et Combinatoire. Par conséquent, les auteurs se sont distanciés du modèle lexicographique des FL en aboutissant à un modèle plus adapté aux besoins terminographiques, le modèle de relations lexico-sémantiques (RLS). Même si les fonctions lexicales sont à la base de ce dernier, des adaptations importantes ont été faites afin de rendre compte de la structure conceptuelle du domaine de spécialité en question en décrivant des relations ontologiques (cause, agent, instrument, etc.) ou logiques (générique, partitif, synonymie, antonymie) qui relient un terme à un ensemble d'autres termes du même champ. Comme le souligne Dancette (2006 : 144), les liens que l'on y cherche à formaliser interviennent sur des plans différents : paradigmatique, syntagmatique, dérivationnel.

La méthode consiste à définir un nombre fini de classes de RLS permettant de représenter l'ensemble des termes renvoyant aux notions clés d'un domaine de spécialité.

Dancette (2011b : 284) précise que si certaines classes de relations sont communes à tous les domaines (les relations *générique*, *spécifique*, *partie*, *tout*, *agent*), nombre d'autres sont spécifiques au domaine.

Ainsi, l'extraction des relations sémantiques des articles du *DAD* version papier (environ 7 000), a permis d'établir 28 classes de RLS reflétant la structure du domaine de la distribution. Nous dénombrons 24 relations paradigmatiques et 4 relations syntagmatiques. Parmi les relations paradigmatiques, on peut distinguer celles qui décrivent les relations classiques en terminologies : **Gener** (générique), **Spec** (spécifique), **Contrast** (contraste ou antonyme), **Syn** (synonyme, concept apparenté), **Mult** (ensemble), **Sing** (élément d'un ensemble), **Tot** (entité globale), **Part** (partie d'une entité), **Phase** (phase d'un processus). Il s'agit par ailleurs des relations les plus présentes dans le corpus du *DAD*.

Un deuxième groupe de RLS reflète des relations actancielles et circonstancielles telles que **Ag** (agent), **Obj** (objet visé), **Recip** (récipiendaire), **Instr** (intrument), **Loc** (lieu typique), **Med** (moyen). Comme nous pouvons le constater, ces RLS correspondent aux FL $\mathbf{S_1}$, $\mathbf{S_2}$, $\mathbf{S_3}$, $\mathbf{S_{instr}}$, $\mathbf{S_{loc}}$, $\mathbf{S_{med}}$. Cependant, les auteurs ont opté pour une notation moins symbolique et plus transparente pour l'usager.

Un troisième groupe est constitué des RLS qui diffèrent considérablement des FL traditionnelles : **Prop** (propriété intrinsèque), **Strat** (stratégie), **Mes** (mesure), **Mod** (modèle théorique), **Caus** (cause), **But** (but), **Fonc** (fonction de quelqu'un), **Util** (utilité, sert à), **Result** (résultat). En effet, il ne s'agit pas ici des relations sémantiques proprement dites mais plutôt des classes de RLS retenues pour décrire les liens conceptuels propres au domaine de la distribution. Voici quelques exemples extraits du *DAD* dans sa version électronique.

Util (échantillon) = lancement
But (loterie) = promotion des ventes
But (analyse situationnelle) = évaluer ses forces, évaluer ses faiblesses, évaluer ses menaces
Mes (taxe sur la valeur ajoutée) = taux de taxe
Fonc (gestionnaire de bail) = assortiment de commerces, crédit-bail, grande surface, location

Prop (grande distribution) = volume de ventes
Prop (grande surface) = politique de prix réduits
Caus (commande en souffrance) = rupture de stock
Result (segmentation du marché) = niche

Figure 23. Modèle des RLS dans le *DAD*, consulté le 12.04.2014

Comme nous pouvons le constater, ces RLS fournissent des informations sur les concepts et se basent sur les connaissances d'experts et non pas sur les propriétés linguistiques des termes, c'est-à-dire leur structure actancielle.

En ce qui concerne les quatre RLS syntagmatiques, elles ont été proposées afin de rendre compte des relations que les termes vedettes (dans la plupart des cas, représentés par des noms) entretiennent avec des verbes ou des adjectifs. Là aussi, les auteurs ont décidé de rejeter la formalisation et la codification sophistiquées adoptées par Mel'čuk en introduisant seulement trois RLS décrivant les relations terme-verbe : a) $\mathbf{V_{deriv}}$ de type dérivationnel ; b) **V** de type collocationnel ; c) $\mathbf{V_{ass}}$ verbe correspondant à une action typique associée à un concept ; et une très rare RLS **A** pour la relation dérivationnelle verbe-adjectif.

$\mathbf{V_{deriv}}$ (palette) = palettiser
$\mathbf{V_{deriv}}$ (palette) = palettisation
V (marque) = déposer
$\mathbf{V_{ass}}$ (loterie) = tirer au sort

Figure 24. Relations dérivationnelles dans le *DAD*, consulté le 12.04.2014

Même si le modèle de RLS n'est pas aussi rigoureux que celui des FL de Mel'čuk (comme nous le voyons dans la figure ci-dessus (Figure 24), la même RLS, comme en l'occurrence $\mathbf{V_{deriv}}$ sert parfois à la description de deux phénomènes dérivationnels distincts: verbalisation et nominalisation), et la description des relations syntagmatiques y est quelque peu négligée,

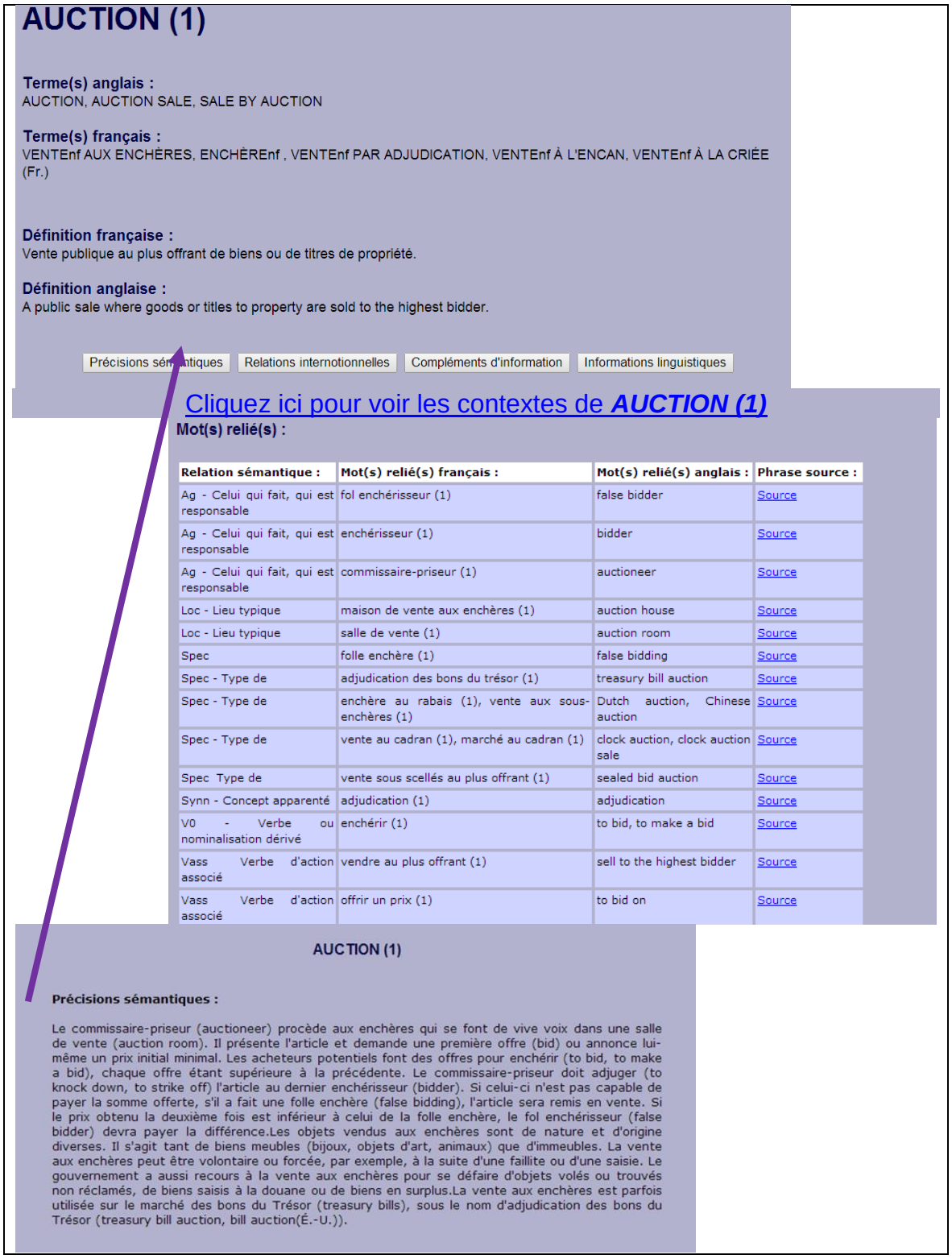


Figure 25. Article *auction* dans le *DAD* version électronique, consulté le 12.04.2014

il constitue un outil original de structuration des connaisssances et de recherche de l'information. Bien adapté à des fins terminologiques, il permet de consigner toute sorte de relations en mettant en évidence des liens de deuxième ou de troisième degré et en montrant

des structures hiérarchiques complexes. Voici le fameux article *auction* présenté plus haut au format papier, et repris ci-dessous dans sa version électronique.

En effet, le *DAD* version électronique propose deux modes de consultation. En cliquant sur les rubriques *Précisions sémantiques*, *Relations internotionnelles*, *Compléments d'information*, *Informations linguistiques* ou *Contexte*, l'utilisateur accède aux données sous forme de descriptions en langue naturelle. En revanche, le tableau *Mots reliés* fournit des explications, en langage semi-formel, sur les relations entre termes. Selon Dancette (285 : 2011), ce passage (possible grâce aux hyperliens) des textes aux représentations semi-formelles et vice-versa s'avère nécessaire pour la compréhension humaine.

« (…) l'utilisateur (humain) a besoin de s'appuyer tant sur les termes que sur les phrases qui les mettent en contexte quand il cherche à augmenter sa compréhension des concepts. Autrement dit, un texte descriptif d'une notion est parfois insuffisant pour faire comprendre les nuances et les rapports de sens entre termes ; et inversement, une liste de termes reliés est peu parlante, même lorsque les relations sémantiques sont explicites. »

(Dancette 2011b : 289)

Le modèle de RLS permet ainsi de saisir la structure conceptuelle du domaine et facilite l'appréhension des nuances de sens alors que les textes (le langage naturel étant le véhicule privilégié de la communication), offrent à l'utilisateur le moyen d'étendre et de valider l'information consignée par le terminologue (Dancette 2011b :285). Le choix d'étiquettes plus parlantes et plus compréhensibles que celles des FL ainsi que la possibilité de naviguer entre les phrases du corpus et les tables des termes associés font du *DAD* un dictionnaire explicite et facile à utiliser.

2.4.2.2 Le *DAMT*

Le *Dictionnaire analytique de la mondialisation du travail*, entrepris en 2004, constitue l'étape suivante dans le travail de formalisation des liens conceptuels. Comme nous pouvons le lire dans sa présentation en ligne (Dancette 2010, http://zedamt.herokuapp.com/more, site consultée le 26.05.2015), le *DAMT* présente, sous forme de mini-articles encyclopédiques, les concepts liés à la mondialisation économique et

sociale en touchant à des disciplines multiples : la sociologie, le droit, l'économie, le commerce et la gestion. Ces articles sont rédigés en anglais, espagnol et français mais restent indépendants les uns des autres. Chaque article est organisé en rubriques distinctes qui offrent les informations suivantes : les équivalents intralinguistiques (synonymes et variantes) et interlinguistiques, les définitions, les descriptions des concepts, les contextes, les sites Web officiels, les références bibliographiques en ligne. La dernière rubrique comporte un tableau des relations sémantiques RS[33] permettant de découvrir le réseau conceptuel du terme vedette.

La différence entre ce nouveau modèle et le modèle de RLS exploité dans le *DAD* consiste à faire la distinction entre les classes de relations sémantiques au sens strict du mot et les classes de relations associatives. Les premières ont été utilisées pour rendre compte des liens hiérarchiques (**Quasi-Synonyme, Générique, Spécifique**, **Contrastif,**) et méronymiques (**Multiple, Singulier, Partie, Totalité**), ainsi que des relations syntagmatiques (**Verb/Nom**, **Adj/Nom**). En revanche les relations associatives (**Acteur/Action, Facteur/Résultat, Instrument/Réalisation, Législation, Lieu typique, Concept proche, Objet, Propriété, Quantificateur**) ont été réservées pour décrire les liens conceptuels propres au domaine de la mondialisation du travail.

Comme le souligne Dancette (2011a : 169), à l'exception des relations indiquées comme **Concepts proches**, qu'il est impossible de qualifier de manière formelle (les liens entre les concepts étant trop instables ou éloignés), toutes les relations sémantiques sont définies et étiquetées. Il s'agit d'une caractéristique héritée des fonctions lexicales de la Lexicologie Explicative et Combinatoire déjà exploitée dans le *DAD*. Cependant, à la différence du modèle des FL et de RLS où les relations établies étaient unidirectionnelles, les relations sémantiques du *DAMT* ne répondent pas toutes à ce principe de directionnalité. En effet, comme nous pouvons le constater plus haut, les relations hiérarchiques et méronymiques relèvent du modèle des FL où le générique conduit au spécifique, le spécifique au générique, le singulier au collectif, etc. En revanche, les relations associatives expriment des rapports de réciprocité, ce qui correspond plutôt à une méthode ontologique ou thésaurale. Ainsi la relation associative **Legislation** renvoie à la fois à un terme désignant une norme, une loi ou une règle ayant effet sur qqch ou qqn et au légiféré, c'est-à-dire à une entité affectée par ce document (2011a : 176). Il en va de même pour les autres classes de relations associatives,

[33] Les RS comprennent aussi bien des relations sémantiques proprement dites que des relations d'ordre conceptuel.

de sorte que toute étiquette correspond à une paire de termes. D'après Dancette (2011b : 294-295), ce principe de bidirectionnalité présente l'avantage de forcer la cohérence. Cependant, l'efficacité de la méthode dépend du bon sens et de la bonne volonté de l'utilisateur, qui doit faire preuve d'une certaine souplesse.

Nous reprenons ci-dessous quelques exemples de relations associatives extraites du *DAMT*.

TERME VEDETTE	RELATION	TERME ASSOCIÉ
Col blanc	**Acteur/Action**	Travail de bureau Travail intellectuel
Foyer transnational	**Facteur/Résultat**	Migration transnationale
Équité en matière d'emploi	**Facteur/Résultat**	Discrimination positive Rémunération égale pour un travail d'égale valeur
Travailleur contractuel	**Législation**	Contrat de travail à durée déterminée
Barrière commerciale	**Instrument/ Réalisation**	Protectionnisme
Flux migratoire	**Quantificateur**	Migration transnationale Travailleur migrant

Tableau 7. Exemples de relations associatives extraites du *DAMT*, consulté le 20.09.2014.

Même s'il est vrai que l'utilisateur est obligé de se baser sur son intuition linguistique ou bien se fier à une sorte d'intuition intellectuelle pour associer un type de relation à un terme donné, le modèle de RS permet une systématisation des liens conceptuels. Le nombre très élevé de relations associatives repérées ainsi que le système de balisage au moyen des étiquettes distingue le *DAMT* aussi bien des dictionnaires de spécialité classiques que des thésaurus qui certes, rassemblent un grand nombre de concepts, sans toutefois préciser et catégoriser les types de relations qui les unissent. Comme le remarque Dancette (2011b : 296), l'implantation des relations sémantiques dans un dictionnaire spécialisé multilingue est un défi pour un terminographe. Nous pouvons ajouter, d'après l'auteur, qu'elle représente une voie nouvelle en terminologie. Un modèle qui permet de refléter l'univers conceptuel d'un domaine en exploitant l'étiquetage des liens entre les unités terminologiques constitue une proposition importante et originale. C'est un outil efficace qui facilite l'organisation et l'extraction des données et favorise l'apprentissage des concepts et du vocabulaire spécialisé.

Le balisage de l'information au moyen des relations sémantiques (comme c'est le cas des modèles de RLS ou RS) est un moyen d'augmenter le potentiel de transfert des connaissances.

Pour résumer, il convient de souligner que chaque projet évoqué dans ce chapitre propose un autre regard sur les possibilités offertes par le formalisme des FL. Tandis que les terminologues de l'OLST, en adoptant les principes de la LEC, décident de se détacher radicalement de l'optique conceptuelle, Dancette ainsi que Mortchev-Bouveret se heurtent à des difficultés d'ordre méthodologique en essayant d'intégrer la dimension conceptuelle dans leur description. En effet, comme le souligne L'Homme (2002 : 40), le modèle des FL demande que l'on se dégage du plan descriptif purement conceptuel. Le terme doit y être considéré comme unité lexicale qui véhicule un sens spécialisé et non comme étiquette de concept. Par conséquent, les relations ne sont plus établies entre des concepts (comme c'est le cas dans les modèles ontologiques), mais entre des unités lexicales. De plus l'application des fonctions lexicales requiert la définition linguistique rigoureuse entraînant l'identification de tous les actants.

Cependant, comme le remarque Mortchev-Bouveret (2007 : 314), l'unité terminologique, de par sa nature, est prise entre propriétés linguistiques et référentielles et il est parfois difficile de proposer sa description uniquement en termes de constituants sémantiques. Elle propose donc d'adapter le modèle des FL pour aboutir à une représentation médiane qui d'un côté décrit les relations sémantiques et d'autre côté met en évidence des liens conceptuels. Cependant, une telle démarche risque de dénaturer le système des FL. Quant à la proposition de Dancette qui s'inspire du modèle des FL pour extraire un nombre fini de classes de relations lexico-sémantiques permettant de représenter l'ensemble des relations entre les termes, elle constitue une piste de réflexion intéressante. En effet, son système de liens dérivés de relations sémantiques permet de décrire à la fois les propriétés linguistiques des termes et de structurer le champ conceptuel du domaine. Dans le cadre de ce travail, nous espérons pouvoir contribuer de manière constructive à ce débat sur la compatibilité du modèle mel'čukien avec un projet terminographique. Mais avant cela, nous proposons d'aborder un autre aspect théorique important du point de vue de cette étude, notamment le rôle du corpus dans la recherche terminologique.

DEUXIÈME PARTIE : Le corpus en linguistique et en terminologie

Le projet *DITerm* se fixe comme objectif de proposer un modèle dictionnairique orienté vers la mise en discours. Il s'agit donc d'une ressource dont l'ambition est de rendre compte des usages observés dans le discours du domaine du droit de l'Internet. Cela situe notre démarche dans un cadre purement descriptif et nous amène à nous tourner vers les « *réalités matérielles accessibles et analysables que constituent les textes spécialisés*. » (Béjoint et Thoiron 2000 : 15). Comme le remarquent L'Homme et Vanaele (2007 : 3), le rapprochement entre la lexicographie et la terminologie qui s'est amplifié au cours des dernières années est dû, entre autres, à une évolution parallèle du rôle des corpus dans chacune de ces disciplines ainsi qu'au développement des outils informatiques. Selon les auteures (*ibid.*) : « *les techniques informatiques forcent les terminologues et les lexicographes à envisager les mots et les termes dans leur environnement linguistique et à appuyer leurs décisions sur ce qui peut s'y observer.* ». En effet, basée sur l'observation et la description des faits réels de la langue, la linguistique de corpus a apporté un vrai changement de perspective en influençant les études menées dans toutes les branches de la linguistique et bien évidement en lexicologie et en terminologie.

« La linguistiques de corpus s'inscrit dans une certaine conception de la langue et des objectifs même de la linguistiques, elle prend sens dès lors que l'on pense la langue non comme UN système désincarné et abstrait mais comme un ensemble vivant, peut-être multiforme, où la description de la variation et de la multiplicité des usages peuvent être aussi fructueux pour la découverte des règles que les raisonnements sur les possibles et les impossibles. »

Jacques (2005 :27)

C’est pourquoi, il nous paraît important de consacrer cette partie du travail à la présentation des différentes approches du corpus en linguistique. En effet, si l’on retrace l’histoire de la linguistique, on peut constater que le corpus (compris ici au sens large comme recouvrant toutes sortes de données linguistiques), a presque toujours été présent dans l’étude de la (des) lange(s). Néanmoins, son statut diffère d’un courant linguistique à l’autre. En effet, il existe différentes conceptions de la notion de *corpus*. Les linguistes ont différentes attitudes à l’égard des données linguistiques et ils développent différentes démarches pour élaborer leurs corpus. Ils portent également différents jugements sur les résultats du traitement des données linguistiques.

Dans les pages qui suivent, nous proposons donc de parcourir les différentes étapes de l’histoire de la linguistique de corpus en évoquant les chercheurs qui ont le plus contribué à son développement. Ensuite, nous nous intéresserons aux principaux apports des corpus à la terminologie. Nous évoquerons ici la terminologie textuelle, une approche qui est le fruit du rapprochement de la linguistique (linguistique de corpus et analyse du discours) et de la terminologie. Finalement, nous ferons un bilan sur la notion de *corpus* et présenterons le point de vue que nous avons adopté dans le cadre de ce travail en exposant la démarche retenue pour constituer notre corpus.

Chapitre 3. La linguistique de corpus

L'histoire de la linguistique de corpus se divise en deux périodes. La première période nommée « early corpus linguistics » (McEnery et Wilson 1996 : 1-4) va jusqu'aux années 50 du XX siècle et englobe tous les travaux basés sur l'observation du corpus menés par les linguistes avant l'avènement de Chomsky. La deuxième période qui démarre dans les années 60 du dernier siècle, correspond au développement de la linguistique de corpus moderne. Cependant, comme le souligne Geoffrey Williams (2005 : 13), c'est au moment de l'arrivée sur le marché d'ordinateurs personnels, c'est-à-dire dans les années 80 et 90 que la discipline a vraiment pris son essor. Malgré son impopularité dans les années 60 et 70 (due aux travaux de Chomsky), la linguistique de corpus a réussi à s'imposer dans différentes branches des sciences du langage. En tant que discipline relevant de la linguistique appliquée qui cherche à observer et à comprendre les mécanismes de la communication et à apporter des solutions à des questions pratiques (Williams 2006 : 13), elle s'est fait une place dans l'enseignement des langues, la traduction et (ce qui nous concerne plus directement), dans la lexicologie et la terminologie. D'autre part, vue par certains comme une théorie et non simplement une méthodologie (nous pensons notamment aux héritiers de l'école contextualiste, par exemple Tognini-Bonelli 2001), la linguistique de corpus tente de proposer une nouvelle approche de la langue.

3.1 *Early corpus linguistics* – les premières études basées sur l'observation du corpus

Comme le remarquent McEnery et Wilson (1996 : 2-3), les premiers exemples de l'utilisation des corpus datent du XIX° siècle. Cependant, à cette époque-là, les chercheurs n'utilisent pas encore le terme *linguistique de corpus*. Les données recueillies sous forme de corpus primitifs (McEnery et Wilson 1996 : 2), étaient considérées comme de simples outils d'investigation permettant d'étudier différents faits linguistiques dans leurs contextes réels. Les études basées sur l'observation des corpus ont été menées dans différentes branches de la

linguistique telles que : acquisition du langage, enseignement des langues étrangères, linguistique comparée, phonétique et phonologie, morphosyntaxe, syntaxe, sémantique. McEnery et Wilson (1996 : 2-3), mentionnent ainsi les travaux consacrés à la description du développement du langage chez les enfants. Ils évoquent entre autres : le cahier d'observation de Preyer (1889), les journaux des Stern (1907, 1924), les études transversales basées sur un large échantillon d'enfants (McCarthy 1930, Day 1932), ou bien de plus récentes études longitudinales sur un groupe d'enfants de Brown (1973).

Dans le domaine de la linguistique appliquée, McEnery et Wilson évoquent, un corpus impressionnant de par sa taille (11 millions de mots), celui du linguiste allemand Käding (1897-1898), réalisé grâce à la collaboration de nombreux analystes. Le but de ce travail était d'étudier la fréquence d'apparition des lettres et des séquences de lettres en allemand afin d'améliorer les compétences des sténographes. Il faut souligner que la taille du corpus (qui peut être comparée à celles des corpus modernes) ainsi que l'envergure du projet étaient révolutionnaires pour son époque. De l'autre côté de l'Atlantique, Edward Thorndike (1921), psychologue et pédagogue américain a compilé un corpus de 4,5 millions de mots afin de recenser les mots les plus courants de l'anglais sous forme d'un dictionnaire en proposant de baser l'enseignement du vocabulaire sur ces listes de fréquences. A la même époque, le linguiste britannique Palmer (1933) a entrepris des études du lexique dans le but de créer un vocabulaire contrôlé pour l'apprentissage de la langue anglaise. Comme le souligne Williams (2006 :153), il s'est surtout intéressé aux phénomènes phraséologiques, notamment aux combinaisons figées et semi-figées qu'il a appelées des « comings-together-of-words », des rassemblements de mots. Ses travaux ont conduit à la rédaction du fameux dictionnaire de langue générale pour apprenants, le Learner's Dictionary of Current English de Hornby.

En ce qui concerne la grammaire descriptive, on peut également citer de nombreux exemples des travaux basés sur l'observation des corpus. Tognini-Bonelli (2002 :51) évoque par exemple les études de Bréal, linguiste français du XIX° siècle, considéré comme fondateur de la sémantique : « […], in the work of Bréal, the study of language was simply equated with the *observation of data* ; the laws that governed the historical development of meaning could only be discovered by looking at specific and observable phenomena. ». Selon Bréal, le sens doit être étudié à travers les faits linguistiques observables. L'observation du corpus est essentielle car elle permet d'analyser les distinctions sémantiques que font les

locuteurs impliqués dans des activités langagières. Évoquons à ce propos un extrait de *l'Essai de sémantique* de Bréal (1897 :27), cité dans Tognini-Bonelli (2001 :167) :

«Il n'y a de bonnes distinctions que celles qui se font sans préméditations, sous la pression des circonstances, par inspiration subite et en présence d'un réel besoin, par ceux qui ont affaire aux choses elles-mêmes. Les distinctions que fait le peuple sont les seules vraies et seules bonnes. »

Dans cette époque, nous pouvons également mentionner les travaux de Jespersen, linguiste danois qui a étudié la grammaire de la langue anglaise en s'appuyant sur des textes journalistiques et littéraires.

Les travaux des structuralistes américains s'appuient également, sur l'observation des données linguistiques. Il est nécessaire de souligner que la notion de corpus est fondamentale dans la linguistique structurale. Cet intérêt pour l'approche empirique s'explique en partie par la diversité linguistique (un grand nombre de langues amérindiennes peu décrites) à laquelle étaient confrontés les linguistes sur le continent américain. Le développement du structuralisme américain est lié à deux figures majeures : Leonard Bloomfield et Edward Sapir (ce dernier était disciple de Franz Boas, anthropologue qui a également mené des études sur le langage). Les deux s'intéressaient aux langues et cultures amérindiennes. Chacun, pour sa part, a réalisé de nombreuses recherches de terrain en menant des enquêtes auprès des locuteurs natifs. Sapir a surtout étudié le langage en tant que fait culturel.

Quant à Bloomfield, grâce aux informations extraites des corpus constitués lors de ses investigations, il a pu proposer des descriptions grammaticales (surtout au niveau phonétique, phonologique et morphologique) des langues en question. Dans ses études, il a en particulier mis l'accent sur l'analyse formelle du langage sur des bases inductives, c'est-à-dire à partir de l'observation des faits réels. En effet, le travail empirique auprès des populations indigènes lui a permis d'édifier des grammaires descriptives et d'accéder aux structures de la langue. Ses travaux ont eu de nombreux successeurs aux États-Unis. On parle même de la période post-bloomfieldienne qui a marqué les années 1940 et 1950. Parmi ces disciples, on peut compter Z.S. Harris (*Methods in Structural Linguistics*, 1951), Ch. C. Fries (*The structure of English*, 1952) et A. A. Hill. Comme le souligne Leech (1991 :8), pour ces linguistes, qui restaient

sous une grande influence du positivisme et du béhaviorisme, le corpus jouait un rôle essentiel.

« This was when linguists […] regarded the 'corpus' as the primary explicandum of linguistics. For such linguists, the corpus - a sufficiently large body of naturally occurring data of the language to be investigated - was both necessary and sufficient for the task in hand, and intuitive evidence was a poor second, sometimes rejected altogether. »

(Leech 1991: 8)

Selon Leech (1992 : 105), le terme *linguistique de corpus* n'est pas apparu à l'époque du structuralisme car pour les linguistes post-bloomfieldiens la linguistique en elle-même se résumait à l'étude des faits observables : *« […] for those who espoused this approach, corpus "linguistics" was simply linguistics – to them, no other linguistics deserved the name. »* Le recours au corpus était si évident et naturel qu'une autre linguistique n'existait pas: « *a corpus of authentically occurring discourse was* the *thing that the linguist was meant to be studying.* » (*ibid.*). Les structuralistes américains, dans leur position radicale, s'interdisaient toute démarche introspective. Le recours à l'intuition était donc méthodologiquement banni des procédures d'analyse. Ils considéraient l'activité de langage comme une activité régie par le modèle stimulus-réponse dont les caractéristiques étaient observables dans les énoncés, en rejetant ainsi la conception mentaliste du langage. Comme le souligne Tognini-Bonelli (2001 :171), leur objet d'étude se résumait, si on le définit en termes saussuriens, à la *parole.*

« In the division of scientific labour, the linguist deals only with the speech-signal (r…..s); he is not competent to deal with problems of physiology or psychology. »

(Bloomfield, 1932: 332 cité dans Tognini-Bonelli 2001 :170)

Pour les structuralistes, le sujet n'était que le siège de la parole et il ne constituait pas l'objet de la linguistique. L'individu, siège de la pensée, a été écarté de leurs préoccupations. En revanche, ils s'intéressaient au fonctionnement de la langue en tant que système, système qui existe en soi, au-dehors du sujet, et possède ses propres lois qu'il est possible d'étudier à travers l'observation des phénomènes linguistiques.

3.2 « *Corpus linguistics does not exist.* »[34]- Chomsky et le rejet de l'empirisme

Malgré de nombreux adeptes et un grand succès connu dans la première moitié du XX° siècle, la démarche structuraliste et son principe de l'observation des corpus langagiers sont tombés en discrédit. Comme le soulignent McEnery et Wilson (1996 : 4), cela est lié presque exclusivement à une seule personnalité. Il s'agit notamment de Chomsky, qui dans une série de publications ayant un énorme impact sur le milieu intellectuel à la fin des années 50 et au début des années 60, a réussi (en une période de temps remarquablement courte) à imposer sa conception du langage. En renouant avec le rationalisme classique et en faisant revivre la notion du sujet cartésien, c'est-à-dire du sujet qui est le siège de la pensée et des facultés supérieures, il s'est inscrit à contre-courant des tendances dominantes. En effet, en défendant l'idée d'innéisme linguistique selon laquelle l'individu est doté d'une faculté de langage innée constituée d'un ensemble de règles, appelée la *grammaire universelle*, il s'est radicalement opposé à l'empirisme. Comme le souligne Leech (1992: 111):

« […], with Chomsky, the pendulum swung to the opposite extreme of rationalism: it was held that children acquire much of the blueprint of their language genetically, through innate structures. These structures, which determine the kind of "grammar" that we acquire for our native language, are mental and hence non-observable. »

Ainsi, contrairement aux structuralistes, Chomsky a adopté une perspective mentaliste sur le langage, c'est-à-dire qu'il s'est intéressé aux structures mentales, cognitives dédiées au langage dans l'esprit du locuteur. Comme le rappelle Tognini-Bonelli (2001: 173) : « […] Chomsky, throughout his work, has been strongly opposed to the descriptive method ; for him "linguistic theory is mentalistic since it is concerned with discovering a mental reality underlying actual behavior" [Chomsky 1965:4]». Pour lui, la langue n'existe pas en dehors du sujet et ce qu'il faut étudier c'est le fonctionnement de la langue dans le sujet, c'est-à-dire la nature du système linguistique qui permet au locuteur de comprendre sa langue maternelle. Afin de désigner cette capacité langagière innée et universelle, Chomsky a introduit le

[34] Rastier (2005 :40) évoque ce « meurtre symbolique » du corpus en citant la fameuse phrase de Chomsky prononcée lors d'un entretien avec Baas Aarts en 1999.

concept de *compétence*[35]qui est au cœur de sa théorie, une théorie qui perçoit et étudie le langage comme une capacité cognitive. De ce fait, la linguistique est considérée par Chomsky comme une partie de la psychologie (1975).

Cependant, ce qui nous intéresse le plus du point de vue de cette étude, c'est l'attitude de Chomsky face aux corpus. Comme nous l'avons mentionné plus haut, le projet chomskyen était de caractériser la compétence, c'est-à-dire de rendre compte de la connaissance que possède tout locuteur de sa langue en étudiant la structure et le comportement de son système linguistique intériorisé. La *compétence* chomskyenne: « *speaker-hearer's knowledge of his language*» (Chomsky 1965: 4) s'oppose à la *performance*[36]*:* « the actual use of language in concrete situations » (*ibid.*). Ce dernier terme renvoie aux réalisations concrètes de la compétence (savoir linguistique intériorisé) dans des situations de communication différentes (qu'il s'agisse de réception ou d'émission). Comme le souligne Rastier (2005 :33), Chomsky renoue ainsi avec l'héritage linguistique de Humboldt et sa dichotomie *energeia* vs *ergon* :

« *Traditionnellement, le rapport entre une grammaire et les productions linguistiques qu'elle règle est conçu comme un rapport entre la* puissance *et l'*acte *(dans la tradition aristotélicienne), ou encore entre* energeia *et* ergon *(selon Humboldt qui la reprend), ou encore entre* compétence *et* performance *(selon Chomsky, qui se recommandait de Humboldt sur ce point). »*

Comme la *parole* (si on se réfère à la terminologie saussurienne), la *performance* constitue les données observables du comportement linguistique. Néanmoins, la performance n'est pas un reflet exact de la compétence car elle n'a pas son caractère d'idéalité: « *A record of natural speech will show numerous false starts, deviations from rules, changes of plan in mid-course, and so on.* » (Chomsky 1965: 4). Comme le soulignent McEnery et Wilson (1996 : 4), la performance du locuteur peut être affectée ou conditionnée par différents facteurs internes ou externes comme par exemple la mémoire, la consommation d'alcool, etc. Ainsi, pour Chomsky, le corpus ne peut pas constituer un outil de recherche sérieux car les

[35] Il faut noter que le concept de *compétence* a évolué. Dans *Knowledge of Language* (1986: 22), Chomsky a introduit un nouveau concept, *I-Language* (langage internalisé): « *some element of the mind of the person who knows the language, acquired by the learner, and used by the speaker-hearer* » qui a en quelque sorte remplacé celui de *compétence*.

[36] Le concept de *performance*, à l'instar de celui de *compétence* a évolué en laissant la place à la notion *E-language* (langage extériorisé).

données qui en sont extraites ne reflètent pas la complexité du système linguistique de l'individu. Tognini-Bonelli (2001 : 51) cite à ce sujet des propos de Chomsky (1962 :159) :

« Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list. »

Comme le rappelle Leech (1992 : 109), selon Chomsky et ses adeptes, la vraie linguistique devrait se concentrer sur l'explication de la nature profonde du langage (*language universals*). Pour ces derniers, la description du corpus basée sur l'observation des données linguistiques (si préconisée par les structuralistes), est une activité banale et superficielle. En effet, ils reprochent aux structuralistes que leur démarche descriptive n'est pas assez théorique et sérieuse car elle ne permet pas de modéliser la compétence linguistique. Comme le remarque Chomsky (1965 : 3) :

« Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech-community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance. »

Comme nous pouvons le constater dans la citation ci-dessus, le sujet chomskyen n'est pas perçu comme un être réel, concret, engagé dans une situation de communication authentique. En effet, la théorie chomskyenne s'intéresse au locuteur idéal qui possède une connaissance intuitive de sa langue et qui appartient à une communauté linguistique homogène. En effet, son système cognitif est doté d'une grammaire universelle qui lui permet de produire et de comprendre un ensemble infini des phrases jamais entendues auparavant. Comme le rappellent McEnery et Wilson (1996 :7-8), cela est possible car cette grammaire est conçue comme un ensemble fini de règles qui sont récursives, c'est-à-dire qu'elles peuvent être réappliquées un nombre infini de fois. Chomsky met ainsi l'accent sur l'aspect créatif du langage humain, sur sa capacité à innover. Pour lui, le langage est un moyen d'expression de la pensée et pas uniquement un simple moyen de communication.

Le caractère infini du langage constitue un deuxième argument de Chomsky contre l'utilisation du corpus dans l'investigation linguistique. Étant donné que le langage est considéré comme un ensemble infini et indénombrable de phrases, un corpus, aussi complexe soit-il, ne pourra pas illustrer tous les cas de figure possibles d'un phénomène linguistique. La description basée sur l'observation des faits de langue authentiques sera toujours incomplète et inadéquate. Citons à ce sujet les propos de McEnery et Wilson (1996 :7-8) qui expliquent la position de Chomsky par rapport aux corpus :

« Observing the recursive nature of phrase structure rules shows clearly how the sentences of natural language are not finite. A corpus could never *be the sole explicandum of natural language. […] Corpora, by their very nature, are incomplete. Language is non-enumerable, and hence no finite corpus can adequately represent language. »*

Ainsi, le linguiste ne devrait pas s'intéresser à la performance. La recherche d'attestations dans des énoncés est une activité vaine et inutile qui prend beaucoup de temps et n'apporte pas de résultats satisfaisants. En revanche, il devrait consacrer toute son énergie à l'explication du phénomène de la compétence en ayant recours à l'introspection : « *[…] Chomsky saw the linguist, or native speaker of a language, as the sole explicandum of linguistics* » (McEnery et Wilson 1996 : 9).

Selon Chomsky, le linguiste, de par sa compétence de sujet parlant, est capable de produire lui-même les données pertinentes sur lesquelles il va travailler par la suite. Il peut ainsi obtenir des phrases-exemples selon les besoins de son étude. Sa connaissance intrinsèque de la langue lui permet par la suite, de porter des jugements de grammaticalité (liée aux contraintes linguistiques qui constituent la grammaire) et d'acceptabilité (liée aux propriétés distributionnelles) sur l'ensemble des énoncés. Il va ainsi juger si les phrases sur lesquelles repose la démonstration d'un point théorique sont conformes à ce qu'autorise la langue. Afin de confronter et vérifier ses réactions, il peut aussi s'adresser à d'autres locuteurs natifs de la langue et connaître leurs appréciations sur les énoncés produits. En effet, dans l'approche chomskyenne, le travail du linguiste suppose le recours à l'intuition et au jugement, seul outil dont il dispose afin de démontrer ce que la langue permet et ce qu'elle interdit. Il faut savoir que les critiques de Chomsky faites à l'encontre des corpus ont poussé un grand nombre de linguistes à abandonner la démarche empirique et se tourner vers la

linguistique introspective. L'approche prônée par les structuralistes et basée sur l'observation des faits de langue authentiques a ainsi été longtemps discréditée.

3.3 Firth, la London School et la tradition empirique britannique[37]- les origines de la linguistique de corpus dans la tradition anglo-saxonne

Malgré une forte influence des travaux de Chomsky dans les années 60 et 70, certains linguistes continuaient à avoir recours à la démarche empirique. Comme le remarquent McEnery et Wilson (1996 : 11), la méthode introspective n'a jamais eu d'impact sur les recherches en phonétique ou bien en acquisition du langage où l'observation des faits authentiques est restée une principale source de données. Cependant, nous ne nous attarderons pas sur la présentation de ces travaux. En revanche, nous nous intéresserons aux représentants de l'empirisme britannique et plus précisément aux linguistes de la *London School* considérés comme les pionniers de la *Corpus Linguistics* (qui a commencé à émerger dans les années 60).

Afin de comprendre le développement de la linguistique de corpus dans le monde anglo-saxon, il est nécessaire de remonter aux origines. En effet, il faut savoir que dans la tradition britannique l'utilisation des corpus (même s'il s'agissait, dans la majorité des cas, de corpus de textes littéraires), était depuis longtemps étroitement liée aux travaux réalisés dans le domaine de la lexicographie et de l'enseignement de l'anglais comme langue seconde. Comme le souligne Léon (2008 : 14), l'intérêt pour la linguistique descriptive (ou empirique), centrée sur l'étude de l'usage, est une caractéristique largement partagée au sein des linguistes britanniques : « *C'est une science appliquée, orientée vers la pratique : enseignement des langues, traduction, confection de grammaires et de dictionnaires, […], etc.* ». Comme le remarque Williams (2006 :152), grâce à l'Empire Britannique, l'anglais était devenu une langue dominante dans les affaires. Il fallait donc que les gens apprennent la langue anglaise d'une manière plus pragmatique pour pouvoir communiquer dans le cadre professionnel.

[37] Le titre vient de l'article de Jacqueline Léon (2008), *Aux sources de la « Corpus Linguistics » : Firth et la London School*, qui retrace les origines de la linguistique de corpus.

Ainsi, on a commencé à s'intéresser à un anglais plus authentique qui permet de mieux refléter la réalité linguistique. C'est à cette époque-là que la linguistique appliquée à l'enseignement de la langue s'est beaucoup développée. Williams (2006 :152) évoque à ce propos les travaux de Sweet (*Pratical study of Languages 1989*) qui a mis l'accent sur le rôle central du lexique et de la phraséologie dans l'enseignement de l'anglais. Pour Sweet, ces éléments doivent être étudiés à l'intérieur de la phrase qui constitue un lien entre le texte et la grammaire. D'où l'importance du contexte (phrases authentiques et non inventées) dans l'apprentissage de la langue.

Williams (*ibid.*) mentionne également un autre linguiste britannique, Harold Palmer, qui a lui aussi beaucoup contribué au développement de la théorie et de la pratique de l'enseignement de l'anglais comme langue seconde (nous avons déjà cité ses travaux plus haut). En étudiant le lexique, il s'est intéressé aux combinaisons de mots et non pas aux mots isolés. Selon lui, il était nécessaire de définir (et ensuite de consigner dans des dictionnaires) ce que l'apprenant devait apprendre comme combinaisons lexicales. Comme le souligne Léon (2008 :17), Palmer a proposé d'utiliser des *construction patterns* et en particulier des *verb patterns* pour favoriser l'apprentissage des groupes de mots. Ainsi, il témoigne d'un souci d'interface entre lexique et grammaire. Il a également insisté sur le fait que ces combinaisons de mots devaient être analysées dans des textes authentiques. En effet, c'est à lui que nous devons le concept *collocation* apparu dans *Second Interim Report on English Collocations* (Palmer et Hornby 1933 cité par Čermáková et Teubert 2007 : 53).

Néanmoins, comme le précise Williams (2006 : 153), l'analyse des collocations en corpus est issue d'une autre tradition de recherche, notamment le contextualisme de Firth (1890-1960), fondateur de la *London School*[38]. Firth est souvent considéré comme le père de la collocation, même si ses écrits sont postérieurs à ceux de Palmer. Comme le remarquent Čermáková et Teubert (2007 : 53), Firth, dans son article *Models of Meaning* datant de 1957, a repris le terme *collocation* introduit par Palmer en 1933. Cependant, quelle que soit l'origine de la notion, il convient de souligner que le mouvement contextualiste, avec sa conception de la collocation, a joué un rôle considérable dans le développement de la linguistique de corpus.

[38] Comme le rappelle Léon (2008 : 14), la *London School* est constituée en grande partie des membres du Department of General Linguistics de la SOAS (School of Oriental and African Studies), d'où un certain nombre des traits communs partagés par les linguistes, commel' intérêt pour les langues non européennes, l'importance accordée à l'étude du langage parlé ou l'attachement à la linguistique appliquée, indissociable de la linguistique théorique.

En effet, l'idée de *meaning by collocation* de Firth (Léon 2007) a marqué de façon décisive l'essor de la linguistique de corpus en préconisant la prise en compte du contexte et l'utilisation des méthodes d'analyse lexicale basées sur les corpus. Il faut toutefois savoir que les pionniers de la linguistique de corpus ont surtout été influencés par l'un des derniers articles de Firth (*Models of Meaning* déjà cité ci-dessus), en négligeant ses principales contributions, notamment dans le domaine de la phonologie. Comme le souligne Léon (2007 : 405), dans son article tiré de *History of Linguistics 2005* de Douglas A. Kibee :

« […] Corpus Linguistics only addressed collocations referring to Firth by quoting very short excerpts from one of Firth's last papers written in 1957 (Firth [1957f] 1968): "You shall know a word by the company it keeps" and "collocation as actual words in habitual company" which have been repeated from paper to paper (see for example Mackin 1978; Sinclair 1991; Stubbs 199; Hanks 1996; Kennedy 1998; Tognini-Bonelli 2001) […] ».

L'originalité de la pensée théorique de Firth a consisté à s'intéresser davantage au sens. Il faut rappeler que dans l'histoire de la linguistique, l'étude du sens est généralement passée au second plan. Comme on l'a vu plus haut, les structuralistes se sont focalisés sur la morphosyntaxe, les chomskyens quant à eux, ont exploré l'aspect génératif et cognitif de la langue. Or selon Firth, la linguistique n'a pas vocation à décrire de simples faits de langue, elle doit surtout en explorer le sens. En partant du principe que l'être humain est avant tout un animal social qui ressent un besoin constant de communiquer, Firth en conclut que ce dernier est continuellement engagé dans des activités de production de sens. Comme le précise Tognini Bonelli (2001: 158), il s'agit de « *meaningful activities through which man expresses himself in order to interact with his fellow human beings and his environment* ».

Ainsi, Firth remet en cause la vision mentaliste de la langue et rejette toutes sortes de dichotomies opposant *langue et parole, compétence et performance* pour insister sur la notion de *contexte de situation.*

« I do not therefore follow Ogden and Richards in regarding meaning as relations in a hidden mental process, but chiefly as situational relations in a context of situation and in that kind of language which disturbs the air and other people's ears, as modes of behaviour in relation to the other elements in the context of situation »

(Firth 1935: 19 cité dans Williams 2006 : 153)

En effet, Firth emprunte le concept de *contexte de situation* à l'anthropologue anglais d'origine polonaise, Bronislaw Malinowski. Ce dernier est considéré comme le père du fonctionnalisme, théorie selon laquelle chaque élément de la culture a une fonction donnée dans la société. Du point de vue de l'étude de la langue, l'approche fonctionnaliste consiste en la prise en compte des éléments culturels dans l'interprétation des phénomènes langagiers. Dans cette approche, le concept de contexte recouvre un ensemble de circonstances dans lesquelles s'inscrit une situation d'énonciation, des circonstances qui sont étroitement liées entre elles et contribuent à la compréhension du sens. En effet, le sens ne peut pas être évalué en dehors du contexte de situation. D'où le nom de contextualisme. Pour Firth, le contexte de situation est une entité abstraite, c'est un «*group of categories, both verbal and non-verbal, which are considered as interrelated* » (Firth 1975 :175 cité dans Tognini Bonelli 2001: 159). Il a même proposé de classer les éléments du contexte en catégories descriptives telles que : 1) actions verbales et non-verbales des participants ; 2) objets correspondant à une situation donnée ainsi que des événements non-verbaux y liés ; 3) effets des actions verbales (voir Tognni Bonelli 2001 : 158). De ce point de vue, le texte est considéré comme partie intégrante du contexte et doit être analysé en relation avec d'autres éléments. En effet, dans l'approche contextualiste, le texte joue un rôle crucial dans la compréhension du sens car il permet d'appréhender le contexte d'une situation d'énonciation.

« We must take our facts from speech sequences, verbally complete in themselves and operating in contexts of situation which are typical, recurrent, and repeatedly observable. Such contexts of situation should themselves be placed in categories of some sort, sociological and linguistic, within the wider context of culture. »

(Firth 1957: 35 cité dans Tognini Bonelli 2001 : 157)

Ainsi, le texte est vu comme une manifestation concrète de l'activité langagière, il reflète la langue en action. Son rôle est donc lié à l'usage et selon les contextualistes, l'usage permet d'accéder au sens. Comme le souligne Léon (2008 : 16), pour Firth, le sens d'un mot est constitué par son usage et dès lors, il doit être analysé en contexte, et plus précisément comme occurrence dans un texte, dans un contexte situationnel spécifique. Afin d'appuyer sa thèse, Firth s'est référé aux travaux du philosophe Ludwig Wittgenstein en le citant à plusieurs reprises (voir Léon 2007 : 407-408). Rappelons ici le passage le plus connu :

« The placing of a text as a constituent in a context of situation contributes to the statement of meaning since situations are set up to recognize use. As Wittgenstein says, 'the meaning of words lies in their use.' (Phil. Investigations, 80, 109). »

(Firth 1968 [1957b] : 179 dans Léon 2007: 407)

Cette citation nous conduit au concept clé de la théorie firthienne (déjà évoqué ci-dessus), notamment celui de la *collocation* et plus précisément *meaning by collocation* qui a eu un impact important sur la linguistique moderne et surtout sur le développement de la linguistique de corpus. Firth conçoit *meaning by collocation* comme « an abstraction at the syntagmatic level » (voir ci-dessous Firth 1957 : 196 cité dans Léon 2008 : 16), ce qui nous renvoie directement au niveau co-textuel, c'est-à-dire à l'environnement purement linguistique d'un fait de langue qui peut être observé directement dans le texte. Firth a distingué ainsi le sens collocationnel du sens contextuel, les deux faisant partie du contexte de situation. L'importance donnée au texte (et à sa structure linéaire) montre que Firth a accordé une priorité absolue à l'étude des faits réels en contexte en clamant la supériorité des phénomènes linguistiques observables au niveau syntagmatique sur les abstractions faites au niveau paradigmatique. Selon Firth, le sens réside dans l'usage et non dans des processus abstraits de catégorisation ou de conceptualisation. Comme le souligne Tognini Bonelli (2001 : 159), pour Firth : « *Both the utterance under observation and the context in which it is embedded are observable rather than ontological or presupposed*. »

Il s'est donc intéressé aux *repeated events*, c'est-à-dire aux cooccurrences récurrentes dans le discours (voir ci-dessus Firth 1957: 35 cité dans Tognini Bonelli 2001 : 157). Comme nous l'avons souligné plus haut, d'après Firth, un être humain est constamment engagé dans un processus de communication. Ce qui est intéressant c'est que, d'un point de vue linguistique, ce comportement discursif est régulier et répétitif. Selon Firth, le locuteur agit d'une façon systématique et la langue est un vecteur de « the continuity of repetition in the social process » (Firth 1957 :183 cité dans Tognini-Bonelli 2001 :160). Il est donc possible de repérer et d'isoler des unités récurrentes (c'est-à-dire des comportements discursifs typiques) pour ensuite en tirer des conclusions théoriques et proposer des modèles de descriptions linguistiques. Dans l'approche firthienne, le linguiste doit « abstract the impersonal from the personal by regarding it as typological» (Firth 1957:188 cité dans Tognini-Bonelli *ibid.*).

Ainsi, au tout début, la *collocation* désignait la cooccurrence de deux éléments linguistiques quelle que soit leur nature. Le phénomène collocationnel ne concernait pas seulement les mots, mais aussi les syntagmes, expressions, phrases entières ou bien des unités plus petites comme les morphèmes ou les phonèmes, bref, n'importe quel fragment de texte permettant une analyse linguistique. Cependant, le concept de *collocation* a évolué dans la théorie de Firth (soulignons que ce dernier n'a jamais proposé de définition formelle du terme). Dans ses travaux ultérieurs, Firth s'est plutôt focalisé sur l'étude des relations de cooccurrence entre les mots. Analysée de ce point de vue, la notion de *collocation* renvoie au sens lexical qui, comme le rappelle Léon (2008 :15), constitue l'un des cinq niveaux où se détermine le sens selon l'approche polysystémique de Firth (les autres étant les niveaux phonétique, morphologique, syntaxique et sémantique).

Il est nécessaire de souligner que selon Firth, l'analyse linguistique devrait être polysysémique, c'est-à-dire qu'elle devrait se dérouler à plusieurs niveaux en prenant en compte toutes les dimensions du sens. En effet, Firth a mis l'accent sur le caractère complexe du langage[39]. Pour lui, ce n'était pas un système homogène, uniforme et cohérent que l'on pourrait étudier à l'aide des mêmes procédés. Il a donc proposé une division en niveaux analyse, chaque niveau correspondant à un autre aspect de la langue (phonétique, morphologique, syntaxique et sémantique) et relié à un contexte différent. L'analyse du sens doit donc être basée sur l'analyse des contextes relatifs à chaque niveau (*component function*), sachant que tous les niveaux sont interdépendants. Comme le souligne Léon (2008 : 15), le sens d'une unité à un certain niveau est la fonction que joue cette unité au niveau supérieur.

Ainsi, le sens lexical (perçu comme une des dimensions du sens), réside dans l'usage des mots en contexte, et l'analyse de ces combinaisons lexicales contribue à la compréhension globale de l'énoncé au même titre que les composants phonétiques, morphologiques ou bien les éléments extralinguistiques relatifs à la situation d'énonciation. Dans l'analyse du sens lexical qui correspond à l'étude des collocations, Firth a mis un accent particulier sur le phénomène d'« attraction mutuelle[40] » (*mutual expectancy*) qu'un mot fait porter sur un autre mot. Comme nous pouvons le constater dans les citations ci-dessous, selon Firth, la

[39] Comme le remarque Léon (2008 : 15), Firth s'opposait aux structuralistes, surtout au monosystémisme de Meillet pour lequel le langage était un système homogène « où tout se tient ».

[40] Nous avons repris la traduction française de *mutual expectancy* proposée par Longrée et Mellet (2013 : 67).

collocation ne devrait pas être perçue comme une simple juxtaposition de mots mais comme une interaction de deux éléments qui sont interdépendants :

« The collocation of a word or a 'piece' is not to be regarded as mere juxtaposition, it is an order of mutual expectancy. »

(Firth [1957f] 1968 : 181 cité dans Léon 2007 : 406)

« Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of night is its collocability with dark, and of dark, of course, collocation with night »

(Firth 1957: 196 cité dans Léon 2008: 16).

« The day-to-day practice of playing language games recognizes customs and rules. It follows that a text in such established usage may contain sentences such as 'Don't be such an ass !', 'You silly ass !', 'What an ass he is !' In these examples, the word ass is in familiar and habitual company, commonly collocated with you silly-, he is a silly-, don't be such an-. You shall know a word by the company it keeps! One of the meanings of ass is its habitual collocation with such other words as those above quoted. »

(Firth [1957f] 1968 :179 cité dans Léon 2007 : 407).

Dans la théorie firthienne, les collocations jouent un rôle essentiel dans l'interprétation linguistique car elles définissent le contexte d'un mot pivot et permettent d'établir son sens. La collocation est même considérée comme une propriété fondamentale du langage. Comme le souligne Legallois (2012 : 37), Firth était avant tout un phonologue et un spécialiste de prosodie et cette dernière qualité, selon l'auteur, explique la conception continuiste, non discrète, des faits de langue chez Firth. De plus, selon Legallois (*ibid.*), une des grandes originalités de ce linguiste a été d'étendre les solidarités entre unités lexicales aux dimensions plus abstraites des catégories grammaticales, en introduisant le concept de *colligation* :

« Grammatical relations should not be regarded as relations between words as such – between "watched" and "him" in "I watched him" – but between a personal pronoun, first person singular nominative, the past tense of a transitive verb and the third person singular in the oblique or objective form. »

(Firth 1957 : 13 cité dans Legallois 2012 : 38)

« Collocations are actual words in habitual company. A word in a usual collocation stares you in the face just as it is. Colligations cannot be of words as such. Colligations of grammatical categories related in a given structure do not necessarily follow word divisions or even subdivisions of words. »

(Firth *ibid.* : 14 cité dans Legallois *ibid.*)

Cependant, comme le souligne Tognini-Bonelli (2001 :162), à l'époque de Firth, l'analyse du sens lexical à travers l'observation des collocations ou des colligations en langue générale ne paraissait pas possible, et cela faute de moyens techniques (corpus informatisés). Firth a donc proposé de circonscrire l'étude des collocations aux langages restreints (*restricted langages*). Selon Léon (2008 : 16), l'idée des *restricted langages* était cohérente avec la conception polysystémique du langage. Rappelons que Firth insistait sur sa nature hétérogène et s'opposait à la description du langage dans son ensemble. Pour lui, l'analyse des collocations était plus facilement appréhendable et accessible au travers des textes authentiques et intégraux appartenant à des genres différents (textes techniques, scientifiques, littéraires) ou même limité à un seul auteur permettant ainsi la description des structures sous-forme de mini-grammaire ou mini-glossaire. Comme le précise Tognini-Bonelli (2001: 162) :
« Hence Firth saw the usefulness of collocation lying mainly in the description of restricted languages and in the stylistic analysis of selected texts ; while he saw that collocation was the type of meaning that characterised a variety of communication, he did not fully recognise it as a mechanism for creating textual meaning. »

Il est intéressant à noter que la pensée de Firth (et en particulier l'intérêt porté aux textes de spécialité ainsi que l'idée de collocation qui contribue à caractériser différents types de communication), s'articule bien avec des projets terminologiques et notamment avec notre étude. Cependant, comme le souligne Williams (2006 : 153), il faut savoir que Firth était un homme de son époque : ses sources étaient authentiques, mais largement littéraires (l'analyse des collocations lui ont permis de mener des études stylistiques de l'œuvre de Swinburne, poète de l'ère victorienne). Par ailleurs, les moyens techniques disponibles de son vivant n'étaient pas assez adaptés pour que sa conception puisse vraiment être mise en pratique. Ainsi, Firth était surtout un théoricien du langage. Ses idées n'ont été développées que par ses disciples, Halliday et Sinclair, considérés comme pionniers de la linguistique de corpus.

3.4 L'essor de la linguistique de corpus dans le monde anglo-saxon

La section précédente a été presque entièrement consacrée à la présentation des travaux de Firth, représentant de l'empirisme britannique, co-fondateur de la *London School* et père du contextualisme. En effet, l'analyse de l'approche firthienne nous a paru nécessaire pour mieux comprendre les origines de de la linguistique de corpus. Comme le souligne Williams (2005 :13), C'est une discipline qui est largement issue de la tradition contextualiste ; elle s'intéresse à la langue en contexte sous la forme de grands ensembles de textes, les corpus. Dans cette partie de notre travail, nous proposons donc d'analyser les travaux des héritiers de Firth, représentant de la deuxième et de la troisième génération de linguistes de la *London School* (M.A.K. Halliday et John Sinclair d'une part, Randolph Quirk et son disciple Geoffrey Leech, d'autre part), qui sont considérés par la communauté comme des pionniers de la linguistique de corpus.

Rappelons que la discipline a connu un essor considérable dans les années 80 et 90, lié à l'arrivée sur le marché des ordinateurs personnels permettant un recours généralisé aux corpus informatisés et un développement de méthodes d'investigation linguistique de données textuelles à grande échelle. D'après certains auteurs (comme Léon 2008, Leech 1991), c'est la parution de l'ouvrage collectif *Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research* (Aarts et Meijs :1984) qui a marqué un tournant symbolique dans le domaine en donnant de la visibilité à toute une communauté des linguistes travaillant sur des corps informatisés. Il est nécessaire de souligner que malgré l'importance de l'ordinateur, pour la plupart des linguistes de corpus, la linguistique de corpus reste une discipline (ou une méthodologie ?) des sciences du langage et non de l'informatique. Fondée sur l'utilisation de grands corpus informatisés de données authentiques et de méthodes probabilistes, elle se distingue de l'Ingénierie linguistique ou de l'Intelligence Artificielle (cette dernière étant plutôt liée à la grammaire générative), courants logico-déductifs qui font appel à l'utilisation de règles, d'inférences logiques, et de bases de connaissance.

Cependant, comme le souligne la majorité des auteurs cités (Williams 2005, Léon 2008, Tognini-Bonelli 2001, Rastier 2005), la linguistique de corpus ne constitue pas un domaine homogène et unifié. En effet, on peut y distinguer deux courants : *corpus-based* dirigé par Geoffrey Leech et *corpus-driven* constitué autour de John Sinclair (Léon 2008 :12). Même si les représentants des deux mouvements (issus tous deux de la tradition britannique) se côtoient et collaborent en utilisant parfois les mêmes corpus ou les même méthodes d'investigation, la distinction entre les deux approches est bien présente (bien que certains auteurs comme Léon (*ibid.* :13) se posent la question de savoir s'il s'agit d'une distinction ayant des fondements théoriques ou bien d'un choix d'ordre purement conjoncturel). En effet, *corpus-based* signifie une approche déductive qui fait appel aux données textuelles afin de confirmer des hypothèses théoriques tandis que *corpus-driven* fait référence à une approche inductive qui explore les données sans *a priori*. On doit chercher l'origine de ces divergences méthodologiques et théoriques à la deuxième et troisième génération de linguistes de la *London School*, notamment chez M.A.K. Halliday, Randolph Quirk, John Sinclair et Geoffrey Leech. En effet, il faut savoir que les idées de Firth ont été différemment appréhendées et réinterprétées par ses héritiers. Tandis que les travaux de Sinclair s'inscrivent dans la continuité de la pensée firthienne (ses choix méthodologiques et théoriques se situent directement dans la tradition contextualiste), Leech ne fait pas ouvertement référence à l'approche contextualiste (ses positions théoriques ayant subi des influences différentes).

3.4.1 Le courant « corpus-based linguistics »

La différence entre l'approche contextualiste *corpus-driven* et d'autres méthodologies réunies sous le nom commun de *corpus-based approach* a été décrite en détail par Tognini-Bonelli (2001), elle-même appartenant au courant Sinclair. D'après l'auteur (*ibid.* : 65), en fonction de l'approche, le corpus peut être utilisé soit pour valider ou illustrer une hypothèse théorique, soit pour construire une théorie concernant la langue. Ainsi, le terme *corpus-based* renvoie à une démarche qui consiste à exploiter les données textuelles en vue d'étudier, expliquer, prouver, confirmer ou nier les théories linguistiques déjà existantes. Le corpus y est considéré comme un réservoir d'exemples permettant de vérifier ou illustrer les suppositions théoriques ou bien de fournir des informations supplémentaires concernant le système linguistique dont les règles sont déjà bien définies. Comme le souligne Tognini-Bonelli (*ibid*.), la relation entre les données et la théorie dans l'approche *corpus-based* est assez

classique. Les linguistes analysent le corpus à partir des modèles préétablis qu'ils jugent pertinents et appropriés. Afin de valider une hypothèse, ils passent au tamis les données en gardant celles qui correspondent à leurs recherches.

« We could say, therefore, that corpus-based linguists adopt a 'confident' stand with respect to the relationship between theory and data in that they bring with them models of language and descriptions which they believe to be fundamentally adequate, they perceive and analyse the corpus through these categories and sieve the data accordingly »

Tognini-Bonelli (2001 : 66)

Ainsi, le corpus sert à valider un phénomène linguistique en termes quantitatifs, il permet également d'apporter certaines modifications au modèle adopté préalablement mais il ne joue pas de rôle déterminant dans la définition de nouvelles catégories linguistiques.

« In this case, however, corpus evidence is brought in as an extra bonus rather than as a determining factor with respect to the analysis, which is still carried out according to pre-existing categories; although it is used to refine such categories, it is never really in a position to challenge them as there is no claim made that they arise directly from the data. »

Tognini-Bonelli (*ibid.*)

En effet, selon Tognini-Bonelli, dans l'approche *corpus-based*, il n'y a pas de place pour la découverte de nouveaux paradigmes. La relation entre un élément étudié et son contexte n'est pas systématiquement prise en compte. D'un point de vue théorique et méthodologique, il existe un vide entre les positions conceptuelles préexistantes et les résultats de l'observation des données.

« A 'received' theoretical statement pre-exists corpus evidence. It might be based on no textual evidence at all, or on some, but almost certainly on less comprehensive and representative evidence than that provided by corpus. This fact may be at the root of a mismatch often observed between received categories and corpus data. »

Tognini-Bonelli (*ibid*)

Cependant, un problème apparaît lorsque les données extraites du corpus ne correspondent pas aux modèles préétablis. D'après Tognini-Bonelli (2001 :68-81), face à ce type de

difficultés, les tenants du courant *corpus-based* ont développé trois stratégies qu'elle appelle respectivement : *isolation, standardisation, instanciation*.

3.4.1.1 Isolation - les travaux du groupe des linguistes de corpus de l'Université de Nimègue

La première stratégie consiste à isoler les données qui ne cadrent pas avec le système préétabli, c'est-à-dire à mettre à l'écart les exemples qui ne sont pas assez fréquents ou qui ne peuvent pas être acceptés par la majorité des locuteurs. Ainsi, dans cette méthode qui laisse quand même une grande place à la démarche hypothético-déductive, le linguiste observe le corpus, mais il recourt en même temps à son intuition pour porter des jugements de grammaticalité et d'acceptabilité sur l'ensemble des énoncés étudiés. Le corpus sert à pallier le manque d'intuition et à compléter certaines informations sans toutefois remettre en question les règles préétablies.

Pour illustrer ce type d'approche, Tognini-Bonelli cite les travaux des linguistes de l'Université de *Nimègue* représentés par Jan Aarts (1991) et portant sur la création d'une grammaire formelle appliquée à un corpus informatisé. En effet, Aarts oppose deux types de grammaires : *intuition-based grammar* et *observation-based grammar* en accordant la priorité à la grammaire basée sur l'intuition. On y retrouve la fameuse dichotomie chomskyenne *compétence-performance* où la compétence correspond à *intuition-based grammar* et la performance à *observation-based grammar*. Selon la démarche proposée par Aarts, le linguiste doit d'abord faire des hypothèses théoriques sur un aspect linguistique donné en les modélisant à travers une grammaire formelle. Ensuite, les hypothèses sont confrontées aux faits réels observés dans le corpus. Si les résultats de cette confrontation s'avèrent satisfaisants, c'est-à-dire quand les données empiriques correspondent aux modèles préétablis, *intuition-based grammar* devient *observation-based grammar* - autrement dit une grammaire attestée. Si néanmoins, le linguiste constate des divergences entre la théorie et les faits observés, il doit se fier à son intuition et non pas au corpus (bien que le corpus puisse apporter certains éclaircissements et certaines modifications). Comme le souligne Aarts (1991: 46-47) :

« […], the first version of the grammar is written on the basis of the linguist's intuitive and explicit knowledge of the language and whatever is helpful in the literature. Basically, this is an intuition-based grammar and as such it can be looked upon as an explicitation of the facts of competence rather than an account of 'language use'. I shall call the products of

competence 'grammatical sentences' and the products of language use 'acceptable sentences'. […]
Ideally, the intuition-based grammar, through its confrontation with corpus data, becomes an observation-based grammar […]. »

Il faut souligner que dans cette approche, l'analyse porte sur un aspect donné du langage, en l'occurrence la syntaxe. Il est donc plus facile de circonscrire les données d'observation.

3.4.1.2 Standardisation – Leech et les corpus annotés

La deuxième stratégie, appelée *standardisation* (Tognini-Bonelli 2001: 71-74), consiste à schématiser et à normaliser les données rencontrées dans le corpus en les réduisant à un ensemble de catégories méthodiquement décrites, donc plus faciles à aborder et à manier. La modélisation se réalise à l'aide des procédés d'annotation et permet un enrichissement du système descriptif existant. Cette méthode, tout comme celle de l'*isolation*, essaie de trouver un compromis entre l'intuition et la démarche inductive, mais d'un point de vue théorique et méthodologique, elle est considérée comme plus empirique et orientée vers le corpus. D'après Tognini-Bonelli (*ibid.* : 71), ce sont les travaux de Leech qui représentent le mieux ce type d'approche.

Rappelons que Leech, étant considéré par la communauté des linguistes comme le chef de file du courant *corpus-based* est issu de la *London School*. Pourtant, il n'a jamais ouvertement revendiqué l'héritage firthien. Il faut néanmoins noter que ses travaux ont exercé une forte influence sur le développement de la linguistique de corpus. Comme le souligne Léon (2008 :25), il a même postulé la création d'une nouvelle linguistique (une convergence de disciplines telles que : études sur corpus, études de l'anglais et linguistique computationnelle), en s'opposant successivement à plusieurs autres modèles : *« I wish to argue that computer corpus linguistics […], defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject.»* (Leech 1992:106)

Pourtant, afin de comprendre l'approche de Leech, il est nécessaire d'analyser les travaux de Randolph Quirk (représentant de la deuxième génération de la *London School*),

dont Leech a été l'élève. Tout d'abord, il faut souligner que les activités de recherche de Quirk ainsi que celles de Leech sont davantage orientées vers la grammaire que vers le lexique, ce qui explique l'absence de références directes à la théorie de Firth. Rappelons que les préoccupations principales des néo-firthiens étaient plutôt centrées sur le sens, notamment le sens lexical en contexte (*meaning by collocation*) tandis que l'objectif de Quirk est pratique ; il s'agit de concevoir une grammaire descriptive qui pourrait répondre à une demande accrue en matière d'enseignement de l'anglais comme langue étrangère. Il faut noter que dans son projet, Quirk se rattache à la tradition britannique par l'importance qu'il donne à la recherche empirique. En effet, sa grammaire est conçue à partir d'un corpus d'usage, c'est-à-dire à partir des données enregistrées ou transcrites qu'il appelle *« a corpus of natural usage »* (Quirk 1958 :166 cité dans Léon 2008 :21) ; *« a body of full and objective data »*, *« a copious body of actually recorded usage »* (Quirk 1960 : 40 cité dans Léon 2008 :21). Comme le précise Léon (*ibid.*) : *« L'objectif de Quirk est la construction d'une grammaire descriptive et prescriptive, fondée sur l'usage grammatical permettant d'obtenir des structures descriptives (*descriptive patterns*) aussi systématiques que possible afin d'établir des prescriptions nouvelles objectivement fondées. »*.

Ainsi, en 1958, Quirk commence à rassembler les données de son *Survey of English Usage*. Selon Léon (*ibid.* : 23), le projet initial était de compiler 200 textes de différents genres (textes littéraires, techniques, scientifiques, juridiques, politiques, religieux, journalistiques), de 5 000 mots chacun, pour un total d'un million de mot. Ce qui est intéressant, c'est que le corpus contient, en proportion égale, des données provenant de corpus écrits et oraux. Comme le soulignent Čermáková et Teubert (2007: 51) : « *Quirk's Survey was a mixture of spoken and written data ; there were about 500 000 words of spoken English within a total of one million words* ». Cependant, il faut savoir que conjointement à l'analyse des données attestées extraites des corpus de langage naturel, Quirk exploite d'autres types de ressources, notamment des données artificielles produites en situation expérimentale (tests de substitution). Ainsi, comme nous pouvons le constater, ses références scientifiques sont variées. On y reconnaît l'influence des structuralistes, en particulier celle de Z.S. Harris (en ce qui concerne le recours aux tests de substitution et aux informateurs) et celle Ch. C. Fries (quant à son intérêt pour les corpus oraux). L'utilisation de données d'origine diverse lui permet de faire la distinction entre l'usage observé dans les corpus, les normes prescrites par les grammairiens et l'intuition du locuteur. En effet, comme le souligne Léon (*ibid.* : 22), Quirk s'interroge beaucoup sur l'écart existant entre les croyances des locuteurs et leur usage

réel (n'oublions pas qu'à cette époque, on est en plein débat sur les notions chomskyennes de *grammaticalité, d'acceptabilité* et *d'intuition*). Cependant, devant le choix de telle ou telle forme, il tend à donner la préférence à l'usage plutôt qu'aux règles préétablies.

Le *Survey of English Usage* a été considéré comme un des premiers corpus à grande échelle, à usage général. Il faut savoir qu'il a constitué une référence et une source d'inspiration pour les linguistes intéressés par la démarche empirique et computationnelle. Pourtant, il n'a été informatisé que dans les années 80 (Čermáková et Teubert 2007 : 51). En outre, comme le remarque Léon (*ibid.*: 24), on peut présumer que le *Survey of English Usage* a servi de modèle au *Brown Corpus*[41], projet entrepris dans les années 1960 par Henry Kučera et W. Nelson Francis, chercheurs de l'Université Brown à Providence aux États-Unis et présenté comme le premier corpus informatisé dans l'histoire de la linguistique de corpus. En effet, selon Léon (*ibid.*), le *Corpus Brown* constitue une forme de l'informatisation du *Survey of English Usage* car il a été réalisé sur le modèle initialement prévu pour ce dernier. Par ailleurs, toujours d'après Léon (*ibid.* : 25), c'est sur le modèle du *Survey of English Usage* et du corpus *Brown* que se poursuivent les travaux de compilation des variation de l'anglais, avec les grands corpus développés essentiellement en Scandinavie par les élèves de Quirk : Leech de l'Université de Lancaster, Svartvik de l'Université Lund, et des Norvégiens d'Oslo et de Bergen entreprennent en 1975 la construction du *London-Lund Corpus of Spoken English* et en 1978 celle du *Lancaster-Oslo-Bergen Corpus of British English*.

Ainsi, les travaux de Leech sont déjà menés sur des corpus informatisés de taille importante.

« The Brown Corpus […] can be thought of as a 'first-generation' corpus; its million-word bulk seemed vast by the standards of the earlier generation of corpus linguistics. But this size was massively surpassed by a 'second generation' of the 1980s represented by John Sinclair's Birmingham Collection of English Text (Renouf 1984, Sinclair 1987) and the Longman/Lancaster English Language Corpus […]. And perhaps the title 'third generation' may be given to those corpora, measured in hundreds of millions of words, almost all in commercial hands, exploiting the technologies of computer text processing […] whereby huge

[41] Publié en 1967, le *Corpus Brown* est considéré comme le premier corpus informatisé. Il est composé d'un million de mots correspondant à 500 échantillons de 2 000 mots extraits uniquement de textes écrits appartenant à 15 catégories. Comme le souligne Leech (1991 :9), la priorité donnée aux données écrites est liée aux difficultés de transcription et de traitement informatique des corpus oraux. Comparé aux corpus modernes, il est de relativement petite taille mais il constitue une étape importante de l'histoire de la linguistique de corpus. Léon (2008 : 24) insiste sur la filiation entre le *Corpus Brown* et le *Survey of English Usage* en mettant l'accent sur de nombreuses rencontres qui ont eu lieu entre Quirk et l'équipe américaine.

amounts of machine-readable text become available as a by-product of modern electronic communication systems. Machine-readable text collections have grown from one million to almost a thousand million words in thirty years, so it would not be impossible to imagine a commensurate thousand-fold increase to one million million word corpora before 2021. »

(Leech 1991 :10)

Rappelons que pour Leech l'approche par corpus fondée sur l'utilisation des techniques informatisées de construction et d'exploitation des bases de données textuelles n'était pas considérée comme une simple méthodologie mais comme une nouvelle linguistique (voir la citation ci-dessus). Fasciné par les possibilités qu'offre l'informatisation, Leech place l'ordinateur au cœur de sa démarche: « *In the case of corpus linguistics, the new master is the computer, without which corpus linguistics could scarcely be having its come-back, sporting the new image and new confidence […].* » (Leech 1992: 105). Même si les méthodes de compilation et d'exploration des corpus auxquels il a recours relèvent d'une autre discipline, notamment du traitement automatique des langues[42], Leech revendique une position théorique en linguistique en parlant de *Computer Corpus Linguistics*.

En effet, dans son article de 1992 (Leech 1992 : 105 - 126), Leech décrit les fondements théoriques de cette « nouvelle discipline » en l'opposant méthodiquement à la vision chomskyenne. Il se focalise plus précisément sur quatre aspects : 1) l'analyse de la performance linguistique plutôt que de la compétence ; 2) la description linguistique plutôt que la recherche des universaux ; 3) l'utilisation aussi bien des modèles qualitatifs que quantitatifs ; 4) l'empirisme plutôt que le rationalisme. Ainsi, Leech critique Chomsky pour la priorité que ce dernier a accordé à l'analyse des potentialités d'un système déconnecté de la réalité. Il considère que la *Computer Corpus Linguistics*, qui se fonde sur l'étude des usages authentiques, permet de dévoiler et de mettre en évidence les phénomènes langagiers qui pourraient échapper à l'intuition d'un locuteur natif.

« Chomsky has made clear that his preference is in the study of I-language, even though this study inevitably relies on highly indirect, if not speculative inferences about what goes on in the human mind. However, in contrast to this, it can be argued on behalf of CCL that

[42] Comme le soulignent Habert *et al.* (1997 :3), la tradition des linguistiques de corpus a reçu dans les années 80 et 90 un appui vigoureux et inattendu de la communauté du TALN, qui a donné un nouvel essor à la constitution et à l'utilisation de corpus annotés.

language performance is abundantly observable, that its study is more obviously useful than that of competence to most applications of linguistics […]. »

Leech (1992 : 108)

Il reproche également à Chomsky d'avoir trop insisté sur l'antinomie entre performance (E-language) et compétence (I-language). Pour Leech, ces deux sphères de la langue sont complémentaires et tout l'enjeu consiste à trouver leur juste articulation, c'est-à-dire de s'intéresser à la langue dans ce qu'elle a de systématique en rendant compte des situations réelles dans lesquelles ces règles se réalisent.

« A description of an individual language, e.g. a grammar of that language, has as its domain (from an E-language point of view) all utterances in that language. It is a "lower-order-theory", i.e. a theory (or model) of a single language. A theory of language universals (as Chomsky's universal grammar) is a "higher-order" theory, applying to language in general, as a human faculty. [..] Descriptive linguistics, as commonly understood, is a linguistics concerned with a lower-order theory, a theory of language, whereas theoretical linguistics, as commonly understood, is a "higher-order" theory linguistics. Both types of linguistics are valid in their own terms, and should be regarded as mutually contributory. […] A "lower order" theory is not of lower order of importance. »

Leech (*ibid.* : 109)

Ainsi, il réhabilite la description linguistique basée sur l'observation en reconnaissant la valeur et l'utilité des méthodes quantitatives si discréditées par les chomskyens. Il rappelle, en évoquant à ce propos le terme « pseudo-procedure » utilisé par David Abercrombie (1965 :114-115 cité dans Leech 1992 :106), que l'étude de corpus telle qu'elle se pratiquait dans les années 40, 50 et 60, était considérée par ses opposants comme superficielle et peu sérieuse : *« The term "empiricism" has become something of a negative word in linguistics, where it has been associated with the rather naive behaviourist and mechanist tendencies of Bloomfield and his successors in the 40s and 50s. »* (Leech *ibid.* : 111). Selon lui, cette méfiance à l'égard de l'approche empirique (vue comme peu théorique) doit s'expliquer par l'insuffisance des outils d'investigation disponibles à l'époque. Cependant, à partir des années 80, l'accessibilité de grands corpus, le développement des méthodes statistiques et la mise à disposition de techniques informatiques d'analyse des données textuelles ont complétement changé la donne, en revalorisant la démarche empirique et en redonnant toute sa place à la

linguistique de corpus. Comme le souligne Tognini-Bonelli (2001 :71), Leech privilégie donc les informations extraites des corpus en mettant l'accent sur le principe d'exhaustivité : *« nothing will be selected in advance and nothing will be deliberately ignored as irrelevant »*.

Il faut tout de même souligner que, même si Leech a opté pour l'approche empirique (*« CCL takes us back towards the empiricist end of the spectrum, where observation contributes to theory more than theory contributes to observation. »* Leech: *ibid.*), il n'a jamais adopté une posture orthodoxe niant l'existence des règles universelles précédant toute expérience:

« The data of a corpus, more thoroughly than we have grown to expect in linguistics, are independent of the tenets of the theory they are required to test. CCL does not, however, deny the more rationalist principle that the way we construct our theory determines the way we categorise and interpret our data. » (Leech *ibid.* : 111).

Comme nous pouvons le constater dans la citation ci-dessus, d'après Leech (et malgré l'importance accordée aux données attestées), la théorie préexiste à la description linguistique et guide en quelque sorte la catégorisation des informations extraites des corpus. Selon l'auteur (2001 :71), Leech établit une distinction nette entre *« the model »* et *« the quantitative values of the model »* et sa démarche consiste à vérifier l'adéquation du modèle préétabli avec la réalité linguistique. Pour ce faire, il fait appel aux procédés d'annotation de corpus qu'il définit comme *« the practice of adding interpretative linguistic information to a corpus. »* (Leech 2005 : 17). Ainsi, l'annotation constitue un acte interprétatif et exige le recours à l'intuition, aux connaissances et à l'expérience. Le linguiste part d'un modèle abstrait pour aboutir à une représentation formelle des réalités linguistiques. On peut dire que l'annotation permet de présenter les intuitions sous forme codifiée facilitant ainsi l'analyse des faits réels. Son objectif est de proposer un corpus enrichi. Pour Leech (2005 : 17), l'enrichissement d'un corpus par étiquetage ou parsage constitue une plus-value : « adding annotation to a corpus is giving "added value" ».

Si l'on retient la définition de Habert *et al.* (1997: 10), le corpus *enrichi* ou *annoté* (par opposition au corpus *nu* ou *brut*) est un corpus dans lesquels les séquences de caractères qui constituent les mots sont assorties d'autres informations : lemmes, étiquettes morpho-syntaxiques, étiquettes sémantiques, arbres syntaxiques, etc. Pour Leech, l'annotation joue un rôle essentiel. Selon lui, les corpus bruts explorés à l'aide d'un simple concordancier ne

fournissent pas suffisamment d'informations pour permettre une analyse fiable et efficace des données: *« […], the concordancer cannot tell the difference between* I *(personal pronoun) and* I *(roman numeral) ; between* minute *(noun) and* minute *(adjective) ; or between* lying *(telling untruths) and* lying *(in a recumbent posture) »* (Leech 1991 : 12). Précisons que ce procédé ne se limite pas à l'étiquetage grammatical (appelé en anglais *POS tagging*) et qu'il existe différents types d'annotation : phonétique, lexicale, pragmatique, stylistique, etc, (Leech 2005 : 17). Par ailleurs, l'annotation peut être manuelle ou automatique, c'est-à-dire effectuée par un programme (étiqueteurs, parseurs, etc.). Cependant, comme le remarquent Habert *et al.* (1997 : 9), l'annotation n'est jamais vraiment « manuelle » car des programmes spécifiques ont pour objectif de vérifier partiellement la cohérence des informations fournies et inversement, l'annotation automatique est souvent précédée ou suivie d'interventions humaines. Actuellement, il existe toute une variété de systèmes d'annotation qui sont en constante évolution afin de répondre aux besoins multiples des linguistes.

D'après Tognini- Bonelli (2001 :72), la méthode proposée par Leech peut se résumer à la formule « annotate – test – revise ». En effet, (comme nous l'avons déjà évoqué plus haut), la procédure consiste à tester l'adéquation du modèle grammatical préexistant avec les faits de langage observés dans le corpus. À cet effet, il propose de travailler sur deux sortes de corpus : un *training corpus* annoté conformément aux règles de grammaire préétablies et un *test corpus* dont l'annotation est modifiée suite à la confrontation avec des données réelles. Selon Leech, cette démarche basée sur l'adaptation du système d'annotation en fonction des résultats des tests menés sur le corpus permet d'aboutir à un modèle de grammaire empirique tout à fait satisfaisant. Il faut aussi savoir que Leech met l'accent sur l'importance de l'interaction homme-machine en évoquant une sorte de « partenariat » entre linguiste, corpus et outils informatiques. En effet, les dispositifs techniques permettent de dégager des régularités et de mesurer leur fréquence tandis que le linguiste intervient comme expert afin d'analyser et éventuellement d'ajuster les résultats : *« Such experiments appear to show that the machine can discover some, but not all, of the truth ; they provide reassurance to those, like myself, who believe that successful analysis depends on a division of labour between the corpus and the human mind. »* (Leech : 1991 : 15).

Comme le remarque Tognini- Bonelli, pour Leech, l'annotation constitue une sorte d'interface entre le chaos de la langue réelle et les règles qui la décrivent. Elle facilite la

manipulation et la catégorisation des faits observables, ce qui n'est pas toujours aisé vu le caractère parfois opaque du discours.

« *The problem that ultimately lies behind the issue of annotation is that raw data does not appear to be tractable unless it is reduced to a set of systematic parameters. Annotation, therefore, is needed as kind of interface between the chaotic imprecise and variable side of language on the one hand, and a formalisable set of parameters on the other. […] Annotation makes the structures of the language manageable and it allows processing at a level of abstraction that would not otherwise be possible using current techniques, by bringing out the likeness of like events*. »

Tognini- Bonelli (2001 :72)

Cependant, il parait nécessaire de souligner que l'annotation présente des inconvénients. En effet, elle entraîne une simplification d'informations de la description linguistique (rappelons que Tognini-Bonelli appelle la stratégie *standardisation*) ; et cette simplification provoque une perte d'informations. C'est pour cette raison que certains linguistes (comme par exemple Sinclair dont nous présenterons les travaux dans le chapitre suivant) préfèrent travailler sur des corpus bruts, non annotés.

3.4.1.3 *Instanciation* – Halliday et l'approche probabiliste (basée sur les propriétés statistiques du langage)

D'après Tognini-Bonelli, c'est la démarche proposée par Halliday et nommée *instanciation,* qui illustre le mieux la troisième stratégie. Il faut savoir que, même si l'on associe souvent les travaux de Halliday et de Sinclair en parlant du courant Halliday-Sinclair[43], leurs approches sont différentes, aussi bien du point de vue méthodologique que théorique. En effet, tandis que Sinclair est considéré comme le chef de file du courant *corpus-driven* (nous lui consacrerons le chapitre suivant), les recherches de Halliday, selon Tognini-Bonelli, se situent toujours dans le courant *corpus-based*. De plus, comme le souligne Williams (2006 : 154), c'est la grammaire descriptive qui occupe une place de choix dans les études des corpus menées par Halliday, alors que Sinclair est plutôt attiré par l'aspect lexical.

[43] Voir à ce propos l'article de Léon (2008 :17).

« Sinclair is by nature a lexicographer, whose aim is to construct the grammar out of the dictionary. I am by nature a grammarian, and my aim (the grammarian's dream, as I put it in 1961) is to build the dictionary out of the grammar. »

(Halliday 1992: 63)

Néanmoins, il est nécessaire de souligner qu'aussi bien Halliday que Sinclair revendiquent l'héritage firthien, et que leurs travaux respectifs s'inscrivent explicitement dans la tradition contextualiste. Comme nous l'avons précisé plus haut, Firth était avant tout un théoricien du langage qui a encouragé un modèle linguistique empirique sans toutefois décrire la notion de corpus, ni proposer les critères de sa constitution. Ce sont ses disciples, notamment Halliday et Sinclair, qui ont développé ses idées. De plus, il faut savoir que les deux linguistes britanniques font mutuellement références à leurs travaux. Ils ont même collaboré dans le cadre du projet OSTI : comme le rappelle Léon (2008 :18), Halliday soutient et supervise ce projet gouvernemental entrepris par Sinclair, alors jeune chercheur à l'Université d'Édimbourg qui consiste à étudier à l'aide de l'ordinateur, des *modèles* collocationnels dans des enregistrements d'anglais oral et écrit.

Ainsi, Halliday, fondateur de la grammaire systémique et fonctionnelle, est-il considéré comme un continuateur direct de l'œuvre de Firth et un des représentants les plus importants du contextualisme britannique. Son modèle vise avant tout à expliquer le fonctionnement de la langue dans le contexte d'une situation donnée (d'où l'approche *fonctionnelle*) :

« […], language operates in context. In terms of linguistic theory, we recognize this important principle by developing an 'ecological' theory of language – one in which language is always theorized, described and analysed within an environment of meaning; a given language is thus interpreted by reference to its semiotic habitat. »

Halliday (2014 :32)

En évoquant le terme *contexte*, il fait référence au concept de *contexte de situation* introduit par Malinowski et développé par Firth selon lequel les éléments culturels jouent un rôle très important dans l'interprétation des phénomènes langagiers. Cependant, pour sa part, Halliday parle de *context of culture*, défini comme le potentiel culturel d'une communauté donnée, c'est-à-dire un ensemble de ressources sémiotiques dont disposent tous les membres de cette communauté afin de générer du sens: *«The context of culture is what the members of a*

community can mean in cultural terms ; that is, we interpret culture as a system of higher-level meanings […] – as an environment of meaning in which various semiotic systems operate, including language, paralanguage […], tempo, and other systems of meaning accompanying language and expressed through the human body […]. » (*ibid.* : 32-33).

D'après Halliday, le contexte culturel se manifeste dans les différentes situations de communication (*contexts of situation*). Les individus, membres d'une communauté donnée, étant engagés dans le processus de communication (dans la production et la compréhension du sens) et interagissant dans des environnements particuliers et précis, produisent des textes. Le *texte* est compris ici au sens large du terme et correspond à toute production langagière écrite comme orale. Selon Halliday:

« The term 'text' refers to any instance of language, in any medium, that makes sense to someone who knows the language; we can characterize text as language functioning in context […]. Language is, in the first instance, a resource for making meaning; so text is a process of making meaning in context. »

Halliday (2014:3)

Ainsi, le texte est vu comme une occurrence (*instance* en anglais) du système linguistique observée dans un contexte de communication particulier. Si la langue sert avant tout à la transmission de la pensée ou de l'information (c'est-à-dire, à la production du sens), le texte doit être considéré comme un processus de production du sens dans une situation donnée. Remarquons que l'approche de Halliday est essentiellement sémantique et sociale : sémantique, parce que sa théorie est centrée sur le sens (« *Linguistics is about how people exchange meanings by 'languaging'* » Halliday 1985 :13) ; sociale, parce que l'on se place dans une situation sociale particulière. Selon Halliday, le contexte culturel détermine et définit la nature du code linguistique mais en même temps ce dernier contribue à la construction de la culture. En effet, c'est la situation sociale qui génère le discours mais en parallèle, le discours fait partie de la situation et la modifie. La production du texte (c'est-à-dire, le processus de communication), est donc un constant va et vient entre le contexte et la langue.

De plus, Halliday (2014 :4) souligne qu'il y a plusieurs types de contextes : politique, juridique, littéraire, clinique, académique, etc. et les textes peuvent appartenir à des registres différents. Chaque contexte peut être analysé à l'aide de 3 fonctions : le *champ*, la *teneur* et le

mode (respectivement en anglais : *tenor, field* et *mode*). Le champ reflète la nature de l'activité qui se déroule (y compris le contenu du texte), la teneur décrit les participants et leurs relations et finalement, le mode concerne le rôle de la langue ou d'autres codes sémiotiques dans la situation de communication. Halliday renoue ici avec l'idée des *restricted langages* de Firth (voir plus haut) en soulignant que les textes appartenant à divers registres pourront présenter des caractéristiques différentes : *« for example in predicative texts like weather forecasting, where future leaps over past and present and becomes the most frequent tense. »* (Halliday 1991 : 38). Cet aspect s'avère essentiel dans l'analyse des langues de spécialité.

Comme nous l'avons précisé plus haut, la langue fait partie du système social. Elle fonctionne dans une situation sociale pour générer du sens. Halliday a distingué trois fonctions sémantiques de la langue et les a nommées *métafonctions : idéationnelle, interpersonnelle, et textuelle.* La première, métafonction idéationnelle (en anglais *ideational metafunction*) permet aux individus de représenter le monde et d'organiser leurs expériences en catégories et en taxonomies. La métafonction interpersonnelle (en anglais *interpersonal metafunction*), quant à elle, concerne la façon dont les locuteurs établissent et maintiennent des rapports sociaux avec d'autres personnes. Pour ce qui est de la métafonction textuelle (en anglais *textual metafunction*), elle permet aux locuteurs de construire le texte, c'est-à-dire d'organiser l'énoncé en veillant à la cohésion, à la cohérence et à la pertinence du discours par rapport à la situation donnée. Il est tout de même important de souligner que dans la théorie systémique et fonctionnelle de la langue proposée par Halliday, le mot *fonctionnel* revêt deux sens. En effet, le terme renvoie aussi bien au fonctionnement de la langue dans le contexte social (au fonctionnement externe), qu'à son fonctionnement interne. La grammaire systémique et fonctionnelle définit plusieurs niveaux d'analyse (Halliday renoue ici avec l'approche polysystémique de Firth), où les différents composants (à chacun sa fonction), contribuent à la création ou à la compréhension globale du sens.

Nous pouvons ainsi aborder la deuxième caractéristique de l'approche développée par Halliday, notamment son côté systémique. Comme l'a souligné Williams (2006 : 154), la grammaire de Halliday est systémique et multi-niveaux, c'est-à-dire qu'il y a une interaction entre tous les constituants qui forment le texte, mais aussi entre le texte et son environnement. En effet, la langue, doit être considérée comme une entité. En analysant un aspect de la

langue, le linguiste doit toujours se référer à l'ensemble du système (y compris le contexte). Selon l'approche systémique et fonctionnelle (et contrairement à la traditionnelle vision compositionnelle), la langue ne se réduit pas à une collection de sous-systèmes qui peuvent être étudiés de façon isolée. Comme le remarque Halliday (2014: 20) :

« A characteristic of the approach we are adopting here, that of systemic theory, is that it is comprehensive*: it is concerned with language in its entirety, so that whatever is said about one aspect is to be understood always with reference to the total picture. At the same time, of course, what is being said about any one aspect also* contributes to *the total picture […].*

Pour Halliday, la production du sens n'est pas un phénomène linéaire mais au contraire un phénomène ayant une architecture complexe aux dimensions multiples. Parmi ces dimensions, on peut distinguer : *structure* (ordre syntagmatique), *système* (ordre paradigmatique), *stratification* (différents niveau de réalisation), *instactiation* (apparition des occurrences), *métafonctions* (présentées ci-dessus), chacune ayant ses règles et sa propre organisation. Il s'agit là des concepts-clés qui sont à la base de la théorie systémique et fonctionnelle de Halliday.

Ainsi, quand on parle de l'approche systémique, on parle d'une conception de la langue qui la voit comme un système ou des réseaux de choix disponibles pour le locuteur au niveau grammatical : *« the grammar is seen as a network of interrelated meaningful choices. In other words, the dominant axis is the paradigmatic one: the fundamental components of the grammar are sets of mutually defining contrastive features. »* (Halliday 2014:49). Le système représente donc un aspect du potentiel sémantique de la langue - *meaning potential* (Halliday 1985 : 193), et il est considéré comme un ensemble abstrait d'alternatives offertes au locuteur à un moment donné de l'énoncé (du texte). En s'intéressant à l'ordre de sélections, aux « *patterns in what* could go instead *of what »* (Halliday 2014 :22), la grammaire systématique et fonctionnelle est indubitablement paradigmatique. Pour mieux comprendre le concept de système, évoquons le cas de la négation (Halliday 2014 : 22-23). Afin de nier un événement, le locuteur doit opérer son premier choix entre la forme négative et affirmative. Ensuite, le système propose d'autres options supplémentaires: *« A text is the product of ongoing selection in a very large network of systems – a system network. »* (Halliday : *ibid.*). Au fur et à mesure que les choix progressent, on devient de plus en plus fin et précis :

« The relationship on which the system is based is 'is kind of'. […] negative clauses may be either generalized negative, like they didn't know*, or some specific kind of negative like* they never know *or* nobody knew. *Here we have recognized two paradigmatic contrasts, one being more refined than the other […]. The relationship between these two systems is one of* ***delicacy****: the second one is 'more delicate than' the first. »*

(Halliday: *ibid.*)

Il est important de souligner que le dernier choix fait dans le système est un choix lexical. En effet, l'originalité de l'approche de Halliday consiste en une non-séparation entre lexique et grammaire. Selon lui, la grammaire et le vocabulaire sont un même phénomène, mais observés et analysés sous deux angles différents.

« There is in every language a level of organization – a single level – which is referred to in everyday speech as the 'wording'; technically it is lexicogrammar, the combination of grammar and vocabulary […]. The point is that grammar and vocabulary are not two different things; they are the same thing seen by different observers. There is only one phenomenon here, not two. But it is spread along a continuum. At one end are small, closed, often binary systems, of very general application […]. At the other end are much more specific, looser, more shifting sets of features, realized not discretely but in bundles called 'words' […]. »

(Halliday 1992: 63)

Ainsi, en introduisant le concept de *lexique-grammaire*, il propose l'idée d'un continuum entre la grammaire et le vocabulaire[44]. D'un côté de ce continuum, nous retrouvons donc un système grammatical avec un nombre restreint de mots très fréquents dans les textes (comme *of* ou *a* en anglais), qui se combinent librement avec un très grand nombre d'unités lexicales et leur combinatoire peut être décrite à l'aide des catégories grammaticales. De l'autre côté, nous observons un grand nombre d'unités lexicales, chacune dotée d'un ensemble de propriétés lexico-grammaticales qui la relient au système. Cependant, contrairement aux unités grammaticales, leur combinatoire reste limitée. En effet, chaque forme lexicale a sa lexique-grammaire spécifique et associée à une série restreinte d'autres formes lexicales. Selon Halliday (1992 : 63), il est possible d'étudier le lexique-grammaire de

[44] Léon (2008 :17) souligne que l'idée d'un continuum entre lexique et grammaire était déjà présente dans la tradition britannique, notamment dans les travaux de Palmer (évoqués plus haut).

deux points de vue, soit celui du grammairien, soit celui du lexicologue, mais les deux approches demandent deux types d'analyse différents. En effet, les données lexicales sont beaucoup plus accessibles que les données grammaticales et leur observation ne nécessite pas de méthodes ni d'instruments sophistiqués. Pour comprendre cette différence, il est nécessaire d'expliquer un autre concept introduit par Halliday, notamment celui de *structure*.

Si l'on parle de la *structure*, on pense à l'aspect compositionnel de la langue, c'est-à-dire à la cooccurrence linéaire. En effet, il s'agit de l'organisation des éléments linguistiques sur l'axe syntagmatique et de leurs relations. Comme le souligne Halliday (2014 : 22) : *« Structure is the syntagmatic ordering in language : patterns, or regularities, in what* goes together with what. *System, by contrast, is ordering on the other axis: patterns in what* could go instead of *what* ». Il est nécessaire de souligner que la structure est facilement observable et analysable alors que le système, (considéré par Halliday (1985 : 194) comme « le cœur de la langue ») présente un caractère abstrait et virtuel. Cependant, les deux phénomènes sont étroitement liés car tous les choix proposés par le système et faits par le locuteur à chaque point de son discours contribuent à la formation de la structure : *« […] each system – each moment of choice – contributes to the formation of the structure. »* (Halliday 2014 : 24). Ainsi, la structure reflète le système : les opérations réalisées au niveau structurel (insertion, combinaison ou organisation des éléments linguistiques) doivent être interprétées comme une réalisation des choix faits au niveau du système. Cela vient du fait que la langue a une architecture stratifiée.

En effet, selon Halliday, la langue est composée de différents niveaux appelés *strates* où la lexique-grammaire (autrement dit le système) joue un rôle essentiel, c'est-à-dire qu'elle sert d'interface entre la pensée (et son contexte socio-sémantique) et la parole.

« We use language to make sense of our experience, and to carry out our interaction with other people. This means that the grammar has to interface with what goes on outside language: with the happening and conditions of the world, and with the social processes we engage in. But at the same time it has to organize the construal of experience, and the enactment of social processes, so that they can be transformed into wording. »

Halliday (2014:25)

Comme nous le voyons dans la citation ci-dessus, la lexique-grammaire permet d'organiser la pensée en la transformant en parole. Ce processus se réalise en plusieurs étapes. D'abord, les expériences et les relations humaines insérées dans un contexte particulier sont liées à un sens. On passe ainsi du plan contextuel au plan sémantique. Ensuite, le sens se transforme en texte : on se situe alors au niveau lexico-grammatical. L'ultime étape consiste en l'expression, autrement dit en la réalisation de la parole à l'aide des ressources physiques de l'être humain. Ainsi, un choix sémantique correspondant à une situation particulière est exprimé par une série de composants grammaticaux et lexicaux, puis réalisé phonétiquement. Nous pouvons donc distinguer 4 strates : *sémantique, lexique-grammaire, phonologie* et *phonétique,* réparties elles-mêmes en deux plans ; celui du contenu et celui de l'expression.

Dans l'approche de Halliday, la langue est multidimensionnelle et correspond à plusieurs niveaux : au texte (autrement dit à la production langagière dans sa forme orale ou écrite), mais en même temps au système (c'est-à-dire aux choix qui s'offrent au locuteur sur l'axe paradigmatique) ou à la structure linéaire de la parole. Pour pouvoir décrire ces différents niveaux et expliquer leurs relations, et notamment le rapport entre le système et le texte, Halliday a introduit le concept d*'instanciation*. Comme nous l'avons souligné plus haut, le système est un potentiel sous-jacent de la langue ; *« a meaning-making resource »* (Halliday *ibid.* : 27). En revanche, le texte est vu comme une occurrence (*instance* en anglais) du système linguistique observée dans un contexte de communication particulier et doit être considéré comme un processus de production du sens dans une situation donnée. Il ne s'agit donc pas de deux phénomènes isolés mais d'un seul phénomène analysé de deux perspectives différentes. Halliday fait ici l'analogie avec le temps et le climat :

« […] the instance-observer is the weatherman, whose texts are the day-to-day weather patterns displaying variations in temperature, humidity, air pressure, wind direction and so on, all of which ca be observed, recorded and measured. The system-observer is the climatologist, who models the total potential of a given climatic zone in terms of overall probabilities. »

(Halliday 1992: 66)

L'analyse des textes est comparée à la météorologie, c'est-à-dire à l'observation et à la description des phénomènes dont les valeurs sont instantanées et locales tandis que l'étude du système correspond à la climatologie qui regroupe et analyse des résultats obtenus sur une

longue période de temps. La relation entre le système et le texte est décrite au moyen du continuum d'instanciation (*cline of instantiation)* où le système et le texte constituent deux pôles du schéma (celui du potentiel et celui d'une occurrence particulière). Le système et le texte se rencontrent au milieu sous forme d'une situation type exprimée par une occurrence type. Le système est donc « exemplifié » à travers le texte :

« Systemic theory accepts the Saussurean concept of how the system is represented by the observed actes de parole*. But, as I see it at least, this has to be interpreted as Hjelmslev interpreted it: […] in the framework of system and process, where the process (text) instantiates the system […]. »*

(Halliday 1985 [2003]: 195)

Ainsi, il développe l'approche probabiliste de la langue, c'est-à-dire une approche qui est basée sur les propriétés statistiques de la langue.

« It had always seemed to me that the linguistic system *was inherently probabilistic, and that frequency in text was the instantiation of probability in the grammar. »* (Halliday 1991:31)

Il propose donc d'intégrer le calcul de probabilité à l'analyse des phénomènes linguistiques : *« frequency information from the corpus can be used to establish the probability profile of any grammatical system. »* (*ibid.* : 35). Cette équation entre la fréquence des occurrences dans les textes et la probabilité dans la grammaire est à la base même de la théorie de Halliday. Il considère que les faits observés dans de grands corpus de textes reflètent le système de la langue et peuvent mener à une généralisation aussi bien au niveau grammatical que lexical :

« […], the transformation of instance into system can be observed only through the technology of the corpus, which allows us to accumulate instances and monitor the diachronic variation in their patterns of frequency. » (*ibid.* : 34)

Néanmoins, comme nous l'avons déjà signalé plus haut, les données lexicales et les données grammaticales demandent des méthodes d'analyses différentes: *« […] grammar is the 'deeper' end of the continuum, less accessible to conscious attention, and this may be why the treatment of grammar (in any form) always engenders more resistance »*. Même s'il s'agit du même phénomène, les unités considérées plutôt comme grammaticales seront analysées

selon les catégories grammaticales et feront partie des systèmes fermés. En revanche, l'analyse lexicale se concentrera sur les ensembles ouverts d'unités lexicales ayant la même probabilité d'occurrence (ce dernier point sera plutôt développé par Sinclair).

Afin de comprendre l'approche probabiliste de Halliday, il est nécessaire de savoir qu'en établissant des *patterns* à partir des textes, on ne peut pas établir un phénomène de façon absolue, mais plutôt comme une tendance probable. Halliday insiste sur le fait qu'en interprétant des faits de langue à partir des données extraites du corpus, il faut toujours prendre en compte leurs propriétés statistiques : *« the meaning of negative is not simply 'not positive' but 'not positive, against odds of nine to one'.* » (*ibid.* : 33). On ne peut pas établir *a priori* que la probabilité d'apparition de telle ou telle forme sera d'ordre 0,5 : 0,5. A titre d'exemple, Halliday (1991 : 33) évoque les travaux de Svartvik sur la voix active et passive en anglais. En déterminant que la probabilité d'apparition de la forme active contre la forme passive y est à la hauteur de 0,88 : 0,12, il remarque que cette valeur peut varier en fonction du registre (la voix passive est beaucoup plus présente dans les textes scientifiques que dans les textes publicitaires). Ainsi, c'est l'analyse du corpus qui fournit ce type de données.

Cependant, bien que le corpus joue un rôle essentiel dans l'approche probabiliste, le linguiste y étudie les données toujours par rapport à un cadre théorique préétabli (*le système*). En effet, les textes permettent de fournir des informations supplémentaires concernant le système linguistique dont les règles sont déjà bien définies. C'est pour cela que Tognini-Bonelli (2001 : 74-77) définit cette approche à l'aide de l'adjectif *corpus-based* par opposition à *corpus-driven approach*, approche à laquelle nous consacrons la partie suivante de notre travail.

3.4.2 Le courant « corpus-driven linguistics »

Comme le souligne Léon (2008 : 12), la distinction entre les courants *corpus-based* (traditionnellement associé à Leech) et *corpus-driven* (constitué autour de Sinclair) est revendiquée essentiellement par les tenants de la tendance Sinclair (Tognini-Bonelli 2001, Francis 1993). Ainsi, comme nous l'avons déjà vu, Tognini-Bonelli (2001 : 85) oppose les deux courants par leur méthodologie : alors que l'approche *corpus-based* utilise le corpus comme un réservoir d'exemples permettant d'appuyer ou vérifier les théories linguistiques

déjà existantes, l'approche *corpus-driven* postule qu'aucune position théorique *a priori* ne préside aux observations sur corpus. Il est nécessaire de préciser que les tenants du courant *corpus-driven* tiennent à se démarquer de la tendance Leech afin de souligner l'aspect innovateur et révolutionnaire des travaux de Sinclair. D'après Tognini-Bonelli (*ibid.* : 84), l'approche *corpus-driven* bouleverse le statu quo au sein de la communauté en modifiant complétement l'attitude des linguistes vis-à-vis des corpus. En même temps, l'auteur déplore que la démarche sinclairienne ne soit pas suffisamment exploitée malgré le réel changement qu'elle apporte en termes de qualité d'analyse des données linguistiques.

En effet, l'originalité de cette approche consiste à mettre le corpus au premier plan en se détachant de ce que Sinclair (1991 : 2) appelle *pre-corpus-beliefs*, c'est-à-dire des suppositions théoriques préexistant à la recherche empirique. Les données extraites du corpus ne sont plus considérées comme de simples instruments permettant de tester des hypothèses abstraites et préconstruites par un mode de raisonnement typiquement grammatical. Selon Sinclair, le linguiste devrait s'en remettre complètement aux faits observables et les analyser tels quels : « accept the evidence » (*ibid.* : 4). Ainsi, le corpus est au cœur même de son modèle. Le recours à l'intuition et au jugement prôné par les chomskyens fournit, d'après lui, des résultats incomplets :

« Indeed, the contrast exposed between the impressions of language detail noted by people, and the evidence compiled objectively from texts is huge and systematic. It leads one to suppose that human intuition about language is highly specific, and not at all a good guide to what actually happens when the same people actually use the language »

Sinclair (1991 : 4)

Il ne s'agit pas de rejeter complètement le recours à l'intuition : *« The way a person conceptualizes language and expresses this conceptualization is of great importance »* (*ibid.*), car le rôle du linguiste n'est pas d'examiner les données de façon purement formelle et mécanique mais d'interpréter les faits en faisant appel à ses connaissances, son expérience et son intelligence (Tognini-Bonelli 2001 :86). Il ne s'agit pas non plus, de démontrer que toutes les catégories prédéfinies selon les méthodes traditionnelles sont erronées et inadéquates. En effet, l'objectif des tenants de l'approche *corpus-driven* est de prouver que leur démarche permet de saisir tous les phénomènes, même ceux qui échappent aux linguistes les plus compétents et les plus rigoureux. Selon eux, l'analyse des corpus basée essentiellement sur

l'étude des patrons récurrents et de la fréquence de distribution dévoile la langue (ou plutôt le langage) dans sa globalité et pas seulement dans ce qui est attendu. Le caractère surprenant de ces découvertes montre que l'intuition ne peut pas être considérée comme une source fiable.

« In our approach, the data comes first. The corpus is the major informant, providing the raw information we need in order to describe the language, and intuition is considered to be of secondary importance. Intuition may be useful to linguists in a number of ways, but for the purposes of saying exactly how language is used, it is notoriously unreliable. […] The corpus has a lot of surprises in store for us at every turn, and contains some threats to our accepted views of language, but it is the only reliable authority and must be treated with respect. »

Francis (1993 : 138)

Ainsi, les concepts théoriques développés par les linguistes doivent être en totale conformité avec les occurrences relevées dans le corpus, ils doivent refléter fidèlement la réalité : *« reflect the evidence »* (Sinclair 1991 : 4).

« There is just no reason or motivation to invent an example when one is knee-deep in actual instances. »

Sinclair (*ibid.* : 5)

L'approche *corpus-driven*, qui selon ses adeptes (Francis 1993, Tognini-Bonelli 2001), contribue à une importante amélioration de la description linguistique, exige donc de la part des chercheurs un véritable changement de pratiques. Les linguistes sont notamment amenés à observer le corpus dans son ensemble et à analyser les données *verbatim*, c'est-à-dire telles qu'elles y apparaissent. Cela exclut toute modification ou ajustement des faits observés par rapport au modèle linguistique pré défini. Cependant, comme le remarque Sinclair (1991 : 5), cette nouvelle approche du corpus est parfois difficile à accepter car elle bouleverse complétement les habitudes d'investigation :

« The stance with respect to real examples still appears to be controversial. For many applied linguists, to abandon the practice of inventing or adapting examples would mean a big change. »

Sinclair (1991 : 5)

Rappelons à ce propos les arguments d'Aarts (1991 : 45-46, vu plus haut) selon lequel, le linguiste doit d'abord faire des hypothèses théoriques sur un aspect linguistique donné pour ensuite confronter ces hypothèses aux faits réels observés dans le corpus.

D'après Sinclair, la théorie n'existe pas en elle-même, indépendamment de tout fait observable. Elle n'est pas une entité abstraite. Bien au contraire, la théorie émerge du texte, elle est induite du corpus. Comme le souligne Tognini-Bonelli (2001 : 85), la démarche méthodologique dans l'approche *corpus-driven* peut se résumer à la formule suivante : *« observation leads to hypothesis leads to generalisation leads to unification in theoretical statement. »*. Comme nous pouvons le constater, cette formule correspond très bien à l'idée de Firth (déjà évoquée plus haut), selon laquelle le linguiste doit *« abstract the impersonal from the personal by regarding it as typological»* (Firth 1957:188 cité dans Tognini-Bonelli *ibid.*. : 88). Rappelons que pour Firth le locuteur agit d'une façon systématique et que son comportement discursif est régulier et répétitif ; la langue est un vecteur de *« the continuity of repetition in the social process »* (Firth *ibid.* : 183 cité dans Tognini-Bonelli *ibid.*). Il est donc possible d'y repérer des « *repeated events »*, c'est-à-dire des unités récurrentes (ou des comportements discursifs typiques), pour ensuite fournir des modèles de descriptions linguistiques. C'est exactement la démarche que prônent les adeptes du courant *corpus-driven* (ces derniers revendiquent bien évidement l'héritage firthien). En effet, leur approche consiste à explorer le corpus en isolant des suites de mots récurrentes qui sont mises en évidence sur l'axe vertical des concordances afin d'aboutir à la théorisation des phénomènes observés dans le discours :

« The main "added value" of a corpus is this vertical dimension, which allows a researcher to make generalities from the recurrences » (Sinclair 2005)

Bien évidement, ce renversement théorique n'aurait pas été possible sans les avancées technologiques et l'apparition de nouveaux outils donnant plus de pertinence et plus de crédibilité à ce type de démarche. En effet, les nouvelles méthodes d'exploration de corpus permettent de traiter de grands volumes de données textuelles (ce qui contribue à leur représentativité) et d'obtenir des résultats objectifs et quantifiables. Cependant, comme le remarque à juste titre Williams (2006 :155), en linguistique *corpus-driven*, l'outil informatique n'est qu'une loupe permettant de mieux voir. Le but de cette approche n'est pas le développement de techniques sophistiquées de traitement automatique des langues mais

l'emploi des outils informatiques pour observer les mots en contexte. Contrairement à Leech, dont les travaux étaient orientés vers la création d'outils d'annotation et des applications typiques du TAL, Sinclair défend l'idée d'annotation zéro. Selon lui, l'enrichissement d'un corpus par étiquetage ou parsage ne constitue pas une plus-value pour la recherche. Au contraire, le recours aux corpus étiquetés ou arborés entraîne la perte d'informations. Comme le rappelle Tognini-Bonelli (2001 :73), pour Sinclair, un texte annoté syntaxiquement, c'est-à-dire transformé en chaînes d'étiquettes morpho-syntaxiques, réduit l'accès à l'information et par conséquent déforme la description des faits langagiers. D'après lui, les mots catalogués dans les classes morpho-syntaxiques perdent leur caractère distinctif sur le plan sémantique. Ainsi, l'objectif est d'explorer les corpus nus, ce qui permet de relever toutes sortes d'informations et de fournir une description globale et objective des réalités linguistiques. Cette démarche reflète très bien le postulat de Sinclair selon lequel aucune position théorique *a priori* ne préside aux observations sur corpus (alors que l'annotation est vue comme une sorte de position prise *a priori* car elle se fonde sur des données antérieures à l'observation).

Ainsi, le linguiste qui traite les données étiquetées, observe en réalité des unités isolées, unifonctionnelles et définies *a priori* sans pouvoir étudier leurs propriétés contextuelles, alors que pour Sinclair, dont les travaux s'inscrivent explicitement dans la continuité de la tradition contextualiste, seuls les mots pris dans leur contexte ont du sens.

« Any instance of language depends on its surrounding context. The details of choice shown in any segment of a text depend - some of them - on choices made elsewhere in the text. »

Sinclair (1991 : 5)

« *(…) the choice of one word conditions the choice of the next, and of the next again. The item and the environment are ultimately not separable (…).* »

Sinclair (2004 : 18)

Il est donc impossible de proposer une description linguistique d'une unité lexicale isolée, extraite de son environnement discursif, car ce qu'elle représente et signifie est déterminé par les éléments (aussi bien linguistiques qu'extralinguistiques) qui l'entourent. On y retrouve le concept firthien du *contexte* qui renvoie aussi bien aux paramètres linguistiques que situationnels et culturels d'un acte de communication. Cependant, ce qui intéresse le plus Sinclair et ses disciples est d'analyser le *co-texte*, c'est-à-dire l'environnement linguistique

d'un fait de langue, des éléments concrets de discours qui sont tous liés entre eux par une toile de relations d'interdépendances. Ces éléments linguistiques observables à travers le corpus reflètent à leur tour le contexte pris au sens large. C'est pourquoi Sinclair (1991 : 5) rejette catégoriquement le recours aux exemples inventés, car bien que plausibles, ils ne sont pas insérés en contexte : *« However plausible an invented example might be, it cannot be offered as a genuine instance of language in use »*. Le manque de co-texte rend toute analyse impossible.

Abordons maintenant les notions centrales de l'approche *corpus-driven*, notamment celles de *collocation, colligation, pattern* (*patron* en français) et finalement celle d'*extented lexical unit* (unité lexicale étendue). Comme le remarque Tognini-Bonelli (2001: 89), si l'on parle des *repeated events* (rappelons que le terme a été emprunté à Firth), on pense soit à la fréquence d'emploi des mots isolés, soit aux suites de mots récurrents ou autrement dit *segments répétés*[45]. En effet, ce que proposent les linguistes se revendiquant du courant *corpus-driven* est de regarder les mots en contexte à travers le mot-clé en contexte (*KWIC* ou *node* en anglais), c'est-à-dire d'analyser les segments répétés (suites d'unités textuelles statistiquement marquées) comportant l'unité nodale choisie. Ces segments répétés (constitués d'un mot-clé et de ses cooccurrents) sont définis ici comme *collocation*. Comme nous l'avons déjà vu dans la première partie de ce travail, le terme *collocation* tel qu'il est utilisé par Sinclair devrait être compris dans un sens plus large que par exemple chez Hausmann ou Mel'čuk[46], c'est-à-dire comme cooccurrence récurrente entre les mots - « *a frequent co-occurrences of words* » (Sinclair, 2003 : 28). Il s'agit là des patrons les plus évidents car le plus facilement repérables et quantifiables. Précisons que c'était Halliday qui a proposé une description de l'analyse de ces combinaisons :

« If we consider n occurrences of a given (potential) item, calling this item the ***node****, and examine its* ***collocates*** *up to m places on either side, giving a* ***span*** *of 2m, the 2mn occurrences of collocates will show a certain frequency of distribution. For example, if for 2,000 occurrences of* sun *we list the three preceding and three following lexical items, the 12,000 occurrences of its collocates might show a distribution beginning with* bright, hot,

[45] Habert, Nazarenko et Salem (1997 : 199), définissent *segments répétés* comme toute suite d'unités textuelles reproduites sans variation à plusieurs endroits d'un corpus. Comme le soulignent Legallois et Tutin (2013 : 9), la notion de *segments répétés* ou *n-grammes* de la linguistique informatique a évolué depuis grâce à l'extension disciplinaire. Ainsi, les auteurs évoquent les *paquets lexicaux* (*lexical bundles*) de Biber (2006) ou le concept de *motif* récemment développé par Longrée et Mellet (2013).

[46] Legallois et Tutin (2013 : 7) précisent que l'« approche continentale » des collocations, représentée par Hausmann ou Mel'čuk (2003), essentiellement basée sur la lexicologie, propose une définition plus étroite et plus formelle du phénomène.

shine, light, lie, come out *and ending with a large number of items each occurring only once. The same number of occurrences of* moon *might show* bright, full, new, light, night, shine *as the most frequent collocates. »*

(Halliday 1966:168)

Cependant, comme le soulignent Legallois et Tutin (2013 : 7), Halliday donne au phénomène une définition[47] essentiellement statistique et textuelle dans la mesure où ces expressions concourent à la cohésion textuelle. En revanche, Sinclair introduit une dimension contextuelle selon laquelle (comme nous l'avons vu plus haut) l'emploi d'un mot est déterminé par les emplois co-textuels dans lesquels il apparaît, qu'il s'agisse de l'environnement lexical, sémantique ou syntaxique. En effet, en développant la notion de *collocation*, Sinclair (1991) défend l'idée d'un *principe idiomatique* (*idiom principle* en anglais) selon laquelle les locuteurs disposent d'un stock d'expressions disponibles pour un « prêt-à-parler » (Longrée et Mellet 2013 : 67), qu'ils puisent dans leur mémoire au fur et à mesure de l'énonciation.

« The principle of idiom is that a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments. […] At its simplest, the principle of idiom can be seen in the apparently simultaneous choice of two words […]. »

Sinclair (1991 : 108)

Comme le souligne Béjoint (2007 : 138), selon ce principe, une proportion importante des discours que nous produisons, variable selon les registres, est constituée d'éléments lexicaux qui ont tendance à être utilisés ensemble dans des schémas syntaxiques récurrents, et qui forment donc des « blocs » plus ou moins figés. Ces pans de la langue préconstruits, complexes et relativement stables intègrent à la fois des éléments du lexique et de la grammaire. Ainsi, afin de distinguer les phénomènes d'associations entre mots lexicaux et grammaticaux, Sinclair oppose la notion de *collocation* à celle de *colligation*.

[47] Les auteurs (Legallois et Tutin 2013: 7), citent la définition de la *collocation* proposée par Halliday (1961 : 267) : « *Collocation is the syntagmatic association of lexical items, quantifiable, textually, as the probability that there will occur, at n removes (a distance of n lexical items) from an item x, the items a, b, c...* ».

« Colligation is similar to collocation in that they both concern the cooccurrence of linguistic features in a text. Colligation is the occurrence of a grammatical class or structural pattern with another one, or with a word or phrase. "Negative", "possessive" and "modal" are the kinds of largely grammatical categories that figure in colligation. The term was first used by J. R. Firth, and has been widened a little for corpus work. »

(Sinclair 2003 : 145 cité dans Legallois 2012 : 39)

Effectivement, comme nous l'avons vu plus haut, la notion de *colligation* a été introduite par Firth pour caractériser un phénomène d'association entre catégories grammaticales. Dans le cadre de la linguistique de corpus de Sinclair, le concept a été élargi pour rendre compte des contraintes syntaxiques ou des préférences sélectionnelles d'un mot en contexte. Le terme *colligation* renvoie alors à l'environnement grammatical privilégié d'un mot. Ce dernier y est plutôt considéré comme une unité lexicale à sens unique et non pas comme un membre d'une catégorie grammaticale comme c'était le cas chez Firth. Soulignons, d'après Tognini-Bonelli (2001 :89), que si les collocations sont facilement visualisables et identifiables sur l'axe vertical de la concordance, les colligations constituent des éléments plus abstraits, donc difficilement repérables.

Cependant, malgré leurs différences formelles, les collocations et les colligations sont intimement liées et interdépendantes et forment ensemble la base de l'analyse des *repeated events* permettant ainsi l'accès au sens. Pour les tenants de l'approche *corpus-driven,* il existe une forte corrélation entre les propriétés sémantiques et les régularités syntaxiques. Les mots ayant leurs préférences, chaque choix lexical entraîne donc une cascade de restrictions aussi bien au niveau grammatical que sémantique. Comme le souligne Francis (1993 : 143) :

« […], we take the view that syntactic structures and lexical items (or strings of lexical items) are co-selected, and that it is impossible to look at one independently of the other. Particular syntactic structures tend to co-occur with particular lexical items and – the other side of the coin – lexical items seem to occur in a limited range of structures. The interdependence of syntax and lexis is such that they are ultimately inseparable, and it becomes merely a methodological convenience to regard them as different perspectives from which to view language use. »

Cet aspect a été développé plus tard par Hunston et Francis (2000) dans le cadre de *Pattern Grammar* (*grammaires de patrons* en français). Les auteurs y ont proposé la notion de *pattern* (patron) qui peut être considérée comme le concept-clé de l'approche *corpus-driven :*

« The patterns of a word can be defined as all the words and structures which are regularly associated with the word and which contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it. »

(Hunston and Francis, 2000:37)

En effet, le concept du *pattern* (*patron* en français), qui renvoie aux phénomènes de solidarités syntagmatiques entre les unités et correspond aussi bien aux collocations qu'aux colligations doit être vu comme un point de rencontre entre lexique et grammaire. Le concept met en évidence la relation mutuelle entre syntaxe et sémantique et leur importance pour l'analyse du sens. Comme le souligne Tognini-Bonelli, cette dimension a été occultée dans la perspective traditionnelle. Elle ajoute également que la démarche *corpus-based* (traditionnellement associée à Leech) ne permet pas non plus d'étudier les relations entre grammaire et lexique car le fait de recourir à l'annotation empêche les linguistes d'accéder à toute la richesse de l'information :

« What is lost, therefore, is the ability to analyse the inherent variability of language which is realised in the very tight interconnection between lexical and grammatical patterns. »

Tognini-Bonelli (2001 :73)

Rappelons que le lien étroit entre lexique et grammaire est déjà bien présent chez d'autres contextualistes britanniques comme par exemple le concept *meaning by collocation* de Firth ou celui de *lexique-grammaire* proposé par Halliday où à toute structure grammaticale, on peut associer un paradigme lexical et à toute unité lexicale, un ensemble unique d'éléments syntagmatiques. Du côté de la linguistique française, nous avons le modèle du lexique-grammaire de M. Gross et de G. Gross qui a pour objectif de décrire le fonctionnement du lexique à travers les structures syntaxiques. Comme le rappellent Legallois et Tutin (2013 : 8), M. Gross développe la notion de construction à verbe support en soulignant le fonctionnement régulier de ces structures alors que G. Gross propose le modèle des classes d'objet dans lequel il systématise l'approche distributionnelle du sens lexical.

Cependant, comme le souligne Tognini-Bonelli (2001 : 99), Sinclair, de façon plus tranchée que ses prédécesseurs, s'oppose à la séparation entre lexique et grammaire en prônant une forte connexion entre ces deux réalités linguistiques qui contribuent au même titre à l'analyse du sens. Ainsi, il remet en cause le lien qui traditionnellement existe entre les lemmes et leurs différentes formes linguistiques. A titre d'exemple, il propose une analyse en contexte des différentes formes des lemmes *decline* (Sinclair, 1991 : 44-51), *yield* (*ibid.* 53-56) ou celles du verbe à particule *set in* (73-74). Dans ces études, il prouve que chaque forme fléchie est associée à un environnement linguistique spécifique, c'est-à-dire qu'elle apparaît dans des structures grammaticales qui lui sont propres et possède des propriétés collocationnelles particulières. Rappelons rap*ibid.*ent que les observations d'un autre représentant du courant *corpus-driven,* Gill Francis vont dans la même direction. Legallois (2012 :41) en évoquant ses travaux (Francis 1991), rappelle que selon ce dernier, il existe une sorte de *déterminisme grammatical*. Pour Francis le système lexico-grammatical est intrinsèquement probabiliste. Ainsi toute unité lexicale possède sa propre grammaire et chaque acception induit des comportements syntaxiques différents.

Quant à Sinclair, il postule que chaque forme linguistique est dotée de son propre profil lexico-grammatical et que mise en contexte, elle doit être analysée individuellement.

« There is a good case for arguing that each distinct form is potentially a unique lexical unit, and that forms should only be conflated into lemmas when their environments show a certain amount and type of similarity. »

Sinclair (1991 : 8)

En effet, (comme nous l'avons signalé plus haut), un mot n'est pas considéré comme un membre d'une catégorie grammaticale mais comme une unité lexicale à sens unique et son sens est intimement lié au contexte dans lequel il apparaît. Ainsi, nous pouvons constater que pour Sinclair, il existe une relation directe entre la forme (vue comme une unité à sens unique) et le sens. Il va même jusqu'à prétendre qu'il n'y a pas de distinction entre ces deux réalités, qu'il s'agit de deux aspects du même phénomène, notamment de la langue en action :

« […] each meaning can be associated with a distinct formal patterning […] There is ultimately no distinction between form and meaning. »

Sinclair (1991: 6-7)

Chaque ligne de la concordance constitue donc un cas particulier, un événement unique, un exemple d'emploi vivant de la langue dans un contexte singulier. La généralisation d'un fait de langue (qui mène à la théorisation), n'est possible que si des parallélismes ou des similarités sont observés dans d'autres contextes de la même forme (rapprochées et alignées sur une colonne de la concordance). Comme le souligne Tognini-Bonelli (2001 :98), si l'on reprend la dichotomie saussurienne *langue/parole*, une ligne de concordance correspond à une instance de *parole* alors que les patrons repérés sur l'axe vertical peuvent être considérés comme des modèles de la *langue*.

Toutes ces observations nous amènent finalement à aborder la notion clé de la théorie sinclairienne, notamment celle d'*extended lexical unit* ou *extended unit of meaning*. En effet, l'intérêt porté au sens (qui est considéré comme un résultat de combinaisons de plusieurs éléments linguistiques en contexte), conduit Sinclair à développer un modèle d'*unité lexicale étendue*[48]. Comme le soulignent Legallois et Tutin (2013 : 12), l'idée est que l'analyse lexicologique ne doit pas porter véritablement sur le mot, mais plutôt sur son environnement linguistique. Une *unité lexicale étendue* (ULE) s'articule donc autour d'un noyau (*core* en terminologie sinclairienne) et s'étend à des unités proches liées entre elles à des degrés différents et sélectionnées en fonction des critères d'affinité. Chaque ULE est décrite à l'aide de quatre paramètres : la collocation, la colligation, la préférence sémantique et la prosodie sémantique et doit être analysée à ces quatre niveaux qui restent étroitement associés. Legallois et Tutin (*ibid.* 13) précisent que la notion d'*unité lexicale étendue* se définit autour de propriétés phraséologiques, propriétés conçues en termes de préférence marquée en discours: préférence lexicale, préférence pour des classes sémantiques, préférence grammaticale et attitude énonciative marquée. En effet, le concept d'ULE peut être vu comme un élargissement de celui de collocation où les dimensions grammaticale, sémantique et pragmatique sont abstraites à partir de l'analyse collocationnelle.

Pour illustrer son modèle d'ULE, Sinclair (2003 : 30-35) propose d'étudier l'environnement syntagmatique de l'expression *naked eye*, noyau qui en lui-même est considéré comme une collocation. Les données provenant de *The Bank of English* ont permis de repérer 154 occurrences de *naked eye* dont 3 sont identiques. L'analyse des concordances a montré qu'il y avait un plus grand nombre de patrons réguliers en contexte gauche. Ainsi, la première position à gauche (N-1) est occupée à 95 pour cent par l'article *the* qui constitue de

[48] La traduction a été proposée par Legallois et Tutin (2013).

ce fait un composant intrinsèque de l'expression *the naked eye*. Quant à la deuxième position (N-2), on remarque deux cooccurrents principaux : *with* et *to*, ainsi que d'autres prépositions moins fréquentes comme *by, from, as, upon*. La prédominance des prépositions à la position N-2 (elles constituent 90 pour cent des cas), permet de dégager une structure grammaticale générale (*PREP + DET*) préférentiellement utilisée avec l'expression *naked eye*. On passe alors d'un critère d'analyse (réalisée au niveau collocationnel) à l'autre, celui des propriétés colligationnelles. En ce qui concerne le troisième critère, à savoir la préférence sémantique, on observe à la position N-3, une attirance pour les unités qui expriment une idée de « visibilité » et ce, quelle que soit leur catégorie grammaticale : *see, seen, visible, invisible, spot, spotted, hiden, perceived*, etc. Finalement, le dernier critère, celui de la prosodie sémantique, caractérisée par Legallois et Tutin (2013 : 12) comme *l'attitude énonciative associée à la forme* rend compte, selon Sinclair, d'une fonction d'un énoncé dans une situation de communication donnée : *« It expresses something close to the function of the item – it shows how the rest of the item is to be interpreted functionally. »* (Sinclair 2003: 34). Dans l'exemple analysé ci-dessus, la suite *'visibilité + préposition + déterminant + naked + eye'* doit donc être interprétée comme une expression de difficulté dans le domaine de la perception visuelle. Il faut souligner que le sens de l'*unité lexicale étendue* se construit d'abord au niveau prosodique (le plus abstrait) pour ensuite pouvoir se concrétiser aux autres niveaux : sémantique, grammatical et lexical. En effet, la prosodie sémantique est un phénomène lié à l'attitude du locuteur. C'est lui qui sait pour quelle raison il crée la phrase et son choix influence les choix réalisés aux autres niveaux :

« The initial choice of semantic prosody is the functional choice which links meaning to purpose; all subsequent choices within the lexical item relate back to the prosody. »

Sinclair (*ibid.*)

Comme nous avons pu le constater dans les pages précédentes, les travaux pionniers de Sinclair et d'autres représentants du courant *corpus-driven* (pour ne citer que Gill Francis) ont apporté un réel changement en termes de qualité d'analyse des données linguistiques. De plus, l'approche contextualiste qu'ils ont développée a complétement changé le regard sur le processus de création et d'interprétation du sens. Cependant, il est nécessaire de souligner que Sinclair n'était pas qu'un pur théoricien. Ses travaux sur l'environnement syntagmatique des éléments linguistiques et les influences réciproques de l'un sur le sens de l'autre ont trouvé une application pratique dans deux projets : OSTI et COBUILD. Le premier a démarré en

1963 à l'Université d'Edimbourg. Son objectif était de créer un corpus initialement basé sur l'oral afin d'explorer les unités lexicales dans la perspective contextualiste. Faute de moyens de traitement automatique, inexistants dans les années 60, Sinclair a interrompu les travaux après avoir publié le rapport OSTI en 1970. Comme le remarque Williams (2006 : 154), outre la problématique de la création d'un corpus, ce rapport constitue un véritable programme de recherche contextualiste, où les collocations jouent un rôle central et elles sont explorées en relation avec des données issues du corpus. Williams (*ibid.*) souligne également que malgré son caractère novateur, le rapport OSTI a été oublié par la suite mais l'approche qui y était élaborée a servi de base à un projet encore plus ambitieux, le projet COBUILD (entrepris dans les années 80 à l'Université de Birmingham en collaboration avec la maison d'édition Collins). COBUILD a été conçu comme une base de données lexicale destinée à confectionner des dictionnaires à partir de textes authentiques. Pourtant, comme le remarque Williams (*ibid.*), COBUILD était plus qu'un dictionnaire et un corpus car il a donné naissance à d'autres applications : des grammaires, des méthodes d'apprentissage, des études linguistiques. Le projet a pris une envergure à laquelle ses auteurs ne s'attendaient pas :

« The initial aims were modest and no one anticipated that the project would have such a wide ranging effect »

(Sinclair 1991: 2).

Comme le souligne Béjoint (2009 : 137), la date de parution de la première édition du *Collins COBUILD Dictionary of the English Language* a marqué les débuts de la lexicographie de corpus. En effet, la description lexicographique qui y était proposée était entièrement basée sur l'exploitation de textes authentiques, ce qui a permis de se rapprocher au plus près de l'usage réel et de décrire l'état de la langue telle qu'elle est (avec toutes les découvertes surprenantes que cela implique) et pas telle que l'on s'imagine : *« the major novelty was the recording of completely new evidence about how the language is used »* (Sinclair *ibid.*). Deuxièmement, le dictionnaire COBUILD a révolutionné les pratiques lexicographiques avec son mode de définition contextuelle (qui bien évidement s'inscrit dans le droit fil des travaux de Sinclair). Selon Sinclair (1991 : 7), les dictionnaires traditionnels dont l'organisation est abstraite et purement conceptuelle, proposent des listes de mots ayant plusieurs sens mais n'aident pas l'usager à les distinguer et à les utiliser en contexte. Dans les définitions du COBUILD, les mots sont insérés dans leurs contextes habituels rencontrés en discours,

entourés de leurs cooccurrents typiques. Ainsi, l'usager peut découvrir la nature grammaticale et sémantique de l'unité lexicale à employer. Cependant, comme le souligne Béjoint (2007 : 138 – 140), ces définitions paraissent parfois peu éclairantes voire incompréhensibles à cause de l'accumulation d'informations extraites de différents contextes. Mais, il ajoute aussitôt que malgré ces vicissitudes, la méthodologie de COBUILD a connu un grand succès auprès du public et a inspiré de nombreux lexicographes.

Ainsi, nous avons terminé cet état de l'art des travaux en linguistique de corpus par la présentation de l'approche *corpus dirven*, qui nous paraît la plus proche et la plus adéquate aux objectifs de notre recherche car elle propose d'analyser une unité lexicale dans sa globalité et en fonction du contexte. De plus, à travers ce chapitre, nous avons pu voir que la linguistique de corpus a déjà une longue histoire et a réussi à influencer des études dans d'autres domaines des sciences du langage, notamment en lexicographie. Béjoint (2007 : 120) remarque que la linguistique et la lexicographie ont toujours eu des rapports étroits. Selon lui, tous les travaux des linguistes sont susceptibles de donner lieu à des applications, un jour ou l'autre, dans un dictionnaire, et tout dictionnaire utilise et transmet des points de vue sur le langage. Il cite à ce sujet une réflexion de Quemada : *« Toute œuvre lexicographique reflète une théorie que l'auteur applique plus ou moins consciemment. »* (Quemada 1972 : 427, cité dans Béjoint 2007 : 120). Nous avons pu le constater dans la première partie de notre travail, où tous les projets présentés se sont inspirés, bien év*ibid.*ent à différents degrés, des travaux de la linguistique de corpus. Ainsi, avant de passer à la présentation de notre propre méthodologie, nous proposons maintenant de voir rap*ibid.*ent les apports de la linguistique de corpus en terminologie.

Chapitre 4. Le corpus et la (les) terminologie(s) nouvelle(s)[49]

Comme nous avons pu le voir plus haut, ces dernières décennies, l'usage de corpus (même si son statut diffère d'un courant à l'autre), a révolutionné les sciences du langage en apportant un renouveau aussi bien théorique que méthodologique dans les différents champs de l'étude linguistique. Dans les pages qui suivent, nous proposons d'examiner de plus près les apports des corpus en terminologie et leur impact sur les objectifs et les méthodes de l'activité terminographique.

En effet, une des conséquences les plus importantes de la prise en compte des textes dans la pratique terminographique est que la terminologie en tant que discipline scientifique s'est rapprochée de la linguistique en générale, et en particulier de la lexicologie. Rappelons que nous avons déjà proposé une réflexion sur le statut de la terminologie moderne dans la perspective linguistique afin de justifier le recours aux méthodes lexicographiques dans la constitution de notre modèle de la base de données terminographique (voir l'Introduction et le deuxième chapitre). Nous y avons abordé les questions des rapports entre le terme et le mot en essayant d'établir les frontières entre langues spécialisées et langue générale. Nous nous sommes également focalisés sur le traitement des relations lexico-sémantiques entre termes. Dans cette section, nous nous intéresserons plutôt aux relations conceptuelles et aux stratégies mises en place par des linguistes-terminologues qui, face à l'irruption de la linguistique de corpus dans la pratique terminographique, *« tentent de desserrer l'étreinte des postulats logicistes »* (Slodzian 2000 : 62) et *« de sortir d'une sémiotique du signe fondée sur la triade terme/concept/référent qui la rend inapte à aborder le texte »* (Bourigault et Slodzian 1999 : 32). Nous regarderons notamment comment l'approche conceptuelle et onomasiologique fondée sur les principes de la Théorie Générale de la Terminologie a dû être reconsidérée sous un angle beaucoup plus pragmatique afin de répondre aux nouvelles réalités. Nous terminerons cette section par la présentation de méthodes d'extraction et de structuration de données terminologiques à partir de corpus spécialisés.

[49] En référence à Béjoint 2007

4.1 La terminologie traditionnelle et son rapport au texte

Rappelons que la terminologie classique telle qu'elle a été définie par Wüster (1976) s'est érigée par opposition à la linguistique, et plus particulièrement à la lexicologie, en se distinguant de cette dernière par un ensemble de caractéristiques telles que : approche conceptuelle, exigence de monosémie référentielle, analyse onomasiologique, démarche normative, point de vue synchronique, etc. (Sager 1990: 8, Cabré 1998 : 74-86, L'Homme 2004 : 25-29). Cependant, l'élément qui semble avoir le plus contribué à éloigner les travaux sur la terminologie de la linguistique est le fait de considérer les termes uniquement comme des étiquettes d'éléments de la réalité. Comme le souligne Condamines (1994 : 38), dans la doctrine wüsterienne, fortement référentielle : *« Seule la fonction de représentation du terme est alors considérée ; l'étiquetage du référent est fixé hors discours et de façon permanente. Une telle approche suppose que le terme-étiquette peut apparaître dans n'importe quel contexte sans que son "sens" en soit affecté. »*

Ainsi, selon l'approche classique, la terminologie a pour objet principal de représenter les connaissances d'un domaine donné en proposant des dénominations univoques de ses constituants. D'après la définition de Felber (1987 : 1), c'est une discipline *« ayant trait aux notions et à leurs représentations »*. Cette double dimension (conceptuelle et linguistique) est représentée par le terme qui en tant que signe linguistique renvoie à un contenu conceptuel, c'est-à-dire sert à communiquer les concepts qui sont considérés comme des représentations mentales des objets du monde réel. Comme le remarque Slodzian (2000 :66) : *« le terme […], est supposé présenter une double face : celle de l'expression, la dénomination, et celle du contenu, le contenu auquel renvoie la dénomination »*. Cependant (et même si Wüster insiste sur l'existence d'un lien étroit entre la terminologie et la linguistique (Felber 1987 : 88)), la fonction du terme en tant qu'élément d'un système linguistique se réduit à un simple acte de dénomination des concepts. La conception classique de la terminologie est donc purement nominaliste et ne considère pas le signe dans sa globalité. En effet, la doctrine wüsterienne ne retient de la langue que le lexique et ceci sous un angle très limité car restreint à un ensemble d'unités lexicales isolées, décontextualisées, dépourvues de leurs aspects connotatifs : *« Comme les noms propres […], les termes sont jugés dépourvus de connotation, et donc censés réaliser la dénotation parfaite. »* (Rastier 1995 : 42). Par ailleurs, comme le soulignent Depecker et Roche (2007 : 107), pour parler de la face linguistique de l'unité terminologique, la terminologie a généralement évité, voire rejeté *signe* (notamment dans son acception

saussurienne), au profit d'autres appellations comme *désignation* (ISO 1087-1, 2001 : 6), *forme, symbole linguistique* (Felber 1987 : 88), ou *terme* (*ibid.* : 1) :

« *Constitué d'un signifiant et d'un signifié, le signe, par son cortège d'évocations et de connotations, forme brouillage et altère la communication industrielle ou scientifique* ».

Depecker et Roche (2007 : 107)

« Désignation *a l'avantage d'apparaître comme une simple étiquette sur l'objet, dépourvue de résonance, d'idéologie et d'arrière-monde.* Désignation, *en tant que côté linguistique de l'unité terminologique, est voué à renvoyer purement et simplement à l'objet désigné.* »

(*ibid.*)

En réduisant l'unité terminologique à une étiquette apposée sur un concept, la terminologie traditionnelle privilégie donc sa dimension conceptuelle. Comme le remarque Rastier (1995 : 35), les notions, entités conceptuelles, priment en terminologie sur leurs expressions, linguistiques ou non, considérées en fait comme des variables, certes importantes, mais inessentielles. Rappelons que l'approche de l'unité terminologique proposée par la terminologie classique est représentée par un modèle sémiotique triadique de tradition aristotélicienne. Comme le souligne Lerat (1995 : 38) : *« Les terminologues de l'école de Vienne ont adopté le triangle sémiotique (objet, concept, signe). C'est effectivement la figuration la plus adéquate pour rendre compte des terminologies comme de systèmes relativement autonomes. »*, où la dimension extralinguistique joue un rôle essentiel. Ainsi, les termes renvoient aux objets du monde réel et plus précisément aux objets de connaissance[50], c'est-à-dire, à une réalité extralinguistique partagée. Ils sont reliés entre eux par des relations notionnelles hors discours en permettant ainsi de refléter la structure conceptuelle d'un domaine donné. Rastier (1995 : 39), remarque pour sa part que l'aristotélisme du triangle sémiotique est durci ici par le positivisme logique qui exprime un idéal de correspondance entre un mot, un concept et un objet. L'auteur (*ibid.*) met également l'accent sur le caractère stable et universel de ces éléments :

« Il est bien entendu dans la tradition aristotélicienne que les concepts pas plus que les choses ne varient avec les langues, et ce n'est pas sans effet sur la terminologie : les notions

[50] Comme le soulignent Depecker et Roche (2007 : 106) : *« Objet n'est pas non plus à entendre comme ponctuel ou statique : ce peut être en terminologie une procédure, un processus, une action, une manière de faire, etc. ».*

(ou concepts) qu'expriment les termes « ne sont pas liées aux langues individuelles » (ISO 3.1). »

Ainsi, selon la vision idéaliste et universaliste de la terminologie wüsterienne (qui est considérée comme une sorte de représentation parfaite d'un système conceptuel sous-jacent (Condamines 1994 : 37), le terme est une unité de connaissance à contenu stable, univoque, monoréférentielle, et indépendante de tout contexte. Comme nous l'avons déjà vu dans la première partie de ce travail, c'est justement sur cet aspect non-contextualiste de la signification que se fonde, selon l'approche classique, la différence entre termes et mots. Slodzian cite à ce sujet les propos de Felber :

« […] la signification du mot est donnée par le contexte ; elle est dépendant du contexte. […] La signification du terme qui est le concept est dépendante de la position du concept dans le système conceptuel correspondant »

(Felber 1984 cité dans Slodzian 2000 : 67)

En effet, la terminologie s'appuie sur la définition logique considérée comme : *« énoncé qui décrit une notion et qui, dans un système notionnel, permet de différencier d'autres notions. »* (Norme ISO 1087, 1990 cité dans Slodzian 200 : 64). Cette définition, qui garantit le caractère stable et univoque du terme, est une construction volontaire qui s'articule sur le schéma hiérarchique et prend en compte les traits formels du concept. Le terme est ainsi défini par les relations qui le lient aux termes apparentés et non pas par le contexte. Comme le remarque Béjoint (2007 : 61), l'unité terminologique est privée de ses relations syntagmatiques, au bénéfice de ses relations paradigmatiques telles qu'elles peuvent être présentées dans l'arbre du domaine.

Ainsi, Wüster croyait en l'existence d'une langue parfaite : *« d'une langue scientifique universelle et épurée (en tout cas épurable) de ce qu'il considérait comme les éléments nuisibles à une communication transparente. »* (Condamines 2005 : 42). Il considérait que la langue dans des domaines spécialisés peut être un moyen de communication parfait favorisant les échanges entre spécialistes et permettant le transfert sans équivoque des connaissances. Pour ce faire, la construction de terminologie doit se faire à l'écart de la réalité des discours professionnels, en dehors de la diversité des usages. Le terme doit être isolé, soustrait à ses *« accidents textuels »*, décontextualisé pour *« être défini par lui-même, indépendamment des*

variations qui pourraient affecter ses occurrences : variations de position dans le texte, de niveau de style, de ton, de mode d'énonciation représentée, d'évaluation. » (Rastier 1995). Comme l'ajoute l'auteur :

« *Perfectionner la langue, c'est par ces voies diverses la soustraire à l'interprétation, soumise à des variations individuelles et historiques, et ainsi lui permettre de refléter sereinement la vérité dans sa permanence »*

(Rastier, 1995 : 49).

D'où la nécessité de clarification, de standardisation et de normalisation qui permettent, selon les défenseurs de la doxa wüsterienne, d'éviter l'ambiguïté dans l'échange d'information entre les membres d'une communauté scientifique donnée.

Ainsi, comme nous avons pu le voir ci-dessus et comme le remarque à juste titre Henri Béjoint (2007 : 61), la vision classique de la terminologie ne laisse aucune place au discours. Le discours et tout ce qui s'attache aux contextes dans lesquels le terme fonctionne sont considérés comme une source de déformations potentielles du sens. Condamines cite à ce sujet les propos de Wüster, qui met en garde contre l'utilisation de productions réelles pour constituer des terminologies :

« […] jusqu'à une date récente, la linguistique n'a fait valoir que l'évolution libre, non dirigée, de la langue. C'est l'usage effectif de cette dernière qui, dans la langue commune, sert de norme. On peut appeler cette norme la norme descriptive. En revanche, en terminologie, fertile en notions et en termes, cette évolution libre de la langue mène à une confusion inacceptable…»

(Wüster, 1981, 65 cité dans Condamines 2005 : 42).

En effet, dans l'approche classique, la construction de terminologie passe par l'interrogation de l'expert, qui est considéré comme un dépositaire du système conceptuel de son domaine de compétence (les terminologues font appel à la démarche onomasiologique en plaçant le concept à l'origine de leur activité). L'étude des termes à travers les textes est vue ainsi comme une menace au dogme wüsterien. Avouer que le terme pourrait avoir un sens contextuel remet en cause les principes fondamentaux (déjà évoqués ci-dessus) de la

terminologie classique tels que : prééminence du concept, perspective onomasiologique, objectif d'univocité et de monosémie, nécessité de normalisation.

Néanmoins, depuis quelque temps, la pratique terminographique basée sur l'approche onomasiologique s'avère non productive (Bourigault et Slodzian 1999 : 30), entraînant un réexamen des concepts de la terminologie classique. Il est nécessaire de souligner que ce réexamen théorique et méthodologique se fait dans une perspective interdisciplinaire, fruit de la rencontre avec la linguistique de corpus, le TAL et l'Intelligence Artificielle. Nous proposons d'examiner ce processus de plus près dans les pages qui suivent.

4.2 La terminologie textuelle – le texte comme source des connaissances

Rappelons que la terminologie traditionnelle est née dans un contexte particulier : marquée par le néo-positivisme et l'aspiration à un idéal d'universalité, elle devait constituer une réponse aux problèmes de compréhension dans la communication scientifique et technique internationale. De ce point de vue, elle semble bien adaptée aux objectifs initialement visés. Cependant, ses postulats n'ont pas résisté à l'épreuve du temps et paraissent peu compatibles avec notre époque. Comme le souligne Béjoint (2007 :65-66), la méthodologie traditionnelle ne correspond plus ni aux besoins, ni aux moyens de la société moderne. En effet, dans la première partie de ce travail, nous avons déjà présenté les travaux qui visent à un renouvellement théorique et méthodologique de la terminologie. Cependant, le point de vue qui y est adopté est essentiellement linguistique et prône le rapprochement entre les deux disciplines. En s'intéressant à l'aspect linguistique de la terminologie et en définissant les termes comme des unités lexicales ayant un sens spécialisé, ces approches se concentrent plutôt sur l'étude du fonctionnement linguistique des termes en contexte et non pas sur la construction d'une terminologie proprement dite. En effet, en considérant les termes comme des éléments du discours à part entière, les terminologues-linguistes ont ainsi inséré la dimension textuelle dans l'analyse de ces unités. C'est la démarche proposée, entre autres par Kocurek (1991a, 1991b) et adoptée par exemple dans le cadre de l'approche lexico-sémantique de L'Homme (2004).

Cependant, comme le souligne Béjoint (2007 : 64), d'autres développements de la terminologie moderne sont apparus, qui tournent plus au moins délibérément le dos aux

modèles anciens. Ainsi, depuis le début des années 90 (Hamon et Nazarenko 2002 :8), on assiste à l'émergence d'une terminologie nouvelle[51], nommée selon les auteurs soit *terminologie computationnelle* (Béjoint 2007, Hamon et Nazarenko 2002) soit *terminologie textuelle* (Slodzian 2000, Bourigault et Slodzian 1999, L'Homme 2004, Condamines 2005) dont l'objet d'étude est la construction de terminologies à partir de corpus de données textuelles, assistée par l'ordinateur (nous proposons d'utiliser ici la seconde appellation, car mettant l'accent sur l'aspect textuel, elle est plus proche de nos intérêts scientifiques). En effet, contrairement aux projets terminologiques inspirés par la sémantique lexicale (évoqués plus haut), qui s'intéressent à l'aspect linguistique des termes pour décrire leur comportement en discours, les recherches menées dans le cadre de la terminologie textuelle analysent la dimension linguistique des unités terminologiques afin d'accéder à leur dimension conceptuelle. Cependant, même si les deux courants n'affichent pas les mêmes objectifs, ils ont un point d'intérêt commun : le terme inséré dans un discours et étudié en contexte. Par conséquent, ils ont recours aux mêmes méthodes d'investigation et utilisent les mêmes outils d'interrogation des textes en s'inspirant largement de la linguistique de corpus.

En effet, il nous paraît important de consacrer ce chapitre à la présentation des travaux qui s'inscrivent dans le courant de la terminologie textuelle, car ils proposent une voie intéressante pour aborder la dimension conceptuelle des termes via leur fonctionnement linguistique. Rappelons que le but de notre étude est de proposer un outil terminographique qui procure toutes sortes d'informations : ce seront aussi bien des informations de nature linguistique (sur le fonctionnement du terme dans son univers discursif) que des renseignements concernant la dimension cognitive des termes (sur les liens conceptuels que le terme entretient avec d'autres termes du domaine en question). Nous espérons donc que les recherches menées dans ce domaine nous donneront des pistes de réflexion pour notre propre travail.

Ainsi, la terminologie textuelle s'intéresse à la construction de terminologies à partir de corpus de données textuelles. Il est nécessaire de souligner que cette nouvelle approche est profondément pluridisciplinaire ; elle est le fruit de la rencontre entre plusieurs disciplines telles que informatique, intelligence artificielle, TAL et bien évi*bid.*ent la linguistique par le biais de la linguistique de corpus. Nous attirons également l'attention sur le caractère collaboratif des projets menés dans ce domaine. Offrant un champ d'investigation

[51] Béjoint (2007 : 64) parle de *terminologies nouvelles* en insistant sur l'apparition de plusieurs versions de ce que l'on appelle la *terminologie computationnelle*.

multidimensionnel, la terminologie textuelle rassemble les chercheurs venant de différents horizons. Comme le remarque Béjoint (2007 : 65) : « *ces nouveaux terminologues […] sont des linguistes ou des documentalistes qui savent se servir de l'informatique, ou des informaticiens qui ont décidé d'appliquer leurs travaux à des objets linguistiques*. » Vu son caractère pluridisciplinaire et collaboratif, la terminologie textuelle ne peut pas être résumée à une seule approche.

4.3 La terminologie et la gestion de l'information – du linguistique au formel

En effet, ce changement radical de la pratique terminologique est dû à plusieurs facteurs, tous liés au développement de l'informatique. Pourquoi l'informatique ? En fait, l'informatique permet d'automatiser ou, tout au moins, d'aider à la réalisation de nombreuses tâches et ceci dans presque tous les domaines de recherche. Rappelons que la terminologie, en tant que discipline à visée essentiellement pragmatique, doit avant tout répondre aux exigences des professionnels et des entreprises. Or, comme le soulignent de nombreux auteurs (Bourigault et Jacquemin (2000 : 216), Slodzian (2000 : 67), Hamon et Nazarenko (2002), depuis les années 90, suite à l'utilisation généralisée des outils de bureautique, à l'internationalisation des échanges et à la montée en puissance d'Internet, les besoins en terminologie en milieu industriel et institutionnel se sont multipliés. Le développement des échanges de données informatisés a entraîné l'accroissement de la production textuelle à caractère technique et scientifique. La gestion de la documentation est donc devenue un enjeu stratégique permettant de structurer des connaissances et de diffuser des savoir-faire. Ainsi, devant la masse de données textuelles nécessitant un traitement rapide et efficace, les entreprises ont été obligées de mettre en place ou d'améliorer leur systèmes de gestion de l'information ainsi que leurs outils de communication interne et externe.

Cependant, comme le soulignent Bourigault et Jacquemin (2000 : 216), pour exploiter les différents documents, les outils de gestion de l'information ont besoin de ressources terminologiques qui garantissent l'efficacité de ces systèmes. Par conséquent, l'un des résultats directs de cette demande en hausse de la part des entreprises est l'élargissement de la gamme des produits terminologiques, qui diffèrent en fonction des applications visées. Aussenac-Gilles *et al.* (2002 : 291) définissent le produit terminologique comme un ensemble plus ou moins structuré de termes et/ou de concepts qui est le résultat d'une analyse de corpus

dans un processus assisté par des outils de TAL mais validé manuellement. Comme le précisent Grabar et Hamon (2004 : 237), les produits terminologiques ont vocation à décrire la connaissance d'un secteur d'activité, et cette description peut être plus ou moins détaillée et fine en allant d'une simple liste de termes à un réseau de termes structurés sémantiquement. Soulignons d'après Aussenac-Gilles *et al.* (*ibid*.) que le projet terminologique devient une ressource terminologique lorsqu'il sert à créer une application. Parmi les applications qui utilisent les ressources terminologiques on peut citer, d'après Bourigault et Slodzian (1999 : 29) et Aussenac-Gilles *et al.* (2002 : 292) : outils d'aide à la traduction (utilisation de lexiques et de terminologies), systèmes d'indexation automatique (utilisation de thésaurus), documentations techniques hypertextuelles (utilisation d'index), systèmes de gestion de données techniques (utilisation de référentiels terminologiques), mémoires d'entreprise (utilisation d'ontologies), etc. Ainsi, les travaux terminologiques peuvent avoir différents objectifs : extraire des termes / mots-clés en vue de l'indexation des documents, extraire des contextes pouvant servir de base à des définitions, localiser l'information, générer des ontologies, fournir à des traducteurs des informations conceptuelles (nous nous intéresserons plus particulièrement à ce dernier point).

Comme nous pouvons le constater, ces travaux se situent à la jonction de différents domaines de recherche. En effet, les types d'utilisation de la terminologie évoqués ci-dessus rendent compte des liens qui s'établissent entre cette dernière et les courants de recherche travaillant sur la représentation des connaissances (l'intelligence artificielle et en particulier l'ingénierie des connaissances), et cela par le biais de la linguistique. Ce rapprochement est dû à une évolution parallèle du rôle des corpus dans chacune de ces disciplines ainsi qu'au développement des outils informatiques[52], notamment des outils TAL. Mais, si les travaux en terminologie, au moins à l'origine, étaient plutôt guidés par les besoins de communication et concernaient la traduction et la normalisation, en documentation, ils sont déterminés par la nécessité de créer un lien entre documents et connaissance. Ainsi, comme le souligne Condamines (2005 :37), l'objet majeur de l'ingénierie des connaissances concerne la constitution d'outils pouvant assister l'homme dans son raisonnement :

[52] Aussenac-Gilles *et al.* (2002 : 293) soulignent que La France est particulièrement bien placée dans l'identification des convergences entre ces trois disciplines, notamment linguistique, terminologie, IA. Depuis 1993, le groupe de recherche TIA (Terminologie et Intelligence Artificielle1) a beaucoup contribué à cette identification en permettant aux chercheurs de ces différentes communautés de se rencontrer régulièrement pour approfondir la connaissance de leurs travaux et amorcer des collaborations précises.

« Il s'agit d'élaborer des systèmes qui représentent la connaissance au plus près de la façon dont elle se manifeste, c'est à- dire en utilisant les éléments langagiers propres au domaine couvert par l'outil. »

(*ibid.*)

Ces systèmes permettant la modélisation et la représentation formelle des connaissances sont appelés *ontologies*[53] et, comme le souligne Roche (2005 : 56) : « *connaissent aujourd'hui un succès considérable qui s'explique principalement par la poursuite d'un mythe, celui d'une représentation du monde et du sens des mots pour en parler qui soit compréhensible, et donc partageable, aussi bien par des acteurs humains que par des agents logiciels* ». Définie par T. R. Gruber (1993) comme : *« an explicit specification of a conceptualization*[54] », l'ontologie est une description structurée de concepts et de relations d'un domaine, une description formelle et exploitable par un système informatique. Étant donné que les ontologies sont construites pour mieux gérer les connaissances et les savoir-faire, faciliter leur échange et assurer l'interopérabilité des systèmes qui les utilisent, elles doivent refléter les connaissances partagées par une communauté, c'est-à-dire les connaissances qui sont obtenues et exprimées de manière consensuelle et présentées dans un langage formel simple et lisible. Comme le souligne Roche (2005 : 57), les ontologies, ce sont des « langages de représentation ». Elles n'ont pas pour objectif de comprendre le monde mais d'en représenter les objets à des fins de manipulation informatique. Ainsi, les connaissances y sont représentées à l'aide des formalismes de type relationnel qui, comme le remarque Condamines (2005 : 37) s'inscrivent dans une parenté revendiquée avec les réseaux sémantiques de Quillan[55]. Ces réseaux fournissent une représentation graphique d'une conceptualisation de connaissances sous la forme de graphes orientés constitués de nœuds (sommets) reliés par des arcs (flèches). Les nœuds représentent des concepts tandis que les arcs traduisent des relations entre concepts. Les nœuds, tout comme les arcs, sont étiquetés. Toujours d'après Condamines (*ibid.*), les premiers sont habituellement étiquetés par des noms alors que les seconds le sont par des formes prédicatives (noms ou verbes). Remarquons d'après Condamines (1995 : 43), que les ontologies font ainsi appel aux termes qui, en tant que signes linguistiques, participent à l'étiquetage. Actuellement, le formalisme le plus couramment utilisé (de par sa simplicité et

[53] En philosophie, l'ontologie est une branche qui s'intéresse à l'essence des choses, c'est-à-dire à l'étude de l'*être,* de ses propriétés.

[54] Pour Gruber (1995), la *conceptualization* est « an abstract, simplified view of the world that we wish to represent for some purpose ».

[55] En effet, Ross Quillian, informaticien et chercheur en intelligence artificielle qui s'est intéressé aux phénomènes d'association dans la mémoire humaine, a mis au point un modèle de l'organisation des connaissances représenté par des réseaux sémantiques.

son expressivité) est le graphe conceptuel proposé par Sowa. Cependant, les ontologies ont aussi recours à d'autres systèmes de représentation de connaissances tels que *frames*, langages-objets ou logiques de description. Nous n'allons pas développer ici cette problématique car elle ne relève pas de notre champ de compétence. Néanmoins, nous voudrions souligner que les ontologies sont très proches des réseaux terminologiques en permettant aux deux disciplines, à savoir l'intelligence artificielle et la terminologie, de trouver un terrain d'étude commun et de contribuer mutuellement à leur développement respectif.

D'après Bachimont (2000), définir une ontologie est une tâche de modélisation menée à partir des données empiriques, à savoir l'expression linguistique des connaissances. Ce travail s'effectue donc à partir de documents attestés dans la pratique d'un domaine et rassemblés en un corpus. L'objectif final de cette opération est de dégager les primitives de représentation qui seront utilisées pour la modélisation formelle. Il s'agit donc de passer du linguistique au formel. Et quel est le rôle de la terminologie ? En effet, comme le précisent Després et Szulman[56] (2008 : 17), les méthodes de construction d'ontologies à partir des textes comportent une phase de conceptualisation au cours de laquelle s'effectue le passage du terme au concept. En fait, l'ontologie s'intéresse aux concepts car ils renvoient aux connaissances et peuvent de ce fait constituer de futures primitives nécessaires à la modélisation. Et c'est là que se croisent l'intelligence artificielle et la terminologie. Cette phase de conceptualisation évoquée par Després et Szulman nécessite des traitements se situant à la fois aux plans linguistiques et ontologiques Selon les auteurs (*ibid.*), on peut distinguer trois étapes : « *l'étude linguistique du corpus qui aboutit à la construction d'un réseau terminologique, l'étape termino-ontologique qui permet de construire un réseau termino-ontologique et l'étape ontologique dont la finalité est l'élaboration de l'ontologie* ».

En effet, l'étude linguistique réalisée à l'aide des outils de traitement automatique des langues (TAL) permet d'extraire des unités linguistiques (termes) et de les organiser en réseaux reflétant la structure lexicale (et non conceptuelle) des textes. L'interprétation des termes et de leurs relations lexicales mène à la création d'un réseau termino-ontologique constitué de concepts terminologiques et des relations sémantiques les liant. Comme le soulignent Després et Szulman (2008 : 28), la dernière étape - l'étape ontologique - consiste à

[56] La méthode proposée par Després et Szulman a été mise en œuvre dans le cadre de la construction d'une ontologie du domaine de l'organisation de la gestion de l'hygiène, de la sécurité et de l'environnement (HSE) des entreprises (2008) ainsi que de celle du droit communautaire (2005).

traduire le modèle conceptuel obtenu à la fin de l'étape termino-ontologique dans un langage formel qui permet de s'affranchir des problèmes liés à la langue naturelle et au contexte. L'ontologie est donc constituée de concepts ontologiques (à ne pas confondre avec les concepts terminologiques) décrits dans un langage formel et reliés par des relations ontologiques (hiérarchiques et descriptives). Les concepts ontologiques sont assimilés à des classes représentant un ensemble d'individus partageant les mêmes propriétés.

Comme nous avons pu le constater, les différentes étapes de la construction d'ontologies relèvent respectivement de la terminologie, puis de l'intelligence artificielle : la structuration terminologique, la modélisation des connaissances et enfin la représentation des connaissances permettent la transition du plan du discours au plan conceptuel, puis au plan formel. Cela témoigne de la compatibilité et de la complémentarité des deux domaines. Ainsi, l'intelligence artificielle peut trouver, dans la terminologie (et dans la linguistique), des méthodes d'acquisition et de représentation des connaissances à partir des données textuelles. En revanche, la terminologie peut puiser dans l'intelligence artificielle des modèles ou des outils qui peuvent l'aider dans la constitution des réseaux notionnels. En fait, le renouvellement des pratiques d'analyse terminologiques à partir de corpus spécialisés s'inscrit dans une perspective applicative. La question du statut théorique du terme y est rarement abordée. Comme le soulignent Bourigault et Jacquemin (2000 : 216), l'adhésion à une théorie linguistique particulière n'est pas nécessaire pour concevoir un modèle d'acquisition de terminologie, la priorité étant donnée à l'efficacité. Cependant, à la lumière des travaux récents menés en terminologie et en intelligence artificielle, certains chercheurs ressentent la nécessité d'une nouvelle réflexion théorique. Ainsi, Christophe Roche (2007) propose d'introduire la notion d'*ontoterminologie,* qui replace le concept et sa dénomination au centre du débat tout en préservant la dimension linguistique. L'ontoterminologie met l'accent sur le caractère épistémologique de la terminologie, qui se traduit par la recherche d'une compréhension du monde. Elle insiste également sur la nécessité d'une distinction nette entre les dimensions linguistique et conceptuelle en différenciant trois types de pratiques : intellection, usage, représentation. Par la prise en compte des usages réels ainsi que des principes épistémologiques centrés sur la notion d'ontologie, elle offre, selon Roche (2007 : 12) de nouvelles perspectives pour la construction de systèmes notionnels et leur représentation.

4.3.1 Bases de connaissances terminologiques – entre le linguistique et le conceptuel

Comme nous l'avons vu plus haut, le rapprochement entre la terminologie et l'ontologie, possible grâce au double statut du terme (qui fonctionne comme signe linguistique et comme clé d'accès aux connaissances spécialisées) ainsi que la prise en compte des données textuelles ont incité les terminologues à s'intéresser de plus en plus aux relations entre les termes et à leur structuration. Comme le remarquent Aussenac-Gilles et Séguéla (2000 : 177), en intégrant les relations sémantiques, on a complété les traditionnelles définitions en langue naturelle en enrichissant ainsi la description terminologique d'un niveau conceptuel. On parle alors des ressources termino-ontologiques (RTO).

Cette nouvelle tendance de la pratique terminologique s'est manifestée tout d'abord dans la création de ce que l'on appelle les *Bases de Connaissances Terminologiques*. On ne peut pas en parler sans évoquer les travaux pionniers d'Ingrid Meyer. Ses recherches menées au confluent de la terminologie et de l'intelligence artificielle ont contribué au rapprochement entre les membres de ces deux communautés. En effet, elle a été la première à proposer ce concept de *Base de Connaissances Terminologiques* (BCT), à savoir un modèle de stockage des données qui rend compte à la fois du fonctionnement linguistique des termes et de leur lien avec les concepts du domaine. Son prototype appelé COGNITERM a été conçu comme une amélioration d'une Base de Données Terminologiques classique enrichie par des relations conceptuelles organisées en réseaux très complexes (qui reflètent aussi bien les liens taxinomiques et méronymiques que d'autres types de relations : cause, conséquence, fonction, etc.). Ainsi, une BCT (TKB en anglais) relève en même temps du domaine de la langue et de celui de la connaissance :

« […] a TKB must represent what a native speaker who is also a subject field expert knows about both concepts and their corresponding terms. We feel that this assumption is in principle fully compatible with traditional terminological practise, not only as regards the "terms" component (obviously, term banks contain terms and other strictly linguistic information), but also as regards the "concepts" component. »

(Meyer, Bowker and Eck 1992: 159).

Comme nous pouvons le constater, Ingrid Meyer ne rejette pas les principes de la terminologie classique mais elle considère que les deux types de données (à savoir linguistiques et notionnelles) peuvent et doivent être représentées ensemble.

En parlant de bases de connaissances terminologiques, rappelons rap*ibid.*ent les travaux de Gabriel Otman, qui ont porté sur un modèle théorique de représentation formelle des unités terminologiques baptisé « réseau sémantico-terminologique » (RST) en écho au modèle relationnel des réseaux sémantiques de Ross Quillian (Otman 1997 : 245). Ainsi, Otman a dénombré 6 relations sémantiques qui unissent les termes au sein d'un système notionnel, réparties en deux groupes : 4 relations associatives : sorte-de ; partie-de ; fonction-de ; proximité-de et 2 relations distinctives : contraste-avec ; équivalent-de. Le RST est organisé en hiérarchie, chaque nœud étant relié aux autres par des relations verticales à vocation classificatoire (auxquelles s'applique la propriété d'héritage) et des relations horizontales à vocation descriptive (auxquelles la propriété d'héritage ne s'applique pas). L'ensemble des liens et des propriétés attachés à ces liens constitue la définition du terme (Otman 1996, chapitre 4). L'objectif de son travail était l'implantation de cette forme de modélisation sur des gestionnaires de bases de données terminologiques afin de développer une véritable base de connaissances terminologiques dotée de modes d'interrogation, de circulation, de mise à jour. Considérant que la terminologie joue un rôle de premier plan dans l'acquisition et le transfert de connaissances scientifiques et techniques, Otman (1996) a situé ses recherches terminologiques dans le domaine de la gestion des connaissances.

Ces modèles de stockage des données qui rendent compte à la fois du fonctionnement linguistique des termes et de leur lien fort avec les concepts du domaine posent le problème de l'adéquation entre les expressions linguistiques repérées dans les textes et leur pendant conceptuel. Cette problématique a été notamment développée par l'équipe de Recherche en Syntaxique et Sémantique de Toulouse dirigée par Anne Condamines (1994, 1999, 2000, 2005). En effet, depuis les années 90, dans le cadre de sa collaboration avec différentes entreprises (industrie spatiale, EDF) et en partenariat avec l'Institut de Recherche en Informatique de Toulouse (entre autres Nathalie Aussenac-Gilles ou Patrick Séguéla), l'ERSS mène une réflexion sur la définition d'un modèle de BCT à partir des données textuelles. Il s'agit d'une entité semi-formelle qui constitue un reflet modélisé et fidèle du corpus en établissant un lien entre le linguistique et le conceptuel. L'intérêt d'une BCT est qu'elle puisse être réutilisée pour construire une entité formelle, gérable par une machine telle qu'une

Base de Connaissances ou bien fournir des données à d'autres systèmes comme les logiciels d'aide à la traduction ou les systèmes d'indexation. Sa principale caractéristique consiste à faire la distinction entre données linguistiques et données conceptuelles, en accordant une égale importance au fonctionnement linguistique du terme (information sur la nature, le genre, les variantes de forme) et à son fonctionnement comme indice de concept (sa définition et ses relations avec les autres concepts). Ainsi, une BTC telle qu'elle est définie par Condamines (1994, 1999, 2000), est constituée de quatre champs :

- le *terme* (T) comportant des données proprement linguistiques relatives au terme (T) qui est une manifestation linguistique du concept (C)
- le *concept* (C) comportant les données qui concernent le concept dénommé par le terme (T). Les concepts (C) sont reliés par des relations conceptuelles telles que *est-un, partie-de, cause, conséquences*
- le *lien terme-concept* qui rend compte des conditions d'usage (groupe de locuteurs qui utilisent tel terme) et indique le statut du concept dénommé par le terme au regard de normes
- le *texte* qui permet le passage au corpus afin de rendre compte des liens entre un terme et ses occurrences et de justifier le choix des relations conceptuelles

Ainsi conçue, la BTC est une modélisation (et non pas, comme le souligne Condamines (1999 : 106), une formalisation) du corpus qui, orientée vers la représentation des connaissances, présente l'avantage de conserver le lien avec la langue.

À travers cet exemple, nous voulions attirer l'attention sur l'intérêt des produits (ressources – voir plus haut) terminologiques qui servent d'intermédiaire entre les dimensions linguistique et conceptuelle. En effet, les ontologies et d'autres systèmes de gestion des connaissances rencontrent des difficultés dans le traitement des problèmes liés à la nature linguistique des termes. La terminologie textuelle, avec ses méthodes d'interrogation des corpus (empruntées par ailleurs à la linguistique de corpus) et d'interprétation des données recueillies, apporte des solutions permettant le passage du niveau linguistique au niveau conceptuel. Comme le soulignent Aussenac-Gilles *et al.* (2002 :292) :

« Quel que soit le point de vue disciplinaire (TALN, IC, Sciences de l'information, terminologie, linguistique de corpus) duquel on se place, les produits terminologiques jouent le rôle de médiateurs entre des utilisateurs dans un contexte d'application donné, et des connaissances dont l'expression exhaustive se trouve dans les textes. ».

Ainsi, dans la section qui suit, nous proposons de présenter des méthodes d'acquisition terminologique à partir des textes pour pouvoir s'en inspirer dans la constitution de notre propre base de données terminographique.

4.4 Méthodes d'extraction et de structuration des données terminologiques à partir de corpus spécialisés

L'activité de construction d'une terminologie est devenue une tâche d'analyse de données textuelles (Bourigault et Slodzian (1999 : 29) qui vise une description des structures lexicales à l'œuvre dans un corpus de référence. Les spécialistes sont donc obligés de se tourner vers les techniques de la linguistique de corpus afin de mener à bien leurs projets. Comme le souligne Slodzian (2000 : 61), ce virage méthodologique a créé une onde de choc qui ébranle les fondements de la doctrine wüsterienne :

« […], la terminologie classique se trouve déstabilisée par les techniques de linguistique de corpus qui imposent des ensembles de textes comme unité d'analyse, coupant ainsi court à des considérations idéalisantes sur la qualification du terme a priori. *»*

(*Ibid.* :82)

Ainsi, alors que les adeptes de Wüster considèrent le corpus comme une menace pour le statut monosémique, univoque, non connoté du terme, selon les tenants de la terminologie textuelle (Bourigault et Slodzian 1999, Slodzian 2000), la prise en compte des données textuelles est une pratique tout à fait naturelle. Bourigault et Slodzian (1990 : 30) remarquent que les applications terminologiques sont le plus souvent des applications textuelles (traduction, indexation, aide à la rédaction). Par conséquent, la terminologie n'est jamais déliée du texte : elle *« doit " venir " des textes pour mieux y " retourner " »*. L'approche proposée afin d'accéder aux données terminologiques est donc résolument linguistique. Pour les terminologues qui ont rompu avec la TGT de Wüster (Bourigault et Slodzian 1999, Slodzian

2000, Condamines 2005), les textes sont considérés comme des sources de connaissances, les expressions linguistiques étant la seule réalité concrète permettant de les atteindre. Cependant, comme le souligne Condamines (2005 : 43), il ne s'agit pas de considérer les textes comme des attestations d'un réseau terminologique préexistant et de voir comment les concepts se mettent en mots en discours. Bourigault et Slodzian (1990 : 31) remarquent que *« les termes ne sont pas des "unités de connaissances" qui viendraient "habiter la langue" »*. En effet, l'analyse du corpus constitue : « *le point de départ de la chaîne de procédures linguistiques et sémantiques qui permettront de faire émerger les termes »* (Slodzian 2000 : 72).

Ainsi, la signification du terme n'est pas définie par la position du concept dans le système conceptuel d'un domaine mais suite à un processus d'analyse des données textuelles. Comme le soulignent Bourigault et Slodzian (1990: 30), pour chaque unité retenue, l'analyste construit une signification (type) à partir des sens (occurrences) attestés dans le corpus. La terminologie textuelle part donc du syntagmatique, c'est-à-dire des occurrences manifestées en texte pour constituer (à partir d'un travail d'interprétation) une sélection de termes dotés d'une forme et d'une signification stables, décontextualisables, liés à d'autres unités sémantiques et susceptibles de recevoir des définitions issues de leur contexte d'origine (Slodzian 2000 : 77).

« Partir des données réelles du texte, prendre en compte les effets liés à la contextualité du sens conduit à réinterpréter les "concepts" comme des signifiés normés par des pratiques discursives et gnoséologiques dans le champ d'activités professionnelles où se croisent le plus souvent plusieurs domaines. » (Slodzian 2000 : 72)

Un mot acquiert le statut de terme en bout de chaîne : *« c'est en bout de chaîne, en normalisant le terme, qu'on lui prescrit une référence. »* (Bourigault et Slodzian (1990 : 31). Dans cette démarche purement sémasiologique qui va à rebours de la conception apriorique de la théorie wüsterienne, le terme est vu comme le résultat d'une analyse terminologique, l'analyse qui n'est pas considérée comme un exercice de redécouverte d'un système notionnel préexistant mais comme un processus de construction. Comme le précisent Bourigault et Jacquemin (2000 : 223) le terminologue *« procède à un travail de* construction *d'*une *terminologie du domaine, et non de* découverte *de* la *terminologie du domaine »*.

Cette citation met aussi l'accent sur une autre caractéristique de l'approche textuelle, à savoir la variabilité des terminologies. Il faut savoir que les connaissances évoluent dans des contextes dynamiques et qu'elles ne sont pas liées à des expressions linguistiques bien stabilisées : *« On est loin de la conception idéalisée du domaine comme fragment de connaissances bien structurées, permanentes et clairement circonscrites. »* (Bourigault et Slodzian (1990 : 31). De plus, il est nécessaire de souligner que le traitement des données textuelles a des finalités différentes, qui donnent lieu à des ressources terminologiques différenciées en fonction de l'application :

« Il n'y a pas une terminologie, qui représenterait le savoir sur le domaine, mais autant de terminologies que d'application dans lesquelles ces terminologies sont utilisées. *Selon l'application, ces terminologies peuvent différer sensiblement quant aux unités retenues et à leur description* ».

Bourigault et Jacquemin (2000 :220).

Ce constat remet en cause le principe de l'unicité, de la fixité et de l'universalité de la terminologie, si cher aux adeptes de Wüster. En effet, l'étude de l'usage des termes dans les textes pousse les terminologues à abandonner la perspective normative et à adopter une démarche descriptive qui prend en compte le réel, c'est-à-dire les locuteurs mais aussi les objectifs précis. Ainsi, toujours d'après Bourigault et Jacquemin (*ibid.*), le travail du terminologue analyste doit être guidé par une double contrainte de pertinence : pertinence vis-à-vis du corpus (il faut retenir et décrire des structures spécifiques au domaine et stables dans le corpus) ; pertinence vis-à-vis de l'application visée.

Ce travail ne peut pas être envisagé sans le recours à des outils du traitement automatique des langues. De l'autre côté, la démarche est nourrie par les travaux sur l'analyse de corpus en provenance de la linguistique qui fournissent des éléments théoriques et méthodologiques. Ainsi, la collaboration entre ces disciplines permet de développer des techniques automatisant partiellement ou complètement le processus d'acquisition d'une terminologie à partir de textes. Actuellement, il existe de nombreux logiciels d'analyse de corpus qui interviennent à différentes phases de la construction d'une terminologie en systématisant le passage des usages en corpus vers la délimitation du sens en réseau. Notre but n'est pas de recenser les différents systèmes existants mais de présenter de manière générale les méthodes utilisées pour accéder aux données terminologiques.

Ainsi, Bourigault et Jacquemin (2000 :224) proposent une typologie fonctionnelle des outils d'aide à la construction de ressources terminologiques. En effet, cette tâche comporte deux étapes essentielles, notamment l'acquisition des termes et leur structuration. Comme le soulignent les auteurs, une première classe regroupe les outils dont la visée est le repérage et l'extraction à partir du corpus de candidats-termes, c'est-à-dire d'unités lexicales susceptibles d'être retenues comme termes. Ces logiciels d'acquisition de termes sont classés à leur tour en trois catégories, en fonction des approches mises en œuvre, à savoir l'approche statistique, linguistique ou mixte. Le fonctionnement des méthodes statistiques, basées sur des algorithmes, se fonde sur le critère de la fréquence. En effet, les outils procèdent au repérage des régularités dans le corpus. Il s'agit d'identifier les schémas (segments de texte répétés) les plus productifs permettant de relier les termes partageant les mêmes têtes ou les mêmes extensions. Parmi les avantages de tels systèmes, on peut énumérer leur performance, c'est-à-dire la capacité de traiter un grand volume de textes et leur indépendance vis-à-vis des ressources linguistiques. L'extracteur de termes conçu sur un modèle statistique que l'on cite souvent comme référence est le logiciel ANA, développé au début des années 90 par Chantal Enguehard.

Une deuxième catégorie d'outils d'acquisition terminologique s'appuie sur une analyse linguistique. Les systèmes qui entrent dans cette catégorie exploitent des informations syntaxiques, morphologiques ou lexicales telles que lacatégorie grammaticale, la variation morphologique, le groupe syntaxique ou la relation syntaxique, afin de proposer des candidats-termes à partir des patrons caractéristiques. En effet, la méthode linguistique consiste à repérer dans un corpus préalablement étiqueté et désambiguïsé des unités polylexicales (il s'agit le plus souvent des syntagmes nominaux) qui correspondent aux schémas sur lesquels se forment les termes. Comme le soulignent Bourigault et Jacquemin (*ibid.*), l'outil TERMINO est considéré comme une application pionnière de l'acquisition automatique de termes basée sur la méthode linguistique. Les candidats-termes extraits par TERMINO sont appelés *synapsies* d'après les travaux de Benveniste. LEXTER est un autre logiciel dédié à l'extraction de candidats-termes exploitant des critères syntaxiques. Il a été conçu par Didier Bourigault dans le cadre d'un projet de gestion de la documentation au sein de la direction des Études de Recherches d'EDF (Condamines et Rebeyrolle 2000 : 230). Son originalité consiste à repérer les syntagmes nominaux maximaux en s'appuyant sur les patrons morpho-syntaxiques. Plus précisément, il s'agit de marquer les frontières syntaxiques de

groupes nominaux à l'aide d'éléments grammaticaux (déterminants, verbes conjugués, pronoms) ou extralinguistiques (ponctuation) qui ne font pas partie de ces syntagmes. Cet outil peut être utilisé en association avec FASTR, logiciel développé par Jacquemin (Bourigault et Jacquemin 2000 : 226) et dédié à la reconnaissance des variantes terminologiques qui peuvent être syntaxiques, morphosyntaxiques ou sémantico-syntaxiques.

La dernière catégorie comprend les systèmes basés sur les modèles mixtes (ou hybrides), qui combinent des techniques linguistiques et statistiques. L'ordre d'application des composants varie selon le système. Parmi les premiers logiciels conçus selon ce modèle, on cite communément (Bourigault et Jacquemin 2000 : 225) deux outils, notamment ACABIT de Daille et XTRACT développé par Smadja. Le premier travaille à partir d'un corpus pré-étiqueté et procède d'abord à une analyse syntaxique permettant, à l'aide des transducteurs, d'extraire des séquences nominales ramenées à des candidats-termes binaires. Les résultats produits à cette étape sont triés au moyen de mesures statistiques. En revanche, XTRACT utilise d'abord une technique statistique pour détecter les cooccurrences et ensuite fait appel aux critères syntaxiques pour les classer.

Comme le soulignent Hamon et Nazarenko (2002 :8), les travaux de terminologie computationnelle (il vaut mieux utiliser ce terme si l'on parle des méthodes automatisant la construction terminologique) ont d'abord été centrés sur le repérage et l'extraction des termes, le choix des candidats-termes étant la première étape de ce processus. Cependant, depuis une décennie, les recherches en ingénierie terminologique se concentrent davantage sur l'organisation des listes de termes et sur le repérage des relations qui existent entre eux. Comme le remarquent Bourigault et Jacquemin (2000 : 222), les ressources terminologiques se présentent rarement sous la forme d'une simple liste. D'après les auteurs, « *construire une terminologie, c'est* structurer *un ensemble de termes* », c'est-à-dire élaborer un réseau conceptuel à partir des données textuelles. Ainsi, les terminologues ne font pas seulement appel aux outils d'aide à l'identification de termes mais aussi aux outils d'aide à leur structuration.

Comme nous avons pu le voir dans la première partie de ce travail, l'éventail des relations que l'on peut trouver dans les ressources terminologiques est très vaste. Grabar et Hamon (2004 : 238) recensent 4 types de relations sémantiques qui peuvent servir à structurer une liste de termes, notamment : relations taxinomiques, relation partitive, relations lexicales

(dont les relations de synonymie et d'opposition) et relations transversales (domaniales). La question qui se pose est de savoir dans quelle mesure ces relations peuvent être identifiées de manière automatique. En effet, l'idée de base est que les connaissances relatives à un domaine sont ancrées dans des éléments textuels. Ainsi, comme le soulignent Condamines et Rebeyrolle (2000 :42), le principe qui guide le travail de structuration est que les données conceptuelles sont implicitement exprimées au niveau linguistique. Cependant, il est nécessaire de souligner d'après Hamon et Nazarenko (2002 :14) qu'il n'existe pas une méthode unique permettant de repérer tous les types de relations terminologiques. Les auteurs (Hamon et Nazarenko (*ibid.*), Grabar et Hamon 2004 : 239) distinguent ainsi deux ensembles de démarches, celles qui reposent sur la structure interne des termes (approches structurelles ou endogènes) et celles qui permettent de repérer des relations à partir des contextes d'emploi (approches contextuelles). Les premières exploitent généralement les informations morphosyntaxiques. L'analyse de la forme des termes, composés de têtes et d'extensions, donne ainsi des indications sur des liens d'hyponymie, hyperonymie ou de méronymie permettant la mise en évidence des relations hiérarchiques et partitives. Il faut savoir que les travaux portant sur les relations hiérarchiques occupent traditionnellement une place privilégiée. Par ailleurs, beaucoup d'outils d'extraction de candidats-termes proposent déjà ce type de structuration. En ce qui concerne l'extraction d'autres types de relations et surtout des relations transversales, la tâche paraît plus difficile. Ce travail repose principalement sur les approches contextuelles qui consistent soit en un repérage de marqueurs de relations permettant d'isoler des segments de textes candidats représentant des relations, soit en l'interprétation des contextes distributionnels. En effet, si l'on parle des approches en contexte, on pense le plus souvent aux deux méthodes permettant de relever des informations sémantiques explicitées dans les textes, notamment : analyse distributionnelle et projection de patrons lexico-sémantiques sur le corpus.

La première méthode s'intéresse aux contextes dans lesquels apparaissent les termes, et plus particulièrement elle vise à repérer les contextes terminologiques communs et récurrents. Elle se fonde sur l'hypothèse formulée par Z.H. Harris, selon laquelle la distribution d'un mot, c'est-à-dire l'ensemble des contextes dans lesquels il apparaît, détermine son sens. Ainsi, les termes qui ont des distributions comparables ont souvent un élément de sens commun. Il est donc possible, sur la base de l'analyse distributionnelle, de mettre en évidence les propriétés sémantiques des termes extraits du corpus et de retrouver par la suite les relations sémantiques qui les relient. En effet, en France, cette technique

d'analyse a été largement exploitée pour des applications de construction de ressources terminologiques ou termino-ontologiques (RTO) à partir de corpus spécialisés. Parmi ces travaux, on peut citer, entre autres, les projets menés par Bourigault (2002a, 2002b) autour des outils LEXTER, SYNTEX et UPERY. Il s'agit d'une méthode d'analyse dite *étendue* qui permet, à partir d'un corpus analysé syntaxiquement, de rapprocher les termes ainsi que les contextes syntaxiques en se basant sur des mesures de proximité distributionnelle. En effet, comme le précise Bourigault (2002a : 75), cette technique demande l'utilisation, en amont, d'un analyseur syntaxique de corpus (en l'occurrence, il s'agit de SYNTEX), qui construit un réseau de mots et de syntagmes calculé sur la base des relations de dépendance identifiées dans les phrases du corpus. À partir de ce réseau, le module d'analyse distributionnelle UPERY construit pour chaque terme du réseau l'ensemble de ses contextes syntaxiques. Les termes sont ensuite rapprochés en se basant sur les contextes identiques. Il faut souligner queles outils de ce type se basent sur les critères statistiques de regroupement et mettent en œuvre une méthode *ascendante* (Condamines et Rebeyrolle (2000 : 228), qui vise à faire remonter toutes les informations du texte, sans *a priori* sur le type de données :

« Pour ces outils, on considère plutôt les corpus spécialisés comme relevant d'un sous-langage (au sens harissien du terme), c'est-à-dire d'un système autonome qui a des règles propres. »

(*ibid.*)

La deuxième méthode qui est plutôt basée sur l'interprétation *a priori*, part du principe que les relations entre les termes sont exprimées par des *marqueurs de relation* (Condamines 2005). En effet, cette approche repose sur l'idée qu'il existe un système linguistique unique et stable, régi par des règles que l'on peut expliciter. On parle alors de méthode *descendante* (Condamines et Rebeyrolle 2000 : 233), méthode qui met en œuvre des connaissances *a priori* sur les phénomènes et les projette sur les corpus spécialisés. L'observation des formes linguistiques peut donc révéler un rapport de sens entre différents éléments. En terminologie, cette idée a été développée, entre autres, par Ingrid Meyer (Bowker et L'Homme 2004 :184) qui a introduit le concept de *contextes riches en connaissances* ou *patrons de connaissances* (*knowledge-rich contexts* en anglais*)*, c'est-à-dire des portions de textes qui contiennent des informations concernant la structure conceptuelle d'un domaine donné :

« *By* knowledge-rich context, *we designate a context indicating at least one item of domain knowledge that could be useful for conceptual analysis. In other words, the context should indicate at least one conceptual characteristic, whether it be an attribute or a relation* »

(Meyer (2001: 281) cité dans Bowker et L'Homme (2004 :184))

Dans le cadre de ses travaux, Meyer répertoriait des marqueurs linguistiques en français et en anglais permettant de révéler différentes relations telles que hyperonymie, méronymie, ou des relations transversales.

Les marqueurs permettent donc d'indiquer l'existence d'une relation particulière. Cependant, comme le souligne Condamines (2005 : 46), ils fonctionnent plutôt sur un mode indiciel plus que réellement sémantique, apparaissant comme des éléments déclencheurs d'une éventuelle interprétation sous la forme d'une relation. En effet, appliquée à la construction de ressources terminologiques et termino-ontologiques, une telle approche part de l'hypothèse selon laquelle les relations lexicales repérées en corpus peuvent fournir des indices pour définir des relations conceptuelles. Ainsi, les *marqueurs* sont des éléments linguistiques de toute nature (lexicaux, syntaxiques) auxquels on attribue un statut particulier, métalinguistique, qui donne la possibilité de leur associer une relation conceptuelle (Condamines 2005 : 45). Comme le précisent Aussenac-Gilles et Séguéla (2000 :180) *: « un* marqueur *correspond à une* formule linguistique *dont l'interprétation définit régulièrement le même rapport de sens entre des termes »*. Pour une relation sémantique observable dans le corpus, il existe donc des formules linguistiques prévisibles et récurrentes qui expriment cette relation. En pratique, c'est-à-dire dans le contexte de la recherche de relations sémantiques en corpus, ces formules sont composées de mots de la langue mais aussi de symboles (à condition que le corpus soit préalablement étiqueté) renvoyant à des catégories grammaticales, à des classes sémantiques ou à des rôles argumentaux. On parle ainsi des *patrons de fouille* qui à la différence des *marqueurs* : « *identifient la relation recherchée plus précisément en définissant également des contraintes syntaxiques ou typographiques sur le contexte des termes* (Grabar et Hamon 2004 : 72). Projetés sur le corpus, ils permettent de localiser des structures lexico-syntaxiques correspondantes, c'est-à-dire des énoncés pertinents par rapport à une relation. Remarquons d'après Aussenac-Gilles et Séguéla (2000 :182) qu'une même relation peut prendre de multiples formes et donc s'exprimer par différents marqueurs. De plus, comme le souligne L'Homme (2004 : 223), les mêmes marqueurs peuvent expliciter des relations de nature différente d'un corpus à l'autre. Cette

méthode demande donc de la part du terminologue des connaissances fines sur le fonctionnement linguistique des termes.

On peut se demander quelles connaissances linguistiques il faut mettre en œuvre. En effet, comme le remarquent Condamines et Rebeyrolle (2000 : 226), tout l'art du linguiste étudiant des corpus spécialisés consiste à s'appuyer sur ce qu'il connaît d'un fonctionnement « standard », attendu, de la langue, tout en sachant que ce fonctionnement n'est pas toujours suivi dans les corpus spécialisés. Ainsi, suivant les cas, il peut soit prendre en compte des régularités linguistiques afin de mettre en évidence des phénomènes implicites dans les textes spécialisés, soit s'appuyer sur le repérage d'un fonctionnement déviant par rapport à celui qui est attendu dans la langue générale. Dans le premier cas de figure, le terminologue retrouve les marqueurs par introspection en se basant sur sa connaissance du fonctionnement de la langue générale. Comme le précise Séguéla (2001:30) : *« il imagine les éléments linguistiques par lesquels s'exprime une relation puis évalue les principes de régularité de ces éléments pour désigner cette relation ».* Si l'on reprend les termes de Chomsky, le linguiste cherche à modéliser la compétence linguistique plutôt que la performance ; les formules ne dépendent pas directement de preuves linguistiques attestées (*ibid.* : 29). Cela laisse supposer que les marqueurs d'un type de relation sont universels et pourraient être observés sur tout corpus (Aussenac-Gilles et Séguéla (2000 : 182). On parle alors des marqueurs généraux qui sont associés à des relations considérées comme générales telles que les liens d'hyperonymie ou d'hyponymie, les relations d'holonymie ou de méronymie ou encore les liens de cause et d'effet. L'Homme (2004 : 156) répertorie un certain nombre de marqueurs généraux pour la langue française : par exemple les verbes *contenir* ou *incorporer* peuvent indiquer une relation d'hyperonymie alors que *permettre de* ou *servir à* fournissent des renseignements sur la fonction.

Cependant, certains marqueurs sont étroitement liés à un domaine ou à un corpus donné (ou bien encore, comme le prétend Condamines (2005 : 44-45) à un genre précis) et par conséquent, ils ne sont pas prédictibles par introspection. C'est notamment le cas des relations transversales, souvent dépendantes des domaines. Ces relations ne sont pas attendues, et il est donc difficile de savoir à l'avance quels marqueurs insérer dans une requête afin de ramener les contextes appropriés. Comme le souligne L'Homme (2004 : 223), pour contrer ce type de problèmes, les terminologues font appel à des systèmes d'extraction de relations qui font d'abord émerger les marqueurs productifs d'un corpus avant de procéder à l'identification

proprement dite. Parmi les outils de ce genre, on cite traditionnellement *Prométhée* de Morin (1999) et *Caméléon* de Séguéla (2001). Ce dernier permet l'acquisition de marqueurs spécifiques par recherche de récurrences à partir de couples de termes entre lesquels la relation a été identifiée dans le corpus ou dans un modèle du domaine. Néanmoins, malgré l'automatisation du processus, le repérage des relations spécifiques reste une tâche difficile qui demande de la part du linguiste un va-et-vient constant entre le corpus et les résultats fournis par les outils d'analyse. D'après Condamines et Rebeyrolle (2000 : 233), il s'agit d'un processus interactif où les patrons de fouille sont ajustés au fur et à mesure des réponses fournies en permettant l'identification de nouvelles relations.

En pratique, afin d'obtenir des résultats satisfaisants, les terminologues font appel à différentes approches (basées aussi bien sur l'interprétation *a priori* qu*'a posteriori*). Les méthodes d'analyse décrites ci-dessus (approche structurelle, approche distributionnelle, approche par marqueurs) s'avèrent donc complémentaires et peuvent être utilisées parallèlement En effet, comme le démontrent Condamines et Rebeyrolle (2000), l'interprétation se fait toujours par un ajustement entre connaissances langagières *a priori* et données du corpus, et ce jusqu'à ce que cette interprétation se stabilise. D'où l'importance du choix d'un outil qui devrait être suffisamment souple pour permette de formuler différents types de requêtes. C'est notamment le cas des concordanciers, qui laissent aux utilisateurs beaucoup de liberté en leur proposant de combiner des critères afin de constituer une interrogation adaptée à leurs objectifs.

Après ce passage en revue rapide des diverses approches d'analyse des données terminologiques, nous proposons, dans la partie suivante de notre travail, de présenter la méthode et les outils que nous avons décidé d'utiliser afin de constituer notre modèle de base de données terminographique. Cette démarche s'attache plutôt à mettre en œuvre les connaissances linguistiques sur le fonctionnement des termes qu'à explorer des techniques statistiques. Cependant, comme nous avons pu le constater en analysant les différents projets de construction de ressources terminologiques et termino-ontologiques, il n'est pas possible de se baser sur une seule méthode. Nous sommes consciente que l'analyse de corpus spécialisés consiste en une interaction permanente entre les données textuelles et les résultats obtenus. Cela nécessite le recours à différentes techniques d'investigation.

TROISÈME PARTIE : Le corpus spécialisé, un habitat privilégié des termes : constitution et traitement du corpus *DITerm* en vue d'extraction d'unités terminologiques

« […], un corpus doit "être aimé" : s'il ne correspond pas à un besoin voire à un désir intellectuel ou scientifique, il se périme et devient obsolète. »

Rastier (2005 : 32)

Comme nous l'avons vu dans les chapitres précédents, ces trois dernières décennies, on a assisté à un réexamen des concepts fondamentaux de la théorie classique de la terminologie. L'unité terminologique n'est plus considérée comme une unité que l'on observe à l'état isolé mais comme un élément qui doit être analysé en contexte. Le terme a ainsi retrouvé sa place dans le discours et même si l'utilisation des corpus en terminographie a été plus tardive qu'en lexicographie, on ne peut plus nier les influences de la linguistique de corpus ni les acquis de la terminologie textuelle. Le texte est à la base de tout travail terminographique, il est devenu une valeur confirmée. Comme le remarque L'Homme (2004 : 119) :

« La recherche terminographique repose principalement sur le contenu de textes de spécialité. La collecte d'une documentation représentative du domaine dont on souhaite

décrire la terminologie et son exploitation constituent les premières étapes d'une recherche en bonne et due forme. »

L'objectif principal de ce travail étant une réflexion sur un modèle de dictionnaire thématique orienté vers la mise en discours, une analyse de données textuelles s'impose donc comme une étape essentielle et incontournable de notre démarche. En effet, afin de décrire les usages d'une langue de spécialité, en l'occurrence celle du droit de l'Internet, nous avons été amenée à construire un corpus de textes spécialisés en vue de l'extraction de données terminographiques. Rappelons que selon L'Homme (2004 : 46), une recherche terminographique thématique (et c'est le cas de notre étude) comprend les sept tâches énumérées ci-dessous :

- la mise en forme d'un corpus
- le repérage des termes,
- la collecte de données,
- l'analyse et la synthèse des données,
- l'encodage des données,
- l'organisation des données terminologiques,
- la gestion des données terminologiques,

Les pages qui suivent se concentrent sur la méthodologie adoptée pour réaliser les quatre premières tâches. Nous proposons de présenter ici les techniques et les outils que nous avons mis en œuvre. Mais avant de passer à la description de notre méthode de travail, il nous semble important de commencer par la définition du terme *corpus*.

Chapitre 5. Elaboration du corpus *DITerm*

5.1 Définition du terme *corpus*

En effet, il faut savoir que les corpus sont exploités dans différentes communautés professionnelles et scientifiques : les sociologues, les littéraires, les anthropologues et les psychologues ont tous recours aux données textuelles. Cependant, comme le souligne Williams (2005 : 14), dans le domaine de la linguistique de corpus, le terme *corpus* revêt un sens bien particulier défini dans la littérature de la discipline : « *Dans cette vision stricte, un corpus est forcément constitué suivant des critères de choix sociolinguistiques.* ». Ainsi, à partir des années 90, marquées par la popularité croissante de la linguistique de corpus, on observe une volonté de mieux délimiter et définir le concept. Tognini-Bonelli (2001 : 52-53) et Pearson (1998 : 42-43) proposent, chacune pour sa part, de passer en revue les définitions qui font autorité dans le domaine et qui montrent l'évolution de la notion. Commençons donc par celle proposée par Francis en 1982 (1982 : 7) et citée dans son article de 1992 (1992 : 17) selon laquelle le corpus est :

« *a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis.* »

Pearson (1998 : 43) reproche à Francis de ne pas avoir assez insisté sur la question de la représentativité des données qui est pourtant une question primordiale pour la linguistique de corpus. Selon Pearson, le corpus *n'est pas supposé être* représentatif, il *doit* être représentatif. Cependant, il faut savoir qu'à l'époque où la vision chomskyenne (pouvant se résumer à la fameuse phrase *Corpus linguistics does not exist)* était encore dominante, le simple fait d'évoquer la question controversée de *représentativité* méritait une attention particulière.

Aarts (1991 : 45), quant à lui, définit le corpus comme :

« *a collection of samples of* running text. *The texts may be in spoken, written or intermediate forms, and the samples may be of any length.* »

Ainsi, le corpus peut être composé de documents écrits ou oraux, de longueur différente à condition qu'ils constituent le corps du texte. Cela exclut à notre avis, de simples listes de mots, des glossaires, des thésaurus ou des recueils d'expressions.

Selon Atkins *et al.* (1992), cités respectivement dans Pearson (1998 : 42) et dans Tognini-Bonelli (2001 : 53), le terme *corpus* renvoie à :

« *a subset of an ELT (Electronic Text Library) built according to explicit design criteria for a specific purpose* »

Les auteurs soulignent que le corpus doit être constitué sur la base de critères explicites et ceci en adéquation avec le projet visé. Ils supposent également que les données textuelles doivent être rassemblées sous forme électronique. La question de la forme a aussi été abordée par Leech (1991 : 106) pour qui le corpus est « *a helluva lot of text, stored on a computer*. » C'est en effet le point de vue adopté par la communauté du TAL qui utilise les collections de documents sous forme électronique pour mettre au point ses traitements.

Passons maintenant à la définition qui fait référence dans la communauté des linguistes, à savoir celle proposée par Sinclair dans le cadre du projet EAGLES (1996):

« *A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.* »

Cette définition a été adoptée entre autres par Habert *et al.* (1997 : 146) qui en proposent la traduction suivante :

« *Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage.* »

Comme le remarque Pearson (1998 : 42), Sinclair dans une publication antérieure, notamment dans son fameux ouvrage *Corpus, Concordance, Collocation* de 1991 a défini le corpus comme étant :

« a collection of naturally-occuring language text, chosen to characterize a state or variety of a language »

Sinclair (1991 : 171)

Selon Pearson, en décidant de remplacer l'expression *« naturally-occuring language text* » par *« pieces of language* » Sinclair voulait souligner qu'il ne s'agit pas forcément de textes complets mais aussi de fragments d'énoncés. Cette conception du corpus vu comme « *vaste ensemble de mots* » est pourtant remise en cause par Rastier (2005 : 31) pour qui l'objet empirique de la linguistique est fait de textes qui restent relatifs à un genre et à un discours. Selon l'auteur :

« *Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, et (ii) de manière pratique en vue d'une gamme d'applications.* »

Rastier (*ibid.* : 32)

Pour clôturer cette liste de citations, nous proposons de mentionner la définition de Tognini-Bonelli (2001 : 55) qui selon nous, résume de façon claire et concise ce qui a été dit par ses prédécesseurs:

« a corpus is taken to be a computerised collection of authentic texts, amenable to automatic or semi-automatic processing or analysis. The texts are selected according to explicit criteria in order to capture the regularities of a language variety or a sub-language. »

Alors, quelles sont les principales caractéristiques du corpus telles qu'elles se dégagent des définitions évoquées ci-dessus ? Les auteurs sont d'accord sur le fait que le corpus est un ensemble structuré de textes ou de fragments de textes authentiques, écrits ou oraux, recueillis et stockés sous forme électronique. Le caractère authentique des documents est une condition nécessaire à la constitution d'un corpus car il permet au linguiste de se rapprocher au plus près de l'usage réel et de décrire l'état de la langue telle qu'elle est. Pour Tognini-Bonelli (2001 : 55), le corpus est un « reservoir of evidence » qui doit rendre compte des situations de communication authentiques :

« All the material included in a corpus, […], is assumed to be taken from genuine communications of people going about their normal business. »

(*ibid.*)

De plus, comme le soulignent la plupart des auteurs cités plus haut, la sélection de ces données textuelles doit reposer sur des critères explicites, définis en amont. Comme on le verra plus haut, les critères pris en compte peuvent être aussi bien linguistiques qu'extralinguistiques. Il est également important de noter que le choix des critères dépend de la (ou des) application visée(s). D'après Rastier (*ibid.* : 32) :

« Tout corpus suppose en effet une préconception des applications, […], en vue desquelles il est rassemblé : elle détermine le choix des textes, mais aussi leur mode de « nettoyage », leur codage, leur étiquetage ; enfin la structuration même du corpus. »

Soulignons au passage que le recours à des critères bien précis ainsi que la dépendance à l'égard d'une application différencie les corpus des assemblages arbitraires de textes comme, par exemple le web :

« *[…], it is important to note that a corpus in not simply a random collection of texts, which means that you cannot just start downloading texts haphazardly from the Web and call your collection a 'corpus'* »

(Bowker et Pearson 2002: 10)

La dernière caractéristique du corpus dont nous voudrions parler ici (et peut-être la plus importante), est sa représentativité. En effet, la question de la représentativité est étroitement liée à celle de l'authenticité des textes (évoquée ci-dessus). Selon L'Homme (2004 : 125), le corpus doit constituer un ensemble représentatif de données linguistiques observables dans leur environnement « naturel ». Il est censé être un échantillon de la langue qui rend compte de l'usage effectif, un échantillon à partir duquel on peut établir des généralisations sur fonctionnement de la langue. Tognini-Bonelli cite à ce sujet Leech (2001 : 27) pour qui un corpus est représentatif si les observations menées sur les textes peuvent être généralisées à l'ensemble du langage étudié : *« the findings based on its contents can be*

generalized to a larger hypothetical corpus[57] ». Comme le précise Jacques (2005 : 27), l'enjeu n'est pas tant de rendre compte du phénomène étudié par le linguiste dans le corpus analysé que de dégager des règles d'une portée plus générale. C'est en effet dans l'échantillonnage (Tognini-Bonelli (2001 : 59) utilise le terme anglais *sampling*), que réside l'accès à la représentativité. Comme nous l'avons déjà précisé, la nature et le nombre de textes choisis dépendent des objectifs qui ont été fixés au préalable et plus précisément de l'application visée. Biber (1993 : 243) remarque que les linguistes se concentrent d'habitude sur la taille du corpus en considérant qu'un nombre élevé d'occurrences peut garantir son caractère représentatif. Cependant, la représentativité ne dépend pas uniquement du nombre de mots, mais aussi (et surtout) de ce que Biber appelle *target population* (*ibid.*), expression qu'Habert (2000 : 16) traduit par « population d'évènements langagiers ». Le corpus doit donc représenter des phénomènes qu'on souhaite observer, c'est-à-dire qu'il doit être assemblé en fonction de l'élément à étudier (L'Homme 2004 : 124).

Cependant, il faut savoir que les linguistes sont systématiquement confrontés à des problèmes de représentativité, celle-ci demeurant une question épineuse de la linguistique de corpus. En effet, il est impossible de mesurer de manière objective si un corpus constitue un échantillon représentatif d'évènements langagiers. Habert (2000 : 16) cite à ce sujet Biber, pour qui un corpus est passible de deux types d'erreurs qui menacent les généralisations à partir de lui : « l'incertitude » (*random error*) qui survient quand un échantillon est trop petit pour représenter avec précision la population réelle et la « déformation » (*bias error*) qui se produit quand les caractéristiques d'un échantillon diffèrent de celles de la population que cet échantillon a pour objectif de refléter. Quant à Rastier (2005 : 32), il considère qu'aucun corpus ne représente vraiment la langue. En revanche, pour l'auteur, un corpus peut être adéquat ou non à une tâche en fonction de laquelle on peut déterminer les critères de sa représentativité : *« La linguistique de corpus peut ainsi être objective, mais non objectiviste, puisque tout corpus dépend étroitement du point de vue qui a présidé à sa constitution. »* Cela permet, selon lui, de dédramatiser le problème récurrent de la représentativité.

En abordant la question de la représentativité, nous avons terminé notre tour d'horizon des caractéristiques principales du corpus. Cependant, il est nécessaire de remarquer que tout ce qui a été dit précédemment concerne les corpus « généraux ». En effet, il faut savoir que la

[57] Selon Tognini-Bonelli (2001 : 63) l'expression *hypothetical corpus* renvoie au langage vu dans son ensemble et considéré comme une somme de manifestations individuelles.

linguistique de corpus s'est tout d'abord intéressée au domaine de la langue générale, par conséquent, la plupart des chercheurs abordent la thématique sous cet angle. Ceux qui travaillent sur des corpus de langues spécialisées (ou de spécialité) sont donc amenés à adapter les procédés établis pour des projets portant sur la langue générale. Cependant, il existe des exceptions. Certains auteurs comme Pearson (1998), Bowker et Pearson (2002), L'Homme (2004) apportent un important cadre de réflexion théorique et méthodologique concernant la constitution des corpus spécialisés.

En effet, comme le soulignent Bowker et Pearson (2002 : 11), il existe autant de modèles de corpus que de projets de recherche : « *Language is so diverse and dynamic that it would be hard to imagine a single corpus that could be used as a representative sample of all language.* ». Tout de même, si on se base sur la classification établie par Sinclair dans le cadre du projet EAGLE (1996) ainsi que sur les typologies proposées respectivement par Bowker et Pearson (2002 : 11 – 13) et Habert *et al.* (1997 : 146-147), on peut distinguer les différents genres de corpus suivants :

- corpus de référence et corpus spécialisé
- corpus écrit et corpus oral
- corpus monolingue et corpus multilingue
- corpus multilingue comparable et corpus multilingue parallèle
- corpus synchronique et corpus diachronique
- corpus ouvert (corpus de suivi) et corpus clos (ou corpus fermé)

Ce qui nous intéresse le plus dans le cadre de ce travail, c'est la distinction entre *corpus de référence* et *corpus spécialisé*. Rappelons que selon Sinclair (1996 cité dans Habert *et al.* 1997 : 146) :

« Un corpus de référence est conçu pour fournir une information en profondeur sur une langue. Il vise à être suffisamment étendu pour représenter toutes les variétés pertinentes du langage et son vocabulaire caractéristique, de manière à pouvoir servir de base à des grammaires, des dictionnaires et d'autres usuels fiables. »

Un corpus de référence est donc censé refléter une langue dans son ensemble en permettant de faire des observations d'ordre général. Ainsi, il doit représenter une grande variété de textes

écrit ou oraux, de sources et d'auteurs différents. En revanche, comme le souligne Sinclair (Sinclair 1994 cité dans Pearson 1998 : 46), les corpus spécialisés sont :

« those which do not contribute to a description of ordinary language, either because they contain a high proportion of unusual features, or their origins are not reliable as records of people behaving normally ».

D'après l'auteur, le terme *special corpora* renvoie à tout ensemble de textes qui rend compte de pratiques langagières considérées comme inhabituelles, soit parce qu'elles présentent des caractéristiques linguistiques spécifiques à un usage, soit parce que le comportement discursif des locuteurs est déviant par rapport à la norme. Il s'agit donc d'une définition large qui englobe toutes sortes de corpus représentant des situations de communication atypiques comme le langage enfantin, des énoncés de locuteurs non-natifs ou (ce qui nous intéresse dans le cadre de cette étude), des échanges entre spécialistes dans un domaine technique ou scientifique concerné. Bref, un corpus spécialisé porte sur une situation de communication donnée ou sur un domaine de connaissance particulier. Afin de mieux caractériser la notion de corpus spécialisé, Pearson (1998 : 48) propose de forger un nouveau terme : *special purpose corpus* qu'elle définit comme : « *a corpus whose composition is determined by the precise purpose for which it is to be used.* ». Cette définition met l'accent sur l'importance de la prise en compte de l'objectif visé ce qui est probablement plus important si on mène des recherches en langue de spécialité qu'en langue générale. Plus tard, Bowker et Pearson (2002 : 12) clarifient davantage le concept en soulignant que le « corpus à objectif spécifique » est dédié à l'étude d'un aspect particulier de la langue :

« *[…], a special purpose corpus is one that focuses on a particular aspect of language. It could be restricted to the LSP of a particular subject field, to a specific text type, to a particular language variety or to the language used by members of certain demographic groups (e.g. teenagers).* »

Comme nous pouvons le constater, il s'agit toujours d'une définition relativement large qui ne se limite pas forcément aux langues spécialisées (ou de spécialité). En effet, il est nécessaire de remarquer qu'à part les grands corpus de référence comme le Corpus of Contemporary American English, la plupart des corpus constitués actuellement peuvent être considérés comme appartenant à la catégorie des *special purpose corpora*, c'est-à-dire qu'ils sont dédiés à l'étude d'un aspect particulier de la langue. Rappelons les propos de Rastier

cités plus haut selon lesquels tout corpus dépend étroitement du point de vue qui a présidé à sa constitution. Et même si l'on mène des recherches dans le domaine de la langue générale, le corpus sera assemblé pour répondre à des objectifs spécifiques. Ainsi, afin de bien situer notre étude dans le domaine de la langue de spécialité, il nous paraît important de restreindre le concept de *corpus spécialisé* à des situations de communication spécialisée qui se manifestent par le biais des textes spécialisés (et non pas à des situations de communication atypiques ni à un aspect particulier de la langue). Dans le cadre de ce travail, un *corpus spécialisé* (ou de spécialité) doit donc être considéré comme un ensemble de textes spécialisés, authentiques, assemblés selon les critères déterminés en fonction de l'objectif visé. Avant de passer à la description des critères qui ont présidé à la constitution de notre corpus, il nous semble nécessaire de préciser ce que nous entendons par le terme *texte spécialisé*.

5.1.1 Texte spécialisé – vecteur de la pensée spécialisée

Comme le souligne Cabré (2008 : 37), pour construire un corpus textuel de spécialité, la première question qui se pose est de savoir comment nous pouvons identifier les textes spécialisés. En effet, ce qui nous intéresse ici, ce n'est pas la notion de *texte*, mais plutôt celle de *spécialité*. Précisons que dans notre domaine de recherche, la notion de *spécialité* (ou bien l'adjectif *spécialisé*) peut s'associer aux différents concepts tels que *langue*, *discours* et finalement *texte* et, selon nous, revêt le même sens quel que soit le terme avec lequel elle apparaît. En effet, tous ces concepts partagent une même caractéristique, exprimée par la référence explicite au *spécialisé*, qui constitue leur propriété distinctive commune par rapport à la langue *générale*, discours *généra*l ou texte *général*. Pour cela, il est possible, afin de caractériser la spécialité des textes, de s'appuyer sur les définitions de la langue spécialisée (ou langue de spécialité) ou celles du discours spécialisé.

Comme le remarque Petit (2010 : 3), du point de vue notionnel, il y a une forme de proximité naturelle entre le discours spécialisé et la langue de spécialité (ou langue spécialisée), les deux concepts étant souvent associés (voire parfois assimilés) l'un à l'autre. De l'autre côté, on observe une certaine tendance à assimiler les termes *texte* et *discours*. Néanmoins, il est nécessaire de souligner que la séparation entre ces trois notions a été exprimée en sciences du langage par un certain nombre de définitions classiques. Ainsi, tandis que la langue est considérée comme un système dont les éléments sont actualisés en discours, le discours est assimilé à la parole et conçu comme « le produit de l'exploitation des

ressources qui sont instituées en langue » (Neveu 2004 : 105 cité dans Petit 2010 : 3). En ce qui concerne l'opposition discours/texte, comme le souligne Rastier (2005), selon la tradition greimassienne, le texte est de l'ordre de l'expression alors que le discours est de l'ordre du contenu (la première assimilée au langage, le second à la pensée). En effet, le texte est vu comme un produit, une substance, un objet qui représente la matérialisation d'une situation de communication. Le discours, quant à lui, peut être considéré comme un événement langagier, un ensemble d'énoncés analysés du point de vue du mécanisme de leur production. En effet, les acceptions du *discours* diffèrent d'une école linguistique à l'autre. Par exemple, selon Benveniste (1966 : 242) le terme en question renvoie à *« toute énonciation supposant un locuteur et un auditeur, et chez le premier l'intention d'influencer l'autre en quelque manière* ». Le discours implique donc un acte langagier d'où émergent un texte, un contexte et une intention. Il s'agit d'une entité complexe ayant des dimensions linguistique et situationnelle. Remarquons par ailleurs la grande extension du concept *discours* qui le rend difficile à appréhender : tantôt, il est synonyme de la parole au sens saussurien, tantôt il désigne un message pris globalement.

Ces quelques précisions théoriques ayant été faites, nous pouvons désormais nous concentrer sur la notion de spécialité des textes qui, comme nous l'avons souligné plus haut, peut être définie par le biais du concept de langue de spécialité ou bien de celui du discours spécialisé. En effet, le texte spécialisé est considéré comme un habitat privilégié des termes. Comme le souligne Cabré (1998 : 147), la terminologie est l'aspect le plus frappant des textes spécialisés. Pearson (1998 : 62) utilise, quant à elle, le terme *specialised text* pour désigner les textes caractérisés par une grande densité terminologique : « *We were using the term specialized texts to describe texts with a high density of terms*. ». Pour Bowker et Pearson (2002 : 26), la langue spécialisée (et plus précisément *language for special purposes* dans la tradition anglophone) : *« typically contains a number of specialized terms […].* ». Selon Kocourek (1991b : 71), *« les textes savants saisissent et expriment le contenu savant, dont les unités sémantiques dominantes sont les termes* ». Rappelons que la terminologie a été traditionnellement présentée comme le critère définitoire quasi-exclusif de la langue, du discours ou des textes spécialisés. Comme le souligne Cabré (1998 : 119), certains auteurs ont poussé cette position à l'extrême en définissant les langues de spécialité comme de simples variantes lexicales de la langue générale. Elle cite à ce propos Rondeau (1983 : 23-24 dans Cabré 1998 : 119) selon qui : *« Il faut noter que les expressions "langue de spécialité"*

(langue spécialisée) et "langue commune" ne recouvrent qu'un sous-ensemble de la langue, celui des lexèmes. » En effet, les caractéristiques lexicales des langues, discours ou textes spécialisées ont reçu beaucoup d'attention de la part des spécialistes, primo en raison de l'importance des besoins référentiels, secundo car elles sont facilement repérables. Cependant, le concept de spécialité ne peut pas se réduire à la composante terminologique. Le fait que les éléments de nature lexico-terminologiques spécialisés puissent être rapportés au discours (textes) ou saisis à partir d'eux ne suffit pas toutefois à caractériser ces derniers en tant que spécialisés. L'un des premiers auteurs qui ont refusé de mettre en avant la terminologie au profit d'autres traits définitoires était Lerat (1995 : 21) selon qui :

« *Une langue spécialisée ne se réduit pas à une terminologie : elle utilise des dénominations spécialisées (les termes), y compris des symboles non linguistiques, dans des énoncés mobilisant les ressources ordinaires d'une langue donnée.* »

Comme nous l'avons prouvé dans la partie précédente de ce travail, les langues spécialisées ne sont pas des sous-systèmes linguistiques autonomes. Rappelons que d'après Lerat (*ibid.* : 24), une théorie des langues spécialisées ne peut se fonder que sur une théorie générale de la langue. Cependant, chaque langue (discours, texte) spécialisée se caractérise par l'utilisation d'un certain nombre de moyens d'expression linguistiques particuliers qui englobent bien évidement une terminologie mais aussi des traits stylistiques ou syntaxiques spécifiques. Parmi les aspects linguistiques permettant de distinguer un texte (langue, discours) spécialisé, Cabré (1998 : 136) énumère entre autres : la sémantique générale du texte, la présence ou l'absence de certaines unités, l'usage d'autres codes. Lerat (1995 : 28-29), quant à lui, parle d'une morphologie composite ainsi que d'une syntaxe présentant des prédilections en matière d'énonciation. Nous pouvons donc reconnaître un texte spécialisé par un agencement particulier des éléments de la phrase, des formules stéréotypées, un style particulier, etc. Il est tout de même important de souligner que les moyens linguistiques mis en œuvre en discours spécialisé ne sont pas différents de ceux mis en œuvre en discours ordinaire. La différence est d'ordre non pas qualitatif, mais quantitatif, car elle se manifeste par la fréquence d'emploi :

« *[…] phenomena observed in general language occur with equal, higher or lower frequency in specialized discourse* »

(Gotti 2003 : 308 cité dans Petit 2010 : 5).

Soulignons néanmoins qu'il est difficile de définir le caractère spécialisé d'un texte (discours, langue) en termes uniquement linguistiques. Comme le remarque Petit (*ibid.*), il existe des textes (discours) manifestement spécialisés dont les caractéristiques formelles sont immédiatement reconnues par la grande majorité des interlocuteurs (l'auteur pense plus particulièrement aux textes scientifiques, techniques, médicaux). Cependant, les critères linguistiques ne s'appliquent pas de manière évidente à tous les types de textes spécialisés (notamment historiques, sociologiques, politiques). Il s'avère donc nécessaire de se baser sur d'autres indices. En effet, si l'on analyse les différentes définitions de la *langue spécialisée* (ou *de spécialité*), on s'aperçoit que les auteurs recourent tous à des éléments extralinguistiques et communicationnels pour déterminer la notion de spécialité. Cabré cite à ce sujet Sager et al (1980 : 2 dans Cabré 198 : 120) pour qui :

« Les langues de spécialité sont facilement identifiables comme des sous-divisions pragmatiques ou extralinguistiques d'une langue donnée. On rencontre des difficultés lorsque l'on tente d'expliquer les langues de spécialité en termes uniquement linguistiques ».[58]

Pour les auteurs (Sager *et al.* 1980 : 21 cités dans Lerat 1995 : 20), la langue de spécialité (et plus précisément *language for special purposes*) peut être définie comme l'ensemble des *« moyens de communication linguistique requis pour véhiculer de l'information spécialisée parmi les spécialistes d'une même matière* ».

Hoffmann (1979 : 16) cité dans Cabré (1998 : 118), entend par langue de spécialité (et plus précisément *language for special purposes)* : « *un ensemble complet de phénomènes linguistiques qui se produisent dans une sphère précise de communication et sont limités par des sujets, des intentions et des conditions spécifiques*[59] ».

Cabré rappelle également la position de Picht et Draskau (1985 : 3 dans Cabré 1998 : 120) d'après qui :

[58] « *Special languages are readily recognized as pragmatic and extra-linguistic subdivisions of language. Certain difficulties arise when we attempt to explain special languages satisfactorily in linguistic terms* ».

[59] « *By LSP we understand a complete set of linguistic phenomena occurring within a definite sphere of communication and limited by specific subjects, intentions and conditions.* ».

« *La langue de spécialité (LSP) est une variété linguistique formalisée et codifiée, employée pour des besoins spécifiques et dans un contexte approprié, c'est-à-dire dans le but de communiquer des informations de nature spécialisée à quelque niveau que ce soit*[60]. »

Quant à Lerat (1995 : 20), il affirme que la notion de langue spécialisée rend compte techniquement des connaissances spécialisées :

« […] c'est une langue naturelle considérée en tant que vecteur de connaissances spécialisées. »

Terminons ce tour d'horizon par la présentation des caractéristiques du discours spécialisé (mais cela s'applique également à la langue et au texte spécialisés) considéré du point de vue de l'analyse du discours. Il s'agit notamment de la définition de Gotti qui fait actuellement référence dans le domaine. Ainsi, selon ce dernier, un discours spécialisé est:

« *[…] the specialist use of language in contexts which are typical of a specialized community stretching across the academic, the professional, the technical and the occupational areas of knowledge and practice. This perspective stresses the type of user and the domain of use, as well as the special application of language in the setting*. »

(Gotti 2003 : 24 dans Petit 2010)

Comme nous pouvons le constater en analysant les citations évoquées ci-dessus, la raison d'être des langues (discours, textes) spécialisées est leur rôle dans la transmission de connaissances. Toutes les définitions reconnaissent le fait que les fonctions communicatives et cognitives ont une importance exceptionnelle. On se situe alors dans des perspectives fonctionnelle et pragmatique en déplaçant l'intérêt du niveau linguistique vers le niveau extralinguistique sous-jacent à la communication spécialisée. Envisagée de ce point de vue, la notion de spécialité se trouve donc élargie et enrichie.

Ainsi, parmi les critères définitoires qui se dégagent des définitions évoquées plus haut et que l'on peut retenir afin d'identifier les textes spécialisés, on cite :

- le sujet,

[60] « *LSP is a formalized and codified variety of language, used for special purposes and in a legitimate context – that is to say, with the function of communicating information of a specialized nature at any level.* ».

- le lien étroit avec un domaine particulier (technique, scientifique ou d'activité),
- l'appartenance des interlocuteurs à une communauté des spécialistes,
- la fonction d'usage spécialisé,
- les circonstances particulières de la situation de communication.

Comme le souligne Cabré (1998 : 121), il est possible d'appréhender la notion de spécialité seulement par le sujet. Elle considère que les sujets spécialisés sont ceux dont les contenus notionnels ne sont généralement pas partagés par l'ensemble des locuteurs d'une langue et qui nécessitent un apprentissage particulier (*ibid.* 125). Néanmoins, d'après l'auteure il parait difficile d'identifier un texte spécialisé uniquement par ce biais car les sujets spécialisés peuvent intervenir dans la vie de tous les jours et subir ce que l'on appelle une « banalisation ».

Quant au lien étroit que les textes spécialisés entretiennent avec les différents domaines, Petit (2010) y voit un critère définitoire important en jugeant nécessaire de mettre l'accent sur le statut théorique du concept, dont il propose la définition suivante:

« […], *nous appellerons domaine spécialisé tout secteur de la société constitué autour et en vue de l'exercice d'une activité principale qui, par sa nature, sa finalité et ses modalités particulières ainsi que par les compétences particulières qu'elle met en jeu chez ses acteurs, définit la place reconnaissable de ce secteur au sein de la société et d'un ensemble de ses autres secteurs et détermine sa composition et son organisation spécifique*. »

Petit (2010 : 10)

La vision du domaine spécialisé en tant que domaine d'activité socialement reconnu est assez large et va à l'encontre du découpage traditionnel en disciplines et sous-disciplines universitaires selon lequel on classe habituellement les langues spécialisées. Nous faisons ici référence à la dimension verticale qui caractérise les langues spécialisées par rapport à la thématique. Rappelons d'après Cabré (1998 : 127-128) que l'on établit couramment deux axes de caractérisation des différentes langues spécialisées (ou de spécialité) : l'axe vertical mentionné ci-dessus et l'autre, horizontal permettant, à l'intérieur de chaque langue spécialisée déterminée par rapport au sujet et à la discipline, de distinguer différents degrés d'abstraction qui conduisent à différents niveaux, ou « styles » discursifs.

En effet, la dimension horizontale est liée à un aspect concret de la communication spécialisée, notamment le type de locuteurs. Comme nous avons pu le voir dans la définition de Sager *et al.* citée plus haut, on suppose que les utilisateurs des langues (discours) spécialisées sont des experts qui connaissent les contenus notionnels des domaines donnés. Cependant, comme le soulignent de nombreux auteurs, le concept du spécialisé ne devrait pas se limiter à des situations de communication entre experts. Selon Lerat (1995 : 20), on ne devrait pas exclure les textes à l'usage des non-spécialistes afin d'éviter de creuser un fossé artificiel entre les moyens d'expression des experts et ceux de l'usager (client, justiciable, citoyen, consommateur, lecteur, téléspectateur). Dans le même esprit, Gotti estime que « *the mere presence of a specialist is not sufficient to ensure specialized use of a language, and this in turn is not limited to peer-communication alone* » (Gotti 2003 : 27 dans Petit 2010 : 8). Ainsi, Pearson (1998 : 35-38) propose de distinguer quatre types de situations qui entraînent des degrés variables de spécialisation : communication d'expert à expert, communication d'expert à initié, communication d'expert à non initié, communication à visée didactique. Un bout de ce continuum correspond à la communication entre spécialistes (qui se caractérise par un haut niveau de technicité et par conséquent, une forte densité terminologique), l'autre, à la communication de vulgarisation destinée au grand public (où l'on trouve beaucoup de paraphrases et de données d'ordre définitoire). Cependant, il est important de souligner d'après Cabré (1998 : 125) que si la grande majorité des interlocuteurs moyens sont en mesure d'être récepteurs de communications spécialisées, seuls les individus qui possèdent la connaissance spécifique d'un sujet, connaissance acquise par apprentissage, peuvent produire des communications scientifiques, techniques ou professionnelles. Ainsi, nous supposons qu'un texte spécialisé est forcément produit par un spécialiste.

En ce qui concerne les deux derniers critères définitoires évoqués ci-dessus, notamment l'usage spécialisé et les circonstances particulières de la situation de communication, rappelons que la fonction fondamentale des langues/discours/textes spécialisés est d'informer et d'échanger de l'information sur un sujet spécialisé et ceci dans un contexte particulier. La majorité des textes spécialisés sont donc de nature référentielle et leur forme est déterminée par des critères professionnels ou scientifiques.

En résumant, il semble nécessaire de souligner qu'un texte spécialisé est un objet langagier complexe qui met en jeu un grand nombre d'éléments de caractère plus en moins

linguistique. Pour les identifier, Cabré (*ibid.* : 135-140) propose de tenir compte de trois types d'aspects : linguistiques, pragmatiques et fonctionnels. Nous les retrouverons dans la définition suivante :

« Les textes spécialisés sont les productions linguistiques, orales ou écrites, qui se manifestent dans le cadre des communications professionnelles et dont la finalité est exclusivement professionnelle. On reconnaît les situations professionnelles par les interlocuteurs qui interagissent, par le sujet évoqué qui relève du domaine ou des domaines concernés par la profession, et par la finalité essentielle de rechercher l'information auprès du récepteur, bien que pour ce faire on utilise des stratégies discursives différentes. »

Cabré (2008 : 38)

Cependant, même si, selon nous, la définition de Cabré propose une caractérisation complète du concept *texte spécialisé,* elle ne pourra pas être adoptée telle quelle dans le cadre de cette recherche. Etant donné que notre étude se situe dans le domaine du droit de l'Internet, nous ne pouvons pas parler des communications professionnelles mais plutôt des communications formelles. De plus, la finalité essentielle du discours juridique, comme le dirait Petit (2010 : 7) paraît être d'une nature plus déontique qu'épistémique. En effet, le rôle principal des textes juridiques n'est pas la transmission de l'information ni de la connaissance mais plutôt l'établissement ou l'explication des règles. Nous proposons donc de modifier la définition de Cabré en l'adaptant en fonction des besoins de notre étude. Ainsi, selon nous, les textes spécialisés sont les productions linguistiques, orales ou écrites, qui se manifestent dans le cadre des communications formelles liées à l'exercice d'une activité socialement reconnue et qui mettent en jeu des compétences particulières chez les participants. Cette définition nous servira de guide lors de la constitution de notre propre corpus spécialisé.

5.2 Objectifs visés

Comme nous l'avons souligné plus haut, un *corpus spécialisé* (ou de spécialité) doit être assemblé selon les critères établis en fonction des objectifs visés. Il nous semble donc important de faire un bref rappel des objectifs que nous nous sommes fixés dans le cadre de ce travail de recherche. Ainsi, l'objectif principal de notre travail est de proposer un modèle de description complète des unités terminologiques du domaine du droit de l'Internet qui doit

servir de base à la conception d'un dictionnaire spécialisé destiné aux traducteurs dont la langue de travail est le français. Le projet, baptisé *DITerm*, tente avant tout de répondre aux besoins de compréhension et d'autonomie discursive de ces derniers. Il s'agit d'un modèle qui cherche à rendre compte des usages observés en discours spécialisé afin de permettre aux traducteurs de reconnaître et de générer l'ensemble des emplois. En effet, son ambition est de refléter aussi bien la dimension linguistique des termes (leur nature linguistique, le comportement en langue, leurs relations lexico-sémantiques, les combinaisons lexicales typiques dans lesquelles ils se trouvent) que leur dimension cognitive (la place des termes dans la structure conceptuelle). Le but est de fournir pour chaque unité un grand nombre d'informations de nature linguistique et conceptuelle qui permettront aux traducteurs d'insérer les termes correctement dans les textes spécialisés. Il s'agit donc d'une ressource explicitement dédiée à la mise en discours, une ressource qui avant tout fournit des données nécessaires à l'encodage.

Le *DITerm* doit donc permettre de :

- trouver un répertoire des termes fondamentaux dans le domaine du droit de l'Internet ;
- trouver des descriptions sémantiques fines facilitant la compréhension des notions;
- trouver, pour chacun des termes, l'ensemble des autres termes ou unités lexicales partageant avec lui une relation sémantique ou un lien conceptuel, car la mise en relation des termes du même champ permet de rendre compte de la structure conceptuelle et sémantique du domaine et guide le traducteur dans son approche d'un nouveau domaine ;
- trouver, pour chaque terme, l'ensemble des autres termes ou unités lexicales se combinant de façon privilégiée car la mise en lumière de la combinatoire lexicale permet de refléter la structure lexicale du domaine.

Afin d'atteindre ces objectifs, il faut mettre en œuvre deux stratégies (souvent considérées comme concurrentes ou bien incompatibles, notamment :

- la description détaillée du fonctionnement linguistique des termes dans leur univers discursif basée sur l'observation des usages dans le corpus ;

- la structuration des connaissances relatives au droit de l'Internet extraites du corpus en établissant des réseaux internotionnels entre certaines séries de termes liés entre eux.

Les attentes ainsi définies, le corpus se place donc au cœur de notre démarche. Pour mener à bien notre projet, il est donc essentiel de passer le temps nécessaire à bien le structurer en déterminant les critères en adéquation avec les objectifs visés. La section suivante sera consacrée à la description des critères qui ont présidé à son élaboration.

5.3 Choix des critères

« The decisions that are taken about what is to be in the corpus, and how the selection is to be organized, control almost everything that happens subsequently. The results are only as good as the corpus. »

Sinclair (1991 : 13)

« La valeur d'une recherche terminologique est directement fonction de la qualité de la documentation qui la fonde. »

Dubuc (2002 : 51) dans L'Homme (2004 : 125)

Comme le souligne L'Homme (*ibid.* : 126) en reproduisant la citation de Dubuc (voir ci-dessus), une sélection rigoureuse des textes est garante de la qualité de la recherche menée par la suite. Le choix des critères qui précède le processus de construction d'un corpus spécialisé s'impose donc comme une étape incontournable. En effet, d'après Pearson (1998 : 52), il existe deux types de critères : des critères internes ou linguistiques et des critères externes ou extralinguistiques :

« *Many corpus linguists distinguish between two categories of criteria for the classification of texts in corpora. These categories are 1) external criteria which concern the participants, the communicative function, the occasion and the social setting and 2) internal criteria which concern the recurrence of language patterns within the piece of language […].* »

Rappelons qu'en déterminant les critères permettant d'identifier un texte spécialisé, nous nous sommes concentrée sur la dimension extralinguistique en retenant les éléments suivants : domaine spécialisé, sujet spécialisé, type d'interlocuteurs, fonction d'usage spécialisé, circonstances particulières de la situation de communication spécialisée qui, selon nous, sont

plus opérationnels que les critères linguistiques. Bien évidement, il est possible de reconnaître un texte spécialisé en se basant sur ces derniers mais cela demanderait une analyse linguistique préalable sur un corpus déjà existant. Or, dans le cadre de notre étude, le corpus doit être créé de toutes pièces ; le recours aux éléments extralinguistiques s'avère donc la seule solution possible. On peut se demander toutefois s'il est légitime du point de vue de la linguistique de corpus de délimiter un ensemble de textes uniquement de manière externe. Kocourek (1991a : 19) apporte à cette question une réponse qui nous paraît fort judicieuse :

« *Ce sont les critères externes à la linguistique – c'est-à-dire les critères du domaine – qui décident le premier découpage de la langue, des textes de spécialité. C'est là un point de départ modeste mais essentiel et fructueux, un préalable qui permet de lancer une analyse linguistique des textes spécialisés.* »

Il est donc théoriquement fondé de structurer un corpus en faisant appel aux critères extralinguistiques, en commençant par la délimitation du domaine spécialisé (ou de spécialité).

5.3.1 1er critère – appartenance au domaine du droit de l'Internet par le biais du sujet

Le choix d'une langue spécialisée à analyser nécessite de circonscrire avec précision le domaine que le terminographe souhaite étudier. Comme nous l'avons vu plus haut, il n'est plus possible de s'en tenir à la division historique de la connaissance car : primo, on considère comme spécialisés des champs d'activité qui traditionnellement n'ont pas été caractérisés comme tels (le sport, les loisirs), secundo en raison de l'aspect pluridisciplinaire de nombreux domaines. La pluridisciplinarité (ou bien l'interdisciplinarité) caractérise notamment le domaine dans lequel nous avons choisi de mener notre recherche terminographique, à savoir celui du droit de l'Internet. Plus précisément, nous pouvons parler ici du caractère interdisciplinaire et multidisciplinaire du domaine. Interdisciplinaire parce que, comme le souligne Dimeglio (2006 : 4), on peut considérer le droit de l'Internet comme un ensemble des règles de droit applicables aux activités qui mettent en œuvre l'Internet. Nous avons donc deux disciplines qui entrent en interaction, à savoir le droit et les nouvelles technologies de l'information et de la communication. Par conséquent, le droit de l'Internet peut être considéré comme le produit d'une combinatoire de concepts à partir de ces deux champs de

connaissances. Multidisciplinaire parce que, comme le souligne Céline Castets-Renard (2010), le droit de l'Internet est le droit « de tout ». Il ne peut pas être considéré comme un nouveau droit à part entière car les normes qui le structurent sont tirées d'autres domaines de droit. Les sources du droit de l'Internet sont ainsi multiples et se trouvent dans toutes les branches, comme celle du droit des données personnelles, du droit des contrats, de la consommation, de la concurrence, de la responsabilité, de la propriété intellectuelle, du droit pénal ou du droit international privé. En effet, il paraît nécessaire de souligner que le droit de l'Internet est une matière extrêmement vaste et transversale et traite des facettes les plus variées du web dont, notamment, le commerce électronique, les créations intellectuelles en ligne, la publicité virtuelle, les régimes de responsabilités des grands acteurs techniques de l'Internet (fournisseurs d'accès, hébergeurs), etc. On parle dans ce cas de corpus pluridisciplinaire, notion que Cabré (2008 : 39) oppose à celle de corpus monodisciplinaire.

Compte tenu de ce caractère pluridisciplinaire, il a été très difficile de délimiter le champ de recherche ainsi que de définir les sources à partir desquelles récolter les textes. Ainsi, afin de faciliter cette tâche, nous avons été amenée à réaliser une étude bibliographique concernant la législation française en la matière. Pour ce faire, nous avons sélectionné quatre ouvrages à visée didactique[61], destinés aux étudiants, professionnels et usagers dont la caractéristique commune est d'aborder les différents aspects du droit de l'Internet d'une manière globale. Nous avons ensuite analysé et comparé (manuellement) les structures des quatre documents afin de dégager les axes thématiques principaux du domaine. Soulignons qu'à cette étape de notre travail, il ne s'agissait pas d'effectuer une analyse approfondie et analytique des textes mais de se faire une idée du domaine en repérant les indices externes constituant le paratexte. En effet, nous nous sommes concentrée avant tout sur le contenu des tables des matières, qui constituent selon nous le reflet thématique du domaine. Cela nous a permis d'obtenir une esquisse de la structure conceptuelle du droit de l'Internet nécessaire au repérage et à la collecte de la documentation la plus représentative. Nous sommes consciente que cette démarche va à l'encontre de l'approche sémasiologique, qui ne présuppose pas une pré-structuration de connaissances et ne considère pas les textes comme les attestations d'un

[61] Voici nos références bibliographiques:
CASTETS-RENARD (Céline), 2010, *Droit de l'Internet*, Lextenso éditions, Paris
DIMEGLIO (Arnaud), 2006, *Droit de l'Internet,* Réponses-Questions. Éditions Leyus, Paris.
FAUCHOUX (Vincent) et DEPREZ (Pierre), 2008, *Le droit de l'Internet*, Litec (LexisNexis), Paris
LARRIEU (Jacques), 2010, *Droit de l'Internet*, 2ème édition Ellipses, Paris.

réseau terminologique déjà existant. Cependant, comme nous l'avons souligné plus haut, dans la mesure où aucun corpus ne préexiste à la recherche terminographique, la délimitation du domaine est une étape incontournable qui nécessite par ailleurs une intervention directe ou indirecte (comme c'est le cas dans cette étude) d'un expert.

Ainsi, l'analyse des ouvrages cités ci-dessus nous a permis de dégager 6 axes thématiques principaux. Comme le souligne Castets-Renard (2010 : 8) :

« *Si de nombreux thèmes peuvent être envisagés dans l'analyse du droit de l'Internet, il a paru évident d'en retenir six : donnée personnelles, contrats, propriété intellectuelle, responsabilité délictuelle des acteurs, Internet et international, la sécurité* »

Étant donné que la problématique *Internet et international* aborde les cinq autres aspects, mais à l'échelle européenne et internationale, nous avons décidé de ne pas l'incorporer à notre corpus et de nous concentrer sur les autres thématiques :

Droit de l'Internet

- l'Internet et les données personnelles,
- l'Internet et les contrats
    - les contrats conclus pour l'Internet
    - les contrats conclus par l'Internet – le commerce électronique),
- l'Internet et la propriété intellectuelle,
- l'Internet et la responsabilité délictuelle des acteurs,
- l'Internet et la sécurité
    - la sécurité du réseau
    - la sécurité via le réseau - la lutte contre la cybercriminalité)

Figure 26. Axes thématiques du domaine du droit de l'Internet

Soulignons entre parenthèses qu'il est important d'identifier les sous-domaines du champ à étudier car cela permet non seulement de bien délimiter le domaine, mais aussi de construire des sous-corpus de textes et de veiller à ce que chaque sous-domaine soit

représenté de façon équilibrée. Ainsi, au moment de la sélection des textes, il faut déterminer avec précision si le texte traite du domaine du droit de l'Internet en général et s'il porte sur une des thématiques énumérées ci-dessus. De plus, cette thématique doit être abordée du point de vue juridique et non pas informatique. Nous attirons l'attention sur ce dernier point car il est parfois difficile de distinguer les textes qui traitent de la problématique propre à une branche de droit de ceux qui portent sur les aspects plutôt techniques propres à l'Internet (appartenant selon nous au domaine de l'informatique et non pas à celui du droit de l'Internet). En effet, il n'est pas évident de reconnaître un texte censé appartenir au domaine du droit de l'Internet en se basant uniquement sur le sujet. Cela nous amène à prendre en considération un autre critère, celui des sources.

5.3.2 2ème critère – sélection des sources en adéquation avec le projet visé

Comme nous l'avons vu plus haut, le recours au critère thématique nous paraît essentiel (mais pas suffisant) dans l'identification des textes appartenant au domaine du droit de l'Internet. Le choix des sources en adéquation avec le projet visé est un autre aspect qui doit être pris en compte lors de la constitution du corpus. En effet, la question de l'hétérogénéité thématique a été évoquée comme une difficulté importante dans la construction de notre corpus. Cependant, nous avons repéré d'autres facteurs qui rendent son élaboration difficile, notamment la profusion des textes et la pluralité des sources. Il paraît nécessaire de souligner que le droit de l'Internet est un « nouveau droit », un droit en construction qui attire l'attention de plusieurs communautés. Etant donné qu'il s'agit d'un domaine dont les frontières s'élargissent et changent constamment, on assiste à la prolifération législative : nouveaux textes, circulaires, décrets, jurisprudences, arrêtés. Chaque problématique suscite de surcroît de nombreux commentaires et interprétation de la part des juristes, professionnels, internautes. Tout cela contribue au foisonnement de la production écrite et à la multiplication des sources d'où proviennent les différents documents.

En outre, il convient de souligner que la plupart des textes sont disponibles uniquement sur Internet qui constitue, comme nous le savons, une ressource sans limites : pour les uns – une mine d'or (L'Homme 2004 : 131) pour d'autres (Rastier 2005 : 32) - une décharge publique. En tout état de cause, en utilisant le web comme source de données textuelles, il faut être extrêmement prudent et prendre toutes les précautions méthodologiques. Comme le remarque L'Homme (*ibid.*), en incorporant les documents provenant d'Internet à un corpus, il importe de procéder à une évaluation serrée. Selon l'auteure, l'évaluation a

toujours constitué une étape importante de la collecte de documentation en terminologie. Cependant, elle est plus critique pour les documents puisés dans Internet. Parmi les problèmes auxquels on peut être confronté, Maniez (2008 : 162) souligne entre autres « *The problem of the relative anonymity of the sources and the absence of control mechanisms, such as peer review panels* […]. » En effet, les documents publiés sur Internet n'obéissent pas aux normes classiques d'édition : il est parfois impossible d'identifier l'auteur, l'information n'est pas toujours juste et le contenu n'a pas forcément été relu et corrigé par un comité éditorial. L'Internet est une tribune où toute personne disposant des moyens et des compétences techniques appropriés peut s'exprimer, quel que soit son lien avec le domaine. C'est un espace d'expression en mouvement constant, où les informations sont régulièrement mises à jour. Ce côté dynamique présente toutefois l'inconvénient d'être difficilement contrôlable. Comme le remarque Maniez (*ibid.*), « *[…] the makeup of the Web is, by its very nature, constantly evolving and cannot be controlled […]* ». D'après l'auteur, d'un point de vue purement sociolinguistique, l'ensemble des textes accessibles sur Internet peut être considéré comme représentatif d'une langue donnée. Cependant, on ne peut pas en dire autant de l'ensemble des textes repérés sur Internet et traitant d'un domaine spécialisé car on n'a pas la certitude que ces derniers aient été produits par des spécialistes. Ainsi, en faisant des recherches sur Internet, nous avons repéré une multitude de sites consacrés aux questions juridique du droit de l'Internet: portails généralistes, portails de veille juridique spécialisés dans le domaine, sites consacrés aux nouvelles technologies de l'information et de la communication, blogs, etc.. Nous avons donc décidé de nous tourner vers les sites gérés par des autorités confirmées (comme ceux des institutions européennes, organismes publics et gouvernementaux, associations) qui mettent à disposition du grand public leurs fonds documentaires. L'avantage est que les contenus publiés sur ce type de sites sont diffusés librement et que leur utilisation n'est pas restreinte par le droit d'auteur. Ainsi, aucune autorisation n'a été nécessaire pour incorporer les documents sélectionnés à notre corpus. Voici la liste de nos sources Internet :

1. *Journal Officiel de l'Union Européenne* –

Site : http://eur-lex.europa.eu/homepage.html / http://old.eur-lex.europa.eu/

Le Journal officiel de l'Union européenne (JO) est le recueil officiel de la législation de l'UE et des autres documents officiels des institutions, organes et agences de l'UE. Le site EUR-Lex est un répertoire de la législation consolidée de l'Union européenne et

contient les versions électroniques de toutes les éditions du Journal officiel depuis la première édition. Il donne accès gratuitement et dans les 24 langues officielles de l'Union européenne aux différents types de documents

- droit de l'Union européenne (traités, directives, règlements, décisions, législation consolidée, etc.);
- actes préparatoires (propositions législatives, rapports, livres verts, livres blancs, etc.);
- jurisprudence de l'Union européenne (arrêts, ordonnances, etc.);
- accords internationaux;
- autres documents publics.

La base de données, actualisée quotidiennement, contient plus de 3 millions d'éléments, certains textes remontant à 1951. Chaque document est affiché avec des métadonnées analytiques (référence de publication, dates, mots-clés, etc.) organisées en différents onglets: notice bibliographique; texte; documents liés; procédure (le cycle de vie d'un document).

Un document peut être affiché en deux ou trois langues simultanément. Il existe la possibilité d'afficher et/ou de télécharger des documents à différents formats: PDF, HTML, DOC et TIFF)

source : http://eur-lex.europa.eu/content/welcome/about.html

Soulignons que dans le cadre de cette étude, nous avons utilisé l'ancienne version du site (old.eur-lex.europa.eu) qui est maintenant archivée et est conservée à titre purement informatif. À partir du 1er avril 2014, le contenu du Journal officiel est disponible dans la nouvelle version d'EUR-Lex : eur-lex.europa.eu

2. *CURIA – Recueil de la jurisprudence de la Cour de justice de l'Union européenne –*

Site : http://curia.europa.eu/

Le Recueil est la publication officielle de la jurisprudence des juridictions qui composent la Cour de justice de l'Union européenne. Ce Recueil de jurisprudence est composé d'un Recueil général, reprenant la jurisprudence de la Cour de justice et du Tribunal, et d'un

Recueil de la fonction publique, reprenant la jurisprudence en matière de fonction publique du Tribunal et du Tribunal de la fonction publique. Le Recueil de jurisprudence est publié dans les langues officielles de l'Union. Les données et documents publics relatifs aux affaires portées devant les trois juridictions peuvent être consultés dans la base de données relative à la jurisprudence.

3. *Legifrance*

Site : http://www.legifrance.gouv.fr/

Légifrance est le site web officiel du gouvernement français pour la diffusion des textes législatifs et réglementaires et des décisions de justice des cours suprêmes de droit français. C'est un portail généraliste, grand public, d'accès libre, son slogan est : *Le service public de la diffusion du droit*. Le site renvoie à la totalité des institutions et administrations concernées et à tous les textes encore en vigueur.

4. *LEGALIS – portail du droit des nouvelles technologies*

Site : http://www.legalis.net/

C'est un site d'information et de réflexion consacré aux aspects juridiques d'Internet: une riche base de données de la jurisprudence française en matières des nouvelles technologies de l'information et de la communication, actualités, commentaires, liens juridiques (praticiens, ressources documentaires, conférences, formations)

5. *FORUM DES DROITS SUR INTERNET (FDI)*

Site : www.forumInternet.org

Fondé en 2001 avec le soutien des pouvoirs public, sous la forme d'une association loi de 1901, Le Forum des droits sur Internet a été un organisme de conseil, de dialogue et de réflexion sur les questions juridiques posées par le développement d'Internet. Organisme indépendant, il avait pour mission :

- d'identifier les problématiques liées à l'Internet et à l'utilisation du réseau et d'y répondre efficacement en consultant les principaux acteurs du secteur.
- contribuer à l'élaboration de codes et de chartes de conduite relatifs au développement des réseaux numériques et à leur usage.
- D'informer et de sensibiliser les internautes en les accompagnant dans leur découverte et leur maîtrise du monde numérique

Le FDI a animé un portail d'information à destination du grand public et des acteurs du secteur. Il a aussi mis en place un service DroitDuNet.fr destiné à informer les internautes sur leurs droits. Il a émis plus de 250 fiches pratiques à destination des particuliers. Faute de subventions, l'association a été dissoute en 2010.

Source : http://fr.jurispedia.org/

Figure 27. Sources Internet ayant servi à la constitution du corpus *DITerm*

Outre ces portails destinés à la diffusion de l'information juridique en matière du droit de l'Internet, nous avons également récolté nos documents dans une revue spécialisée qui fait référence dans le milieu des juristes, à savoir la Revue Lamy Droit de l'Immatériel :

RDLI – Revue Lamy Droit de l'Immatériel

La Revue Lamy Droit de l'Immatériel est une revue spécialisée dans le droit des nouvelles technologies de l'information. Elle présente l'actualité législative, réglementaire et jurisprudentielle du droit des "créations immatérielles" (propriété littéraire et artistique, propriété industrielle) et du droit des "activités de l'immatériel" (audiovisuel, presse, informatique, télécommunications, réseaux, commerce électronique) commentée par les experts. Il propose des études, réflexions, opinions, commentaires, illustrations pratiques signés par des spécialistes du domaine. Elle est adressée aux juristes d'entreprise, aux avocats, aux chefs d'entreprise, aux étudiants.

Figure 28. RDLI – revue spécialisée

Nous tenons à souligner qu'afin de garantir l'équilibre de notre corpus, nous avons veillé à ce que tous les sous-domaines soient représentés de façon égale. Par conséquent, nous avons privilégié les sources qui abordent tous les aspects du droit de l'Internet sans se concentrer sur une thématique précise et écarté celles dédiées à une problématique particulière comme par exemple le site de la CNIL consacré à la protection des données personnelles.

5.3.3 Autres critères de représentativité et d'équilibre

Comme nous l'avons souligné au début de ce chapitre, le corpus doit constituer un ensemble *représentatif* de données linguistiques. Nous considérons que la variété et la pertinence des sources sont des éléments importants contribuant à la représentativité de notre corpus. Comme le souligne Sinclair (1991 : 18) : « *The diversity of sources is an essential safeguard* ». Cependant, il est aussi nécessaire de prendre en considération d'autres paramètres comme : la taille, la date de publication, la variété des auteurs, des niveaux de spécialisation, des genres et des situations communicatives.

5.3.3.1 Taille

Comme le souligne L'Homme (2004 : 128), il n'existe pas de véritable consensus en ce qui concerne la taille idéale d'un corpus spécialisé. Rappelons les propos de Sinclair (1991 : 18), qui préconise l'utilisation de corpus aussi grands que possible : « *The only guidance I would give is that a corpus should be as large as possible, and should keep on growing* ». Cependant, ce conseil s'adresse plutôt aux lexicographes qui travaillent sur des corpus de référence. Quant aux corpus spécialisés, leur taille doit rester en adéquation avec les finalités des projets. Il faut savoir que les corpus réunis à des fins terminographiques sont souvent moins volumineux que ceux utilisés en lexicographie. Comme le précisent Bowker et Pearson (2002: 45):

« *It is very important, however, not to assume that bigger is always better. You may find that you can get more useful information from a corpus that is small but well designed than from one that is larger but is not customized to meet your needs.* »

Remarquons qu'à l'heure actuelle, la taille du corpus n'est plus limitée par la disponibilité des textes, ni par les moyens techniques à mettre en œuvre. Au contraire, le problème auquel nous avons été confrontée dans le cadre de cette étude était la façon de réagir face à la multitude des ressources et la définition des critères à adopter pour restreindre le corpus. En effet, vu l'objectif principal, qui vise une description globale de la langue du droit de l'Internet, nous avons décidé de rassembler un maximum de données textuelles liées au domaine. Nous sommes d'accord avec L'Homme (2004 : 129), selon laquelle: « *Un nombre élevé de textes différents constitue un repère plus fiable lorsqu'il est question de décrire des*

usages en cours dans un domaine spécialisé. » Ainsi, initialement notre corpus a atteint le chiffre de 8 millions de chaînes de caractères. Cependant, compte tenu des contraintes pratiques, nous avons été amenés à réduire sa taille pour faciliter son exploitation. Actuellement, le corpus *DITerm* compte 5 111 267 chaînes de caractères et peut être considéré comme un corpus volumineux.

5.3.3.2 Date de publication

Comme c'est le cas pour toutes les recherches synchroniques, nous avons privilégié les textes récents qui reflètent l'état actuel du domaine aussi bien d'un point de vue linguistique que conceptuel. Cependant, nous avons également tenu à intégrer à notre corpus des textes plus anciens qui témoignent des débuts de la discipline. Comme le soulignent Bowker et Pearson (2002 : 52) :

« *[…], older texts can also be valuable. For instance, experts usually provide lots of definitions and explanations when a new concept is developed or a new term is introduced, but these explanations become less frequent as this information becomes part of the experts' general knowledge.* »

En effet, les documents qui correspondent à l'époque où l'on a introduit les concepts fondamentaux du droit de l'Internet contiennent beaucoup plus d'éléments définitoires et de contextes explicatifs que des textes plus récents où les concepts sont déjà bien établis. Ainsi, les documents les plus anciens incorporés à notre corpus remontent aux années 1990: la protection des données à caractère personnel constitue une première étape dans le processus de la régularisation du droit de l'Internet et la directive 95/46/CE – la première (et principale) source[62] en la matière. Soulignons tout de même que ces ressources ne représentent qu'un petit pourcentage des textes sélectionnés. Les autres documents couvrent principalement la période de 2005 à 2011. Il s'agit donc d'un corpus clos (ou fermé) dont le contenu est stable. Compte tenu du caractère très dynamique du domaine choisi, il serait beaucoup plus intéressant de travailler sur un corpus ouvert (ou de suivi), ce qui nous permettrait de tenir à

[62] En effet, il s'agit de la principale source communautaire. Comme le souligne Castets-Renard (2010), la France fut un des premiers pays à envisager de protéger les individus contre l'utilisation de leurs données à caractère personnel avec la Loi n° 78-17 du 6 janvier 1978, et ceci bien avant les premières directives européennes en la matière.

jour les données. Malheureusement, pour des raisons pratiques, nous avons été obligée de renoncer à ce projet.

5.3.3.3 Participants de la situation de communication spécialisée et niveaux de spécialisation des textes

D'après Pearson (1998 : 60), il est essentiel d'utiliser des textes provenant d'auteurs dont les qualifications et le niveau d'expertise sont reconnus dans leur milieu : « *The author(s) must be an acknowledged individual or institution. By 'institution', we mean a body such as a standards institute, an academy or institute of experts. By 'acknowledged', we mean that the authors must be recognized by their peers as having the level of expertise required to write about the subject*. ». Dans le cadre de notre étude, les auteurs sont représentés aussi bien par des spécialistes appartenant à la communauté des juristes que par des institutions. Comme nous l'avons vu plus haut, leur réputation ainsi que leurs connaissances du domaine sont garanties par le choix des sources dites renommées ou confirmées. De plus, il paraît nécessaire de préciser que la plupart de notre corpus est constituée soit par des textes de loi édictés par le législateur, soit par des décisions de justice produites par le juge. Leu autorité en la matière est donc difficile à mettre en cause.

Ainsi, tandis que les auteurs des textes spécialisés sélectionnés sont nécessairement des experts dans le domaine du droit l'Internet, les destinataires (qui constituent le deuxième élément de la situation de communication spécialisée) peuvent avoir soit le même niveau d'expertise que l'auteur, soit un niveau inférieur. Rappelons que Pearson (1998 : 35-38) propose de distinguer quatre types de situations qui entraînent des degrés variables de spécialisation : communication d'expert à expert, communication d'expert à initié, communication d'expert à non initié et communication à visée didactique. En analysant ces quatre types de communication spécialisée, l'auteure suggère d'écarter les textes correspondant à la troisième situation, notamment ceux qui sont adressés au public non initié, c'est-à-dire à un public possédant une formation générale mais aucune connaissance dans le domaine en question.

Cependant, dans le cadre de notre étude, la classification proposée par Pearson ne peut pas être adoptée telle quelle. En effet, les textes de loi (plus de la moitié de notre corpus) constituent un cas à part de par leur caractère général, neutre, abstrait et impersonnel. C'est une situation où le destinataire est mis à distance. Il convient de souligner que la loi ne

concerne pas un individu particulier mais l'ensemble des justiciables. Comme le dit le célèbre adage : *nul n'est censé ignorer la loi*. Le système de justice demande donc à tous les citoyens de connaître la loi ce que signifie, en théorie, que les textes juridiques s'adressent à tous les individus, quel que soit leur niveau d'expertise : spécialiste, initié, profane. Et même si en pratique, la lecture des textes de loi s'avère fastidieuse et nécessite l'interprétation des juristes[63], on ne peut pas considérer cette situation comme un exemple de communication spécialisée entre experts. En ce qui concerne les autres documents incorporés à notre corpus, nous avons veillé à ce que les trois types de situations identifiées par Pearson, à savoir : communication d'expert à expert, communication d'expert à initié, et communication à visée didactique soient représentés afin de donner accès à des variantes de termes ou à des structures utilisées à différents niveaux. De plus, nous avons veillé à ce que les textes proviennent d'une variété d'auteurs afin de neutraliser les particularités stylistiques ou lexicales d'un auteur surreprésenté.

5.3.3.4 Variété de genres, variété de situations de communication – approche du texte juridique

Comme nous l'avons vu plus haut, la variété de sources, d'auteurs, de niveaux de spécialisation assure une certaine représentativité du corpus. À cette liste de critères, on peut ajouter un autre paramètre, notamment la variété de genres qui est liée au but de l'auteur et au type de situation de communication. Comme le soulignent Bowker et Pearson (2002 : 51), la prise en compte de différents types de textes contribue à l'enrichissement de tout projet terminographique : « *In order to ensure a more complete conceptual and linguistic coverage of the LSP in question, you would need to compile a corpus that includes a variety of text types* […]. ». Nous avons donc intégré à notre corpus les documents reflétant différentes situations de communication caractéristiques du domaine du droit de l'Internet. Comme nous l'avons souligné plus haut, la fonction essentielle du discours juridique est normative : le rôle principal des textes de loi est l'établissement des règles. Ainsi, selon Cornu, il existe deux types de discours juridique[64] : le discours législatif, correspondant aux dispositions légales qui

[63] Dans la préface de l'ouvrage, *Lexique de termes juridiques*, outil de travail précieux pour tout étudiant en droit, les auteurs Guillien et Vincent soulignent : « *Le langage des juristes présente pour le non-initié une particularité déroutante […]. En pénétrant dans la sphère du droit, le mot usuel subit une inflexion, parfois même une mutation qui lui confère la précision technique, facteur nécessaire à la sécurité juridique, mais qui l'isole et le rend peu à peu incompréhensible au non-spécialiste.* » (Guillien et Vincent 1970 : préface citée dans Guével 2007 : 82).

[64] En effet, Cornu (2005 : 335 – 337) distingue trois types de discours juridiques: législatif, juridictionnel et coutumier, mais ce dernier n'est pas pris en compte dans la présente étude.

émanent du pouvoir législatif (2005 : 266), et le discours juridictionnel, (*ibid.* : 335) qui est un discours du juge. Bocquet (2008 : 10-11), quant à lui, distingue trois catégories de textes juridiques : les textes normatifs, les textes des décisions qui appliquent ces normes (l'auteur qualifie ces textes de « juridictionnels ») et, enfin, les textes doctrinaux qui exposent le contenu des règles du droit:

« *[…] le droit est lui-même un discours, puisqu'il se définit comme l'énoncé d'un ensemble coordonné de normes. Il existe aussi un discours qui est destiné à concrétiser les normes du droit, à les mettre en œuvre, à les appliquer à des situations imaginaires mais fondées sur leur analogie avec des faits passés (Kalinowski 1974 : 197). Dès lors que le droit est un phénomène observable, il existe aussi un discours qui peut le décrire.* »

Bocquet (*ibid.* : 10)

Dans le cadre de ce travail, nous avons décidé d'adopter la typologie proposée par ce dernier. Ainsi, dans le corpus *DITerm*, la première catégorie de textes (fonctionnant sur le mode performatif[65]), est représentée entre autres par des directives et règlements communautaires ainsi que par des lois, ordonnances, décrets nationaux. La deuxième catégorie de documents (rédigés selon le mode syllogistique) regroupe des arrêts de la Cour de justice de l'Union européenne ainsi que des décisions de justice rendues en France. Finalement, la troisième et dernière catégorie correspondant aux textes doctrinaux (composés selon le mode descriptif) englobe des avis, rapports, recommandations, articles de presse spécialisée, dossiers thématiques. Soulignons que nous avons décidé d'incorporer les textes entiers afin de faciliter les recherches conceptuelles.

Comme nous avons pu le constater, différents paramètres doivent être pris en considération lors de l'élaboration d'un corpus spécialisé. Il paraît nécessaire de s'assurer que le corpus est suffisamment volumineux et diversifié pour être représentatif des usages à décrire.

[65] Bocquet distingue trois modes d'expression du discours juridique – performatif, syllogistique et descriptif.

Chapitre 6. Exploitation du corpus *DITerm* : méthodes et techniques

Rappelons que le projet *DITerm* se fixe comme objectif de proposer un modèle de description globale des unités terminologiques appartenant au domaine du droit de l'Internet, un modèle qui procure toutes sortes d'informations : aussi bien des informations de nature linguistique (sur le fonctionnement du terme dans son univers discursif), que des renseignements concernant la dimension cognitive des termes. Comme nous l'avons vu dans le chapitre précédent, le corpus de textes spécialisés est un vaste réservoir de données terminologiques. Il fournit des attestations des termes ainsi que des renseignements sur leur fréquence d'emploi. De plus, comme le souligne L'Homme (2004 : 120-123), les textes spécialisés renferment d'autres données terminologiques qui permettent de mieux saisir le sens des termes ou de caractériser leur comportement. Ainsi, notre recherche terminographique basée sur le corpus va comporter deux étapes[66], notamment l'acquisition des termes (soulignons qu'il n'existe aucun dictionnaire consacré au droit de l'Internet ; la nomenclature est donc à créer de toutes pièces) et l'extraction de données relatives à chacun des termes sélectionnés. Dans les pages qui suivent nous nous concentrons sur la description de la méthodologie adoptée pour extraire les candidats-termes. Nous proposons de présenter ici les techniques et les outils que nous avons mis en œuvre pour réaliser cette tâche.

Comme nous l'avons vu dans la partie précédente de ce travail, en linguistique de corpus, on peut situer son approche comme relevant soit de la méthode appelée *corpus-based* soit de la méthode que l'on qualifie de *corpus-driven*. Ces deux approches se différencient par rapport à la place que l'on va donner aux hypothèses linguistiques. Rappelons que *corpus-based* signifie une approche déductive qui fait appel aux données textuelles afin de confirmer des hypothèses théoriques tandis que *corpus-driven* fait référence à une approche inductive qui explore les données sans *a priori* dans le but de définir des phénomènes linguistiques directement à partir des textes. Tout comme Condamines et Dehaut (2011 :275), nous n'avons pas voulu opter pour une approche plutôt que pour une autre et nous avons préféré combiner les deux méthodes sachant que chacune apporte des avantages et des inconvénients (décrits

[66] Voir la liste des sept étapes essentielles d'une recherche terminographique proposée par L'Homme (2004 : 46).

précédemment). Bien évidement, tout au long de notre analyse, nous avons essayé de rester au plus près des phénomènes qui apparaissent dans le corpus. Pourtant, nous ne nous sommes pas interdit d'utiliser des connaissances *a priori* sur la langue. Ainsi, en suivant l'exemple de Condamines et Dehaut (*ibid.*), nous avons décidé de nous baser sur deux types d'indices : les indices guidés par ls données et les indices guidés par des hypothèses. Ces indices peuvent être de trois natures : quantitatifs, linguistiques (formels, distributionnels, lexico-sémantiques) et extralinguistiques. En effet, Condamines et Dehaut (*ibid.*) distinguent les indices quantitatifs, formels et distributionnels. D'après les auteures, les indices quantitatifs s'appuient sur les méthodes statistiques et concernent la fréquence d'un phénomène langagier dans un corpus. Les indices formels renvoient à la forme des mots (chaînes de caractères) ou éventuellement des groupes de mots. Quant aux indices distributionnels, ils ont rapport aux contextes dans lesquels apparaissent les unités. À l'instar de L'Homme (2004 : 58), nous proposons d'ajouter à cette liste les indices lexico-sémantiques (permettant d'établir le caractère spécialisé d'une unité lexicale) et extralinguistiques (qui reflètent son appartenance au domaine en question). De plus, selon nous, les indices formels, distributionnels et lexico-sémantiques constituent l'ensemble des indices linguistiques.

6.1 Description des outils d'aide à l'extraction des données terminographiques (*TermoStat* / *NooJ*)

Comme le soulignent la plupart des auteurs cités dans la partie précédente de ce travail (Condamines et Rebeyrolle 2000, Hamon et Nazarenko 2002, Bourigault et Jacquemin 2000), il est tout à fait impossible d'envisager de travailler sur des corpus sans l'assistance d'outils ; les corpus sont trop volumineux pour qu'un traitement manuel soit possible. Dans le cadre de cette étude, nous avons décidé de faire appel à deux outils différents, notamment à *TermoStat*, mis au point par Patrick Drouin, et à *NooJ*, développé par Max Silberztein. Le premier appartient à la catégorie des outils à vocation terminologique construits pour assister le processus d'extraction de candidats-termes. Le second est un outil destiné à l'analyse automatique de corpus à visée plus générale. En effet, nous sommes partie du principe que l'utilisation de ces deux outils devrait correspondre respectivement à deux étapes de notre analyse, notamment au repérage des candidats-termes et à la collecte des données. Cependant, en réalité, *TermoStat* et *NooJ* se sont avérés complémentaires et nous y avons eu recours tout

au long de l'analyse de notre corpus. Il convient également de souligner que chacun des outils, compte tenu de leurs possibilités respectives, a influencé le choix des indices utilisés lors du traitement de notre corpus. *TermoStat*, en tant qu'extracteur de termes, recherche et ramène des données sans l'intervention préalable de l'utilisateur, en se basant principalement sur des critères statistiques. En revanche, *NooJ* fait partie des systèmes qui se caractérisent par la souplesse de l'accès au texte et laissent beaucoup de liberté à l'utilisateur en lui proposant de combiner des critères afin de se constituer une interrogation spécifique, adaptée à ses objectifs propres (Condamines et Rebeyrolle 2000 : 233). Il permet donc de faire des recherches guidées par des hypothèses sur le fonctionnement linguistique des termes.

6.1.1 *TermoStat*

TermoStat est un logiciel d'acquisition automatique de termes développé par Patrick Drouin (2003, 2004, 2006) à l'Université de Montréal. C'est un outil qui a pour objectif l'identification de candidats-termes propres à un domaine spécialisé. Rappelons que les chercheurs travaillant dans le domaine de l'acquisition automatique des données terminographiques préfèrent parler des candidats, c'est-à-dire des unités lexicales susceptibles d'être retenues comme termes. En effet, les extracteurs de termes ne produisent pas des listes parfaites. Comme le soulignent Drouin et Langlais (2006 : 389), l'automatisation du dépouillement est une tâche difficile. L'ordinateur parvient à dresser une liste, mais cette dernière contient toujours certaines unités qui n'ont aucun intérêt terminologique. Les candidats indésirables sont ainsi regroupés sous ce que l'on appelle le *bruit*. Étant donné que les extracteurs de termes ne sont pas en mesure de distinguer les unités terminologiques des autres unités sans risque de se tromper, il revient au terminographe de prendre la décision finale quant au statut et à la nature des unités retenues. Cependant, il est important de souligner que toutes les recherches menées actuellement dans le domaine se concentrent sur l'amélioration des performances des outils d'acquisition automatique. C'est notamment le cas de *TermoStat* qui vise à la réduction du bruit obtenu lors du dépouillement des textes spécialisés : « *Our main goal is to reduce the amount of noise in the list of candidate terms (CTs) by restricting the lexical items that can appear inside candidate terms.* » (Drouin 2003 : 99).

TermoStat s'appuie sur une approche contrastive qui consiste à mettre en opposition le comportement des unités lexicales de corpus de différents niveaux de spécialisation : *« La*

méthodologie proposée repose sur la constitution dynamique d'un corpus hétérogène (technique/non technique) visant à faire ressortir la trace lexicale laissée dans le corpus technique par la terminologie d'un domaine. » (Drouin 2004 : 345). En effet, l'auteur part du principe que les termes spécifiques ont une fréquence « anormalement » élevée dans le texte spécialisé. Sa stratégie consiste donc à comparer la fréquence à laquelle les candidats-termes apparaissent dans le corpus spécialisé avec celle à laquelle ils apparaissent dans un corpus non spécialisé. La démarche proposée requiert donc l'utilisation de deux corpus, un corpus spécialisé nommé *corpus d'analyse* (CA) et un corpus non spécialisé appelé *corpus de référence* (CR). Ce dernier est intégré dans le logiciel et constitue, d'une certaine façon, la norme linguistique sur laquelle sera modelé le comportement des unités lexicales (Drouin et Langlais 2006 : 391). La version disponible en ligne de *TermoStat* prend en charge le français, l'anglais, l'espagnol, l'italien et le portugais. Le corpus de référence français comporte environ 28 500 000 occurrences correspondant à approximativement 560 000 formes différentes. Il est composé d'articles de journaux portant sur des sujets variés tirés du quotidien français *Le Monde* et publiés en 2002. Comme le souligne Drouin (*ibid.* : 347), la diversité des thèmes traités est importante et nécessaire à la démarche puisqu'elle vient minimiser l'uniformité thématique du CR.

TermoStat est un système basé sur un modèle mixte (ou hybride), qui combine des techniques linguistiques et statistiques. Dans un premier temps, les textes qui constituent le corpus d'analyse sont étiquetés par le logiciel d'étiquetage TreeTagger. À partir des textes étiquetés, *TermoStat* extrait une première liste de candidats qui correspondent aux matrices syntaxiques prédéfinies (il s'agit des structures de surface typiques des unités terminologiques). Il est important de souligner qu'un candidat-terme retenu peut être simple (un mot) ou complexe (une suite de mots). Le logiciel réalise ensuite une série des tests statistiques qui ont pour objet de comparer les fréquences des candidats-termes dans les corpus de référence et corpus d'analyse. Pour ce faire, *TermoStat* procède à la constitution dynamique d'un corpus global hétérogène en fusionnant virtuellement le corpus de référence et le corpus d'analyse (Drouin 2004 : 348). Parmi les méthodes statistiques implémentées au sein du logiciel, Drouin et Langlais (2006 : 394) énumèrent : le test x^2, le log-likelihood, le log-odds ratio et le calcul de spécificités. Ce dernier, proposé par *LAF*on (1980) permet de déterminer dans quelle mesure un candidat-terme est spécifique au corpus dépouillé. En effet, tous les tests conduisent à l'obtention d'un poids. Ainsi, un candidat-terme qui obtient un

poids élevé est potentiellement plus intéressant d'un point de vue terminologique qu'un candidat-terme ayant une valeur plus basse (*ibid.* 395).

Ainsi, la liste des candidats-termes générée par le logiciel permet de comparer les scores reçus par les candidats en fonction de la méthode choisie (le calcul de spécificité, le test x^2, le log-likelihood, le log-odds ratio) – la liste est triée par ordre décroissant. Elle comporte également des renseignements supplémentaires tels que la fréquence brute, c'est-à-dire le nombre d'occurrences dans le corpus, les variantes flexionnelles concernant la variation sur le genre et le nombre et la matrice correspondant à la suite de catégories grammaticales de chaque mot constituant le candidat-terme. L'interface du logiciel permet de consulter tous les contextes d'occurrence d'un candidat à l'aide d'un hyperlien. En outre, la page des résultats contient 4 autres onglets, notamment : *Nuage, Statistiques, Structuration, Bigrammes* permettant de compléter les informations sur des termes retenus. Comme nous pouvons le lire sur le site consacré à la présentation du logiciel[67], *Nuage* est la liste alphabétique des 100 termes dont le score est le plus élevé. L'impression de nuage est donnée par la différence de taille de caractère des candidats en fonction du score qui leur a été attribué. La fenêtre *Statistiques* affiche le nombre de candidats sélectionnés pour le texte ainsi que le nombre de candidats pour chaque matrice. Quant à la page des résultats de la structuration, elle permet, pour chaque candidat-terme, d'accéder à la liste des candidats qui l'incluent. De plus, elle donne la possibilité de passer à la page de décomposition où l'utilisateur peut retrouver les mots qui sont en relation avec le terme sélectionné. Il est également possible de générer un graphe des termes ayant des relations syntaxiques partagées. La dernière fenêtre présente les bigrammes les plus forts du corpus analysé qui sont composés d'un verbe et d'un nom (sujet ou objet du verbe).

Le logiciel *TermoStat* version Web est accessible sur le site de l'Université de Montréal et son utilisation est gratuite à des fins de recherche.

6.1.2 *NooJ*

Nous avons présenté *NooJ* comme un outil destiné à l'analyse automatique de corpus. Cependant, il est nécessaire de souligner qu'il s'agit d'un outil très puissant dont les

[67] Il s'agit du guide de l'utilisateur consultable après enregistrement à l'adresse suivante : http://*TermoStat*.ling.umontreal.ca/doc_*TermoStat*/doc_*TermoStat*.html.

fonctionnalités sont multiples. En effet, *NooJ* est un environnement de développement linguistique utilisé comme outil d'analyse de corpus mais aussi et surtout comme outil de formalisation des langues naturelles et de développement d'applications du TAL (voir Silberztein *et al.* 2005)

« *NooJ is a development environment used to construct large-coverage formalized descriptions of natural languages, and apply them to large corpora, in real time.* »

Silberztein (2003: 7)

Ainsi, il permet de construire, de tester et de gérer des descriptions formalisées à large couverture des langues naturelles. Il peut être utilisé afin de formaliser différents phénomènes linguistiques : orthographe, morphologie flexionnelle et dérivationnelle, lexique (mots, mots composés, expressions figées), syntaxe locale, syntaxe structurelle et transformationnelle, sémantique. Ces descriptions formalisées peuvent ensuite être appliquées pour traiter des textes et corpus de taille importante. *NooJ* est aussi utilisé afin de développer des applications du TAL (moteurs de recherche, extracteurs, traduction automatique). Il s'adresse donc à un public varié : aux linguistes, aux informaticiens, aux documentalistes, aux terminologues et même aux enseignants en langues étrangères (voir à ce sujet Silberztein et Tutin 2005). Les utilisateurs de *NooJ* forment une véritable communauté qui développe et partage les ressources dans une vingtaine de langues.

La plateforme *NooJ*, issue de 22 années d'expériences de son auteur, Max Silberztein se caractérise par une structure logicielle complexe basée sur un ensemble de modules linguistiques qui renvoient aux quatre niveaux des langages formels de la hiérarchie de Chomsky-Schützenberger, chaque classe de langage correspondant à un type de grammaire (ou logiciel) : grammaires rationnelles (ou régulières), grammaires algébriques (ou non contextuelles), grammaires contextuelles et grammaires non restreintes. Ces quatre classes de langages sont reconnues respectivement par quatre classes d'automates : automates finis, automates à pile déterministes, automates linéairement bornés, ce qui donne finalement à *NooJ* la puissance d'une machine de Turing. Les quatre types de formalismes, compatibles entre eux, sont graduellement plus puissants au fur et à mesure qu'on monte dans la hiérarchie linguistique, du niveau orthographique vers le niveau sémantique. En effet, *NooJ* impose une approche ascendante, de bas en haut : chaque phénomène doit être traité localement, par des ressources linguistiques locales, (le plus souvent des graphes à états finis). Les ressources sont

ensuite réutilisées, en cascade, pour traiter des phénomènes plus généraux. De plus, le système est équipé d'un ensemble d'outils de développement associés tels que : concordancier, analyses statistiques, éditeur de graphes, débogueur de grammaires. Les utilisateurs disposent donc d'outils de développement et d'analyse adaptés à chaque niveau de description. *NooJ* est un système assez robuste qui prend en charge un grand nombre de langues et répond aux besoins les plus pointus des linguistes et des spécialistes du TAL. Il est nécessaire de souligner que le moteur linguistique *NooJ* a été réécrit à partir de son prédécesseur INTEX (issu des travaux réalisés au sein du LADL) mais présente par rapport à ce dernier des améliorations significatives à tous les niveaux de la formalisation linguistique (Silberztein *et al.* 2005 : 10).

Sur le plan pratique, les différentes ressources linguistiques sont décrites dans deux types de structures : dictionnaires et grammaires (Silberztein 2003 :13-14). Les premiers sont représentés dans des formats textuels, les seconds existent sous forme de graphes. Les uns et les autres sont compréhensibles et accessibles pour un non-informaticien et peuvent être gérées et modifiées sans avoir à maîtriser l'ensemble de l'architecture de *NooJ*. En ce qui concerne le traitement des éléments linguistiques, le système définit et identifie les ALU (*Atomic Linguistic Units*), c'est-à-dire les plus petits composants dont la signification ne peut pas être calculée ni prédite et dont les propriétés doivent être décrites explicitement. (Silberztein 2003 :77). Les ALU sont divisées en 4 classes formelles : mots simples, affixes, mots composés, expressions figées dont chacune correspond à un logiciel de reconnaissance intégré au sein de l'analyseur *NooJ*. Ainsi, pour reconnaître les mots simples et les mots composés ou les expressions figées dont les formes recherchées ne sont pas contigües, *NooJ* effectue des recherches dans les dictionnaires. Afin de découper les chaînes de caractères dans le but d'identifier les affixes, le système fait appel aux descriptions et grammaires flexionnelles et dérivationnelles. Et finalement, pour reconnaître les expressions figées contigües, il applique les grammaires syntaxiques. Toutes les ressources permettant de reconnaître les ALU sont disponibles dans la partie *Preferences* qui contient deux onglets, *Lexical Analysis* et *Syntactic Analysis*, le premier étant divisé en deux zones *Dictionary* et *Morphology*, destinées respectivement à la description lexicale et morphologique. L'utilisateur peut choisir un outillage en fonction du phénomène décrit.

Ainsi, comme nous l'avons mentionné plus haut, les dictionnaires recensent les mots simples, les mots composés, leurs variantes orthographiques, lexicales ou morphologiques. En

effet, les entrées des dictionnaires contiennent les lemmes (Silberztein 2003 :81) accompagnés de toutes sortes d'informations présentées sous forme de codes (un code morpho-syntaxique, des codes syntaxiques et sémantiques, des paradigmes flexionnels et dérivationnels) et de propriétés. Les description des lemmes doivent être exhaustives et ne peuvent contenir aucune règle implicite. Il convient de préciser que les utilisateurs peuvent créer de nouveaux codes et les ajouter à des entrées des dictionnaires *NooJ*. Le but d'une analyse lexicale d'un texte est de mettre en correspondance les unités de ce texte avec le vocabulaire décrit dans des dictionnaires choisis. En ce qui concerne le niveau morphologique, *NooJ* offre deux outils équivalents pour décrire les phénomènes flexionnels et dérivationnels, notamment des descriptions lexicalisées et des grammaires représentées par des ensembles structurés de graphes. Quant à la syntaxe, *NooJ* a recours aux grammaires locales qui sont des grammaires à états finis. Ces dernières sont représentées par des graphes composés d'un ensemble de nœuds connectés et étiquetés incluant un nœud initial et un nœud terminal. Comme nous l'avons souligné précédemment, *NooJ* propose un éditeur de graphe qui permet à l'utilisateur de créer ses propres grammaires, aussi bien morphologiques que syntaxiques.

À part son puissant système d'annotation et les multiples fonctionnalités qu'il offre aux utilisateurs, il est nécessaire de souligner que *NooJ* est un outil très pratique qui peut traiter des ensembles importants de documents codés dans plus d'une centaine de formats, y compris HTML, PDF, XML, MS-Office, etc. Nous voudrions également souligner que l'application qui nous intéresse le plus dans le cadre de cette étude est le concordancier *NooJ*, outil très complexe permettant différents types de fouilles. Nous le verrons plus en détail dans les sections consacrées à l'extraction des unités terminologiques et à la collecte des données.

6.2 Prétraitement du corpus *DITerm*

Comme nous l'avons vu précédemment, l'ensemble des textes constituant le corpus *DITerm* est en format électronique ce qui a considérablement facilité son analyse. Cependant, l'utilisation des outils de traitement automatique, à savoir *TermoStat* et *NooJ*, nous a imposé quelques contraintes d'ordre technique. Premièrement, nous avons été amenée à convertir tous les documents (initialement téléchargés et enregistrés en pdf. ou en doc.), en fichiers txt. En effet, bien que *NooJ* soit capable de traiter une centaine de formats, tous les corpus soumis à *TermoStat* doivent être au format texte brut. Le fait de devoir changer de format de fichier a

eu une conséquence sur le contenu des textes. Plus précisément, nous avons dû procéder au nettoyage des textes en résolvant des problèmes typiquement techniques tels que : problèmes de retour à la ligne, suppression de numérotation, de saut de page. Nous avons également été obligée de supprimer certains images et tableaux ainsi que des références et des notes de bas de page à l'intérieur du corps du texte. Ces deux derniers éléments ont dû être placés en fin de chaque texte (il s'agit surtout des articles provenant de la *Revue Lamy Droit de l'Immatériel*) pour ne pas fausser la recherche de contextes d'un terme. De plus, il est nécessaire de souligner que les textes juridiques sont fortement structurés et possèdent beaucoup d'éléments paratextuels que l'on peut qualifier de stéréotypés. Afin de ne pas alourdir l'analyse, nous avons décidé de les écarter en retenant uniquement les éléments textuels.

Deuxièmement, compte tenu du caractère hétérogène de notre corpus, nous avons décidé de l'organiser en différents sous-corpus. La constitution de ces sous-corpus a été faite en fonction de la thématique (par rapport à cinq sous-disciplines retenues au début de la constitution du corpus) et au regard de la zone de provenance et du type des textes. Ainsi, nous avons tout d'abord divisé le corpus *DITerm* en deux parties :

- la première partie englobe tous les textes communautaires (il convient de souligner que des corpus parallèles existent entre autres en espagnol et en polonais et sont disponibles en ligne)
- la deuxième partie regroupe les textes du domaine du droit de l'Internet qui renvoient au contexte français

Les textes appartenant à ces deux grands sous-corpus ont été eux-mêmes répartis en sous-corpus plus petits. Les résultats de la répartition sont présentés dans le tableau ci-dessous (Tableau 8). Pour plus de détails concernant le type des documents incorporés dans notre corpus, nous renvoyons à l'Annexe 1)

CORPUS	Nombre de documents	Nombre de mots	Type de classification
CORPUS_DROIT INTERNET	**810**	**5 111 267**	**ensemble du corpus**
CORPUS_UE_ ET_JURISPRUDENCE_UE	278	2 023 051	par zone de provenance : ensemble des textes communautaires

		CORPUS_UE	206	1 589 423	par type de textes : textes législatifs communautaires
		CORPUS_DONNÉES	59	557 936	par discipline : textes législatifs communautaires relatifs à la protection des données personnelles sur Internet
		CORPUS_PROPRIÉTÉ INTELLECTUELLE	37	196 368	par discipline : textes législatifs communautaires concernant la question de la propriété intellectuelle sur Internet
		CORPUS_ECOMMERCE	49	477 377	par discipline : textes législatifs communautaires relatifs au commerce électronique et aux contrats conclu sur Internet
		CORPUS_SÉCURITÉ INTERNET	32	190 582	par discipline : textes législatifs communautaires relatifs à la sécurité du réseau et à la cybercriminalité
		CORPUS_STRATÉGIE	29	167 160	par discipline : textes législatifs communautaires concernant la responsabilité délictuelle des acteurs
		CORPUS_JURISPRUDENCE _UE	72	433 628	Par zone de provenance et par type de textes : textes juridictionnels communautaires
	CORPUS_FR		532	3 088 216	Par zone de provenance : ensemble des textes français
		CORPUS_LOI.FR	18	251 252	Par type de texte: textes législatifs français
		CORPUS_LEGALIS.COM	126	457 447	Par type de texte: textes juridictionnels français
		CORPUS_LE FORUM DES DROITS SUR INTERNET	16	217 599	Par type de texte: textes doctrinaux
		CORPUS_RDLI_2007_2012	372	2 161 918	Par type de texte: textes doctrinaux
		CORPUS_RDLI_2007	77	378 267	Par type de texte: textes doctrinaux
		CORPUS_RDLI_2008	79	362 085	Par type de texte: textes doctrinaux
		CORPUS_RDLI_2009	86	433 983	Par type de texte: textes doctrinaux
		CORPUS_RDLI_2010	70	302 735	Par type de texte: textes doctrinaux
		CORPUS_RDLI_2011	78	341 986	Par type de texte: textes doctrinaux
		CORPUS_RDLI_2012	68	342 862	Par type de texte: textes doctrinaux

Tableau 8. Structure du corpus *DITerm*.

La constitution des sous-corpus par sous-discipline, zone et types de texte a eu pour objectif de gagner en flexibilité du point de vue du traitement du corpus, qui est considéré comme volumineux. Cela nous a également permis de repérer des différences et des similitudes lexico-sémantiques entre les données terminologiques provenant des différents documents.

6.3 Extraction des unités terminologiques

Rappelons que nous avons défini deux tâches principales dans le processus de l'exploitation du corpus *DITerm*, à savoir l'acquisition des termes et l'extraction de données relatives à chacun des termes sélectionnés. Dans les pages qui suivent, nous allons présenter les étapes qui nous ont conduite à la constitution de la liste des termes fondamentaux du domaine du droit de l'Internet. Nous aborderons également les difficultés que nous avons rencontrées lors de la réalisation de cette tâche. Soulignons que d'un côté, l'utilisation du corpus volumineux a exclu un dépouillement manuel, de l'autre côté il a été impossible d'automatiser complètement le processus. Comme nous l'avons remarqué plus haut, le repérage de termes est une tâche complexe et laborieuse, et ce malgré l' apport incontestable de la terminologie computationnelle. En effet, un outil informatique doit reconnaître les termes dans une masse au préalable indifférenciée d'unités lexicales (L'Homme 2004 : 167). Pour ce faire, il s'appuie sur un certain nombre d'indices, aussi bien quantitatifs que linguistiques. Parmi ces indices, L'Homme (*ibid.* 168) énumère : la fréquence, la prédominance de termes de nature nominale, la complexité des termes et le nombre fini de séquences pouvant constituer un terme complexe. Cependant, les listes produites par les extracteurs ne sont pas parfaites et doivent être écrémées par le terminographie. On parle alors d'extraction semi-automatisée:

« *[…], some of the candidates on the list may not be terms at all, whereas some actual terms in the corpus may be overlooked by the program and do not appear on the list. Therefore, the list of candidates must be verified by a human, and for this reason, the process is best described as being computer-aided or semi-automatic rather than fully automatic.* »

(Bowker et Pearson, 2002:165)

Il convient de préciser qu'en identifiant un terme, le terminographe se base souvent sur des renseignements qui ne peuvent pas être exploités directement par une machine. Il peut

notamment (ou plutôt il doit !) s'appuyer sur des connaissances extralinguistiques en cherchant à établir un lien entre le sens d'un candidat-terme et le domaine de spécialité en question. Afin de valider ses intuitions concernant le statut terminologique d'une unité, il peut également avoir recours à certains tests lexico-sémantiques (L'Homme 2004 : 168). Dans le cadre de ce travail, nous avons donc été amenée à combiner trois types d'indices : indices quantitatifs, linguistiques et extralinguistiques. De plus, comme nous l'avons vu plus haut, nous avons décidé (à l'instar de Condamines et Dehaut 2011 : 275), de distinguer les indices guidés par les données des indices guidés par des hypothèses. Le but était de constituer une liste d'une centaine de termes centraux du domaine de l'Internet. Nous proposons de décrire notre démarche dans les pages qui suivent.

6.3.1 Les indices guidés par les données

6.3.1.1 Indices quantitatifs

6.3.1.1a La fréquence

Comme le remarque L'Homme (*ibid.*), la fréquence d'une unité dans un ensemble de textes représentatifs constitue un bon indice de son statut terminologique. Un terme significatif est normalement utilisé à plusieurs reprises dans un texte spécialisé. Nous avons donc décidé de commencer par examiner la fréquence d'apparition des unités lexicales dans notre corpus. L'opération d'indexation automatique de l'ensemble du corpus *DITerm* réalisée à l'aide du logiciel *NooJ* a attribué une fréquence à chacune des formes. Il s'agit de la fréquence brute, c'est-à-dire du nombre total d'occurrences de la forme linguistique dans l'ensemble des textes. En effet, *NooJ* découpe les textes en *tokens*[68], chaînes de caractères délimitées par des espaces, considérés comme des éléments de base soumis au traitement. Les *tokens* sont placés dans un index qui peut être organisé soit par ordre alphabétique, soit par fréquence décroissante. Soulignons que l'ensemble de l'index produit à la suite du traitement par *NooJ* du corpus *DITerm* recense 49 375 formes différentes. Vu sa taille importante, nous avons décidé de retenir les mots ayant une fréquence égale ou supérieure à 150 ce qui nous a permis de dresser une liste de 1954 *tokens*. Le tableau ci-dessous (Tableau 9) reproduit les 100 premiers mots de la liste, triés par fréquence décroissante, chaque forme étant assortie d'une indication du nombre d'occurrences.

[68] Les *Tokens* se divisent en *Word Forms* (mots graphiques), *Digits* (chiffres), *Delimiters* (signes de ponctuation et des caractères spéciaux).

1	de, TOKEN + Freq = 330655	35	s, TOKEN + Freq = 16081	68	membres, TOKEN + Freq = 7206
2	la, TOKEN + Freq = 181183	36	sont, TOKEN + Freq = 15706	69	États, TOKEN + Freq = 7173
3	l, TOKEN + Freq = 133676	37	L, TOKEN + Freq = 14772	70	si, TOKEN + Freq = 7173
4	des, TOKEN + Freq = 121268	38	La, TOKEN + Freq = 14010	71	ligne, TOKEN + Freq = 7072
5	à, TOKEN + Freq = 119485	39	son, TOKEN + Freq = 12424	72	caractère, TOKEN + Freq = 7042
6	d, TOKEN + Freq = 111763	40	services, TOKEN + Freq = 12326	73	non, TOKEN + Freq = 6953
7	et, TOKEN + Freq = 106452	41	ces, TOKEN + Freq = 12066	74	traitement, TOKEN + Freq = 6742
8	les, TOKEN + Freq = 96860	42	directive, TOKEN + Freq = 12023	75	ses, TOKEN + Freq = 6686
9	le, TOKEN + Freq = 94085	43	cette, TOKEN + Freq = 11838	76	Commission, TOKEN + Freq = 6666
10	en, TOKEN + Freq = 75841	44	droits, TOKEN + Freq = 11826	77	site, TOKEN + Freq = 6635
11	du, TOKEN + Freq = 74812	45	Le, TOKEN + Freq = 11820	78	autres, TOKEN + Freq = 6609
12	que, TOKEN + Freq = 58258	46	plus, TOKEN + Freq = 11620	79	informations, TOKEN + Freq = 6491
13	un, TOKEN + Freq = 58001	47	société, TOKEN + Freq = 11462	80	entre, TOKEN + Freq = 6458
14	une, TOKEN + Freq = 51283	48	Les, TOKEN + Freq = 11166	81	service, TOKEN + Freq = 6308
15	par, TOKEN + Freq = 49471	49	se, TOKEN + Freq = 10759	82	sans, TOKEN + Freq = 6261
16	est, TOKEN + Freq = 41066	50	peut, TOKEN + Freq = 10130	83	mise, TOKEN + Freq = 6113
17	dans, TOKEN + Freq = 39374	51	protection, TOKEN + Freq = 9688	84	public, TOKEN + Freq = 6087
18	au, TOKEN + Freq = 36940	52	elle, TOKEN + Freq = 9631	85	ci, TOKEN + Freq = 6005
19	ou, TOKEN + Freq = 36788	53	Internet, TOKEN + Freq = 9166	86	cadre, TOKEN + Freq = 5981
20	pour, TOKEN + Freq = 36576	54	leur, TOKEN + Freq = 9109	87	Cour, TOKEN + Freq = 5870
21	sur, TOKEN + Freq = 36286	55	été, TOKEN + Freq = 8965	88	doit, TOKEN + Freq = 5804
22	qui, TOKEN + Freq = 30668	56	fait, TOKEN + Freq = 8711	89	paragraphe, TOKEN + Freq = 5714
23	a, TOKEN + Freq = 27554	57	En, TOKEN + Freq = 8410	90	y, TOKEN + Freq = 5681
24	pas, TOKEN + Freq = 26246	58	Il, TOKEN + Freq = 8340	91	sous, TOKEN + Freq = 5628
25	aux, TOKEN + Freq = 25043	59	comme, TOKEN + Freq = 8337	92	utilisation, TOKEN + Freq = 5545
26	il, TOKEN + Freq = 24800	60	ont, TOKEN + Freq = 8250	93	tout, TOKEN + Freq = 5544
27	qu, TOKEN + Freq = 24491	61	avec, TOKEN + Freq = 8227	94	notamment, TOKEN + Freq = 5504
28	ne, TOKEN + Freq = 23034	62	accès, TOKEN + Freq = 7782	95	application, TOKEN + Freq = 5472

29	n, TOKEN + Freq = 20997	63	même, TOKEN + Freq = 7749	96	soit, TOKEN + Freq = 5458
30	données, TOKEN + Freq = 20941	64	loi, TOKEN + Freq = 7694	97	information, TOKEN + Freq = 5432
31	ce, TOKEN + Freq = 20737	65	ainsi, TOKEN + Freq = 7630	98	responsabilité, TOKEN + Freq = 5415
32	article, TOKEN + Freq = 19585	66	cas, TOKEN + Freq = 7428	99	électronique, TOKEN + Freq = 5375
33	être, TOKEN + Freq = 18897	67	sa, TOKEN + Freq = 7255	100	également, TOKEN + Freq = 5353
34	droit, TOKEN +Freq =16281				

Tableau 9. Liste des 100 premières unités extraites du corpus *DITerm* triées par fréquence décroissante (outil *NooJ*)

Comme nous pouvons le constater, cette première partie de l'index ne revêt pas d'intérêt particulier du point de vue de notre recherche. En effet, la plupart des mots relevés sont des mots grammaticaux (*NooJ* ne permet pas de placer les mots-outils dans ce que l'on appelle une liste d'exclusion). Les autres unités appartiennent soit au vocabulaire général (*utilisation, information, également)*, soit au vocabulaire juridique compris au sens large (*droit, article, cour, responsabilité, paragraphe*), soit à celui de l'informatique (*accès, données, Internet, électronique*). Elles sont donc peu susceptibles de retenir notre attention. Consciente du fait que le résultat brut issu de l'indexation ne nous permettra pas de faire émerger les termes fondamentaux du domaine, nous avons tout de même décidé d'examiner l'ensemble des *tokens* retenus. Il s'est avéré que la technique avait le mérite d'offrir un ancrage dans le corpus. La lecture attentive de la liste (que nous avons également triée par ordre alphabétique) nous a fourni un certain nombre d'informations sur le contenu du corpus. Nous y reviendrons plus tard.

Ainsi, l'analyse de l'index produit à la suite du traitement par *NooJ* s'est avérée peu révélatrice de la présence des termes propres au domaine étudié. D'après L'Homme (2004 : 169), c'est une technique relativement simple à mettre en place qui rend bien des services en terminographie. Toutefois, la fréquence calculée uniquement en fonction du nombre d'occurrences d'une chaîne de caractères ne constitue pas un indice fiable. Comme le souligne l'auteure, un mot peut être fréquent dans un texte spécialisé sans forcément avoir un réel statut terminologique. À l'inverse, un mot peu fréquent peut être un terme central. Nous avons donc décidé de confier le repérage des termes à un programme d'extraction de données terminologiques, à savoir au logiciel *TermoStat*. Ce dernier est basé sur l'approche mixe,

c'est-à-dire que son fonctionnement se fonde aussi bien sur les critères statistiques (la fréquence) que sur les indices linguistiques (la reconnaissance de la structure syntaxique de surface). Comme nous l'avons déjà mentionné, *TermoStat* met en opposition un corpus spécialisé (en l'occurrence le corpus *DITerm* divisé en sous-corpus) et un corpus de référence (le corpus *Le Monde 2002*) afin de faire émerger les unités dont la fréquence dans le corpus spécialisé est proportionnellement plus élevée. D'après L'Homme (2004 : 169), le principe général de cette approche repose sur l'idée voulant que des termes spécifiques aient une fréquence « anormalement » élevée dans le corpus spécialisé.

Ainsi, nous avons soumis à *TermoStat* deux sous-corpus, notamment *CORPUS_FR* et *CORPUS_EU_JURISPRUDENCE*. À la suite de ce traitement, l'extracteur a produit deux listes de candidats-termes regroupant respectivement 14 998 candidats (pour *CORPUS_FR*) et 12 635 candidats (pour *CORPUS_EU_JURISPRUDENCE*). Pour quantifier le potentiel terminologique des unités, le logiciel a utilisé un certain nombre des mesures statistiques : la fréquence brute, le calcul de spécificité, le test du x^2, le log-likelihood, le log-odds ratio. Chaque candidat-terme a reçu un score en fonction de la méthode choisie. En effet, lors de l'affichage des résultats, nous avons pu sélectionner la mesure statistique qui nous intéressait. Le tableau ci-dessous (Tableau 10) montre, à titre illustratif, les 25 premières entrées obtenues à la suite de la comparaison du sous-corpus *CORPUS_FR* et du corpus de référence. Chaque colonne correspond à un type de mesure statistique distinct.

Spécificité	**Score**	**X^2**	**Score**	**Log Likelihood**	**Score**	**Log odds ratio**	**Score**
Droit	264.6	Article	79094.1	Article	42861.7	donnée à caractère personnel	9.11
Société	226.56	Droit	57990.84	Droit	34427.49	mise à disposition	9.09
Donnée	200.9	Donnée	45599.06	Donnée	24665.31	titulaire de droit	8.98
l'article	190.92	Contenu	41184.83	Contenu	22206.88	fournisseur d' hébergement	8.8
Site	175.65	Site	34857.14	Société	20623.78	société ebay	8.8
Service	170.2	Société	32725.56	Site	20076.54	acte de contrefaçon	8.7
Loi	160.18	Disposition	28818.04	Disposition	15891.58	offre légale	8.55
Contenu	159.7	Hébergeur	26468.54	Hébergeur	14770.01	qualité d' hébergeur	8.52
Responsabilité	153.05	Service	21843.46	Service	13937.84	Blog	8.42
Marque	139.05	Contrefaçon	21649.03	Accès	12278.39	Blogs	8.38
Disposition	138.41	Accès	21561.86	Marque	11933.64	mesure	8.29

						technique de protection	
Décision	138.14	Marque	21059.37	Communication	11864.54	fournisseur d'accès à Internet	8.24
Fait	136.52	Prestataire	20871.79	Contrefaçon	11825.29	Demanderesse	8.19
Juge	135.9	Communication	20778.77	Internet	11638.45	accès Internet	8.17
Ligne	135.61	propriété intellectuel	20105.2	Prestataire	11341.55	accès à Internet	8.15
d'une	133.38	Internet	20015.6	propriété intellectuel	10985.16	statut d' hébergeur	8.15
Personne	133.09	Propriété	18090.86	Responsabilité	10640.86	société défenderesse	8.14
Nom	127.25	Responsabilité	18040.95	Loi	9955.29	qualité d'éditeur	8.11
Protection	126.53	Obligation	17362.6	Propriété	9938.79	contenu litigieux	8.05
Mesure	126.36	Logiciel	16728.08	Obligation	9897.36	qualification d' hébergeur	8.05

Tableau 10. Comparaison des mesures statistiques implantées dans *TermoStat*

Comme nous pouvons le constater, les résultats obtenus suite au calcul du X^2 et du log-likelihood présentent une certaine similarité. En revanche, les listes générées à partir du calcul des spécificités et de celui du Log odds ratio sont différentes. Face à ces résultats divergents, il convient de s'interroger sur la performance des différents tests statistiques mis en œuvre par le logiciel. Pour ce faire, nous nous sommes basée sur l'étude de Drouin et Langlais (2006) visant à comparer l'apport des méthodes implantées dans *TermoStat* et dont le but est l'identification des termes au sein de listes produites par l'extracteur. L'expérimentation effectuée par les auteurs a consisté en une évaluation du potentiel des mesures statistiques par rapport à un dépouillement manuel d'un corpus réalisé par un terminologue. Les résultats de ces recherches ont confirmé nos observations concernant une similarité marquée entre les scores générés par le calcul du X^2 et celui du log-likelihood. Quant à la performance dans l'identification des unités terminologique, les mesures testées par Drouin et Langlais ont permis de confirmer l'utilité du calcul de la spécificité qui, avec la fréquence brute, a concentré le plus grand nombre de véritables termes en tête de la liste des candidats-termes.

Compte tenu de ces conclusions, nous avons donc décidé de nous appuyer en premier lieu sur le calcul de la spécificité. Le seuil de spécificité fixé dans *TermoStat* s'élève à 3,18 ce qui veut dire que seules les unités qui ont obtenu un score égal ou supérieur à 3,18 ont été prises en compte. Cela a tout de même permis de faire émerger 14 998 candidats dans

CORPUS_FR et 12 635 candidats dans *CORPUS_EU_JURISPRUDENCE*. Afin de restreindre le nombre des candidats-termes proposé par l'extracteur, nous avons remonté le seuil d'acceptabilité à une valeur de 13,01 ce qui a réduit les deux listes respectivement à 2 224 (*CORPUS_FR*) et à 2 342 (*CORPUS_EU_JURISPRUDENCE*) entrées. Le tableau ci-dessous (Tableau 11) présente 50 premiers candidats-termes extraits à partir de ces deux sous-corpus.

	CORPUS_UE_JURISPRUDENCE_UE				**CORPUS_FR**		
	Candidat de regroupement	**Fréquence**	**Score (spécificité)**		**Candidat de regroupement**	**Fréquence**	**Score (spécificité)**
1	Donnée	13422	365.59	1	Article	13385	264.6
2	Directive	8604	300.49	2	Droit	16585	226.56
3	Paragraphe	5856	252.16	3	Donnée	7550	200.9
4	Service	10829	227.42	4	Contenu	6501	190.92
5	l'article	4559	227.29	5	Site	8562	175.65
6	Protection	6266	218.91	6	Société	12525	170.2
7	Droit	10515	207.84	7	Disposition	5559	159.7
8	Traitement	5266	201.94	8	Hébergeur	3445	153.05
9	Membre	7278	178.9	9	Service	8885	139.05
10	communication	4367	165.76	10	Contrefaçon	2927	138.41
11	Article	4501	164.51	11	Accès	4991	138.14
12	Disposition	3885	163.89	12	Marque	4750	136.52
13	Caractère	3797	162.62	13	Prestataire	2876	135.9
14	Utilisateur	2321	149.79	14	Communication	4878	135.61
15	Information	5226	149.23	15	propriété intellectuel	2716	133.38
16	communication électronique	1806	142.9	16	Internet	5154	133.09
17	caractère personnel	1771	140.99	17	Propriété	3403	126.53
18	Commission	5158	136.26	18	Responsabilité	4985	126.36
19	Règlement	2533	133.27	19	Obligation	4040	123.96
20	consommateur	2535	129.59	20	Logiciel	2870	121.66
21	Prestataire	1606	128.39	21	Fichier	2568	120
22	d'une	1437	127.26	22	Utilisateur	2457	119.28
23	personne concernée	1483	126.54	23	Information	5610	116.01
24	Fournisseur	1850	123.37	24	Loi	7375	115.38
25	consentement	1438	121.4	25	Fournisseur	2420	112.04
26	caractère personnel	1260	119.44	26	Caractère	3335	111.64
27	vie privée	1497	117.97	27	communication électronique	1828	111.54
28	donnée à caractère personnel	1224	117.77	28	nom de domaine	1783	110.07
29	Utilisation	2115	117.4	29	Code	3010	109.87
30	Paiement	1970	116.55	30	Protection	3550	109.1
31	l'union	1161	114.24	31	Application	3061	108.97
32	Mesure	4310	114.13	32	Copie	2322	108.74
33	Réseau	3157	111.87	33	Titulaire	2075	108.32

34	Application	2145	110.57	34	droit d'auteur	1878	108.02
35	Titulaire	1359	109.19	35	Directive	2283	107.5
36	Contenu	1834	108.92	36	Internautes	1907	105
37	Législation	1706	108.91	37	Ligne	4311	99.9
38	Internet	2832	108.9	38	Utilisation	2420	97.85
39	prestataire de service	1085	108.28	39	Editeur	2859	97.41
40	Signature	1912	107.42	40	Internaute	1534	97.34
41	données à caractère	988	105.8	41	Atteinte	1865	96.48
42	Autorité	3665	105.64	42	Ebay	1356	95.17
43	l'utilisation	973	104.99	43	site Internet	1273	91.43
44	Ligne	3161	103.39	44	mot clé	1228	91.34
45	l'Internet	940	103.19	45	Auteur	4589	90.76
46	signature électronique	930	101.57	46	Téléchargement	1233	89.44
47	Cadre	3387	101.18	47	Juge	3734	89.24
48	Procédure	2357	98.53	48	fournisseur d'accès	1224	87.27
49	base de données	978	96.38	49	Usage	2427	86.5
50	l'information	784	94.23	50	Traitement	2435	85.63

Tableau 11. Comparaison des résultats obtenus à partir du calcul des spécificités (*TermoStat*)

6.3.1.1b La répartition

Comme le souligne L'Homme (2004 : 58), la répartition des candidats-termes dans les différents textes formant le corpus constitue, avec la fréquence, un indicateur précieux en terminographie :

« *Une forme peut être extrêmement fréquente, mais dans un seul texte, ce qui diminue son intérêt. Toutefois, si cette fréquence s'observe dans plusieurs textes spécialisés différents, l'unité risque fort d'être significative pour la terminographie.* ».

C'est pourquoi nous avons décidé de diviser notre corpus en différents sous-corpus. Comme nous l'avons déjà vu, l'organisation en sous-corpus a été réalisée en fonction de la thématique (par rapport à cinq sous-disciplines retenues au début de la constitution du corpus) et au regard de la zone de provenance et du type des textes (pour une description plus détaillée, voir le Tableau 8). Cela nous a permis de calculer les spécificités de chacun des sous-corpus, c'est-à-dire de faire émerger les candidats-termes les plus caractéristiques. La comparaison des résultats a également permis d'évaluer la répartition des formes dans l'ensemble du corpus et de repérer des termes communs à tous les sous-corpus. Étant donné que l'objectif de ce travail

est de répertorier les termes centraux du domaine du droit de l'Internet, la recherche des candidats-termes communs à plusieurs sous-corpus a particulièrement attiré notre attention. Il est nécessaire de souligner que l'analyse de la répartition a été réalisée manuellement.

6.3.1.2. Indices formels

6.3.1.2a La prédominance de termes de nature nominale

On admet couramment que la partie du discours privilégiée dans les travaux de terminologie est celle du nom. Comme le souligne L'Homme (2005a : 1119), l'examen des dictionnaires spécialisés révèle que les unités terminologiques de nature nominale représentent entre 84% et 98% de l'ensemble des entrées. D'après l'auteure, ce déséquilibre constitue un reflet de l'optique conceptuelle puisque les noms dénotant des entités sont les meilleurs candidats pour étiqueter un concept. Par conséquent, un très grand nombre d'extracteurs de termes sont conçus pour rechercher uniquement des unités de nature nominale. Ce n'est cependant pas le cas de *TermoStat* qui permet d'évaluer le statut terminologique des unités lexicales appartenant à quatre catégories : adjectifs, adverbes, noms et verbes. Rappelons par ailleurs que l'objectif de notre étude est la description détaillée du fonctionnement linguistique des termes dans leur univers discursif. Nous ne devrions donc pas nous limiter au repérage des unités nominales mais chercher à représenter toutes les parties du discours. Ainsi, en soumettant à *TermoStat* les deux sous-corpus, nous lui avons demandé de prendre en compte l'ensemble des quatre catégories. Les deux premières listes de candidats-termes générées par *TermoStat* étaient donc censées contenir, à côté des noms, des adjectifs, des verbes et des adverbes. Cependant, comme le montre la rubrique *Statistique,* leur nombre dans les deux sous-corpus n'est pas du tout significatif : ils représentent respectivement 5%, 3% et 1% des candidats-termes sélectionnés par *TermoStat*. Quant à la fréquence d'apparition, elle n'est pas non plus révélatrice du potentiel terminologique : les unités appartenant à ces trois catégories n'ont généralement pas été placées en tête des listes produites par le logiciel. Le tableau ci-dessous (Tableau 12), reproduit une partie des données statistiques proposées par le logiciel.

CORPUS_EU_JURISPRUDENCE Nombre de termes : 12 635		CORPUS_FR Nombre de termes : 14 998	
Nom Adjectif	3 832 (30%)	Nom Adjectif	4 184 (28%)

Nom Préposition Nom	2460 (19%)	Nom Préposition Nom	3 209 (21%)
Nom	1470 (12%)	Nom	1 972 (13%)
Nom PPAdj	1296 (10%)	Nom PPAdj	1 598 (11%)
Nom Préposition Nom Adjectif	673 (5%)	Adjectif	797 (5%)
Adjectif	621 (5%)	Nom Préposition Nom Adjectif	700 (5%)
Verbe	366 (3%)	Verbe	475 (3%)
PPAdj	320 (3%)	PPAdj	425 (3%)
Nom Adjectif Adjectif	291 (2%)	Nom Nom	332 (2%)
Nom Préposition Nom Préposition Nom	250 (2%)	Nom Préposition Nom Préposition Nom	282 (2%)
Nom Nom	221 (2%)	Nom Adjectif Préposition Nom	216 (1%)
Nom Adjectif PPAdj	204 (2%)	Adverbe	194 (1%)
Nom Adjectif Préposition Nom	163 (1%)	Nom Adjectif PPAdj	193 (1%)
Adverbe	133 (1%)	Nom Adjectif Adjectif	183 (1%)

Tableau 12. Candidats-termes selon leur appartenance aux parties du discours.

Compte tenu de ces résultats, et dans un souci de clarté, nous avons préféré nous concentrer dans un premier temps sur les unités de nature nominale en écartant les autres parties du discours. Bien évidement, nous sommes consciente du fait que la sélection exclusive de noms est incompatible avec ce qui peut être observé dans les textes spécialisés (L'Homme 2004 : 60). Cependant, en excluant les résultats obtenus pour les adjectifs, les verbes et les adverbes de nos deux premières listes de candidats-termes, nous n'avions pas du tout l'intention de les omettre. Bien au contraire, nous avons décidé de les traiter séparément afin de ne pas les perdre de vue dans la masse écrasante des unités de nature nominale.

6.3.1.2b La complexité des termes

On considère communément (L'Homme 2004 : 168) que les unités terminologiques, sont, dans leur grande majorité, des unités complexes, c'est-à-dire des unités qui se caractérisent par une structure syntaxique complexe et possèdent un sens compositionnel. D'un point de vue technique, un terme complexe est constitué de deux ou plusieurs entités graphiques séparées par des blancs ou par des caractères comme, par exemple, le trait d'union ou l'apostrophe. La plupart des extracteurs ne recherchent donc que les termes complexes en tenant pour acquis qu'ils dégageront ainsi la plus grande partie de la terminologie d'un domaine. Afin de les repérer, ces logiciels mettent en œuvre des techniques statistiques qui permettent de faire émerger les chaînes de caractères qui ont tendance à apparaitre ensemble de façon récurrente dans un corpus spécialisé. Quant à *TermoStat*, il présente l'intérêt de repérer aussi bien les termes complexes que les termes simples, ce qui est très important d'un point de vue terminologique car ces derniers sont, comme nous venons de le voir, bien

souvent laissés de côté par les logiciels d'extraction automatique : « *The system identifies not only complex terms, but also simple terms, which often tend to be ignored by automated systems.* » (Drouin 2003 : 100). Ainsi, en analysant les premières parties des listes produites par *TermoStat* à partir des sous-corpus *CORPUS_FR* et *CORPUS_EU_JURISPRUDENCE* et basées sur le calcul de la spécificité (Tableau 11), nous constatons que les candidats-termes retenus y sont majoritairement représentées par des unités simples. Cependant, il est intéressant d'observer que plus on descend dans la liste, plus les candidats-termes complexes deviennent nombreux. Ceci confirme la thèse de Drouin et Langlais (2006 : 398) selon laquelle le calcul de la spécificité (tout comme le X^2 et le log-likelihood) favorise les termes plus courts.

Le fait de donner la priorité aux termes simples présente quand même un inconvénient. En effet, en examinant les listes produites par *TermoStat,* on ne peut pas savoir si les mots relevés sont des termes simples ou plutôt des parties de termes complexes[69]. Si nous regardons de plus près les résultats présentés dans le Tableau 11, nous pouvons constater par exemple que les candidats-termes : *donnée, caractère, caractère personnel* retenus par *TermoStat* servent (au moins pour une partie des occurrences de chacun d'entre eux) à former un terme complexe, notamment *donnée à caractère personnel* qui pour sa part, occupe la 132ème place dans la liste extraite du *CORPUS_FR*. Il en va de même pour d'autres termes tels que : *fournisseur d'hébergement, fournisseur d'accès à Internet, communication électronique, consentement préalable*, etc. De plus, une analyse rapide des candidats-termes extraits à l'aide du calcul de la spécificité montre que ce sont des termes appartenant plutôt au vocabulaire juridique compris au sens large. Or, ce qui nous intéresse dans le cadre de cette étude, ce sont les termes spécifiques au domaine du droit de l'Internet. Nous devons donc prendre en compte le fait que le droit de l'Internet, de par son caractère pluridisciplinaire, est considéré comme le produit d'une combinatoire de concepts à partir de différents champs de connaissances. Par conséquent, la plupart des termes qui appartiennent à ce domaine ont forcément un sens compositionnel.

Compte tenu de ces observations, nous avons décidé de générer des listes supplémentaires, cette fois-ci faisant appel au calcul du log-odds-ratio. En effet, en comparant les différents scores dont une partie est reproduite dans le Tableau 10, nous nous sommes aperçue que le calcul du log-odds-ratio permet d'extraire beaucoup plus de termes complexes

[69] Le problème a déjà été soulevé par L'Homme (2004 : 173).

que les autres méthodes implantées dans *TermoStat*. Le log-odds-ratio est une technique statistique qui cherche à mesurer le degré d'association entre les mots, c'est-à-dire le caractère non accidentel de la combinaison de deux entités graphiques. Il est nécessaire de souligner que ce type de mesure ramène des termes complexes, mais aussi d'autres séquences, comme par exemples des collocations. Cependant, comme le souligne L'Homme (2004 : 179), même si les différentes données ne sont pas distinguées les unes des autres, cette technique présente l'avantage d'écarter les suites dont l'association est faible. Le tableau ci-dessous (Tableau 13) compare les résultats obtenus à la suite du calcul de la spécificité et du log-odds-ratio.

	CORPUS_FR				**CORPUS_FR**		
	Candidat de regroupement	**Fréquence**	**Score (spécificité)**		**Candidat de regroupement**	**Fréquence**	**Score Log-odds-ratio**
1	Article	13385	264.6	1	donnée à caractère personnel	586	9.11
2	Droit	16585	226.56	2	mise à disposition	575	9.09
3	Donnée	7550	200.9	3	titulaire de droit	512	8.98
4	Contenu	6501	190.92	4	fournisseur d'hébergement	429	8.8
5	Site	8562	175.65	5	société ebay	430	8.8
6	Société	12525	170.2	6	acte de contrefaçon	388	8.7
7	Disposition	5559	159.7	7	offre légale	334	8.55
8	Hébergeur	3445	153.05	8	qualité d' hébergeur	326	8.52
9	Service	8885	139.05	9	Blog	294	8.42
10	Contrefaçon	2927	138.41	10	Blogs	281	8.38
11	Accès	4991	138.14	11	mesure technique de protection	258	8.29
12	Marque	4750	136.52	12	fournisseur d'accès à Internet	246	8.24
13	Prestataire	2876	135.9	13	Demanderesse	232	8.19
14	Communication	4878	135.61	14	accès Internet	228	8.17
15	propriété intellectuel	2716	133.38	15	accès à Internet	223	8.15
16	Internet	5154	133.09	16	statut d' hébergeur	223	8.15
17	Propriété	3403	126.53	17	société défenderesse	222	8.14
18	Responsabilité	4985	126.36	18	qualité d'éditeur	215	8.11
19	Obligation	4040	123.96	19	contenu litigieux	202	8.05
20	Logiciel	2870	121.66	20	qualification d' hébergeur	202	8.05
21	Fichier	2568	120	21	téléchargement illégal	203	8.05
22	Utilisateur	2457	119.28	22	communication électronique	1828	8.05
23	Information	5610	116.01	23	dénomination sociale	200	8.04
24	Loi	7375	115.38	24	contenu mis	198	8.03
25	Fournisseur	2420	112.04	25	exception de copie	198	8.03

Tableau 13. Comparaison des résultats obtenus à la suite du calcul de la spécificité et du log-odds-ratio (*TermoStat*).

Ainsi, le dépouillement du corpus *DITerm* à l'aide du logiciel *TermoStat* s'est fait de plusieurs manières en prenant en compte différents critères tels que : mesure statistique, catégorie grammaticale ou distinction par type de sous-corpus. Cette démarche nous a permis d'obtenir 18 listes de candidats-termes différentes en fonction des critères choisis. Bien évidement, il s'agit d'une première approximation du statut terminologique car les formes retenues ne correspondent pas toutes à des termes. Comme le souligne Van Campenhoudt (2010 : 2), les logiciels d'extraction de candidats-termes véhiculent une certaine idée particulière du terme :

« Il pourra ainsi être tout à tour conçu comme un syntagme conforme à un patron morphosyntaxique particulier, comme une suite de caractères qui se situe aux frontières de certaines parties du discours, comme un figement quantifiable par la lexicométrie …] ».

Cependant, l'auteur ajoute à juste titre que les indices statistiques ou morphologiques de figement ne peuvent suffire à justifier le statut de terme pour une expression récurrente « […] *il n'y a guère d'adéquation entre les notions de* terme *et de segment répété […].* » Ainsi, bien que les données quantitatives constituent un indicateur précieux, elles ne peuvent être utilisées aveuglément sans l'application de critères additionnels (L'Homme 2004 : 58). En effet, elles ne permettent que de révéler des tendances. Les résultats proposés par *TermoStat* doivent donc être considérés comme autant d'indices qu'il faudra évaluer, comparer, interpréter et compléter.

Outre un nombre important de candidats-termes, un examen minutieux des résultats proposés par *TermoStat* nous a permis d'esquisser un portrait thématique du corpus. En effet, le logiciel a donné la possibilité de nous familiariser assez rap*ibid.*ent avec le domaine du droit de l'Internet L'affichage des candidats-termes selon leur fréquence ou leur spécificité nous a permis de prendre connaissance, par le biais des unités extraites par le logiciel, des notions les plus importantes du domaine. Nous y reviendrons.

6.3.2 Les indices guidés par des hypothèses

La méthode mise en œuvre lors de l'étape précédente de ce travail correspond plutôt à l'approche ascendante dans laquelle on récolte et exploite les données textuelles sans *a priori* théorique. Il s'agit des résultats quantitatifs qui constituent un indicateur important de la

présence des termes dans le corpus. Rappelons que de manière générale, les termes ne se distinguent des autres unités lexicales ni sur le plan formel, ni dans leur comportement dans les phrases : certaines unités lexicales spécialisées se conduisent de manière semblable aux unités du lexique général. Le travail d'identification terminologique basé sur les critères linguistiques paraît donc complexe et difficile, mais pas impossible. En effet, Krieger (2002) en décrivant les propositions théorico-méthodologique adoptées lors de l'élaboration d'un glossaire multilingue du droit international de l'environnement a attiré l'attention de la communauté scientifique sur le rôle des spécificités des univers de discours dans la constitution de leurs terminologies. En effet, selon l'auteure, le statut terminologique d'une unité se lie fortement à des particularités des contextes discursifs où l'unité s'insère. Une recherche de données terminologiques requiert donc une série de considérations sur le fonctionnement d'un discours spécialisé donné.

6.3.2.1 Sur la piste des indices morpho-syntaxiques - quelques caractéristiques du discours juridique

Ainsi, nous sommes partie de l'hypothèse selon laquelle la terminologie du droit de l'Internet s'inscrit dans ce que l'on appelle communément le *discours juridique* pris au sens large (rappelons que le langage du droit englobe plusieurs genres discursifs – la plupart étant représentée dans notre corpus). Ce dernier est un type de communication spécialisée qui se différencie bien évidement par un vocabulaire spécialisé mais aussi par des structures discursives propres. En effet, comme le soulignent de nombreux auteurs (Sourioux et Lerat 1975, Cornu 2005, Gémar 1991, Bocquet 2008), la lecture d'un texte juridique nous met devant un ensemble discursif particulier, où l'on retrouve de nombreux éléments spécifiques. Comme l'ont montré ces différentes études, il est possible de dégager des traits caractérisant les textes juridiques aussi bien sur le plan morphosyntaxique qu'aux niveaux sémantiques ou stylistiques : il y a « quatre éléments généralement reconnus comme constitutifs du langage du droit : le sens, la syntaxe, le lexique et le style » (Gémar, 1991 : 275). Dans cette section, nous nous intéresserons plus particulièrement aux phénomènes morphologiques et syntaxiques inhérents au langage du droit qui par ailleurs reflètent ses spécificités stylistiques. Il ne s'agit pas ici de relever et de décrire des paramètres morphosyntaxiques en se basant sur l'observation de notre corpus (leur statut est suffisamment établi pour faire l'objet d'une nouvelle étude), mais de nous appuyer sur les recherches existant déjà dans le domaine afin de

faire émerger des unités terminologiques. En effet, nous supposons que la prise en compte des marques morphologiques et syntaxiques offre des critères pour l'identification du répertoire terminologique du domaine du droit de l'Internet. Pour ce faire, nous devons partir d'un certain nombre de considérations *a priori* sur le fonctionnement de la langue juridique.

Selon nous, une des caractéristiques principales du discours juridique est la condensation nominale. Comme nous l'avons vu, la place prépondérante donnée au nom est une particularité commune à toutes les langues spécialisées. Des études quantitatives menées entre autres par Van Campenhoudt (2010 : 5) ont montré que les textes identifiés comme spécialisés tendent à se distinguer d'autres, « plus littéraires », par un nombre supérieur de substantifs et un nombre inférieur de verbes et de pronoms. Les résultats obtenus à la suite du traitement du corpus *DITerm* à l'aide du logiciel *TermoStat* ont confirmé ces tendances. En effet, la prédominance des noms est liée à un autre phénomène très courant dans le discours spécialisé, notamment celui de la nominalisation (Sager 1990, Lerat 2007). Comme le souligne Lerat (2007 : 81), la nominalisation consiste en le remplacement d'une formulation verbale (ou adjectivale) par une formulation nominale. Précisons que les noms résultant de cette opération sont des unités prédicatives (c'est-à-dire des unités auxquelles on peut attacher des arguments). En règle générale, elles représentent les actions, états et processus. Le recours à la nominalisation dans les textes spécialisés s'explique tout d'abord par des raisons référentielles : les noms prédicatifs renvoient aux concepts plus directement que les expressions verbales. L'emploi du style nominal permet également de réaliser une certaine économie de moyens linguistiques et de répondre aux exigences de concision. Van Campenhoudt (2010 : 7) cite à ce sujet Kocourek (1991a : 79-82) selon qui le souci de concision « *constitue un facteur puissant dans la formation des phrases technoscientifiques*. ». En effet, le phénomène de nominalisation (à côté d'autres mécanismes évoqués par Kocourek[70]) fait augmenter la densité d'un texte spécialisé menant à ce que Kocourek appelle une « condensation syntaxique » qui donne une « complexité concise » à la phrase. Comme le souligne Van Campenhoudt, il ne s'agirait pas tant de produire des phrases brèves que de les charger sémantiquement sans les allonger inutilement : « *Grâce à la nominalisation, on condense fréquemment sous la forme d'un nom ou d'un syntagme nominal une idée qui pourrait faire l'objet d'une prédication indépendante* ».

[70] Il s'agit des mécanismes de condensation tels que constructions participiales, gérondives, infinitives (Kocourek 1991a : 81-82.).

Ainsi, la nominalisation facilite la création de syntagmes nominaux qui sont particulièrement fréquents dans les textes spécialisés (nous avons pu le constater plus haut). Van Campenhoudt (2010 : 5) attire l'attention sur l'existence de syntagmes nominaux aux modes de formation ultracomplexes qui, à la différence de ce qui est habituellement reconnu comme critères d'identification des syntagmes figés dans la langue courante, adoptent un comportement original. L'auteur précise que les unités complexes repérées dans des textes spécialisés peuvent par exemple intégrer des déterminants, contenir des prépositions variées, inclure des coordinations ou bien ne pas se conformer aux règles de la syntaxe. En effet, dans le cas des syntagmes nominaux spécialisés, le complément du noyau nominal est souvent exprimé par un syntagme prépositionnel dont le régime est une deuxième nominalisation dont le complément est un autre syntagme prépositionnel dont le régime est une nouvelle nominalisation et ainsi de suite. Ceci permet de placer dans une seule proposition plusieurs relations. Il convient également de souligner que les syntagmes nominaux dont le noyau est un nom déverbal (donc à caractère prédicatif) correspondent aux schémas d'arguments que Lerat a qualifiés de spécialisés (Lerat 2002, 2006). Selon ce dernier (2002 : 157), une expression prédicative se reconnait à ce qu'elle a besoin d'objets pour être syntaxiquement complète, de même qu'un prédicat logique a besoin d'être « saturé » par des arguments.

L'auteur précise aussi que : « […] *l'usage d'un prédicat spécialisé, […], appelle celui d'arguments fortement contraints […]* (Lerat 2006 : 91). En effet, d'après Lerat, la prise en compte des phénomènes de prédication s'avère très utile dans les langues spécialisées car de nombreuses expansions actancielles permettent d'accéder à un nombre important d'informations : « *[…] en s'attachant à des prédications spécialisées (= une expression prédicative et un argument in situ) dans le discours […]; il y a là en effet un grand intérêt, d'une part parce que les expressions prédicatives spécialisées représentent les actions, états, processus et propriétés caractéristiques d'un métier, d'autre part parce qu'autour d'elles se distribuent plus ou moins les arguments typiques* ».

Tout ce que nous venons de dire à propos des langues spécialisée se rapporte bien évidement à la langue juridique. Cependant, il paraît nécessaire de souligner que le discours du droit privilégie le style nominal pour des raisons supplémentaires. En effet, en plus de répondre aux exigences d'économie linguistique et de concision terminologique évoquées ci-dessus, le recours à la nominalisation dans le langage du droit permet d'obtenir un degré important de dépersonnalisation de l'énoncé. Rappelons rap*ibid.*ent que le discours juridique

est le produit d'une instance dépersonnalisée et qu'il se distingue par son caractère général, neutre, impersonnel et abstrait. Or, toute forme verbale exige l'indication de la personne, la spécification du nombre, du temps et du mode, ce qui rend l'énoncé plus concret et plus précis. La transformation de la forme verbale en nom offre en contrepartie la possibilité d'occulter le sujet de l'action. De plus, les rédacteurs de documents juridiques recourent au style nominal car il permet de présenter la réalité de façon neutre et objective.

Quel est le rapport entre ces faits et notre étude ? De fait, nous considérons que les caractéristiques du discours juridique énumérées ci-dessus, à savoir la nominalisation et la présence de syntagmes nominaux aux modes de formation ultracomplexes, peuvent servir d'indices linguistiques importants dans l'identification des termes du domaine du droit de l'Internet. Pour ce faire, nous avons mis en œuvre des techniques consistant en le repérage des fonctionnements considérés comme réguliers. Nous avons défini un certain nombre de patrons morphosyntaxiques sous forme de règles que nous avons projetées sur le corpus afin de localiser les séquences correspondantes.

Lors de cette étape, nous avons eu recours au logiciel *NooJ,* qui se caractérise par la souplesse de l'accès au texte et laisse beaucoup de liberté à l'utilisateur en lui proposant de combiner des critères afin de se constituer des interrogations spécifiques. Plus précisément, nous avons utilisé la fonction *Locate pattern* qui nous a permis d'appliquer au corpus (traité cette fois-ci dans son intégralité), un certain nombre d'expressions régulières propres à *NooJ* et correspondant aux patrons morphosyntaxiques identifiés auparavant. La technique d'interrogation proposée par *NooJ* exploite le corpus préalablement étiqueté. L'opération d'étiquetage est réalisée à l'aide de la fonction *Linguistic analysis,* qui permet un traitement lexical et syntaxique du corpus. En effet, lors de cette analyse, *NooJ* applique les ressources existantes et sélectionnées par l'utilisateur ou bien créées par ce dernier selon les besoins de l'étude (dans le cadre de ce travail, nous avons décidé de faire appel aux dictionnaires et grammaires déjà implantés dans *NooJ*). Le processus se décompose en deux étapes : la première correspond à l'analyse lexicale réalisée à l'aide des dictionnaires et des grammaires morphologiques et flexionnelles ; la deuxième renvoie à l'analyse syntaxique du corpus effectuée par des grammaires locales. Le corpus ainsi étiqueté offre la possibilité de mener des interrogations à un niveau plus général, en se basant sur des annotations d'ordre morphosyntaxique mais aussi sémantique.

Ainsi, la fonction *Locate pattern* permet de réaliser quatre types de recherche : 1) recherche d'unités fixes, de mots particuliers 2) recherche de chaînes de caractères correspondant à un patron codé sous forme d'expressions régulières de *NooJ*, 3) recherche de chaînes de caractères correspondant à un patron codé sous forme d'expressions régulières en Perl, 4) recherche de chaînes de caractères à l'aide des grammaires locales *NooJ*. Dans le cadre de cette étude, nous avons fait appel aux expressions régulières de *NooJ* qui, selon nous, permettent un traitement rapide, souple et fiable des chaînes de caractères. En effet, une expression régulière peut être considérée comme un petit programme avec sa propre syntaxe dont le seul objectif est d'examiner une chaîne de caractères et d'indiquer si elle correspond ou ne correspond pas au modèle (il n'existe pas de correspondance approximative). Autrement dit, une expression régulière est une façon concise de représenter une famille de chaînes de caractères. Les expressions régulières *NooJ* peuvent être composées de lettres, de chiffres, de symboles spéciaux, d'un nombre précis de caractères correspondant aux opérations de base telles que la disjonction, la concaténation, la fermeture, ainsi que de codes associés à des catégories morpho-syntaxiques ou à des propriétés sémantiques (pour une liste complète des opérateurs utilisés dans ce travail – voir Annexe II). En effet, toutes les informations représentées dans les dictionnaires *NooJ* peuvent être utilisées dans nos requêtes (Silberztein 2003 : 28-53). Les résultats des recherches sont donnés dans la fenêtre *Concordance.* Il est nécessaire de souligner que les tables de concordances offrent plusieurs possibilités de traitement : on peut nettoyer manuellement les résultats affichés en excluant le bruit et les hapax, consulter un rapport statistique, exporter les concordances et les index en plusieurs formats, etc.

6.3.2.1a 1ère hypothèse – les termes appartenant au domaine du droit de l'Internet sont des mots composés

Comme nous l'avons vu plus haut, on tient pour acquis que la plupart des termes complexes dans les langues spécialisés sont des syntagmes nominaux. Certains travaux ont également montré que ces derniers se construisent au moyen d'un nombre fini de séquences de parties du discours (L'Homme 2004 : 168). Il est nécessaire de souligner qu'un grand nombre d'extracteurs, comme par exemple *TermoStat*, utilisé dans la première partie de notre étude, font appel à ces indices. Basés sur des méthodes mixtes ou linguistiques, ils s'appuient sur des matrices syntaxiques prédéfinies pour repérer les candidats-termes. Cependant, il s'agit de matrices universelles. Or, nous supposons que chaque langue spécialisée se distingue

par des modalités de composition propres au domaine. Par conséquent, nous avons décidé de dégager, à l'aide du logiciel *NooJ,* des séquences régulières de parties du discours caractéristiques de la langue du droit afin de compléter les résultats proposés par *TermoStat*. Pour ce faire, nous nous sommes appuyée sur les observations faites par Cornu.

Ainsi, selon Cornu (2005 : 168) : « *La complexité croissante du système juridique profite de la composition pour nommer les nouvelles figures de l'ordre juridique.*». Pour l'auteur, la composition est une source vive de néologie. Il remarque que les deux tiers des termes du vocabulaire juridique sont des mots composés. Il attire aussi l'attention sur un grand nombre de combinaisons possibles parmi lesquelles il énumère les suivantes :

- apposition d'un substantif à un autre avec ou sans trait d'union,

- juxtaposition d'un substantif et d'un adjectif (il parle des adjectifs de relation, qui jouent le rôle de complément de nom)

- compositions binaires à cheville qui associent deux substantifs mais avec à l'aide d'une cheville (article, adverbe, préposition)

- composition avec verbe

- ensembles soudés (séquences figées) caractérisées par l'indivisibilité du sens d'ensemble

Les règles de composition évoquées par Cornu nous ont permis d'identifier un certain nombre de patrons lexico-syntaxiques sous formes d'expressions régulières composées de codes syntaxiques de *NooJ* et de mots grammaticaux (prépositions, déterminants, articles). Voici une liste des patrons recherchés accompagnés d'exemples des séquences correspondantes localisées par *NooJ* (les chiffres placés à droite se rapportent à la fréquence d'apparition).

<N><N> (nous avons retenu les occurrences qui apparaissent 25 fois ou plus dans le corpus)			
État membre	1401	sites web	227
site Internet	1043	réseau Internet	193
mots clés	854	lien hypertexte	144
droits voisins	692	connexion Internet	110
Loi type	470	support papier	109

haut débit 354 jeu vidéo 249 accès Internet 233	société Web 72 œuvre multimédia 70
<N>-<N> (nous avons retenu les occurrences qui apparaissent 10 fois ou plus dans le corpus)	
plates-formes 621 plate-forme 548 États-Unis 474 mots-clés 382 dommages-intérêts 290 sous-traitant 272 sous-traitants 186 décision-cadre 167 non-respect 161 mot-clé 118 artistes-interprètes 107 e-commerce 107	e-mail 76 e-accessibilité 67 artiste-interprète 65 contrat-cadre 50 CD-ROM 40 programme-cadre 38 e-paiements 25 e-réputation 24 e-mails 20 décret-loi 20 code-source 19 plate-formes 19 E-commerce 15
<N><A> (nous avons retenu les occurrences qui apparaissent 25 fois ou plus dans le corpus)	
communications électroniques 2448 commerce électronique 1261 données personnelles 1081 copie privée 1047 gestion collective 783 mesures techniques 662 économie numérique 552 signature électronique 550 fichiers électroniques 47 fichiers musicaux 167 infractions pénales 101	autorité compétente 92 acte modificatif 51 fichiers personnels 33 motifs légitimes 42 fracture numérique 44 intermédiaire technique liens commerciaux 500 liens publicitaires 53 logiciels espions 39 intérêt collectif 26 offre illégale 25
<N><A>*de<N><A>* (nous avons retenu les occurrences qui apparaissent 20 fois ou plus)	
nom de domaine 1783 Cour de cassation 1271 base de données 1308 titulaires de droits 828 moteur de recherche 739 noms de domaine 693 traitement de données 668 groupe de travail 664 projet de loi 698 droits de propriété intellectuelle 626 juridiction de renvoi 480 bases de données 444	services de communications électroniques 342 mesures techniques de protection 258 réseaux de communications électroniques 121 forum de discussion 127 données de connexion 105 éditeurs de logiciels 94 contrat de maintenance 49 contrefaçon de marque 41 fournisseurs de réseaux publicitaires stockage de données 39

prestataire de services 310 sociétés de gestion collective 385	fournisseurs de contenus 39 droit de rectification 26
<N><A>*(a|en|au|aux|à la|pour|contre)<N><A>* (nous avons retenu les occurrences qui apparaissent 10 fois ou plus)	
données à caractère personnel 1626 accès à Internet 519 mise en ligne 434 services en ligne 228 pair à pair 102 contenus mis en ligne 133 rémunération pour copie privée 145 données relatives au trafic 144 commerce en ligne 95	communications électroniques accessibles au public 92 accès aux données 68 accès au site 45 diffusion en ligne 37 contenus créatifs en ligne 34 accès public à Internet 33 médias audiovisuels à la demande 31
<N><A>*(de|de la|du|des|à|au |aux|en|pour|contre) <N><A>* (de|de la|du|des|à|au|aux|en|pour|contre) <N><A>*(de|de la|du|des|à|au|aux|en|pour|contre)*<N>*<A>* (nous avons retenu les occurrences qui apparaissent 10 fois ou plus dans le corpus)	
Service de communication au public 369 traitement de données à caractère personnel 313 traitement des données à caractère personnel 280 responsable du traitement des données 67 arrêt de la Cour de cassation 57 protection juridique des bases de données 56 titulaire du nom de domaine 54 jugement du Tribunal de grande instance de Paris 46 enregistrement de noms de domaine 37 fournisseur de moteur de recherche 29 fournisseur de service de référencement 10 fournisseurs de services de communications électroniques 72 générateur de mots-clés 52 plateforme de réseau social en ligne 10	

Tableau 14. Extraction de candidats-termes à l'aide des patrons lexico-syntaxiques (*NooJ*).

Outre les syntagmes nominaux, les textes spécialisés se caractérisent par la présence d'un autre type de formes linguistiques récurrentes. Leur compréhension pose d'ailleurs des problèmes de compétence qui vont au-delà du linguistique. Il s'agit notamment de sigles et d'acronymes définis respectivement comme « *terme complexe abrégé ou non formé des*

lettres initiales de ses éléments » et : *« terme complexe abrégé formé de plusieurs groupes de lettres d'un terme et dont la prononciation est exclusivement syllabique*. » (Lerat 1995 : 58). Selon l'auteur (*ibid.*), c'est justement dans les langues de spécialité que l'on observe la plus grande fréquence d'emploi de ces éléments. Nous avons donc décidé de vérifier la présence de sigles dans notre corpus et d'en recenser les plus fréquents. En examinant leurs contextes, nous avons constaté que certains sigles fonctionnent comme des variantes orthographiques des termes complexes.

<UNK+MP="[A-Z]"> (ou UNK = mots non reconnus par *NooJ*)
UE 1480 (Union Européenne) LCEN 1252 (La loi pour la confiance dans l'économie numérique) IP 1245 (adresse IP) CNIL 714 5 Commission nationale de l'informatique et des libertés) FAI 667 (fournisseur d'accès Internet) HADOPI 510 (Haute Autorité pour la diffusion des œuvres et la protection des droits sur Internet) TGI 451 (Tribunal de grande instance) DVD 395 (« Digital Versatile Disc ») RFID 228 (radio-identification ou « radio frequency identification ») DADVSI 202 (La loi relative au droit d'auteur et aux droits voisins dans la société de l'information) MTP 150 (mesures techniques de protection) BCR 199 (règles d'entreprise contraignantes, ou « Binding Corporate Rules »)

Tableau 15. Extraction de sigles (*NooJ*).

Tous ces résultats proposés par *NooJ* nous permettrons d'enrichir les listes de candidats-termes extraits par *TermoStat*.

6.3.2.1b 2ème hypothèse – le langage du droit de l'Internet a recours à la nominalisation

Comme nous l'avons vu plus haut, il est d'observation courante que les nominalisations dont le contenu conceptuel correspond aux formulations verbales équivalentes sont particulièrement fréquentes dans les textes spécialisés. D'après Lerat (2007 : 79) : *« Dans le cas des textes techniques, on peut tirer parti des nominalisations, qui ont la réputation d'y proliférer*. » Bien évidement, comme le souligne l'auteur (*ibid.* : 82), les nominalisations ne sont pas forcément des termes ; il arrive très souvent que les noms

prédicatifs résultant de nominalisations appartiennent à la langue courante. Il est tout de même important de les prendre en considération lors de la recherche terminologique car leur analyse constitue un accès intéressant (et relativement simple) au contenu du corpus spécialisé. Il s'agit ici de l'approche basée sur l'étude de la structure interne des termes qui suppose l'exploitation des indices morphologiques. En effet, comme le soulignent les auteurs comme Lerat (*ibid.*) ou Grabar et Hamon (2004), il est possible de s'appuyer sur la morphologie dérivationnelle pour repérer des unités terminologiques ainsi que des relations qui les relient. Parmi les opérations dont dispose la morphologie, c'est l'affixation (et plus particulièrement, la suffixation) qui paraît la plus intéressante du point de vue de notre recherche.

Ainsi, afin de faire émerger des candidats-termes issus d'une nominalisation nous avons décidé d'opérer en deux temps. Premièrement, nous avons dégagé un certain nombre de suffixes de dérivation nominale exprimant une action, un résultat, un processus pour constituer un inventaire des formes déverbales présentes dans notre corpus. Ensuite, nous avons procédé au repérage des termes complexes dont le noyau est un nom déverbal. À l'instar de Lerat (*ibid.*), qui a mené une étude sur le phénomène de nominalisation dans des textes technico-administratifs communautaires, nous nous sommes intéressée plus particulièrement aux formes en *–tion*, qui présentent plusieurs avantage pour un traitement automatique. Tout d'abord, cette forme n'est pas du tout bruyante. De plus, le pluriel est rare en cas de nominalisation. Finalement, le suffixe *-tion* évite de distinguer *-ation, -ition* et *-ution*, en les englobant. Étant donné que les formes en *–tion* ne constituent qu'une partie (certes majoritaire) dans les textes spécialisés, nous avons étendu notre investigation à d'autres terminaisons. Nous avons donc effectué une recherche sur les formes nominales se terminant par les suffixes *–age* et *–ment* (les études menées par Van Campenhouldt montrent un pourcentage nettement supérieur de mots en *–age* et *–ment* dans les textes spécialisés par rapport aux textes « littéraires »). Pour terminer, et vu que le domaine du droit de l'Internet est étroitement lié à celui des nouvelles technologies ou prolifèrent les emprunts à la langue anglaise, nous nous sommes intéressée au suffixe *-ing*. Le tableau ci-dessous (Tableau 16) reproduit, à titre d'exemple, quelques résultats de nos recherches (les chiffres placés à droite se rapportent à la fréquence d'apparition)

<table>
<tr><td colspan="2"><N+MP="tion$"> (NooJ a localisé 1 571 formes différentes correspondantes au patron de fouille ; nous avons retenu celles qui apparaissent 25 fois ou plus dans le corpus).</td></tr>
<tr><td>protection 9681
utilisation 5513
application 5468
information 5431
communication 4503
obligation 3821
création 1938
diffusion 1845
reproduction 1479
conservation 1362</td><td>violation 1253
reproduction 1479
notification 1099
conclusion 845
usurpation 448
authentification 471
introduction 398
divulgation 203
navigation 202</td></tr>
<tr><td colspan="2"><N+MP="age$"> (NooJ a localisé 323 formes différentes correspondant au patron de fouille ; nous avons retenu celles qui apparaissent 10 fois ou plus dans le corpus).</td></tr>
<tr><td>usage 2703
filtrage 1096
stockage 1072
partage 692
image 682
piratage 533
affichage 422
nuage 414
blocage 393</td><td>arbitrage 341
cryptage 100
verrouillage 98
visionnage 74
profilage 53
hameçonnage 35
décryptage 34
démarrage 34
traçage 31</td></tr>
<tr><td colspan="2"><N+MP="ment$"> (NooJ a localisé 539 formes différentes correspondant au patron de fouille ; nous avons retenu celles qui apparaissent 25 fois ou plus dans le corpus).</td></tr>
<tr><td>traitement 2782
paiement 2307
règlement 2262
développement 1619
enregistrement 1441
hébergement 1318
téléchargement 1248
jugement 1139
fonctionnement 1087
fondement 1065
consentement 862
établissement 845</td><td>référencement 662
comportement 554
manquement 493
abonnement 477
gouvernement 399
amendement 346
avertissement 285
signalement 119
effacement 118
positionnement 46
dédommagement 30</td></tr>
<tr><td colspan="2"><UNK+MP="ing$"> (NooJ a localisé 278 formes différentes correspondant au patron de fouille ; nous avons retenu celles qui apparaissent 5 fois ou plus dans le corpus. Il est également nécessaire de souligner que le patron de fouille a généré beaucoup de bruit).</td></tr>
<tr><td>streaming 176
computing 101</td><td>emailing 15
grooming 11</td></tr>
</table>

caching	60	pharming	10
phishing	50	hacking	10
cybersquatting	35	uploading	9
spamming	25	watermarking	8
crowdsourcing	21	peering	6
framing	20	browsing	6
typosquatting	20	finger printing	6
fingerprinting	18		

Tableau 16. Extraction des formes en – tion, - ment, - age, - ing (*NooJ*)

Même si les nominalisations à base verbale dominent dans les textes spécialisés, il nous paraît important de s'intéresser également aux nominalisations à base adjectivale. En effet, tandis que les premières renvoient à des noms d'action, de processus et de résultat, les nominalisations adjectivales représentent en règle générale des propriétés. Soulignons que selon Cornu (2005 :121), les propriétés (ou autrement dit les qualités) appartiennent à un groupe de termes juridiques abstraits qui servent à caractériser des choses concrètes. Afin de faire émerger cette catégorie de termes, nous avons donc décidé de faire appel au suffixe *–ité* qui semble être le plus productif dans la langue juridique Voici un extrait des résultats de cette requête.

<N+MP="ité$"> (*NooJ* a localisé 699 formes différentes correspondant au patron de fouille. Nous avons retenu celles qui apparaissent 25 fois ou plus dans le corpus. Il est également nécessaire de souligner que le patron de fouille a généré beaucoup de bruit, c'est-à-dire qu'il a fait émerger des noms qui ne résultent pas de la nominalisation).			
responsabilité	4682	accessibilité	530
sécurité	2874	confidentialité	513
qualité	2305	intégrité	447
autorité	1986	Comité	428
possibilité	1873	validité	383
nécessité	1018	fiabilité	353
finalité	659	criminalité	253
neutralité	575	intégralité	247

Tableau 17. Extraction des formes en – ité (*NooJ*).

Une analyse des listes des résultats (dont, à cause de leur longueur, nous reproduisons seulement une partie), nous a permis de constater que toutes les formes correspondant aux patrons de fouille ne sont pas nécessairement des nominalisations. Les listes extraites par *NooJ* comportent, à côté des noms déverbaux, des noms d'objets ou d'entité comme : *comité, parlement, environnement, nuage, image* ; et des expressions « non saturées » (Lerat 2007 : 83), telles que *conclusion, décision, introduction.* Ainsi, pour réduire le bruit et faciliter le traitement automatique ultérieur, nous avons écarté des résultats produits par *NooJ* tous les mots qui ne sont pas issus d'une nominalisation. Cela nous a amenée à construire une sorte de liste d'exclusion comme par exemple :

<N+MP="age$"-MP="^page$"-MP="^dommage$"-MP="^message$"-MP="^nuage$"-MP="^image$"-MP="^langage$"- MP="^pourcentage$">

Soulignons également que le fait de restreindre notre requête aux formes en *–tion, –age, –ment,* ne nous a pas permis de recenser des unités à formation irrégulière ou des dérivés régressifs comme *mise à disposition, visite, clic, envoi* ou *accès,* pourtant très fréquentes dans notre corpus.

6.3.2.1c 3ème hypothèse – la prédication spécialisée comme source de données terminologiques

Comme nous l'avons vu plus haut, le recours à la nominalisation favorise la formation des syntagmes aux modes de formation ultracomplexes qui peuvent intégrer des déterminants, contenir des prépositions variées, inclure des coordinations, etc. Nous sommes donc partie de l'hypothèse qu'une partie de ces syntagmes sont des termes du domaine. Ainsi, une fois un inventaire des noms déverbaux constitué, nous avons procédé au repérage des séquences formées à partir de ces unités selon les matrices syntaxiques prédéfinies plus haut.

<N+MP="ion$"><A>* (de|de la|du|des|d'|d'une|d'un|à|au|aux|en|pour|contre|sous|sur| dans|avec)<N><A>*(de|de la|du|des|d'|d'une|d'un|à|au|aux|en|pour|contre|sous|sur| dans|avec)*<N>*<A>*
communication au public en ligne 672 application de l'article 64

régulation des communications électroniques 348 protection des données à caractère personnel 242 protection des consommateurs 166 rémunération pour copie privée 98 application de la directive 94 obligation de surveillance 582 protection juridique des bases de données 57 identification des contributeurs à un contenu 32 circulation des données à caractère personnel 22 obligation de conservation des données 16 exécution de l'opération de paiement 8 circulation de l'information 4
<N+MP="ment$"> <A>* **(de|de la|du|des|d'|d'une|d'un|à|au|aux|en|pour|contre|sous|sur|dans|avec)<N><A>*** **(de|de la|du|des|d'|d'une|d'un|à|au|aux|en|pour|contre|sous|sur|dans|avec)*<N>*<A>***
traitement des données à caractère personnel 246 fonctionnement du marché intérieur 104 consentement préalable au traitement des données personnelles 114 référencement sur Internet 74 règlement des litiges 57 enregistrement du nom de domaine 41 abonnement à Internet 16 paiement en ligne 9 consentement préalable au traitement des données personnelles 9
<N+MP="age$"> <A>* **(de|de la|du|des|d'|d'une|d'un|à|au|aux|en|pour|contre|sous|sur| dans|avec)<N><A>*** **(de|de la|du|des|d'|d'une|d'un|à|au|aux|en|pour|contre|sous|sur|dans|avec)*<N>*<A>***
message de données 147 stockage de signaux 72 stockage de données 32 piratage en ligne 26 partage de fichiers 19 partage de vidéos en ligne 18 filtrage des contenus 18 affichage de liens commerciaux 12 filtrage des communications électroniques 7

Tableau 18. Repérage des séquences contenant un nom déverbal.

Un examen rapide des résultats nous a permis de constater qu'effectivement, certains syntagmes de ce type semblent avoir le statut terminologique comme par exemple : *communication au public en ligne*, *obligation de surveillance*, *traitement des données à caractère personnel*, *rémunération pour copie privée*. Cependant, ce n'est pas le cas de toutes

les séquences extraites par *NooJ*. En revanche, à l'instar de Lerat (2007 : 81), nous avons observé que même si les noms prédicatifs résultant de l'opération de nominalisation ne sont pas terminologiques (ils appartiennent à la langue courante), les noms de leurs arguments le sont. Ainsi, dans les syntagmes suivantes : *affichage de liens commerciaux, circulation des données à caractère personnel, filtrage des communications électroniques, enregistrement du nom de domaine,* ce sont les actants en position de complément qui constituent des unités terminologiques. Rappelons que la prise en compte des phénomènes de prédication dans les langues spécialisée s'avère très utile (Lerat 2002, 2006). En effet, les syntagmes nominaux dont le noyau est un nom résultant d'une nominalisation possèdent plusieurs expansions actancielles permettant d'accéder à un nombre important d'informations terminologiques : « *Les nominalisations conduisent à des blocs d'informations quand elles sont accompagnées d'actants.* » (Lerta 2007 : 79). C'est pour cette raison qu'il nous a paru important de nous intéresser à ce type de syntagmes dans notre corpus. L'analyse de l'ensemble des arguments nous a permis : primo de repérer des unités de statut terminologique, secundo d'en tirer profit au moment de l'établissement des liens entre les termes sélectionnés. En effet, nous considérons, tout comme Lerat (*ibid.* 2007 : 89), que les formes issues de l'opération de nominalisation constituent de bons matériaux pour la terminologie.

« *Il est d'observation courante que les nominalisations ainsi comprises sont particulièrement fréquentes dans les textes spécialisés. Ce qui n'a pas été exploité systématiquement, en revanche, et que je voudrais mettre en évidence, c'est qu'autour d'elles gravitent, les saturant dans un contexte étroit, non seulement des actants mais aussi des circonstants, qui désignent des éléments de ce qui est donné comme le réel pertinent, autrement dit une ontologie spécialisée.* »

Lerat (*ibid.* 2007 : 82)

6.3.2.1d 4ème hypothèse – il est possible d'identifier les acteurs typiques du domaine en se basant sur des indices morphologiques

Nous avons vu plus haut que la prise en compte des phénomènes de nominalisation donne accès à un grand nombre de candidats-termes. En effet, l'extraction de ces expressions prédicatives permet de repérer les termes qui renvoient aux actions typiques du domaine ainsi qu'aux objets (leurs actants) et à leurs propriétés. Cependant, l'opération de nominalisation

engendre une conséquence : un prédicat nominal dérivé d'un prédicat verbal attribue aux acteurs des actions des valeurs indéterminées difficilement repérables dans la structure actancielle. Afin de dégager ces derniers (qui selon nous, sont susceptibles d'avoir un statut terminologique) il paraît nécessaire d'avoir recours à d'autres procédés qui s'appuient sur la morphologie dérivationnelle.

Ainsi, nous nous sommes encore une fois tournée vers Gérard Cornu (2005 : 157), qui attire l'attention sur l'importance de certains suffixes dans le langage juridique dont le rôle est d'indiquer les protagonistes du droit, sujets de droit ou organes, considérés dans leurs activités, missions ou fonctions. Selon lui, les suffixes en *–eur* marquent plus spécialement l'action, l'initiative, la position active et les suffixes en *–aire* servent à indiquer soit la réception d'un profit, la jouissance d'un bienfait ou d'une position avantageuse soit la titularité d'un droit ou d'une fonction. Ainsi, afin de vérifier cette hypothèse et d'acquérir une liste des protagonistes du droit de l'Internet, nous avons appliqué au corpus les patrons suivants. Le tableau ci-dessous (Tableau 19) présente une partie des résultats de notre recherche (les chiffres à droite correspondent à la fréquence d'apparition)

<N+MP="eur$">** **\|<N+MP="eurs$"> (nous avons retenu les occurrences qui apparaissent 10 fois ou plus dans le corpus).	
auteur 4982 utilisateurs 3019 consommateurs 2579 utilisateur 2120 ordinateur 1902 fournisseur 1864 éditeur 1758 opérateurs 1490 secteur 1206 auteurs 1202 intérieur 1161 acteurs 1135 législateur 933 vigueur 803	annonceur 708 moteurs 739 serveur 555 producteur 517 faveur 471 défendeur 367 rapporteur 209 erreur 203 ampleur 186 administrateur 136 procureur 125 utilisateur final 110 modérateur 54 routeurs 44
<N+MP="aire$">\|<N+MP="aires$"> (Nous avons retenu les occurrences qui apparaissent 10 fois ou plus dans le corpus).	
titulaire 2553 prestataire 1854 affaire 1492 intermédiaire 969	circulaire 69 préliminaire 63 disciplinaire 62 parlementaire 61

prestataire de services	813	notaire	54
destinataire	783	prioritaire	44
publicitaire	551	secrétaire	31
propriétaire	307	adversaire	15
gestionnaire	285	honoraire	12
bénéficiaire	218		

Tableau 19. Identification des acteurs du droit de l'Internet à l'aide d'indices morphologiques.

Comme nous pouvons le constater, la projection des patrons de fouille définis plus haut sur notre corpus n'a pas ramené uniquement des noms désignant les protagonistes du droit de l'Internet. Parmi les résultats, nous avons repéré de nombreux adjectifs en *–aire* (parlementaire, préliminaire, disciplinaire, prioritaire) ainsi que des noms appartenant à la langue courante (*secteur, ampleur, erreur*) ou au vocabulaire juridique compris au sens large (*faveur, vigueur, affaire*). De plus, l'interrogation nous a permis de repérer de nombreux noms désignant des instruments comme *moteur, ordinateur, routeur, serveur, lecteur*.

6.3.2.2 Sur la piste des indices lexico-sémantiques

Jusqu'à présent, notre méthode d'extraction des candidats-termes a été basée sur des indices morpho-syntaxiques. Cependant, nous considérons qu'il est possible, dans le processus du repérage des unités à statut terminologique, de s'appuyer sur un autre type d'indices linguistiques, à savoir des indices lexico-sémantiques. Comme nous l'avons vu dans le chapitre précédent (Meyer cité dans Bowker et L'Homme 2004, L'Homme 2004, Condamines 2005, Aussenac-Gilles et Séguéla 2000), il existe des marqueurs lexico-sémantiques qui permettent de trouver des contextes porteurs d'informations sémantiques ou conceptuelles sur les termes étudiés. Certains marqueurs sont ainsi étroitement associés à un type de relation précise comme l'hyperonymie, l'hyponymie ou la métonymie. Nous allons développer cet aspect dans la partie suivante de ce travail, qui est consacrée à l'analyse de l'environnement contextuel des termes sélectionnés. Ici, nous proposons d'utiliser des marqueurs lexico-sémantiques afin d'identifier les termes eux-mêmes et non pas les relations qu'ils entretiennent avec d'autres unités (quoique le repérage de certains contextes contenant des candidats-termes permette d'accéder à ces deux types de données). Pour ce faire, nous avons décidé d'explorer quelques pistes proposées par Pearson (1998 : 121-167). Cette

recherche a pour but de compléter et d'affiner les résultats produits lors des étapes précédentes.

6.3.2.2a 5ème hypothèse – le vocabulaire juridique : du générique au spécifique

Nous nous référons ici à la notion d'héritage, principe clé de la terminologie traditionnelle sur lequel repose le modèle hiérarchique d'organisation des données terminologiques (L'Homme 2004 : 254). En effet, ce modèle permet de rendre compte des différentes relations taxinomiques entre les concepts. Depecker (2002 : 159) parle des relations de superordination et de subordination qui dérivent de la propriété de certains concepts de subsumer d'autres concepts, c'est-à-dire d'en englober d'autres sous eux. Parmi ces relations, les plus utilisées en terminologie sont les relations génériques et spécifiques. Rappelons d'après Depecker (*ibid.* 151-152), qu'une relation est dite générique lorsque l'intension d'un concept (l'ensemble des caractères qui le composent) inclut celle d'autres concepts qui lui sont subordonnés. En revanche, une relation est dite spécifique lorsqu'un concept est inclus dans un autre concept et qu'il possède au moins un caractère distinctif supplémentaire. Cela permet de faire la distinction entre concepts de niveau supérieur (concepts générique) et concepts de niveau inférieur (concepts spécifiques).

Nous avons commencé par ce bref rappel théorique car, comme le souligne Pearson (1998 : 128), il est possible de tirer parti du principe d'héritage dans l'identification des unités terminologiques à partir du corpus. En effet, d'après l'auteure, le caractère générique de certains concepts est un critère important à prendre en compte lors du processus de repérage des candidats-termes : « *The first and, we believe, the most important criterion is that of generic reference. Generic reference is one of the key tenets of the traditional theory of terminology where a clear line is drawn between generic concepts and individual objects*. » L'auteure cite à ce sujet la norme ISO/R 704 (1968) révisée par ISO 704/2000: « *It should always be borne in mind that the concepts cannot be taken for the individual object themselves. They are mental constructions serving to classify the individual objects of the inner or outer world by way of a more or less arbitrary abstraction.* »

Ainsi, un terme spécifique qui représente un concept de niveau inférieur est classé sous un terme générique qui englobe d'autres termes spécifiques appartenant à la même classe. Quant au langage du droit, Cornu (2005 : 121) remarque que le vocabulaire juridique

regroupe un certain nombre de termes abstraits génériques qui peuvent être considérés comme les outils essentiels de la pensée juridique. Il s'agit des notions permettant de regrouper diverses choses en un genre, dégageant, à un degré supérieur de généralité, le trait générique qu'elles ont en commun. À titre d'exemple, l'auteur cite : *acte juridique*, *fait juridique*, *droit réel*, *responsabilité civile*, etc.

Compte tenu de ces remarques, nous supposons donc que le repérage et l'analyse des termes génériques propres au domaine du droit de l'Internet peuvent donner accès aux termes spécifiques. Cependant, se pose la question de savoir comment identifier ces termes de niveau supérieur. Selon Pearson (*ibid.* 129), on peut distinguer les termes génériques des termes spécifiques à l'aide de paramètres linguistiques. Ainsi, l'auteure considère que dans des contextes définitoires, s'ils apparaissent ensemble, les termes spécifiques sont d'habitude précédés d'un article défini (elle parle de *flagged term*), alors que les termes génériques sont précédés d'un article indéfini (*unflagged* term). De plus, Pearson, a identifié une série de patrons de fouille basés sur des marqueurs linguistiques qui, combinés aux termes génériques, permettent de repérer des termes spécifiques dans des textes spécialisés. Selon nous, il est aussi possible de se baser sur le critère de fréquence. En effet, nous proposons de nous intéresser surtout à la fréquence des unités simples et de repérer celles qui d'un point de vue sémantique se caractérisent par un degré de généricité plus important. Ainsi, l'analyse des mots simples qui apparaissent en tête des listes de candidats-termes produites par *TermoStat* tout comme celle des mots triés par fréquence proposée par *NooJ* nous a permis de dégager un ensemble de termes génériques susceptibles de renvoyer à des catégories de termes spécifiques. Le Figure 29 reproduit une partie des résultats ; les chiffres entre parenthèses représentent le nombre d'occurrences suite au traitement effectué par *NooJ*.

acte (3619), activité (5457), autorité (5967), cas (7428), communication (9715), contenu (8835), dispositif (1843), droit (29 524), instrument (896), infraction (2249), matériel (1068), mécanisme (1208), mesure (8554), méthode (798), norme (1148), obligation (6 154), organisme (764), outil (726), pratique (3049), prestation (1009), procédure (5078), processus (817), protection (1779), service (18626), stratégie (653), système (5087), technique (5701)

Figure 29. Liste de termes abstraits génériques du domaine du droit extraits du corpus *DITerm*

Comme nous pouvons le constater, la plupart des termes génériques extraits de notre corpus renvoie aux notions juridiques abstraites évoquées par Cornu. À cela s'ajoutent un certain nombre de noms de la langue générale affichant une fréquence élevée qui acquièrent, dans notre corpus le statut juri*DiCo*-technique. Nous considérons qu'une partie de ces termes génériques peuvent entrer dans la composition des unités terminologiques de niveau inférieur. Il est donc possible d'extraire de notre corpus des séries de termes complexes dont une des composantes, notamment la tête, est un terme générique identifié à l'étape précédente. Pour ce faire, nous avons inséré les termes (Figure 29) dans les matrices morpho-syntaxiques définies plus haut (que nous n'allons pas reproduire dans leur intégrité) Voici quelques résultats de cette requête.

acte|actes (3619) acte de communication au public, acte de concurrence déloyale, acte de contrefaçon, acte de copie privée, acte de démarchage, acte de dépôt de la marque, acte de mise à disposition des fichiers, acte de parasitisme, acte de publication, acte de publicité mensongère, acte de reproduction, acte de téléchargement illégal, acte de transfert, acte de vente, actes de parasitisme, actes de piratage, actes de représentation, actes de terrorisme, acte authentique électronique, acte de reproduction provisoire,
système|système (5087), système d'archivage électronique, systèmes d'authentification, système d'échange de données, système de « responsabilité graduée », système de sanction, système d'adressage IP, système de communications électroniques, système de contrôle parental, système de consentement préalable, système de distribution sélective, système de noms de domaine, système de notification et de retrait, système de reconnaissance faciale, système de référencement payant, système de traitement automatisé de donnée, système d'identification électronique
service|services (18626) service de communications électroniques accessibles au public, service d'informatique en nuage, service de référencement sur Internet, services licites de musique en ligne, services Internet, services Web 2.0, service de certification électronique, services de medias audiovisuels à la demande, services de paiement, service de communication au public en ligne, service de filtrage de courrier

électronique
obligation (6154) obligation de collecte des données, obligation de communication de données, obligation de confidentialité, obligation de conseil, obligation de conservation des données d'identification, obligation d'information, obligation de contrôle à priori, obligation de contrôle préalable du contenu, obligation de diligence, obligation de neutralité, obligation de notification, obligation de sécurisation de l'accès Internet, obligation de surveillance de l'accès Internet, obligation générale de surveillance des contenus, obligation particulière de surveillance

Tableau 20. Résultats de la recherche des classes sémantiques à partir de termes abstraits génériques du domaine du droit.

Cette méthode permet de dégager des classes sémantiques, c'est-à-dire de réunir les candidats-termes complexes qui partagent probablement des composantes sémantiques. Comme le souligne L'Homme (2004 : 212), un candidat faisant partie d'une classe comportant de nombreux membres est plus susceptible d'être un terme qu'un autre candidat isolé. Par ailleurs, les résultats de cette investigation pourront être utilisés à une étape ultérieure de notre travail, notamment au moment de l'organisation des données et de l'établissement des liens entre les termes déjà sélectionnés. Malheureusement, cette recherche par mot clé ou mot étiquette ne permet pas d'identifier les termes simples ou complexes sémantiquement apparentés aux termes génériques mais ne partageant pas avec ces derniers de composante formelle.

Ainsi, les recherches menées dans cette partie du travail se sont basées sur l'hypothèse qu'il est possible d'avoir recours à un certain nombre d'indices morpho-syntaxiques ou lexico-sémantiques afin de cibler des candidats-termes du domaine du droit de l'Internet. L'interrogation du corpus a été réalisée à l'aide de la fonction *Locate pattern* de *NooJ* qui a permis de formuler différentes requêtes dans le but de faire émerger les séquences correspondant aux patrons de fouille définis préalablement. Cependant, la prise en compte des critères linguistiques dans l'extraction des candidats-termes, même si elle permet d'affiner et de compléter les résultats des analyses statistiques effectuées à l'aide de *TermoStat*, ne garantit pas que toutes les unités extraites soient des termes du domaine de l'Internet. En effet, pour pouvoir valider le statut terminologique des candidats-termes extraits lors de deux

étapes précédentes et prendre la décision finale quant à la nature des unités à retenir et à leur mode de description, il est nécessaire de faire appel aux indices extralinguistiques tels que l'appartenance au domaine et l'objectif visé. Nous proposons de voir ceci de plus près dans la partie suivante, consacrée à l'analyse et à la synthèse des résultats menant à la description des termes choisis. En outre, nous voudrions attirer l'attention sur le fait que l'extraction des candidats-termes a permis de faire remonter d'autres informations précieuses du point de vue d'un terminographe, notamment des indices linguistiques de relations entre les termes. Dans les pages qui suivent, nous allons donc approfondir cette piste en nous concentrant surtout sur la description des relations lexicales et conceptuelles que les termes sélectionnés entretiennent avec d'autres unités. Cette description sera basée sur l'observation des contextes dans lesquels apparaissent les termes choisis. Ceci va conduire à la proposition d'un modèle de description des unités terminologiques.

Chapitre 7. Analyse des données terminologiques extraites du corpus *DITerm*

Comme nous avons pu le voir dans la partie précédente de notre travail, l'interrogation du corpus *DITerm* réalisée à l'aide de deux outils informatiques, à savoir *TermoStat* et *NooJ* nous a permis d'accéder à un nombre important de candidats-termes du domaine de l'Internet. Rappelons que lors de la réalisation de cette tâche, nous nous sommes appuyée aussi bien sur des critères statistiques (qui visent à faire remonter des informations sans *a priori* sur le type de données) que des indices linguistiques (basés sur des hypothèses qui mettent en œuvre des connaissances sur le fonctionnement de la langue). Cependant, comme nous l'avons déjà signalé, il est difficile de statuer sur le caractère terminologique d'une unité retenue sans avoir recours aux éléments extralinguistiques tels que l'appartenance au domaine et l'adéquation au projet visé. Ainsi, dans les pages qui suivent, nous allons analyser et organiser les données obtenues lors des étapes précédentes en prenant en compte ces deux derniers critères.

7.1 Classification des candidats-termes extraits du corpus – le domaine comme paramètre classificateur des sens[71]

D'après L'Homme (2004 : 54), l'identification d'un terme repose avant tout sur le lien que l'on peut établir entre son sens et un domaine de spécialité ; donc sur des connaissances extralinguistiques. Rappelons rap*ibid.*ent qu'au moment de la constitution du corpus *DITerm* nous avons procédé à une délimitation du domaine du droit de l'Internet. Cette délimitation doit maintenant nous servir de point de référence. Elle nous permettra non seulement de sélectionner les termes à décrire, mais également d'en circonscrire le sens. Cependant, en analysant les données extraites du corpus, il est nécessaire de prendre en compte trois questions, notamment : l'aspect pluridisciplinaire du droit de l'Internet, la nature déontique du langage juridique, l'absence de frontières rigides entre la langue générale et la langue

[71] Nous avons emprunté ce sous-titre à L'Homme (2004 : 53).

spécialisée. Ces trois caractéristiques contribuent à la complexité et à la diversité du vocabulaire étudié.

7.1.1 La complexité et la diversité du vocabulaire du droit de l'Internet

7.1.1.1 L'aspect pluridisciplinaire du droit de l'Internet

Il ne faut pas oublier qu'en raison de l'aspect pluridisciplinaire des sciences et plus particulièrement des sciences humaines et sociales, les frontières entre les différents domaines ne sont pas étanches. En effet, il arrive souvent qu'une discipline ait recours à des concepts d'un autre champ de connaissances. Les concepts ainsi adoptés évoluent en fonction d'un nouveau contexte en reflétant une autre réalité. Par conséquent, les termes qui représentent ces concepts ont un caractère dynamique. Comme le souligne Krieger (2002 :234), ces derniers n'appartiennent pas exclusivement à un domaine, mais y prennent une signification spécifique, ce qui remet en cause la croyance de la terminologie traditionnelle dans l'idéal de l'exclusivité dénominative et dans la monosémie. C'est notamment le cas du droit de l'Internet qui, rappelons-le, est une matière vaste et transversale, considérée comme le produit d'une combinatoire de concepts à partir de différents champs de connaissances. Nous avons déjà évoqué le caractère interdisciplinaire et multidisciplinaire du domaine. Interdisciplinaire parce qu'il y a deux disciplines qui entrent en interaction, à savoir le droit et les nouvelles technologies de l'information et de la communication. Multidisciplinaire parce que les sources du droit de l'Internet sont multiples et se trouvent dans plusieurs branches, comme celles du droit des données personnelles, du droit des contrats, de la consommation, de la concurrence, de la responsabilité, de la propriété intellectuelle, du droit pénal ou du droit international privé.

7.1.1.2 La nature déontique du langage juridique

De plus, la complexité du droit de l'Internet ne tient pas seulement à son caractère pluridisciplinaire mais aussi au fait qu'il ne s'agit pas seulement d'un domaine de connaissance (et ceci est une caractéristique du domaine du droit en général). En effet, comme le souligne Krieger (2002 : 235), le droit est articulé par des objectifs pragmatiques en raison de l'aménagement juri*DiCo*-social qu'il établit. L'auteure attire également l'attention sur sa nature déontique primordiale, ce qui explique la présence d'une série de mécanismes

linguistiques pragmatiques et sémiotiques qui génèrent des effets d'impérativité dans les textes légaux[72]. Elle cite à ce propos Maciel (2001 : 26) selon qui :

« *[…], le premier but de la communication du droit, un domaine humain, social et normatif, est de prescrire des normes de comportement. C'est pourquoi les critères d'attribution du statut terminologique et de reconnaissance des unités lexicales qui composent sa terminologie diffèrent de ceux adoptés dans d'autres domaines de connaissance et d'activité avec des objectifs distincts. Ainsi, c'est dans la communication des normes juridiques que se configure sa spécificité.* »

Compte tenu de ces remarques, il est donc possible de distinguer deux volets ou niveaux linguistiques spécialisés qui se combinent dans la construction de l'espace discursif du droit de l'Internet : d'une part un volet commun à toutes les branches du droit qui englobe les unités et les constructions propres au domaine juridique compris au sens large du terme et, d'autre part, un volet spécifique formé par le vocabulaire et la phraséologie relevant de la matière sur laquelle portent les textes de loi, à savoir les aspects juridiques de l'Internet. Précisons également que le premier volet ne correspond pas non plus à une réalité homogène. Comme le souligne Cornu (2005 :22), le langage du droit est plurifonctionnel et pluridimensionnel : « *Il n'existe pas un langage juridique mais un langage législatif, un langage judiciaire, un langage coutumier, un langage conventionnel, un langage administratif, un langage doctrinal.* » ; et chacun se distingue par des éléments qui lui sont propres. Nous sommes consciente du fait que ceci peut rendre le travail d'identification du répertoire terminologique problématique. Pour faire face à ces difficultés, nous serons donc amenée à établir des critères de sélection des termes qui nous permettront de définir la pertinence thématique d'une unité face au champ du savoir qui nous intéresse.

7.1.1.3 L'absence de frontières rigides entre la langue générale et la langue spécialisée

Il convient également de signaler un autre problème auquel nous pourrons être confrontée au moment de la sélection des unités terminologiques. Il s'agit du fait qu'il est très difficile d'établir des frontières rigides entre le lexique spécialisé et le lexique général et de

[72] Sourioux et Lerat (1975 : 48-50) parlent des marques modales en distinguant les auxiliaires de mode suivant les quatre catégories traditionnelles : l'obligatoire, l'interdit, le permis et le facultatif. Ils font ainsi référence à la logique des normes qui selon Kalinowski doit être considérée comme l'essence du droit.

décider si une unité donnée possède un statut terminologique ou pas. En effet, tout travail d'un terminographe repose sur l'hypothèse selon laquelle : « *il y aurait les mots, essentiellement généraux ou « simples » par vocation sinon par essence, par opposition à d'autres, les termes, qui auraient pour seule fonction de caractériser un type particulier de discours, puisqu'ils lui confèreraient, de par leur charge plus ou moins grande de sens spécialisé, une signification singulière […] propre à un champ d'activité donné* ». (Gémar 1991 : 275). Ainsi, selon Gémar (*ibid.*), l'ensemble des termes d'un domaine, c'est-à-dire sa nomenclature, forme un noyau dur à partir duquel se réaliserait le discours spécialisé. Cependant, même si la terminologie d'un domaine constitue la base d'un discours spécialisé, les termes sont associés à d'autres éléments du discours, notamment à des cooccurrents précis (le vocabulaire de soutien) et à des mots de la langue générale (le vocabulaire général) qui jouent également un rôle important dans la construction du sens spécialisé.

« *En schématisant, la structure du discours spécialisé pourrait être idéalement représentée par une série de cercles concentriques dont le premier, au centre, serait le noyau dur (la nomenclature, soit la charge notionnelle des termes, en nombre limité) ; le cercle suivant, plus large, contiendrait le vocabulaire de soutien (les cooccurrents du domaine), à la fois moins nombreux mais aux possibilités d'association par agrégat naturel néanmoins bien supérieures ; le troisième cercle, le plus éloigné, symboliserait les éléments aléatoires de la langue que sont les « mots » (articles, pronoms, verbes, adjectifs,…)* »

(Gémar, *ibid.* : 276).

Comme le précise l'auteur, la réalisation du discours, et donc l'expression du sens, passe obligatoirement par l'agrégat de ces trois éléments qui contribuent conjointement à la pleine signification du message.

Pourtant, il est nécessaire de souligner que ces trois types de vocabulaire formant le sens en langue spécialisée ne sont pas traités de la même manière selon la démarche adoptée. Ainsi, dans l'optique traditionnelle, les terminographes s'intéressent au noyau dur, à la nomenclature, c'est-à-dire à l'ensemble des termes reflétant la structure conceptuelle dans un domaine donné. Ceci explique la prédominance dans les ressources terminographiques classiques des unités de nature nominale qui ont l'avantage d'apparaître comme de simples étiquettes apposées sur des concepts. Cependant, comme nous avons pu le constater, la

sélection exclusive de noms est incompatible avec ce qui peut être observé dans les textes spécialisés. Le traitement des termes représentant d'autres parties du discours (verbes, adjectifs, adverbes) ainsi que la description du vocabulaire de soutien s'avèrent donc nécessaire si l'on veut donner une image complète d'une langue spécialisée (de spécialité) et répondre par là aux besoins des utilisateurs des dictionnaires spécialisés. Ainsi, comme le remarque L'Homme (2000 : 79, 2004 : 62), les terminographes, souvent tiraillés entre les principes de la terminologie traditionnelle (qui privilégie les entités) et les réalités linguistiques (la nécessité de la description des sens qui renvoient à des activités, des propriétés, différents types de relations), ont recours à différentes stratégies. Certains par exemple, répertorient les verbes, les adjectifs et les adverbes uniquement si leur emploi est exclusif de tout autre domaine spécialisé. Ils peuvent également s'intéressent aux mots appartenant à ces trois parties du discours s'ils ont un sens distinct du sens qu'on leur donne généralement.

Cependant, comme le remarque L'Homme (2004 : 62), cette approche amène les terminographes à mettre de côté des unités qui sont significatives dans le domaine étudié mais leur sens ne se détache pas du sens qu'on leur attribue dans d'autres contextes. Selon l'auteur, une autre pratique consiste à relever les verbes, adjectifs ou adverbes seulement dans le cas où ces derniers sont sémantiquement apparentés à un terme de nature nominale. Cette démarche présente également des lacunes car certaines unités véhiculent un sens défini en fonction d'un domaine spécialisé mais ne sont pas apparentées à un nom. En sélectionnant les termes propres au domaine de l'Internet, nous serons donc obligée de prendre une décision quant au type des unités à retenir et à leur degré de spécialisation.

7.1.2 La classification des candidats-termes extraits du corpus *DITerm*

L'analyse des listes des candidats-termes produites par *TermoStat* et par *NooJ* nous a permis de confirmer les hypothèses évoquées ci-dessus. En effet, il s'est avéré que l'identification du répertoire terminologique du domaine du droit de l'Internet est une tâche très difficile de par la complexité et la diversité de sa structure discursive. Par conséquent, afin d'introduire un peu d'ordre dans les résultats obtenus lors de l'étape précédente de notre travail, nous avons décidé de procéder à une classification des occurrences relevées. Pour ce faire, nous nous sommes inspirée entre autres de la catégorisation du vocabulaire juridique proposée par Gérard Cornu (2005 :60-131).

Rappelons que d'après Cornu (*ibid.*), la nomenclature du droit est constituée de termes d'appartenance juridique exclusive et de termes de double appartenance. Le premier groupe comprend environ 400 termes qui n'ont pas d'autre sens que le juridique. L'autre ensemble, beaucoup plus vaste, regroupe, d'un côté les termes d'appartenance juridique principale et de l'autre côté, les termes appartenant au vocabulaire juridique mais empruntés à la langue courante. Parmi les premiers, on recense les unités qui ont un sens juridique principal et un sens extrajuridique secondaire. Ces termes sont passés dans le langage courant avec un sens dérivé et constituent un apport à la langue commune. Comme le souligne Cornu (*ibid.* : 70), il s'agit, dans la plupart des cas, des termes porteurs des notions fondamentales du droit: « *ils se distribuent dans les départements essentiels du système juridiques et occupent, en chacun, le premier rang*. » Ce sont les mots-clés du vocabulaire de base (*interdiction, permission, sanction, règle, légitime),* les mots-clés du vocabulaire judiciaire (*juge, tribunal, juger, avocat, plaideur, débat, sursis),* les mots-clés désignant les opérations juridiques principales et les actes juridiques courants.

Quant aux termes d'appartenance juridique secondaire, Cornu en distingue deux types : ceux qui possèdent le même sens dans la langue du droit et la langue générale (il s'agit surtout des termes qui se rattachent à des démarches essentielles de l'esprit humain et relèvent de l'ordre de la preuve et de la logique), et ceux qui ont acquis dans la langue juridique un sens particulier. Selon Cornu (*ibid.* : 77), il est fréquent que le langage du droit utilise un terme commun doté d'un sens générique en lui conférant, dans son ordre, un sens spécifique : « *Le sens extrajuridique et le sens juridique sont dans le rapport du genre et de l'espèce.* ». Comme le précise l'auteur, la spécificité juridique de l'emploi charge le terme de tant de particularités que le sens juridique prend relativement au sens commun un caractère très spécial et très technique. Il existe donc un lien étroit entre le langage du droit et la langue générale et cela constitue l'une des grandes difficultés de cette langue spécialisée. En effet, en étudiant le vocabulaire juridique, il faut prendre en considération le fait que les mots courants acquièrent, dans le contexte du droit, une charge de sens juridique, c'est-à-dire qu'ils sont liés à des événements susceptibles de produire des effets juridiques.

Ainsi, en analysant les données extraites du corpus, nous avons pu observer qu'il existe quelques catégories d'unités lexicales spécialisées qui entrent dans la composition du

vocabulaire du droit de l'Internet. Voici notre proposition de classification (les chiffres placés à droite se rapportent à la fréquence d'apparition) :

1. Les termes correspondant à la nomenclature, le noyau dur du vocabulaire du droit de l'Internet constitués des termes spécifiques, propres à ce domaine. Soulignons que le tableau ci-dessous (Tableau 21) reproduit une partie des termes sélectionnés. Pour la liste complète, il faut se reporter à l'Annexe III.

fournisseur d'accès à Internet	246
fournisseur d'hébergement	479
éditeur de service de communication au public en ligne	292
obligation de surveillance générale	234
téléchargement illégal	203
mesure technique de protection	258
procédure alternative de règlement des conflits	75
données à caractère personnel	3913
communication commerciale non sollicitée	63
consentement préalable	114
traitement de données à caractère personnel	1059
cybersquatting	35
offre légale (en ligne)	334
piratage	545
hameçonnage	36

Tableau 21. Exemples de termes extraits du *DITerm* constituant la nomenclature du droit de l'Internet.

2. Les termes relevant d'autres branches du droit : droit de la communication, droit de la propriété intellectuelle, droit du commerce, droit des libertés fondamentales. Il s'agit des termes concrets qui désignent des réalités particulières, des choses matérielles, des personnes, des faits juridiques concrets liés à une branche du droit donnée autre que le droit de l'Internet.

œuvre audiovisuelle 1074,	consommateur 40652
producteur 1102,	consommation 829
phonogrammes 369	marchandises 344
marque 6783	contrefaçon 2997
médias 766	signe distinctif 64
paiement 2762	nom commercial 182
produit 4457	dénomination sociale 170
copie privée 194	concurrence déloyale 623

Tableau 22. Exemples de termes extraits du *DITerm* et appartenant à d'autres branches du droit.

3. Les termes liés aux nouvelles technologies

cookie 674	blog 590
logiciel 3274	réseau social 395
ordinateur 2306	streaming 176
référencement 662	accès à Internet 1112
réseau 286	forum de discussion 349
caching 60	plate-forme/plateforme 1533
moteur de recherche 1415	flux RSS 91
site web/site Internet 9818	hyperlien/lien hypertexte 306
adresse IP 1239	hébergement 1347

Tableau 23. Exemples de termes extraits du *DITerm* et appartenant au domaine des nouvelles technologies.

4. Les termes clés du vocabulaire juridique de base considérés comme étant porteurs des notions fondamentales du droit. Dans la plupart des cas, il s'agit de termes abstraits qui désignent des notions regroupant diverses choses en un genre. On peut parler de mots étiquettes, c'est-à-dire de termes de sens général qui représentent une classe d'objets.

règlement 3069	mécanisme 1148
réglementation 970	victimes 679
règles 2911	autorités 6926
obligation 6121	droit 28881
dérogation 438,	normes 1325
autorisation 1768	acte 3608
litige 855	pratique 3159
responsabilité 5415	violation 1498
mesures 8563	atteinte 3962
dommages 1282	délit 529
dispositif 2477	sanction 1936
procédure 5571	fait 1702

Tableau 24. Exemples de termes juridiques de base.

5. Les termes appartenant au vocabulaire judiciaire

tribunal 1252	Cour 6870
juge 2319	défenderesse 556
affaire 2541	demandeur 940
jugement 1221	huissier 364
audience 324	rapporteur 210
article 24585	requérant 199
chambre 802	requérante 255

Tableau 25. Exemples de termes appartenant au vocabulaire judiciaire extraits du *DITerm*.

6. Les termes appartenant au langage administratif typique de la Commission Européenne

Bruxelles 343	directive 20023
Commission 6861	Europe 1430
communautaire 1632	gouvernement 732

communautaires 515	transposition 434
Communauté 928	UE 1488
États membres 7206	

Tableau 26. Exemples de termes appartenant au langage administratif typique de la Commission Européenne extraits du *DITerm*.

7. Les mots qui possèdent le même sens dans la langue du droit et la langue générale. En l'occurrence, il s'agit des verbes qui représentent les instruments essentiels de la pensée et se rattachent à des démarches essentielles de l'esprit humain.

considérer	1507	énoncer	1130
constater	963	assigner	429
prouver	241	interdire	1536
preuve	1422	conférer	318
permettre	3732	supposer	567
appliquer	1190	recommander	547
procéder	888	incomber	521
ordonner	762	viser	1248
déduire	371	disposer	1835
condamner	1221	veiller	1459
invoquer	847	autoriser	1985
statuer	653	notifier	658
sanctionner	567	induire	171
devoir	6791	présumer	250
imposer	2360	valoir	812
démontrer	915		

Tableau 27. Exemples de verbes juridiques extraits du *DITerm*.

Rappelons que même si la terminologie a longtemps négligé le verbe, ce dernier est doté d'un statut particulier dans le langage juridique. En effet, le verbe dans le domaine du droit a un caractère performatif et y joue un rôle très important. Comme le soulignent Sourioux et Lerat (1975 : 50), en s'appuyant sur les travaux du logicien anglais J.L Austin, le langage du

droit est un langage d'action. Le vocabulaire juridique englobe donc un grand nombre de mots-actes dont l'énonciation « annonce et accomplit en même temps une action » (Austin cité dans Sourioux et Lerat *ibid.*). Il s'agit notamment des performatifs stricts, des constatifs officiels (*ibid.* : 52), des verbes qui désignent les décisions exécutoires : les auteurs (*ibid.* : 54-55) distinguent les décisions normatives considérés comme étant l'expression de la loi et de la règlementation et les décisions judiciaires énoncées par le juge. À cette liste, s'ajoutent aussi les verbes ou les expressions verbales qui expriment l'obligation, la permission, l'interdiction (*ibid.* : 48-50). Pour sa part, Maciel (cité dans Pimentel 2011 : 148) propose de classer les verbes exprimant les différents types d' « actes juridiques »[73] en trois groupes : 1) des verbes qui créent des normes juridiques ; 2) ceux qui confèrent à certains individus ou institutions une partie du pouvoir gouvernemental ; 3) ceux qui gouvernent le comportement dans une société politiquement organisé. Il convient de souligner que l'étude des verbes spécialisés dans la langue juridique a fait l'objet de nombreux travaux parmi lesquels nous tenons à mentionner deux références, notamment la description des unités prédicatives basée sur la théorie des classes d'objet (Chodkiewicz et Gross 2005) et la description des verbes juridiques au moyen de la sémantique des cadres proposée par Pimentel (2011).

8. Les mots appartenant à la langue générale qui correspondent au vocabulaire de soutien au sens défini par Gémar (1991). Il s'agit des cooccurrents qui forment l'environnement lexical de prédilection des unités terminologiques. Ils peuvent constituer des groupements spécialisés avec ces dernières ou bien acquérir eux-mêmes le statut terminologique (certaines unités à fort potentiel terminologique identifiées comme un vocabulaire de soutien font partie de la nomenclature recensée à l'Annexe III où elles apparaissent accompagnées d'un astérisque).

créer 749	accessible 1819
développer 411	interactif/interactive 54
informer 543	circuler 803
utiliser 1745	partager 1053
traiter 478	échanger 1729
collecter 744	accéder 1048

[73] L'expression « acte juridique » renvoie ici à un acte de parole et pas à un acte au sens juridique du terme

stocker 237	visiter 455

Tableau 28. Exemples de mots extraits du *DITerm* appartenant à la langue générale et constituant des groupements spécialisés

9. Les mots appartenant à la langue générale et représentant des concepts génériques regroupant des classes d'objets

service/services 18 931	matériel/matériels 1068
activité/activités 5333	mécanisme/mécanismes 1230
prestation/prestations 1009	technologies 1332
œuvre 8869	risque/risques 2639
offre/offres 1666	système/systèmes 5930
produit/produits 4452	processus 817
instrument/instruments 896	moyen/moyens 5104
outil/outils 1275	

Tableau 29. Exemples de mots de sens génériques extraits du *DITerm*.

10. Les mots appartenant à la langue générale parmi lesquels les mots grammaticaux (ou autrement dit les mots-outils) constituent la majorité et se situent en tête de liste (si l'on prend en considération la fréquence brute).

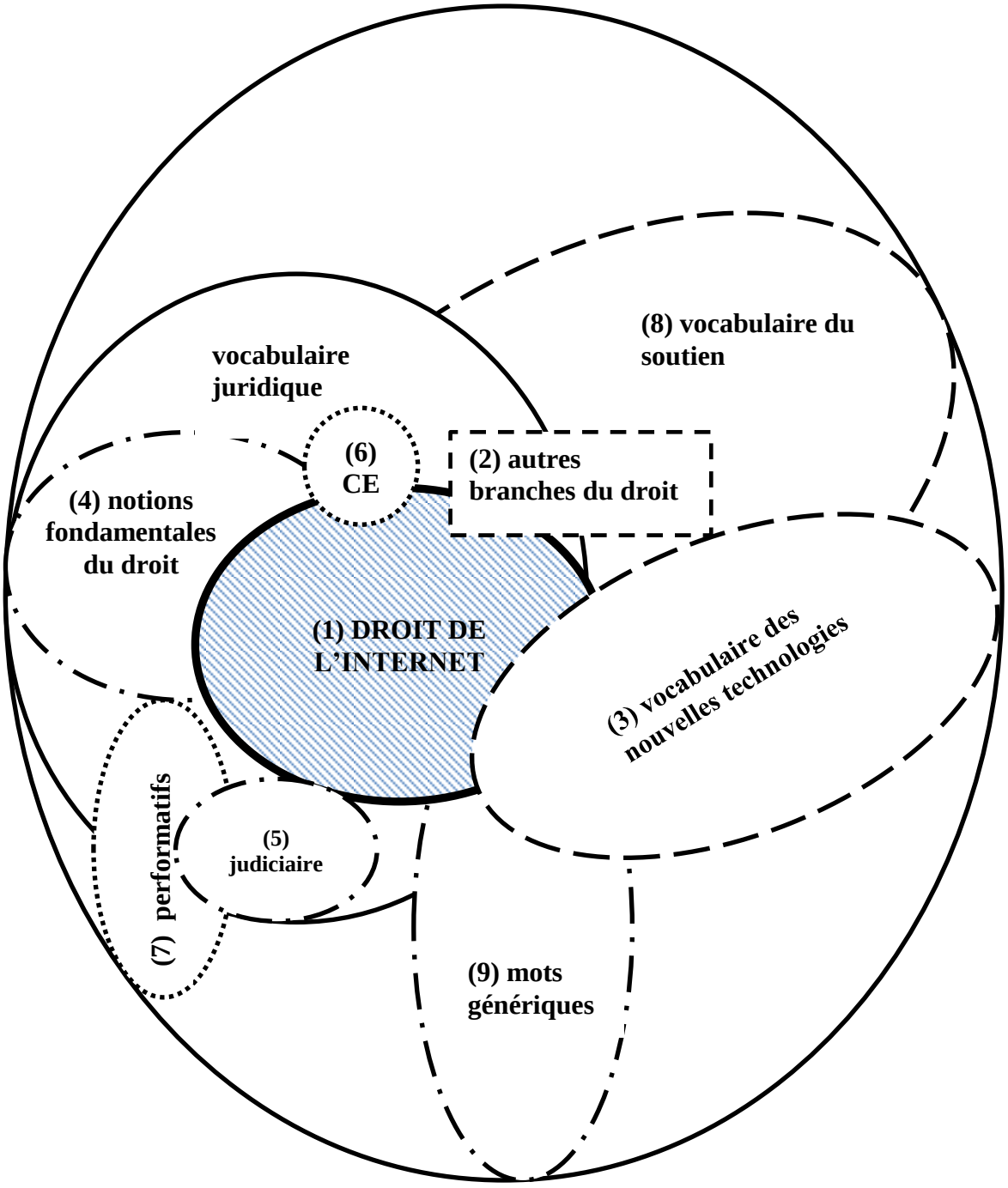
de	330655	un	58001
la	181183	une	51283
des	121268	par	49471
à	119485	est	41066
et	106452	dans	39374
les	96860	au	36940
le	94085	ou	36788
en	75841	pour	36576
du	74812	sur	36286
que	58258	être	18897

Tableau 30. Exemples de mots extraits du *DITerm* appartenant à la langue générale.

L'analyse du corpus nous a donc permis de dégager 10 catégories différentes d'unités qui restent les unes par rapport aux autres dans des relations d'interconnexion, de chevauchement ou d'inclusion. La Figure 30 reflète ce caractère complexe de la structure lexicale du langage du droit de l'Internet. Comme nous pouvons l'observer, la nomenclature du droit de l'Internet n'est pas une réalité linguistique isolée que l'on peut étudier tout en faisant abstraction d'autres éléments. En effet, en proposant notre schéma, nous avons tenté de rendre compte des différents volets lexicaux qui concourent à la construction du sens spécialisé dans le domaine en question. Bien évidement, les unités terminologiques constituant la nomenclature proprement dite du droit de l'Internet (dont les frontières sont par ailleurs fluctuantes et perméables) restent le principal centre d'intérêt de la présente étude. Cependant, il faut savoir qu'autour de ce noyau dur gravitent d'autres unités lexicales et terminologiques qui entretiennent avec les termes clés du domaine de nombreuses relations.

Ainsi, les termes appartenant à la catégorie (1), considérés comme étant le noyau dur du langage du droit de l'Internet feront l'objet de la description et c'est à eux que nous consacrerons les articles de notre dictionnaire. Les unités relevant d'autres branches du droit (2): droit de la communication, droit de la propriété intellectuelle, droit du commerce, droit des libertés fondamentales sont retenues en tant que vocabulaire connexe dont nous nous servirons afin de décrire les termes sélectionnés. Il en va de même pour les mots de la catégorie (3), à savoir les termes du domaine des nouvelles technologies. Cependant, quant à ces derniers, il est parfois difficile de décider s'il s'agit de simples termes techniques ou bien de termes dotés d'un sens juridique, la frontière entre les deux n'étant pas étanche. En effet, il convient de souligner que les termes appartenant au domaine des nouvelles technologies, même s'ils n'ont pas de statut juridique, jouent un rôle très important car ils entretiennent de nombreux liens (aussi bien syntagmatiques et paradigmatiques que conceptuels) avec les termes faisant partie de la nomenclature du droit de l'Internet. La dernière catégorie des mots que nous avons décidé de retenir afin d'enrichir la description des termes clés est constituée de ce que Gémar (1991) appelle le vocabulaire de soutien (8), c'est-à-dire l'ensemble des cooccurrents qui constituent l'environnement lexical de prédilection des unités terminologiques. Il ne s'agit pas là des termes *stricto sensu* (même si certains parmi eux peuvent acquérir le statut terminologique), mais des unités qui contribuent à la construction du sens spécialisé en établissant des relations de type syntagmatique ou conceptuel avec les

termes choisis. Elles ne feront pas l'objet d'entrées à part mais apparaîtront comme des éléments descriptifs rattachés aux termes choisis. Par ailleurs, nous tenons à souligner que l'ensemble des unités appartenant aux catégories considérées comme étant annexes (car elles ne feront pas l'objet des entrées), à savoir les catégories (2), (3), (8) entrent au même niveau que les termes clés (1) dans la composition du schéma conceptuel du domaine.



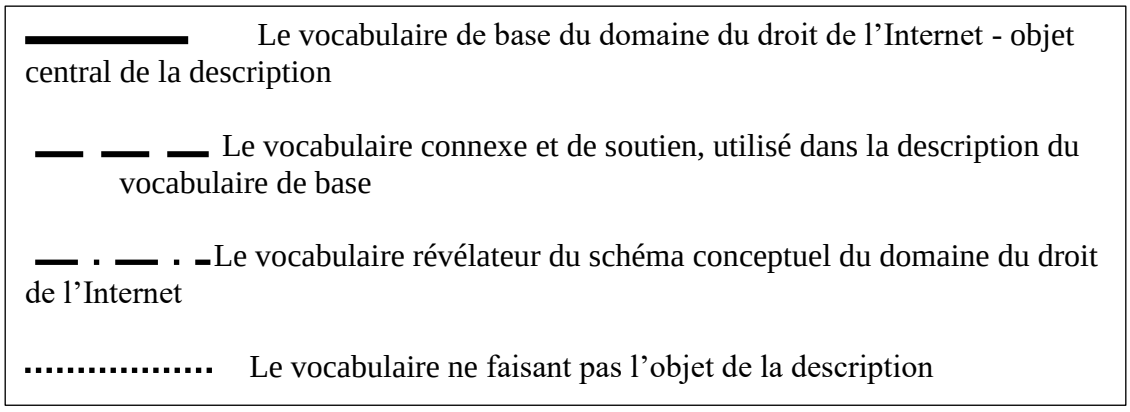


Figure 30. Structure lexicale du droit de l'Internet

Quant aux unités appartenant aux catégories (4), (7) et (9), à savoir : (4) les mots clés du vocabulaire juridique de base (considérés comme étant porteurs des notions fondamentales du droit), (7) les verbes juridiques les démarches essentielles de l'esprit humain et (9) certains mots génériques de la langue générale (dont la fréquence est considérablement élevée dans notre corpus), elles nous intéressent de par leur caractère classificateur. En effet, il s'agit des mots étiquettes qui peuvent servir de points d'accès au domaine du droit de l'Internet. Nous avançons l'hypothèse que l'analyse de ces unités et de leur comportement dans le contexte permettra de structurer le schéma conceptuel du domaine et d'établir les liens entre les termes clés et les unités lexicales décrits ci-dessus. En ce qui concerne les mots rangés dans les catégories (5) - le vocabulaire judiciaire, (6) - le vocabulaire administratif de la Commission Européenne, bien qu'ils appartiennent au vocabulaire juridique, ils ne présentent que peu d'intérêt pour notre étude. Nous avons donc décidé de les écarter de notre analyse.

En résumant, il convient de souligner que l'analyse des résultats obtenus lors de l'extraction des candidats-termes ainsi que leur classification en différentes catégories nous a mise sur deux pistes : 1) celle de l'existence des relations lexico-sémantiques entre les termes clés et d'autres unités extraites du corpus ; 2) celle de l'existence de termes révélateurs du schéma conceptuel du domaine.

1. Premièrement, la présence des différents volets lexicaux qui se chevauchent et s'imbriquent (voir la Figure 30) nous a conduite à reconnaitre l'existence d'un réseau de liens sémantiques (aussi bien sur le plan syntagmatique que paradigmatique), entre

les termes clés et les autres unités extraites du corpus et décrites plus haut. Nous allons donc poursuivre cette piste, en essayant de trouver pour chaque terme sélectionné l'ensemble des autres unités lexicales partageant avec ce premier un type de relations. Nous nous concentrons notamment sur la mise en lumière du phénomène de la combinatoire lexicale.

2. Deuxièmement, l'examen des unités extraites du corpus *DITerm* a mis en évidence l'existence d'un certain nombre de termes génériques qui peuvent être considérés comme révélateurs du schéma conceptuel du domaine du droit de l'Internet. Nous avançons l'hypothèse selon laquelle leur étude nous permettra d'organiser les données extraites du corpus en catégories conceptuelles et d'établir un réseau de liens internotionnels.

Notre travail sera donc réalisé parallèlement sur deux plans, un plan linguistique et un plan conceptuel. L'objectif final de cette recherche est de trouver un modèle de description hybride permettant de rendre compte aussi bien des relations lexico-sémantiques que des liens conceptuels entre les unités appartenant au vocabulaire du droit de l'Internet. Étant donné que la méthodologie adoptée dans ce projet s'inscrit dans le cadre de la linguistique de corpus, le point de départ de cette double analyse consistera en l'étude de l'environnement contextuel des termes constituant la nomenclature du domaine en question. L'observation de leur comportement en contexte nous permettra de mettre en évidence les liens qui existent entre ces derniers et d'autres unités faisant partie de la structure lexicale du droit de l'Internet.

7.2 Analyse de l'environnement contextuel des termes choisis : sur les traces des informations sémantiques et conceptuelles

Même si le sens d'un terme se définit d'abord en fonction d'un point de repère extérieur à la langue, à savoir un domaine de spécialité, sa réalisation concrète a lieu dans un environnement linguistique, à savoir le texte. (L'Homme 2004 : 66). Ainsi, le terme cohabite avec d'autres unités lexicales avec lesquelles il entretient des relations d'ordre lexico-sémantique. Il faut savoir que très souvent, ces dernières nous mettent sur la piste des liens

existant au niveau conceptuel. Comme nous l'avons mentionné plus haut, nous croyons qu'il est possible d'établir ces différentes relations à partir des contextes dans lesquelles apparaissent les termes clé du domaine. Notre méthode d'investigation correspond à la démarche proposée par Pearson (1998 : 190 – 203) qui montre la possibilité d'utiliser un terme : « *as the search node for the retrieval of information about a term's meaning and usage* ». Tout comme l'auteure, nous considérons que l'analyse de l'univers discursif des termes peut donner accès à un grand nombre d'informations permettant la structuration du domaine du droit de l'Internet aussi bien au niveau lexical que conceptuel. Soulignons encore une fois qu'en décrivant les phénomènes propres à une langue spécialisée, on ne peut pas s'intéresser uniquement aux termes constituant sa nomenclature, mais on devrait aussi analyser l'ensemble des unités qui gravitent autour de la terminologie de base. En effet, les unités considérées comme étant connexes et d'appui entrent au même titre que les termes clés dans la composition du schéma lexical et conceptuel du domaine.

Notre méthodologie d'exploration consistera donc en le repérage manuel d'unités significatives apparaissant dans les contextes associés à des termes clés. Par *unités significatives*, nous comprenons toutes sortes de données textuelles récurrentes fournissant des informations sur le sens et l'usage des termes ainsi que sur les liens que ces derniers entretiennent avec d'autres unités appartenant au vocabulaire du droit de l'Internet. La consultation des contextes se fera au moyen de la fonction *Local pattern* de *NooJ*. Le concordancier intégré au système est un outil très souple qui offre beaucoup de possibilités à ses utilisateurs. Il permet de paramétrer la longueur des contextes ainsi que de les organiser par ordre alphabétique. Il combine aussi deux modes d'affichage : les concordances peuvent être affichées dans un index KWIC ou consultées en plein texte ce qui est une solution optimale pour un terminographe à la recherche d'informations contextuelles. Le concordancier *NooJ* permet également l'interrogation de plusieurs sous-corpus et signale la source d'où a été prélevé un contexte.

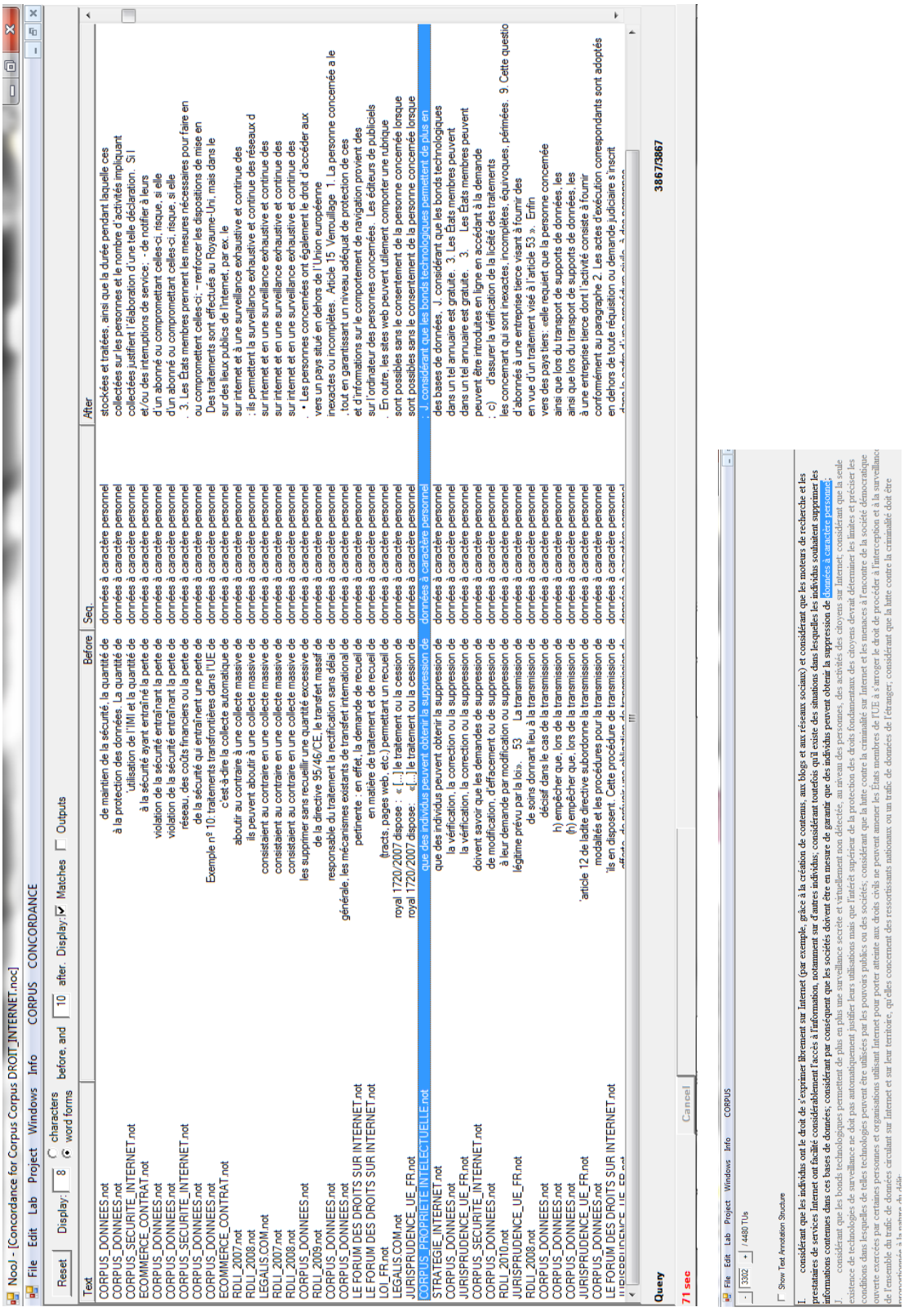


Figure 31. Concordances du terme *données à caractère personnel* générées par *NooJ* (index KWIC combiné à une concordance en plein texte)

Ainsi, pour chaque terme sélectionné, nous avons obtenu une liste des concordances illustrant son comportement, c'est-à-dire la façon dont il est utilisé concrètement dans le corpus. L'analyse de l'ensemble de ces contextes nous a permis de repérer, d'un côté des informations de nature linguistique et de l'autre côté, des indices sur l'organisation conceptuelle du domaine. Plus précisément, l'étude de l'environnement contextuel des termes choisis nous a permis d'effectuer les tâches suivantes :

- extraire les cooccurrents les plus fréquents des termes et observer leur comportement linguistique
- établir la structure actancielle des termes
- dégager les relations circonstancielles
- extraire les relations hiérarchiques ou paradigmatiques que le terme entretient avec d'autres unités
- repérer les relations conceptuelles spécifiques au domaine du droit de l'Internet par le biais des termes génériques identifiés préalablement (voir plus haut)
- repérer des informations encyclopédiques sur le domaine de spécialité et extraire des définitions

À titre illustratif, nous présentons ci-dessous quelques exemples de contextes contenant des informations sur les termes choisis.

7.2.1 Cooccurrents typiques des termes

Les contextes illustrant les cooccurrents les plus fréquents permettent de rendre compte aussi bien des combinaisons lexicales typiques dans lesquelles se trouvent les termes que de leurs affinités conceptuelles. Rappelons que la fréquence est considérée par certains linguistes (nous pensons surtout aux contextualistes) comme un des éléments fondamentaux pour caractériser le phénomène collocationnel. Et même si, dans le cadre de ce travail, la notion de collocation se définit non sur des bases statistiques mais sur des bases fonctionnelles et sémantiques (en renouant avec la théorie Sens-Texte), nous considérons que la fréquence d'apparition figure parmi les premières critères permettant de reconnaitre des collocations. En effet, il convient de souligner le rôle crucial des unités phraséologiques dans

la construction du discours spécialisé. Selon l'idée défendue par Mejri (2011 :129), la phraséologie est un élément définitoire du texte spécialisé. Les langues de spécialités connaissent plus de restrictions d'usage que la langue générale. Par rapport aux mots, les termes limitent davantage leurs possibilités combinatoires en montrant des préférences sémantiques spécifiques au domaine. D'après L'Homme et Meynard (1998 :199), les termes semblent préférer la compagnie de certaines unités lexicales à celles d'autres substituts synonymiques en raison de conventions établies au sein d'un groupe de spécialistes. Ainsi, le discours spécialisé ne se caractérise pas seulement par une haute densité terminologique mais aussi par des constructions linguistiques spécifiques, des regroupements syntagmatiques privilégiés. Ces moyens d'expression conventionnels structurent le discours spécialisé et facilitent la communication entre experts. Il nous paraît donc important de privilégier l'étude des contextes qui permettent de repérer ces affinités sémantiques (mais aussi conceptuelles) entre les termes et d'autres unités.

données à caractère personnel 3913 occurrences
LEGALIS.COM.not … l'ensemble des opérations mises en œuvre par l'agent constituent un traitement automatisé de **données à caractère personne**l entrant dans les prévisions des articles 2 et 25 de la loi du 6 janvier 1978 …
CORPUS_DONNEES.not ….la destruction, la perte, l'altération, l'accès non autorisés ou la divulgation de **données à caractère personnel** transmises, stockées ou traitées d'une autre manière….
RDLI_2012.not … « la collecte, l'enregistrement, la conservation, la consultation et la communication **de données à caractère personnel** doivent être justifiés…
CORPUS_DONNEES.not Un groupe de sociétés doit régulièrement transférer **des données à caractère personnel** de ses sociétés apparentées établies sur le territoire
nom de domaine 1783 occurrences
RDLI_2009.not Les offices sont tenus de bloquer, supprimer ou transférer, selon le cas, des **noms de domaine** : – lorsqu'ils constatent qu'un enregistrement a été effectué en violation des règles
RDLI_2011.not Peuvent demander l'enregistrement d'un **nom de domaine** , dans chacun des domaines de premier niveau, les personnes physiques résidant sur le

territoire…
RDLI_2010.not Il en ressort, d'une part, que le transfert d'un **nom de domaine** n'est plus possible par le biais d'une action en référé …
LEGALIS.COM.not Il apparaissait que ces **noms de domaine** avaient été réservés par la société Web vision.
contenu illicite 638 occurrences
CORPUS_SECURITE_INTERNET.not Les lignes directes sont des points de contact auxquels les utilisateurs peuvent signaler le **contenu illicite** sur Internet.
RDLI_2012.not aucun manquement à l'obligation de promptitude à retirer le **contenu illicite** ou à **en** interdire l'accès ne pouvait être reproché à la société Dailymotion
RDLI_2010.not PERSONNE OU D'UNE ENTREPRISE Lorsqu'un acteur est responsable de la diffusion d'un **contenu illicite** , la victime peut réagir pour faire supprimer et/ou sanctionner des propos diffamatoires, injurieux, outrageants
RDLI_2011.not Dès lors que le prestataire a eu connaissance du **contenu illicite** , il doit promptement **le** retirer conformément à l'article 14 de la directive.
RDLI_2007.not … salariés assignent alors en référé la société Wikimedia afin de faire ordonner la suppression du **contenu illicite** sur le fondement de l'article 6-I-8 de la loi du 21 juin 2001.
RDLI_2009.not Cela se traduit par la création d'une obligation prétorienne de non réapparition du **contenu illicite** déjà notifié.

Tableau 31. Exemples de contextes illustrant les cooccurrents les plus fréquents des termes**.**

7.2.2 Structure actancielle des termes à sens prédicatif

D'après Lyons (1978 : 122-127), les propositions se composent de deux sortes de termes : les *noms* et les *prédicats*. Les noms sont des termes qui font référence à des individus, c'est-à-dire les entités distinctes et identifiables comme personnes, choses, lieux de la vie quotidienne. En revanche, par prédicat *« on entend un terme qui est combiné à un nom dans le but de fournir une certaine information sur l'individu que le nom désigne : c'est-à-dire afin de lui attribuer une propriété quelconque »* (Lyons 1978 : 123). C'est un terme qui dit

quelque chose de l'autre terme. En effet, les prédicats sont considérés comme des opérateurs qui permettent de construire des propositions simples à partir des noms. Une proposition simple est donc une fonction du nom (ou des noms) dont elle se compose, (*le nom* étant un argument dans le sens mathématique). Ainsi, la structure logique d'une proposition peut être exprimée de la manière suivante : *f (x),* où *x* est une variable pour le *nom* et *f* pour le *prédicat*.

Il convient de souligner que la notion logico-mathématique est distincte de la notion linguistique. Comme le soulignent Mel'čuk et Polguère (2008), le concept de prédicat sémantique est une métaphore scientifique construite à partir du concept de prédicat logique : « *[…] les lois de la logique formelle ne s'appliquent que de façon « métaphorique » sur les structures sémantiques prédicatives* ». En effet, les auteurs (*ibid.* : 3) identifient quatre classes majeures de sens lexicaux ou sémantèmes : les prédicats sémantiques, les noms sémantiques, les quasi-prédicats sémantiques et les prédicats non actanciels. Cette classification repose sur deux axes de caractérisation de sémantèmes : 1) une propriété sémantique générale : dénoter un fait *vs* entité ; 2) une propriété de combinatoire : contrôler *vs* ne pas contrôler un nombre donné d'actants. D'après les auteurs, cette classification est primordiale, car elle a une incidence directe sur la façon dont le sens linguistique doit être décrit et modélisé. Ainsi, les noms sémantiques dénotent des entités (« choses qui existent ») par opposition aux faits, (« choses qui ont lieu »). Ils sont des sens non liants (« fermés sur eux-mêmes) : il s'agit d'unités dénuées d'actants sémantiques, c'est-à-dire d'unités aux sens desquelles il est impossible d'assigner une situation qui présuppose des participants clairement identifiés (Mel'čuk *et al.* 1995 : 77). Une entité peut renvoyer à un objet physique, une substance, un nom propre, un phénomène naturel ou à une espèce naturelle. Comme le souligne Polguère (2008 : 133), la partie du discours type des unités dont le sens est un nom sémantique est celle des noms.

Rappelons qu'en réduisant l'unité terminologique à une étiquette apposée sur un concept et en privilégiant une structuration hiérarchique, la terminologie traditionnelle a souvent considéré les termes comme des entités en occultant leur caractère prédicatif. Cependant, comme le remarque L'Homme (2004 : 62), de nombreuses unités terminologiques ne peuvent pas se décrire en utilisant comme seul point de repère l'organisation du monde réel. D'après l'auteure, pour expliquer le sens de ces unités, il faut les mettre en rapport avec d'autres sens. Elle parle alors de termes à sens prédicatif. Précisons d'après Polguère (*ibid.* : 132), Mel'čuk (2005 : 7) , que le sens prédicatif est un sens liant qui, du fait de sa structure interne et du comportement en phrase de l'unité qui le porte, est fait pour se combiner avec d'autres sens

afin de former une sorte de micro-message. En effet, les sens prédicatifs dénotent des faits (des actions, états, événements, processus) et par conséquent impliquent des participants appelés des actants sémantiques. De plus, les auteurs (Mel'čuk et Polguère 2008 : 99) remarquent la présence d'un ensemble très intéressant, et très important en langue, de sens « intermédiaire », qui tiennent à la fois des prédicats et des non-prédicats :

« Comme les non-prédicats, ces sens dénotent des entités, et non des faits. […] Comme les prédicats, ils ne peuvent être modélisés sans tenir compte de « positions sémantiques » qu'ils contrôlent […]. »

Comme leur comportement en langue est très proche de celui des prédicats, ils proposent de les appeler quasi-prédicats. Il s'agit notamment des individus impliqués dans une action donnée ou bien des objets ayant une fonction particulière.

Comme nous pouvons le constater en examinant les exemples présentés ci-dessous (Tableau 32), les contextes dans lesquelles apparaissent les termes permettent d'identifier si le terme que nous voulons décrire est un nom sémantique ou bien un prédicat ou quasi-prédicat sémantique et, le cas échéant, d'identifier le nombre d'actants qu'il contrôle. Ainsi, notre analyse a montré que la plupart des termes appartenant au domaine du droit de l'Internet sont des unités prédicatives (*consentement préalable, hébergement, obligation générale de surveillance*) qui dénotent une situation ou bien des quasi-prédicatives (*fournisseur d'hébergement*, *donnée à caractère personnel*, *nom de domaine*) qui désignent des entités engagées dans une situation. Sachant que les prédicats impliquent un certain nombre de participants obligatoires qui contribuent à leur sens (appelé *actants*), l'analyse des contextes dans lesquels se trouvent les termes à sens prédicatif permet donc d'identifier la structure actancielle de ces derniers. Cependant, nous tenons à souligner que la définition de la structure actancielle des unités terminologiques n'est pas une tâche facile et nous proposons de revenir sur cette question plus tard.

consentement préalable 114 occurrences
CORPUS_DONNEES.not Obligation d'obtenir le **consentement préalable** des personnes concernées pour diffuser des publicités comportementales

POSTES_COMMUNICATIONS_ELECTRONIQUES.not Le **consentement préalable** des abonnés à un opérateur de téléphonie mobile est requis pour toute inscription de données ….. RDLI_2009.not … la proposition de loi obligerait les responsables de traitements à obtenir le **consentement préalable** des personnes concernées avant de stocker des informations dans un équipement terminal de connexion.
obligation de surveillance de l'accès Internet 144 occurrences
LOI_FR.not la loi met à la charge de l'abonné à Internet une **obligation de surveillance** de son accès, prévue à l'actuel article L. 335-12 du code de la propriété intellectuelle. RDLI_2009.not L'article L. 336-3 institue une nouvelle **obligation de surveillance** à la charge de la personne titulaire de l'accès à des services de communication au public en ligne LOI_FR.not … cet article établit clairement, pour tout titulaire d'un abonnement, une **obligation de surveillance** de son accès à Internet.
hébergement 1349 occurrences
LE FORUM DES DROITS SUR INTERNET.not Prestataires de service d' hébergement (**hébergement** de pages personnelles, de sites commerciaux, **hébergemen**t de blogs, etc.) RDLI_2009.not eBay a la qualité d'hébergeur pour son activité d' **hébergement** à l'égard des annonces de ventes aux enchères postées sur ses sites. RDLI_2008.not la Cour d'appel de Paris rappelle les obligations incombant aux fournisseurs d' **hébergement** de sites Internet, tant en matière d'appréciation du caractère manifestement illicite de leurs contenus….

Tableau 32. Exemples de contextes permettant d'identifier la structure actancielle des termes.

7.2.3 Relations circonstancielles que le terme entretient avec d'autres unités

Comme nous l'avons vu plus haut, les actants sémantiques sont des participants obligatoires au sens des unités terminologiques prédicatives. Cependant, il convient de souligner qu'il existe d'autres types de relations sémantiques. En effet, certaines unités sont étroitement associées à des termes prédicatifs sans toutefois entrer dans la combinaison de

leur sens. Il s'agit de circonstants, c'est-à-dire d'unités qui font référence à des circonstances dans lesquelles se déroule la situation représentée par le prédicat (comme le lieu, le temps, le résultat, le moyen, la manière, l'instrument). On parle alors de relations circonstancielles. Dans les exemples présentés ci-dessous, nous pouvons repérer quelques circonstants du terme *téléchargement* comme : 1) *logiciel* qui fait référence à l'instrument utilisé à réaliser l'opération de *téléchargement* ; 2) *P2P* renvoyant soit à un instrument soit à un mode de *téléchargement*, 3) *sur Internet, plates-formes* à un lieu, à une source ou à un support. Par ailleurs, nous sommes d'accord avec L'Homme (2004 : 106) pour qui la distinction entre l'actant sémantique et le circonstant n'est pas toujours facile à réaliser dans les faits. La structure du terme *téléchargement* contient quatre actants (la personne qui télécharge, les données téléchargées, la source d'où les données sont téléchargées et la destination vers laquelle les données sont téléchargées). En analysant les contextes ci-dessous, on peut se demander si *sur Internet* correspond effectivement à une localisation (circonstant) ou bien à la source (actant).

téléchargement 1282 occurrences
LOI_FR.not … le filtrage de protocole permet de bloquer les **téléchargements** illicites utilisant le P2P, mais le P2P n'est plus la seule voie de **téléchargement** illicite de contenus depuis le développement des news groups et des sites de partage vidéo dont le succès va croissant
RDLI_2010.not Les questions juridiques posées par les plates-formes de **téléchargement** payant ne viennent pas du support mais des modalités d'exploitation.
RDLI_2012.not En premier lieu, cette société fit valoir que l'activité de diffuseur de presse se distinguait radicalement de celle de l'éditeur de presse, indépendamment du mode diffusion des titres de presse : support imprimé classique sur papier ou format numérique accessible au **téléchargement** sur Internet ; seul le second étant responsable du contenu des magazines qu'il édite.
LOI_FR.not Les logiciels de **téléchargement** sont encore plus accessibles au grand public, leur utilisation s'acquiert rap*ibid.*ent.

Tableau 33. Exemples de contextes permettant d'identifier des relations circonstancielles que le terme entretient avec d'autres unités.

7.2.4 Relations hiérarchiques ou paradigmatiques que le terme entretient avec d'autres unités

Il convient de souligner qu'un grand nombre de contextes extraits du corpus *DITerm* contiennent des hyperonymes, des hyponymes, des antonymes ou d'autres unités reliées paradigmatiquement avec le terme cible. Les relations présentées ici sont considérées comme étant fondamentales dans la description terminologique. En effet, comme le souligne L'Homme (2004 : 90), elles sont très proches des relations étudiées traditionnellement dans les représentations conceptuelles et permettent au même titre que les relations ontologiques de refléter la structure hiérarchique du domaine. C'est notamment le cas des relations taxinomiques et méronymiques. Les premières sont fondées sur le fait que le sens possède des composantes communes. Une taxinomie comprend donc des relations verticales représentées par les hyperonymes et les hyponymes (les unités qui n'ont pas le même rang) et des relations de type horizontal entre les co-hyponymes (les unités de même rang). Comme le remarque Polguère (2008 : 148), l'hyperonymie et l'hyponymie sont deux relations mutuellement converses qui correspondent à une situation d'inclusion de sens : le sens de l'hyperonyme, plus général est inclus dans le sens de l'hyponyme plus spécifique (comme c'est le cas de *service de communications électroniques* qui est un hyperonyme de *téléchargement* ou bien celui de *données sensibles* qui est un hyponyme de *données à caractère personnel*).

Quant aux co-hyponymes, ils possèdent toutes les composantes de l'hypéronyme mais se distinguent entre eux par une ou quelques composantes (comme c'est le cas des paires de co-hyponymes *téléchargement* et *streaming* par rapport à *service de communications électroniques*; *consentement préalable* (ou *« opt-in)* et *droit d'opposition* (ou *« opt-out »*) par rapport à *mécanisme (de protection)*, ou bien *téléchargement gratuit* et *téléchargement payant* par rapport au *téléchargement* Remarquons que dans le dernier cas, la paire de termes peut être considérée comme étant une paire d'antonymes. Les antonymes sont des termes qui entrent dans une relation d'opposition (L'Homme 2004 : 96), c'est-à-dire dont les sens se distinguent par la négation ou, plus généralement, la mise en opposition d'une de leurs composantes (Polguère 2008 : 152). La relation d'antonymie ne doit pas être confondue avec celle de conversivité. En effet, comme le précise Poguère (*ibid.* : 154), pour bien comprendre la notion de conversivité, il faut utiliser la modélisation des sens lexicaux en tant que prédicats ou quasi-prédicats sémantiques. Selon l'auteur, deux lexies sont conversives si elles

remplissent deux conditions. Elles doivent être des prédicats sémantiques dénotant une même situation (ou des quasi-prédicats sémantiques dénotant deux entités impliquées dans une même situation) et elles doivent s'expriment dans la phrase avec une inversion de l'ordre de leurs actants. C'est notamment le cas des termes *téléchargement descendant* ou (« download ») et *mise à disposition* (« upload »).

En ce qui concerne les relations méronymiques qui selon L'Homme (*ibid.* : 98) sont fondées sur les notions vagues de proximité ou d'association dans l'espace, nous n'en avons repéré que très peu d'exemples dans notre corpus. Dans la plupart des cas, il s'agit des relations de type *élément-ensemble* comme *données à caractère personnel* et *base de données à caractère personnel*. Le Tableau 34 reproduit quelques contextes extraits du *DITerm* contenant des indices des relations paradigmatiques.

téléchargement 1282 occurrences
LOI_FR.not … À côté du **téléchargement** effectué depuis un ordinateur distant (téléchargement descendant ou download) existe également la mise à disposition de fichiers que certains appellent également « téléversement » (transmission de données vers un ordinateur distant ou upload) … JURISPRUDENCE_UE_FR.not … d'autres services tels que, notamment, des services de courrier électronique, de **téléchargement** ou de partage des fichiers… LEGALIS.COM.not … la présence de deux liens permettant à l'internaute d'avoir accès gratuitement au film « Les dissimulateurs » dans son intégralité, en flux continu (streaming) ou en **téléchargement**…. RDLI_2010.not … la dichotomie opérée par le public et les médias entre plates-formes de **téléchargement** payant considérées comme légales et réseaux de **téléchargement** gratuit présentés comme favorisant la contrefaçon.
consentement préalable 114 occurrences
LEGALIS.COM.not les termes de l'alternative entre système de **consentement préalable** (dit "opt-in") ou mécanisme fondé sur le droit d'opposition ("opt-out") sont maintenant bien définis
donnée à caractère personnel 3913 occurrences

CORPUS_RDLI 2011.not … des parlementaires membres de la Cnil ont souhaité du législateur qu'il affirme enfin « sans ambiguïté que l'adresse IP constitue une **donnée à caractère personnel**

CORPUS_DONNEES.not … En vertu de la législation sur la protection des données, le traitement de **données à caractère personnel**, comme en l'espèce le traitement des données relatives au trafic et au contenu, doit reposer sur une base juridique adéquate.

CORPUS_DONNEES.not il existe un danger réel que des niveaux différents de protection des **données à caractère personnel** données sensibles soient autorisés au titre de la directive.

CORPUS_DONNEES.not … les règles de protection des **données à caractère personnel** applicables au traitement de données relatives au trafic et de données de localisation générées par l'utilisation de services de communications.

CORPUS_DONNEES.not Par «données anonymes» au sens de la directive, on entend toute **données** concernant une personne physique lorsque cette personne ne peut être identifiée, ni par le responsable du traitement des données ni par une autre personne…

RDLI_2008.not Si l'inscription au site est gratuite, s'il ne poursuit directement aucun but lucratif, en fait, il s'agit de constituer une base de **données personnelles** comprenant potentiellement tous les enseignants de France, les élèves (souvent mineurs), les parents d'élèves…

Tableau 34. Exemples de contextes contenant des indices des relations paradigmatiques ou hiérarchiques.

7.2.5 Indices sur l'organisation conceptuelle du domaine du droit de l'Internet

Jusqu'à présent, nous nous sommes penchée sur l'étude des contextes reflétant des relations lexico-sémantiques que les termes sélectionnés entretiennent avec d'autres unités réalisées dans le corpus. Maintenant, nous proposons de nous tourner vers un autre type de contextes, à savoir les contextes contenant des indices sur l'organisation conceptuelle du domaine. Rappelons que le concept de *contextes riches en connaissances* ou *patrons de connaissances* (*knowledge-rich contexts* en anglais) a été introduit par Meyer et défini comme « *a context indicating at least one item of domain knowledge that could be useful for conceptual analysis.* (Meyer 2001 : 281 citée dans Bowker et L'Homme (2004 :184). Il est

important de souligner que notre étude ne s'inscrit pas dans une optique classique ni ontoterminologique (Roche 2007) dont la finalité est l'élaboration d'un réseau termino-ontologique constitué de concepts terminologiques et des relations les liant. Cependant, nous considérons que la prise en considération de la dimension conceptuelle est nécessaire pour fournir une description riche et complète des termes du domaine. Et même si nous adhérons à l'approche lexico-sémantique, nous ne devrons pas occulter le fait que les termes reflètent la structure conceptuelle du domaine. Pour pouvoir la dégager, il est donc nécessaire de s'intéresser aux relations spécifiques au domaine en question, appelées parfois des relations associatives (Dancette 2011a et 2011b) ou transversales (Grabar et Hamon 2004). Rappelons les travaux de Sager (1990 : 34-37) qui a défini toute une série de différentes relations conceptuelles (matériau et produits, procédé et instrument, objet et qualité). Cependant, les types de liens dégagés par Sager se révèlent peu pertinents pour notre étude. Comme nous l'avons déjà signalé, le droit est l'ensemble des règles qui régissent la conduite de l'individu en société. Il ne s'agit donc pas d'un domaine de connaissance dont la description se résumerait à l'énumération des caractéristiques des entités qui le forment. En effet, en proposant une représentation conceptuelle d'une discipline juridique, il faut prendre en compte sa logique déontique qui vise à formaliser les rapports existant entre les quatre alternatives : l'obligation, l'interdiction, la permission et le facultatif.

Pour ce faire, nous nous sommes intéressée à un groupe de termes génériques qui peuvent être considérés comme révélateurs du schéma conceptuel du domaine du droit de l'Internet (voir les catégories (4) et (9) définies plus haut). Nous avons décidé de rechercher leur présence dans l'environnement contextuel des termes sélectionnés pour ensuite dégager les liens qui les relient à ces derniers. Comme nous l'avons déjà mentionné, nous considérons que cette analyse nous permettra d'organiser les données extraites du corpus en catégories conceptuelles et d'établir un réseau de liens internotionnels. Ainsi, comme nous pouvons l'observer ci-dessous, le concept *téléchargement* est qualifié comme *acte de reproduction* et *représentation* et peut avoir le caractère *légal* ou *illégal*. Quant au deuxième concept, il est possible de déduire des contextes repérés que les *données à caractère personnel,* font l'objet de mesures de protection légale et que leur traitement est soumis à un certain nombre de règles. Finalement, les contextes associés au terme *fournisseur d'hébergement* rendent compte des obligations qui incombent à ce prestataire de services.

téléchargement 1282 occurrences
LEGALIS.COM.not Considérant qu'il est incontestable que le **téléchargement** constitue à la fois un acte de reproduction, à raison du copiage des œuvres et de leur stockage sur le disque dur de l'internaute, et un acte de représentation à raison de leur communication au public des internautes par télédiffusion RDLI_2010.not … la dichotomie opérée par le public et les médias entre plates-formes de **téléchargement** payant considérées comme légales et réseaux de **téléchargement** gratuit présentés comme favorisant la contrefaçon. RDLI_2011.not Reste donc à espérer que l'aspect dissuasif généré par les nombreux débats autour de ces lois suffira à faire prendre conscience aux pirates des dangers du **téléchargement** illégal et les incitera à recourir massivement au **téléchargement** légal, dont l'offre s'étend et se démocratise progressivement.
données à caractère personnel 3913 occurrences
CORPUS_DONNEES.not …En cas de violation de **données à caractère personnel**, le fournisseur de services de communications électroniques accessibles au public avertit sans retard … CORPUS_DONNEES.not … qui réclament une plus grande sécurité juridique et une harmonisation plus poussée des règles en matière de protection des **données à caractère personnel.** CORPUS_PROPRIETE INTELECTUELLE.not Les titulaires de droits devraient être dûment informés du traitement de leurs données, de l'identité des destinataires de celles-ci, des délais de conservation de leurs **données** dans les bases de données des destinataires, ainsi que des modalités d'exercice de leurs droits d'accès aux **données à caractère personnel** les concernant et de leurs droits de rectification ou d'effacement de **celles-ci**, conformément aux articles 10 et 11 de la directive 95/46/CE.
fournisseur d'hébergement / hébergeur 964 occurrences
RDLI_2009.not D'autre part, le **fournisseur d'hébergement** a l'obligation de conserver les données permettant l'identification des contributeurs à un contenu. RDLI_2008.not Puisque le **fournisseur d'hébergement** n'est pas responsable des contenus qu'il héberge par principe, il est nécessaire de permettre l'identification du responsable de ce contenu

RDLI_2008.not C'est une obligation particulière de surveillance qui est imposée et ce uniquement sur les contenus dont la diffusion a été notifiée comme illicite au **fournisseur d'hébergement**.
LEGALIS.COM.not … les sociétés Google ont maintenu en ligne le contenu illicite après la mise en demeure alors qu'une seule notification oblige l'**hébergeur** à rendre l'accès impossible

Tableau 35. Indices sur l'organisation conceptuelle du domaine du droit de l'Internet.

7.2.6 Énoncés définitoires et contextes explicatifs

L'extraction des contextes associés aux termes nous a également permis d'accéder à de nombreuses informations sur les notions véhiculées par les termes. Rappelons que nous avons décidé d'intégrer à notre corpus les documents qui correspondent à l'époque où l'on a introduit les concepts fondamentaux du droit de l'Internet car ils renferment selon nous beaucoup plus d'éléments définitoires et de contextes explicatifs que des textes plus récents où les concepts sont déjà bien établis. Ainsi, nous avons pu repérer un certain nombre d'énoncés définitoires qui contiennent des définitions du concept plus au moins formelles (comme celle du terme *données à caractère personnel* extraite de la Directive 95/46/ce du 24 octobre 1995 relative à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données), ou bien des contextes explicatifs proposant des informations sommaires et fragmentaires. Dans les deux cas, les renseignements pourront être utilisés pour préparer une définition et serviront à nourrir la description du terme. Précisons qu'un certain nombre de chercheurs, entre autres Pearson (1998 : 121-134), ont élaboré des stratégies visant à isoler ce type de contextes au moyen de marqueurs linguistiques. L'étude des occurrences réalisées dans notre corpus a permis de confirmer l'hypothèse selon laquelle le terme utilisé comme pivot et assorti de marqueurs tels que *on entend par*, *appelé(e)*, *le terme*, *l'expression*, *défini(e)* conduit le terminographie vers ce type de contexte (Pearson 1998 :135 – 190).

CORPUS_DONNEES .not …Aux fins de la présente directive, on entend par: a) «**données à caractère personnel**»: toute information concernant une personne physique identifiée ou identifiable (personne concernée); est réputée identifiable une

personne qui peut être identifiée, directement ou indirectement, notamment par référence à un numéro d'identification ou à un ou plusieurs éléments ….
CORPUS_DONNEES.not … ii) l'accord doit être obtenu avant que le cookie soit placé et/ou que les informations stockées dans l'équipement terminal de l'utilisateur soient collectées, ce qui est généralement appelé «**consentement préalable**»…
RDLI_2007.not D'autre part, l'autre question posée par les termes de la prévention était la portée des actes englobés par l'expression « **téléchargement** ». Selon le prévenu, cette expression ne recouvrait que les actes de reproduction, l'équivalent du terme anglais « download ». Au contraire, les plaignants soutenaient que le terme « **téléchargement** » englobait tant les actes de reproduction de réception (download) que les actes de diffusion d'émission (upload).
RDLI_2009.not La directive « commerce électronique » définit trois types de prestations susceptibles d'être offertes par les intermédiaires techniques : le simple transport, le stockage temporaire dit « caching » et le stockage permanent dit « **hébergement** » (18). Les deux premières activités sont de la compétence du fournisseur d'accès, la troisième est de celle de l'hébergeur. Au niveau national, l'article 6-I-2 de la LCEN définit les **hébergeurs** comme des « personnes physiques ou morales qui assurent, même à titre gratuit, pour mise à disposition du public par des services de communication au public en ligne, le stockage de signaux, d'écrits, d'images, de sons ou de messages de toute nature fournis par des destinataires de ces services (...) » LEGALIS.COM.not l'**hébergement** (formule désignant la fourniture d'un service consistant à stocker des informations fournies par un destinataire du service

Tableau 36. Contextes définitoires fournissant des informations encyclopédiques sur le domaine de spécialité.

L'étude de l'environnement contextuel des termes, nous a permis de confirmer l'hypothèse selon laquelle il est possible d'accéder à des relations lexicales et conceptuelles entretenues par les termes par le biais de ses relations contextuelles. Pour terminer cette analyse, il convient de noter que selon l'approche dite *corpus-driven* chaque ligne de la concordance constitue un cas particulier, un événement unique, un exemple d'emploi vivant de la langue dans un contexte singulier. La généralisation d'un fait de langue (qui mène à la théorisation), n'est possible que si des parallélismes ou des similarités sont observés dans

d'autres contextes de la même forme (rapprochées et alignées sur une colonne de la concordance). Ceci nous conduit à rappeler la notion clé de la théorie sinclairienne (Sinclair 2003 : 30-35), notamment celle d'*extended lexical unit* ou *extended unit of meaning* (déjà évoquée dans ce travail, voir la section 3.3.2.2), l'idée selon laquelle l'analyse lexicologique (et dans notre cas - terminologique) ne doit pas porter sur l'unité, mais plutôt sur son environnement linguistique. Rappelons que pour les contextualistes, il s'agirait d'un phénomène collocationnel, étudié au niveau textuel et défini en fonction de l'apparition de cooccurrences à l'intérieur d'une fenêtre. L'analyse se situe donc sur l'axe syntagmatique et concerne les liens d'enchaînement qui se nouent, dans le discours, entre les unités consécutives qui constituent la ligne de la parole. Une *unité lexicale étendue* (ULE) s'articule donc autour d'un noyau et s'étend à des unités proches liées entre elles à des degrés différents et sélectionnées en fonction de critères d'affinité. Chaque ULE est donc décrite à l'aide de quatre paramètres : la collocation, la colligation, la préférence sémantique et la prosodie sémantique et doit être analysée à ces quatre niveaux qui restent étroitement associés. Pour notre part, nous proposons de considérer une unité terminologique comme une *unité terminologique étendue* qui peut être analysée à l'aide de cinq paramètres : la collocation, la colligation, la préférence sémantique, la prosodie sémantique et l'affinité conceptuelle. En effet, nous considérons qu'aussi bien les propriétés linguistiques que conceptuelles des termes se manifestent dans le discours et qu'il est possble d'y accéder par le biais de leurs relations contextuelles. Citons à ce propos Cruse (1986 : 1) :

« […], *it is assumed that the semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts. […], the relation between a lexical item and extra-linguistic contexts is often crucially mediated by the purely linguistic context […], any aspect of an extra-linguistic context can in principle be mirrored linguistically; […], linguistic context is more easily controlled and manipulated. We shall therefore seek to derive information about a word's meaning from its relations with actual and potential linguistic contexts.* »

Cruse (1986 : 1)

« *The full set of normality relations which a lexical item contracts with all conceivable contexts will be referred to as its contextual relations. We shall say, then, that the meaning of a word is constituted by its contextual relations.* »

Cruse (*ibid. :* 16)

Il est intéressant de noter que l'étude de l'environnement contextuel (c'est-à-dire des rapports syntagmatiques entre les unités terminologiques) permet d'accéder aux relations paradigmatiques (mais aussi conceptuelles), ce qui prouve que ces phénomènes sont étroitement liés.

QUATRIÈME PARTIE : Le terme et la nébuleuse de ses relations - à la recherche d'un modèle de description des unités terminologiques du domaine du droit de l'Internet

Comme le souligne Cornu (2005 : 192), chaque terme, pris comme base, peut être le centre d'une constellation d'où partent des termes coordonnés. La Figure 32 schématise une telle situation. Nous y voyons le terme *données à caractère personnel* accompagné d'unités lexicales ou terminologiques qui partagent avec ce premier un certain nombre de relations. Chaque flèche correspond à une classe de relation. Nous pouvons constater que le terme en question entretient différents rapports sémantiques, notamment :

- relations paradigmatiques comme : a) synonymes : *données personnelles, données relative aux personnes physiques, données nominatives* ; b) antonymes : *données anonymes,* c) hyponymes : *données sensibles, adresse IP, cookies,* d) hypéronymes : *données numériques*
- relations actancielles : *responsable du traitement, personne concernée*
- relations circonstancielles : *réseau des communications électroniques*
- relations collocationnelles : *traiter, collecter, enregistrer*

Le schéma présente également quelques exemples de liens de parenté sémantique faible qui relèvent plutôt de l'ordre conceptuel. Il s'agit notamment des relations indirectes comme celles partagées par le terme avec les unités suivantes :

- *confidentialité, intégrité, sécurité, disponibilité, authenticité* – termes dénotant des propriétés des *données à caractère personnel*
- *G29, CNIL, CIL* - noms des organismes de contrôle chargés de la protection des données à caractère personnel
- *droit d'accès, droit d'opposition, droit de rectification, consentement préalable* – termes renvoyant à une série de dispositifs adoptés pour assurer la protection des données à caractère personnel.

Ces relations permettent de rendre compte de la structure du domaine de l'Internet en fournissant des informations de nature encyclopédique. Néanmoins, il est difficile de les décrire en se basant sur des critères strictement linguistiques.

Ainsi, comme le montre l'exemple ci-dessous, l'analyse des contextes extraits du corpus *DITerm* nous a permis de collecter un grand nombre de données formant autour des termes choisis des réseaux de relations de nature différente. Il est important de souligner que pour certains termes, comme *données à caractère personnel* cité ci-dessus, ou bien *nom de domaine*, *contenu illicite* ou *téléchargement illégal*, nous avons recensé jusqu'à quelques dizaines de relations différentes. En effet, ce qui caractérise, dans leurs rapports mutuels, les éléments de ces regroupements (comme ceux de la Figure 32), c'est leur communauté d'action. Comme le souligne Cornu (2005 : 202), le droit gouverne et sanctionne des activités (licites et illicites). Or, dans chaque opération entrent plusieurs sortes d'éléments qui président à sa formation et à son exécution. Il est donc possible, sur le plan linguistique, de trouver les termes désignant les éléments d'une même opération et de les rassembler dans des ensembles lexicaux formant des familles opérationnelles. Cornu (*ibid.*) parle de groupes d'intervention constitués de termes coacteurs qui sont complémentaires et ont vocation à être associés dans un même énoncé. Les rapports qui les unissent correspondent à des liaisons opératoires.

Il est donc évident que pour rendre compte du comportement du terme dans le discours (et tel est l'objectif final de notre recherche), il faut décrire toutes les relations que ce dernier

partage avec les unités qui ont tendance à apparaître dans son univers contextuel. Cependant, sachant que le seul point commun à l'ensemble de ces unités est leur participation à l'opération évoquée par le terme en question, un certain nombre de questions d'ordre méthodologique doivent être posées. En effet, face à la multitude de données associées aux termes choisis pour la description, nous nous demandons comment expliciter cette variété de relations dans une base de données terminographique. Quel formalisme adopter pour décrire toute la richesse des informations extraites du corpus ? Comment systématiser les données ? L'enjeu principal de notre travail est d'essayer de répondre à ces questions en proposant un modèle de description hybride qui permettrait de rendre compte à la fois des relations conceptuelles et des relations sémantiques (y compris les phénomènes phraséologiques propres à la langue de spécialité donnée). Pour ce faire, nous serons amenée, dans les pages qui suivent, à mettre en œuvre deux stratégies (souvent considérées comme concurrentes ou bien incompatibles), notamment :

- la description détaillée du fonctionnement linguistique des termes appartenant au domaine du droit de l'Internet en se basant sur leurs caractéristiques lexico-sémantiques;

- la structuration du schéma conceptuel du domaine du droit de l'Internet

Ainsi, comme nous l'avons déjà mentionné plus haut, notre travail sera réalisé parallèlement sur deux plans, un plan linguistique et un plan conceptuel. Le premier volet de cette étude visera donc à proposer un modèle de description des propriétés linguistiques des termes. Le deuxième devra nous conduire à trouver une méthode de représentation des connaissances exploitable dans un projet terminologique.

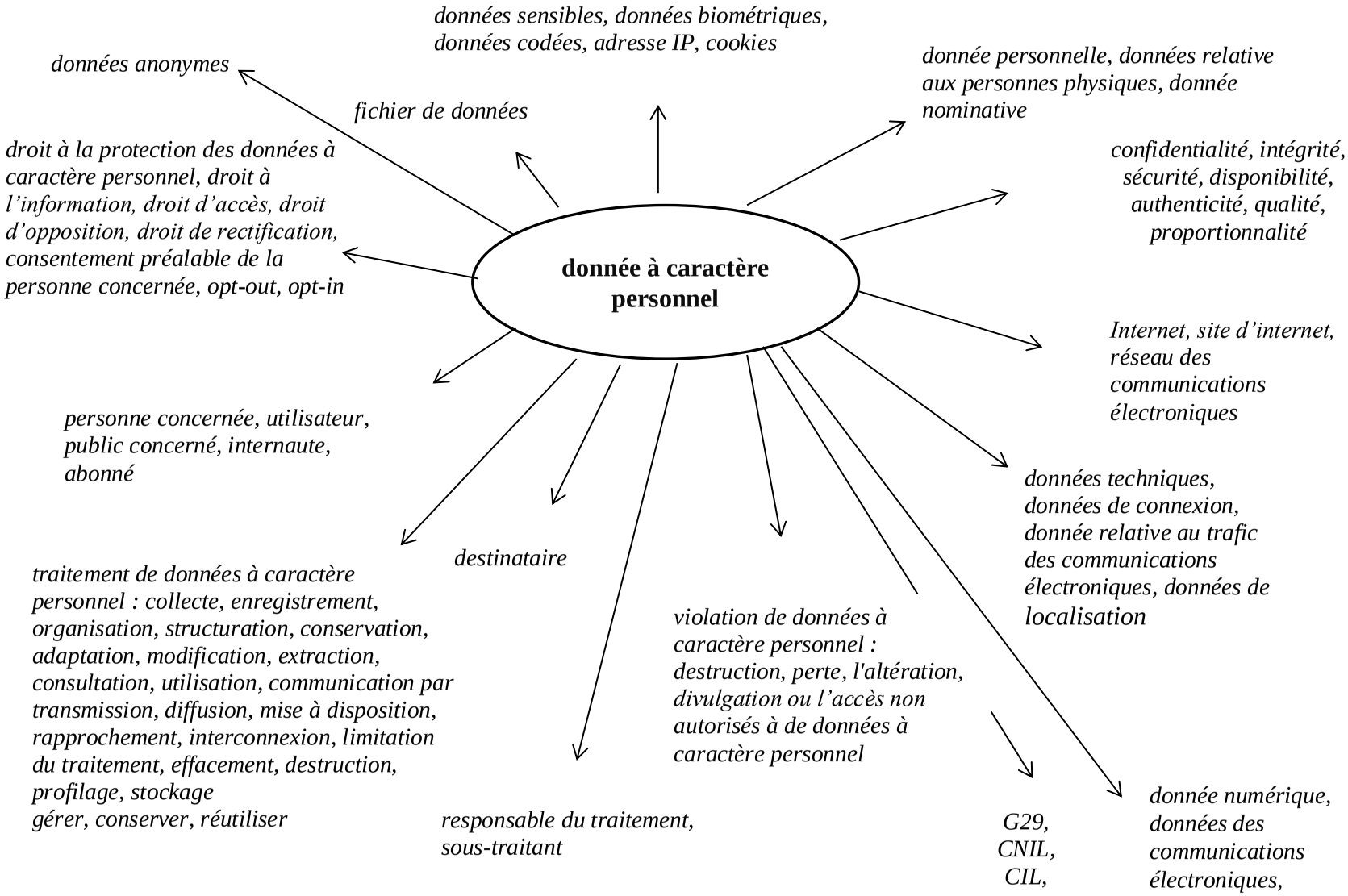


Figure 32. Schéma des relations du terme *données à caractère personnel*

Chapitre 8. Description des propriétés lexico-sémantiques des unités terminologiques – exploitation du modèle des fonctions lexicales et adaptation de celui-ci

La présente partie sera consacrée au premier volet de notre démarche, notamment à la recherche d'une méthode de description des propriétés linguistiques des termes. L'objectif est de rendre compte des relations lexico-sémantiques que ces derniers entretiennent avec d'autres unités terminologiques et lexicales appartenant au vocabulaire du droit de l'Internet. Pour ce faire, nous nous sommes intéressée au modèle des fonctions lexicales (dorénavant notées *FL*), développé par Mel'čuk et ses collaborateurs Alain Polguère et André Clas dans la cadre de la Lexicologie Explicative et Combinatoire (Mel'čuk *et al.* 1995), qui constitue, comme nous l'avons vu dans la première partie de ce travail, une composante d'une théorie plus générale, à savoir la Théorie Sens-Texte (TST) (Mel'čuk 1997). En effet, l'approche mel'čukienne offre une méthode de description globale de l'unité lexicale. L'originalité des FL est de proposer un modèle fonctionnel unique qui permet de rendre compte de façon uniforme de différents phénomènes. Les FL mettent en lumière une multitude de relations qu'une lexie entretient avec d'autres unités aussi bien sur l'axe paradigmatique que sur l'axe syntagmatique. Vu que nous avons déjà consacré une partie importante du premier chapitre à la présentation des travaux de Mel'čuk, nous voudrions, dans les lignes qui suivent, rappeler uniquement quelques détails d'ordre méthodologique.

En effet, l'objectif des FL est de décrire une relation sémantico-lexicale en un encodage synthétique rendant compte du sens et des caractéristiques syntaxiques d'une relation. Ainsi, du point de vue formel, une fonction lexicale ressemble à une fonction mathématique qui peut être représentée de la manière suivante :

$$f(x) = y,$$

où *x* est l'argument de la fonction (ou son mot-clé) et *y* sa valeur. Ces fonctions sont appelées lexicales car elles n'acceptent en tant qu'argument que des lexies, et en tant que valeur que

des ensembles de lexies (Mel'čuk *et al.* 1995 : 126). Autrement dit, une fonction lexicale est une correspondance *f* qui associe à une lexie L (argument de *f*), un ensemble de lexies ou syntagmes figés f(L) – valeur de *f*.

Bien qu'elle soit basée sur une approche lexico-sémantique, la méthode de description des unités terminologiques au moyen des FL a déjà séduit un grand nombre de terminographes. Comme nous l'avons vu plus haut, l'adaptation du modèle mel'čukien à la terminologie a fait l'objet de travaux conduits par Frawley (1988), L'Homme (*DiCoInfo, DiCoEnviro*), Dancette (*DAD, DAMT*), Mortchev – Bouveret (2007), (Faber et Sánchez 2001). D'autres auteurs de dictionnaires spécialisés tels que Cohen (*Bourse et conjoncture économique,* 1986) ou Binon et al (*DAFA*, 2000) s'en sont également inspirés en proposant leurs modèles de description des phénomènes phraséologiques. Le projet *DITerm* cherche à s'inscrire dans cette mouvance. En effet, parmi les projets évoqués ci-dessus, il y en a un qui a particulièrement influencé notre travail. Il s'agit du *DiCoInfo*, dictionnaire spécialisé en ligne, élaboré sous la direction de Marie Claude L'Homme au sein de l'Observatoire de linguistique Sens Texte (OLST) de l'Université de Montréal, décrit en détail dans le deuxième chapitre de notre étude. Rappelons seulement que les auteurs se détachent radicalement de l'optique conceptuelle et envisagent le terme comme une unité lexicale dont le sens peut être associé à un champ disciplinaire préalablement délimité (L'Homme 2005 : 142). La perspective adoptée a d'importantes conséquences sur les descriptions proposées. En effet, les terminologues de l'OLST s'intéressent plutôt au sens que revêtent les formes linguistiques qu'à la place des concepts (que ces formes linguistiques dénotent) dans un système. Cela permet de mettre au jour la structure lexicale observable à l'intérieur d'un domaine spécialisé et de décrire la multitude des liens linguistiques existant entre les termes (voir la section 2.3).

8.1 L'identification de la structure actancielle des termes

Comme évoqué ci-dessus, l'objectif de notre étude est de trouver une méthode de description des propriétés linguistiques des termes qui s'inspire du modèle des fonctions lexicales. Rappelons brièvement que les FL permettent de modéliser deux types de phénomènes : les dérivations sémantiques et les collocations. Cependant, en faisant appel à cet outil, il faut savoir que les relations décrites au moyen des fonctions lexicales sont toutes ancrées dans le contenu sémantique de la lexie. En effet, selon la Lexicologie Explicative et

Combinatoire (décrite en détail dans le premier chapitre de ce travail), la plupart des propriétés de comportement d'une lexie (une unité terminologique dans notre cas) sont sous-tendues ou même carrément déterminées par son sens *dénotationnel,* appelé aussi *sens situationnel* ou *sens propositionnel* (Mel'čuk *et al.* : 1995 : 73). La description lexicale réalisée à l'aide des FL s'appuie donc fortement sur la décomposition du sens, et il faut savoir que dans ce processus, l'identification de la nature prédicative d'une lexie donnée constitue une étape essentielle. Comme le soulignent Mel'čuk & Polguère (2007 : 24), dans la plupart des cas, le sens d'une lexie ne peut être clairement compris, et donc décrit, sans prendre en considération sa structure actancielle (c'est-à-dire sa caractérisation en tant que prédicat sémantique). « *Bon nombre de phénomènes de dérivation sémantique (S_i, A_i, etc.) et de collocation ($Oper_i$, $Func_i$, etc.) ne se comprennent qu'en référence à la notion d'actant et, donc, de prédicat sémantique.* » Polguère (2003 : 130). Ainsi, avant de procéder à une modélisation des phénomènes dérivationnels ou collocationnels au moyen des FL, il est nécessaire d'établir le caractère prédicatif ou non prédicatif des unités terminologiques choisies. Ceci aura une incidence directe sur la façon dont leur sens linguistique sera décrit et modélisé.

Compte tenu de ces remarques, il nous paraît important de nous arrêter sur la question de la prédicativité (Mel'čuk et Polguère 2008, Polguère 2012). Ainsi, d'après Polguère (*ibid.*), la prédicativité est une propriété intrinsèque d'une lexie, propriété qui concerne la structure de son signifié. Autrement dit, il s'agit de la propriété d'un sémantème donné d'être un prédicat sémantique ou de ressembler à un prédicat sémantique. Rappelons que la Lexicologie Explicative et Combinatoire définit un sens prédicatif (pris comme terme logico-sémantique) comme : « *un sens qui a des « trous » pour recevoir d'autres sens ; un sens prédicatif est un sens « liant » - il réunit d'autres sens en des configurations sémantiques tout comme un tube de jonction réunit les pôles d'une tente pour former le squelette porteur de la tente.* » (Mel'čuk *et al.* (1995 : 76). Pour bien saisir la notion de prédicativité, il est donc important de comprendre l'idée de liage. Comme le précise Polguère (2012 : 4), le concept de sens liant a été proposé dans le contexte de la caractérisation formelle des réseaux sémantiques Sens-Texte, afin de particulariser les sémantèmes qui, dans un graphe sémantique, ont la capacité de gouverner d'autres sémantèmes. En effet, étant donné qu'une unité prédicative typique dénote un fait et qu'un fait présuppose un certain nombre de participants, il est approprié de modéliser son sens par une micro-structure qui contient la représentation des participants potentiels du fait en question. Ainsi, comme le soulignent Mel'čuk et Polguère (2008 : 2), cette micro-structure peut être visualisée sous la forme d'un graphe du type *réseau*

sémantique[74] constitué du sens liant lui-même, connecté aux variables qui désignent les « positions disponibles » pour les sens correspondant aux participants. Prenons l'exemple du terme *téléchargement illégal*. La situation dénotée par ce terme à sens prédicatif sous-entend quatre participants: la personne qui pratique un téléchargement illégal, les données qu'elle télécharge, la source à partir de laquelle elle télécharge ces données et la destination des données en téléchargement. La figure ci-dessus (Figure 33) illustre la structure sémantique en question : téléchargement illégal de X réalisé par Y à partir de Z vers W.

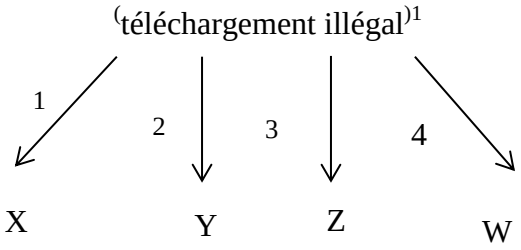


Figure 33. Structure actancielle du terme *téléchargement illégal*

Comme nous pouvons le constater, la structure sémantique du (téléchargement illégal) inclut 4 variables (X, Y, Z, W) qui réfèrent aux participants de la situation dénotée par le terme. En réalité, il est impossible de décrire le (téléchargement illégal) sans prendre en compte ces quatre participants qui, du fait de leur implication dans la situation linguistique, contribuent au sens de l'unité terminologique en question. On dit alors que le sémantème (téléchargement illégal) contrôle quatre positions actancielles.

D'après la théorie Sens-Texte, une position actancielle (*actant slot* en anglais) contrôlée par un sémantème est, dans une structure sémantique que l'on appelle *structure actancielle*, une position destinée à être occupée par un autre sémantème (Polguère 2012 : 4), appelé *actant*. Comme le précise Mel'čuk (2004 : 5) : *« […], an actant slot of L in the Lexicon is an "empty place" or "open position" foreseen in the lexicographic description of L* ». L'auteur souligne en même temps qu'il est important de faire la distinction entre le concept de position actancielle et celui d'actant. Pour illustrer cette différence, Mel'čuk (*ibid.*

[74] Précisons que dans la théorie Sens-Texte, la représentation du sens de la lexie (soit une représentation sémantique [=RSém]), s'écrit sous la forme d'un réseau constitué de nœuds (étiquetés de noms d'éléments sémantiques), reliés par des flèches (munis de numéros de relations « prédicats ~ arguments »), (Mel'čuk *et al.* 1995 : 73).

: 60) évoque l'image proposée par Boguslavskij (1985 : 11), selon laquelle les positions actancielles pourraient être comparées à différents types d'hameçons adaptés à la pêche de différentes espèces de poissons tandis que les poissons attrapés à l'aide de ces différents types d'hameçons symboliseraient, quant à eux, les actants. Mel'čuk rappelle également que la notion d'actant a été introduite pour la première fois par Tesnière (1988 : 107-115), qui a étudié le phénomène dans ses travaux sur la valence. Selon ce dernier (*ibid.* : 105), l'actant correspond aux personnes ou aux choses qui participent à un degré quelconque au procès :

« *Le nœud verbal que l'on trouve au centre de la plupart de nos langues européennes […], exprime tout un petit drame. Comme un drame en effet, il comporte obligatoirement un procès, et le plus souvent des acteurs et des circonstances. Transportés du plan de la réalité dramatique sur celui de la syntaxe structurale, le procès, les acteurs et les circonstances deviennent respectivement le verbe, les actants et les circonstants.* »

(*ibid.* : 102)

La Lexicologie Explicative et Combinatoire, quant à elle, propose la définition suivante du phénomène : « *Un actant sémantique d'une lexie L, dans une structure sémantique donnée, est un sémantème qui remplit une position actancielle sémantique associée à L dans le lexique.* » (Mel'čuk et Polguère (2008 : 4). Il convient de souligner que pour être considéré comme actant sémantique d'une lexie L, un sémantème doit remplir certaines conditions qui ont été décrites en détail par Mel'čuk (2004).

La première condition est d'ordre sémantique et a déjà été évoquée ci-dessus : l'actant sémantique d'une lexie L doit correspondre à un participant de la situation dénotée par la lexie L. Cependant, comme le souligne Mel'čuk (*ibid.* :15), le fait de dénoter un participant de la situation exprimée par L est une condition nécessaire mais pas suffisante : « *Corresponding to a participant of SIT (L) is thus a necessary, but not sufficient condition for a semanteme* $^{(}\delta^{)}$ *to be a SemA of* $^{(}L^{)}$ *:* $^{(}\delta^{)}$ *must also BE EXPRESSIBLE IN THE TEXT IN A PARTICULAR WAY.* » En effet, un sémantème qui occupe la position actancielle doit être exprimable dans la phrase sous le contrôle syntaxique d'une lexie à sens prédicatif. Mel'čuk et Polguère (2008 : 4) parlent alors de l'« exprimabilité » des positions actancielles. En établissant une distinction entre actants sémantiques, syntaxiques profonds et syntaxiques de surface, Mel'čuk (*ibid.* : 5) a attiré l'attention sur le fait que la nature prédicative d'un sens ne doit pas être présentée uniquement dans une perspective sémantique, mais doit aussi être considérée sous l'angle des

structures syntaxiques contrôlées par les lexies. Comme le précise Polguère, bien qu'ancrée dans le signifié, la prédicativité relève, dans les faits, de l'interface sémantique-syntaxe : « *Dire qu'une lexie est prédicative, c'est dire que son signifié possède certaines propriétés qui trouvent leur manifestations dans la combinatoire lexicale et grammaticale de la lexie*. » (Polguère 2012 : 3). Comme le soulignent les auteurs (Mel'čuk et Polguère *ibid.*), l'« exprimabilité » des positions actancielles se manifeste de la manière la plus évidente dans le cas des prédicatifs verbaux où les actants apparaissent comme des dépendants syntaxiques directs des lexies sémantiquement liantes. Quant aux noms, il est nécessaire de considérer d'autres types de structures, notamment les collocations verbales. Nous proposons de revenir à nos exemples. Comme nous l'avons vu plus haut, le terme *consentement préalable* : ⁽consentement préalable de X donnée à Y pour une action Z⁾ est un syntagme nominal à sens prédicatif. L'analyse des contextes extraits du *DITerm* a révélé que le terme en question constitue la base des collocations à verbe support telles que : *donner son consentement préalable* ou *exprimer son consentement préalable*. En effet, les verbes supports jouent ici un rôle fondamental car ils permettent de connecter syntaxiquement l'expression des actants auprès du terme qui les gouverne dans la structure sémantique. Il est nécessaire de souligner que ces verbes collocatifs n'ont pas d'actants sémantiques propres ; ils ne font que réaliser les actants sémantiques du terme *consentement préalable*.

Le terme *obligation générale de surveillance* constitue un autre cas de figure. En effet, il s'agit d'une unité terminologique à sens compositionnel contenant deux prédicats : ⁽obligation⁾ et ⁽surveillance⁾. Le premier contrôle 3 positions actancielles : ⁽l'obligation imposée par X à Y de faire Z⁾. Comme nous avons pu le constater en analysant les cooccurrences du terme en question, les collocatifs (verbes supports ou de réalisation), dans les expressions suivantes : *imposer une obligation générale de surveillance, mettre à la charge de* Y *une obligation générale de surveillance, être tenu à une obligation générale de surveillance, être soumis à une obligation générale de surveillance, avoir une obligation générale de surveillance* permettent de connecter syntaxiquement les actants au sémantème qui les gouverne, à savoir ⁽obligation⁾. De plus, en ce qui concerne la deuxième position actancielle de ⁽obligation⁾, elle est remplie par le sémantème ⁽surveillance⁾ qui, quant à lui, contrôle deux 2 positions actancielles : ⁽surveillance de Y par X⁾. Nous remarquons que le premier actant du sémantème ⁽surveillance⁾ correspond au premier actant de ⁽obligation⁾. Par conséquent, dans la structure actancielle du terme *obligation générale de surveillance*

(l'obligation de surveillance de Z imposée par X à Y), le premier et le deuxième actant sont « hérités » du prédicatif (obligation) tandis que le troisième actant est contrôlé par la composante (surveillance).

L'identification de la structure actancielle paraît encore plus compliquée, si l'on prend le cas des quasi-prédicats. Rappelons que tout comme les noms sémantiques, les quasi-prédicats dénotent des entités et s'expriment par des lexies nominales. En revanche, pareillement aux prédicats, ils ne peuvent être modélisés sans tenir compte des positions actancielles, qu'ils contrôlent. Prenons le cas du terme *hébergeur* : (*hébergeur* X de Y sur W pour Z). En réalité, les actants de ce quasi-prédicat sont les actants d'un *prédicat inhérent*, un prédicat véritable « interne », qui apparaît au sein du sens sans en être le composant central (voir Mel'čuk et Polguère 2008 : 6). Ici, le prédicat inhérent, enchâssé dans le sens quasi-prédicatif correspond à l'activité de l'individu X, notamment à (hébergement). Il en va de même pour les termes à sens quasi-prédicatif dénotant des artéfacts tels que *nom de domaine* : (nom de domaine X de Y utilisé par Z) ou *donnée à caractère personnel* : (donnée à caractère personnel X sur Y utilisée par Z). Dans les deux cas, les prédicats inhérents correspondent à l'utilisation des artéfacts en question. De ce fait, ces derniers contrôlent nécessairement des positions actancielles correspondant aux agents utilisateurs ainsi que celles pour les entités auxquelles les artéfacts sont appliqués. Remarquons entre parenthèses que la prédicativité est un phénomène graduel : certaines lexies ont un caractère prédicatif plus marqué que d'autres (voir à ce sujet Mel'čuk et Polguère 2008).

Ainsi, pour résumer ce qui vient d'être dit, la propriété de liage (autrement dit, la nature prédicative des unités lexicales et, dans notre cas, terminologiques) s'appuie sur la capacité *vs* l'incapacité du sémantème (S) de contrôler une position actancielle, c'est-à-dire une position destinée à être occupée par un sémantème appelé *actant* (Mel'čuk 38 : 2004). Pour considérer un sémantème comme un actant, il doit remplir deux conditions : il doit, d'une part correspondre à un participant de la situation dénoté par (S) et, d'autre part, il doit pouvoir s'exprimer auprès de l'expression lexicale de (S) dans la phrase (Polguère 2012 : 5). Comme nous l'avons vu plus haut, le lexique doit offrir des moyens d'expression privilégiée de l'actant auprès de son prédicat (les structures syntaxiques régies dans le cas des prédicats verbaux, les collocations contrôlées par la lexie liante si cette dernière est un nom ou bien les FL paradigmatiques comme indice de l'existence d'une position actancielle chez les quasi-

prédicats). En effet, la Théorie Sens-Texte insiste beaucoup sur la facette linguistique (lexicale et syntaxique) de la notion de prédicativité. Selon Mel' čuk (2004 : 10), une situation linguistique est une situation telle qu'elle est reflétée par la langue et exprimée dans les emplois de l'unité à sens prédicatif et non pas une représentation schématique d'une situation réelle :

« *What is meant here is by no means a real-life situation, that is, NOT a state of affairs in the universe ; it is rather a situation strictly as it is portrayed by the language, that is, by the LU L, and reflected in possible uses of L.* »

De plus, l'auteur nous sensibilise au fait que les participants de la situation linguistique ne renvoient pas automatiquement aux positions actancielles :

« *SemA-slots will be defined as corresponding to participants of the SIT(L) ; in other words, a SemA-slot in the definition of L necessarily corresponds to a participant of the SIT(L), while the inverse is not true […].* »

(*ibid.*).

Les participants d'une situation linguistique peuvent être obligatoires ou optionnels. Ces derniers donnent lieu aux positions actancielles optionnelles et ne doivent pas être confondus avec les circonstants. Quant aux participants obligatoires, ils correspondent soit aux constantes soit aux variables actancielles. Étant donné que les constantes actancielles qui constituent un élément fixe de la définition lexicographique ne s'expriment que rarement dans les phrases, ce sont plutôt les participants obligatoires variables qui correspondent aux positions actancielles.

Comme nous le voyons ici, l'identification de la structure actancielle est une tâche délicate. Il n'est pas toujours évident de faire la distinction entre les participants obligatoires, optionnels et circonstants et de prendre une décision adéquate quant au nombre d'actants devant apparaître dans la forme propositionnelle du terme vedette (c'est-à-dire dans la représentation de sa structure actancielle). Ceci nous paraît encore plus difficile dans le cas d'un projet terminologique où on est constamment tenté de mettre en évidence d'autres types de relations, notamment des relations de nature conceptuelle. Revenons à l'exemple de *nom de domaine*. Ainsi, le terme en question dénote le concept qui implique plusieurs participants : le site Internet et son adresse IP auxquels correspond un nom de domaine donné, le titulaire

de ce dernier, le bureau d'enregistrement qui gère la réservation des noms de domaine, le serveur DNS qui met en correspondance des adresses IP avec les noms de domaine et finalement, le cybersquatteur qui utilise un nom de domaine illégalement. La formule permettant de représenter cette situation se résumerait à ceci :

Un nom de domaine (X), associé à une adresse IP (Y) d'un site Internet (Z) par l'intermédiaire d'un serveur DNS (W), est choisi par l'utilisateur (V). Pour que le nom de domaine fonctionne, l'utilisateur (V) le fait enregistrer auprès d'un bureau d'enregistrement (U) en devenant ainsi son titulaire (V). Un cybersquatteur (T) est un utilisateur qui utilise un nom de domaine (X) de façon illégale.

Bien évidement, nous présentons ici les participants de la situation réelle. Soulignons que cette dernière devrait trouver son reflet dans une définition encyclopédique du terme. Quant à la situation linguistique dénotée, elle nous permet d'identifier trois actants qui sont exprimés dans la phrase de la façon suivante : X est un actant « incorporé » qui correspond à un nom de domaine, Y renvoie à une ressource Internet et Z englobe les différents acteurs engagés dans le processus de l'exploitation du nom de domaine : titulaire, bureau d'enregistrement, DNS, cybersquatteur, internaute, ce qui correspond à (nom de domaine X de Y utilisé par Z). La représentation de la situation linguistique au moyen de la forme propositionnelle ne rend pas compte de la dimension conceptuelle de ses éléments. En effet, du point de vue de l'analyse linguistique, les différents intervenants qui ne remplissent pas le rôle des actants doivent être considérés comme des participants extérieurs.

Comme nous le voyons, la structure actancielle des termes à sens prédicatif (décrite au moyen d'une forme propositionnelle) ne peut pas être considérée comme une représentation conceptuelle. Certes, une situation linguistique rend en partie compte des états du monde réel. Il est aussi possible de remplacer les variables par des étiquettes sémantiques[75] qui disent plus que ces premières sur la nature des actants (nous comptons par ailleurs explorer cette piste dans de futurs travaux). Cependant, la structure actancielle définie à l'aide de critères linguistiques stricts (Mel'čuk 2004) ne permet pas de mettre en évidence les liens conceptuels. Nous considérons tout de même que l'analyse et la caractérisation des

[75] Nous avons vu précédemment que les actants peuvent être représentés soit au moyen des variables *X, Y, Z* comme c'est le cas pour le DiCo ou le LEC, soit à l'aide d'un système d'étiquettes dont l'objectif est de typer très généralement le rôle des actants par rapport au terme décrit (*DiCoInfo*, DicoEnviro).

participants des situations linguistiques et l'identification de la nature prédicative des termes constituent un point d'entrée au schéma conceptuel du domaine. En effet, la mise en évidence des relations actancielles existant entre les termes et les actants permet de rendre compte de la structure lexicale du droit de l'Internet (aussi bien sur l'axe paradigmatique que syntagmatique) et cette dernière constitue en quelque sorte le reflet du schéma conceptuel du domaine. De plus, l'identification de la structure actancielle est incontournable si l'on veut proposer une description basée sur les FL.

8.2 La modélisation des données extraites du corpus au moyen des FL

Dans les pages qui suivent nous proposerons de représenter les diverses relations observées dans le corpus *DITerm* au moyen des fonctions lexicales. Nous nous intéresserons plus particulièrement aux relations paradigmatiques ainsi qu'aux collocations verbales. Nous nous arrêterons également sur la modélisation des liens actanciels et circonstanciels. Le recours aux FL nous amènera à aborder la question de la compatibilité de ce modèle avec notre projet terminographique.

8.2.1 Encodage des relations paradigmatiques

Commençons par les FL paradigmatiques car elles représentent le groupe de relations lexicales qui ont retenu la plus grande attention en terminologie (voir L'Homme 2002, 2007, Dancette et L'Homme 2002, 2004, Mortchev-Bouveret, 2006). Rappelons que parmi ces dernières les relations taxinomiques, c'est-à-dire les relations d'hyponymie, d'hyperonymie et de co-hyponymie sont traditionnellement considérées comme étant fondamentales dans la description terminologique. Selon L'Homme (2007 : 194), il est possible de les modéliser à l'aide des fonctions lexicales **Syn$_{\supset}$**, **Syn$_{\subset}$**, **Syn$_{\cap}$** qui indiquent respectivement trois types de synonymes : le sens $^{(}$A$^{)}$ est plus spécifique que le sens $^{(}$B$^{)}$ (quand le $^{(}$A$^{)}$ inclut tous les sèmes de $^{(}$B$^{)}$ et au moins un autre) ; le sens $^{(}$A$^{)}$ est moins spécifique que le sens $^{(}$B$^{)}$ (quand le $^{(}$B$^{)}$ inclut tous les sèmes de $^{(}$A$^{)}$ et au moins un autre), les sens A$^{)}$ et $^{(}$B$^{)}$ ont une intersection non vide (quand les sens $^{(}$A$^{)}$ et $^{(}$B$^{)}$ renferment des sèmes communs ainsi que des sèmes différents)[76]. De plus, en ce qui concerne la co-hyponymie, habituellement codée au moyen de la fonction **Syn$_{\cap}$**, il faut savoir qu'il existe d'autres moyens pour capturer ce type de relation.

[76] Voir Mel'čuk *et al.* (1995 : 129-130).

Ainsi, dans le cas où les co-hyponymes ⁽A⁾ et ⁽B⁾ se distinguent par la négation, c'est-à-dire par la mise en opposition d'un de leurs composants (Polguère 2008 : 152), la relation qui les unit peut être notée par la fonction lexicale **Anti.** En outre, pour affiner les relations au sein des séries de co-hyponymes, certains auteurs (Mortchev-Bouveret 2006, Dancette et L'Homme 2004) ont recours à la fonction lexicale **Contr** qui décrit la situation où ⁽A⁾ et ⁽B⁾ ont le même hyperonyme immédiat mais contrastent sur un trait ou plusieurs traits. Cependant, il est nécessaire de souligner que l'utilisation de ces différentes fonctions lexicales demande une analyse rigoureuse en termes de sèmes : la LEC s'intéresse aux relations lexicales fondées sur l'héritage de propriétés sémantiques. Or, dans le cadre d'un projet terminologique (tel que le nôtre), on est obligé de prendre également en compte les propriétés référentielles des unités qui renvoient à des concepts situables en dehors de la langue. Rappelons que d'après Kleiber et Tamba (cités par Mortchev-Bouveret, 2006 : 237), les relations d'hyperonymie et d'hyponymie sont issues de la logique et caractérisent une relation mixte d'inclusion, à la fois référentielle et linguistique.

Ainsi, tout comme Mortchev-Bouveret (2006) et Dancette et L'Homme (2002, 2004), nous considérons que les fonctions **Gener** et **Spec** (cette dernière a été proposée par Grimes 1990 : 358) sont plus appropriées à la description des unités terminologiques que les FL **Syn⊃**, **Syn⊂**, car elles permettent de hiérarchiser les taxinomies et de représenter les relations d'hyperonymie et d'hyponymie sans répondre strictement à l'analyse sémique. Comme nous pouvons le constater en analysant le Tableau 34, le terme *donnée à caractère personnel* renvoie à toute une série d'hyponymes : *donnée directement identifiante, donnée indirectement identifiante, donnée sensible, donnée biométrique, donnée de connexion, adresse IP, donnée de navigation, donnée de localisation.* S'il est possible d'utiliser la fonction **Syn⊃** pour décrire les relations telles que :

Syn⊃ (*donné à caractère personnel*) = *donnée indirectement identifiante, donnée directement identifiante* ;

il nous paraît méthodologiquement injustifié de représenter, à l'aide de la même la fonction, la relation d'hyponymie existant entre *donnée à caractère personnel* et les autres termes de la série :

Syn⊃ (*donné à caractère personnel*) ≠ *adresse IP, donnée de connexion, donnée relative au trafic, donnée sensible, donnée biométrique, donnée de navigation, donnée comportementale, donnée de localisation*

Tout d'abord, il est important de souligner que les hyponymes du terme *données à caractère personnel* n'appartiennent pas strictement à un même niveau de la taxinomie hiérarchisée : *adresse IP* est une *donnée à caractère personnel* qui appartient à la classe des *données de connexion*, considérées, quant à elles, comme des *données indirectement identifiantes*. De plus, le lien qui existe entre ces termes est déterminé plutôt par rapport à la réalité juridique qu'à la réalité linguistique. Les termes évoqués ci-dessus représentent des concepts qui relèvent du même régime juridique, mais ne partagent pas forcément de parenté sémantique (dans la plupart des cas, ils présentent un degré de parenté sémantique très faible). Leur appartenance à la catégorie des *données à caractère personnel* ne repose pas sur des traits linguistiques mais résulte d'un texte législatif ou d'une décision de justice. Ainsi, quand on sort du cadre juridique, les relations qui relient les termes en question changent leur caractère : d*onnée à caractère personnel, donnée relative au trafic ou donnée comportementale* peuvent être considérés comme co-hyponymes, puisqu'étant dans un rapport d'inclusion par rapport à un terme très général : *donnée*.

Tout cela prouve qu'il est difficile, en terminologie, d'hériter de propriétés strictement sur la base d'une description lexicographique en sèmes, composants sémantiques. Ainsi, comme le souligne Mortchev-Bouveret (2006 : 308), l'utilisation de la fonction **Spec** au lieu **Syn⊃** paraît plus souple car les hyponymes notés **Spec** d'une classe n'héritent pas nécessairement de toutes les propriétés de l'hyperonyme **Gener** de la classe. Il en va de même pour les relations d'hyperonymie :

Gener (*fourniture d'accès Internet*) = *service de la société de l'information*

Gener (*hébergement*) = *service de la société de l'information*

Gener (*référencement*) = *service de la société de l'information*

Gener (*vidéo à la demande*) = *service de la société de l'information*

Gener (*service de musique en ligne*) = *service de la société de l'information*

Nous préférons les capturer au moyen de la fonction **Gener** et non pas **Syn⊂** car cela nous permet de nous référer aussi bien à la réalité linguistique qu'extralinguistique, quoique le recours à cette dernière pose parfois de nombreux problèmes de classification liés à l'ambiguïté de la nature juridique de certains faits et acteurs de l'Internet. C'est notamment le

cas des termes : *service de la société de l'information* et *service de communication au public en ligne*. Le premier, propre à la législation européenne, est défini à l'article 1er, paragraphe 2, de la directive 98/34/CE, modifiée par la directive 98/48/CE du Parlement européen et du Conseil du 20 juillet 1998 comme : « *tout service presté normalement contre rémunération, à distance par voie électronique et à la demande individuelle d'un destinataire de services* ». [ECOMMERCE]. Comme le souligne le législateur communautaire dans la directive 2000/31/CE, il s'agit d'une notion qui englobe *« un large éventail d'activités économiques qui ont lieu en ligne. […] Les services de la société de l'information ne se limitent pas exclusivement aux services donnant lieu à la conclusion de contrats en ligne, mais, […] ils s'étendent à des services qui ne sont pas rémunérés par ceux qui les reçoivent, tels que les services qui fournissent des informations en ligne ou des communications commerciales, ou ceux qui fournissent des outils permettant la recherche, l'accès et la récupération des données. Les services de la société de l'information comportent également des services qui consistent à transmettre des informations par le biais d'un réseau de communication, à fournir un accès à un réseau de communication ou à héberger des informations fournies par un destinataire de services.* » [ECOMMERCE] La notion de *service de la société de l'information* apparaît donc comme une catégorie commune à des activités très variées et l'analyse du corpus *DITerm* nous permet de relever un grand nombre de termes, d'un degré de parenté sémantique très faible, correspondant à cette définition (voir ci-dessus).

Quant à *service de communication au public en ligne*, c'est une notion introduite en droit français suite à la transposition des directives citées plus haut. Ainsi, *service de communication au public en ligne* est censé correspondre au terme communautaire *service de la société de l'information*. Cependant, il s'avère que la définition française regroupe uniquement des activités consistant en *« transmission, sur demande individuelle, de données numériques n'ayant pas un caractère de correspondance privée, par un procédé de communication électronique permettant un échange réciproque d'informations entre l'émetteur et le récepteur* » [LOI.FR]. Par conséquent, le terme français ne peut pas être considéré comme une simple variante du terme *service de la société de l'information* venant du droit communautaire. De plus, avec l'arrivée des services de communication au public en ligne dits du web 2.0 qui proposent à la fois des services de stockage et de transmission, il existe un flou autour de la qualification juridique de ces derniers (le législateur français envisage la création d'un statut hybride à mi-chemin entre l'hébergeur et l'éditeur de service

de communication au public en ligne). Le terminologue soucieux de décrire les relations existant entre les termes du domaine étudié ne peut pas faire abstraction de ces éléments extralinguistiques.

Ainsi, à l'instar de Mortchev-Bouveret (*ibid.*), nous avons décidé de proposer une représentation médiane entre relations taxinomiques et composantes linguistiques. Les fonctions **Gener** et **Spec** ont été réservées aux liens hiérarchisants déterminés en termes de référents, tandis que les relations au niveau des co-hyponymes **Syn, Anti, Conv,** (synonymes absolus, antonymes, conversifs) ont été définies en termes de composantes sémantiques. Ainsi, le terme *contenu illicite* (contenu X de Y non autorisé par la loi que Z utilise sur Internet) est l'antonyme de *contenu légal* (contenu X de Y autorisé par la loi que Z utilise sur Internet). Quant à la relation de conversivité, elle peut être illustrée par la paire des termes : *téléchargement :* (téléchargement de X par Y à partir de Z vers W) et *mise à disposition :* (mise à disposition de X par Y sur Z à partir de W)**,** qui dénotent la même situation mais s'expriment dans la phrase avec une inversion de l'ordre de leurs actants. En ce qui concerne d'autres cas de co-hyponymes, nous avons choisi de les coder à l'aide de la fonction **Contr** et non pas **Syn**$_{\cap}$. Cependant, nous nous posons la question de la pertinence d'une telle démarche. En effet, comme nous l'avons vu plus haut, certains co-hyponymes n'ont pas le même hyperonyme immédiat. De plus, ils contrastent sur des traits analysés en termes de référents et non pas en termes de composantes sémantiques. C'est notamment le cas de *donnée à caractère personnel* (analysé plus haut) ainsi que celui de la paire des termes : *communication au public en ligne* et *communication privée*. Ces derniers, regroupés sous un terme commun : *communication électronique* ne se trouvent pas au même niveau de la taxinomie hiérarchisée. En effet, le terme *communication au public en ligne* est un hyponyme direct d'un autre terme, notamment *communication au public par voie électronique* défini comme : *« toute mise à disposition du public ou de catégories de public, par un procédé de communication électronique, de signes, de signaux, d'écrits, d'images, de sons ou de messages de toute nature qui n'ont pas le caractère d'une correspondance privée »* [ECOMMMERCE], qui renvoie aussi bien à une *communication au public en ligne* qu'à une communication audiovisuelle. Néanmoins, il nous paraît pertinent de les rapprocher en décrivant leur relation comme étant une co-hyponymie indirecte.

donnée à caractère personnel 3913 occurrences **STRUCTURE ACTANCIELLE : donnée à caractère personnel X sur Y utilisée par Z**		
FONCTION LEXICALE	**VALEUR**	**TYPE DE RELATION**
Syn (donnée à caractère personnel)	*donnée personnelle, donnée relative aux personnes physiques*	**Synonymie**
Anti (*donnée à caractère personnel*)	*donnée anonyme, donnée anonymisée*	**Antonyme**
Gener (*donnée à caractère personnel*)	*donnée, information*	**Hyperonymie**
Spec* (*donnée à caractère personnel*) *FL proposée par Grimes	*donnée sensible, donnée biométrique, adresse IP, donnée de connexion, donnée comportementale, donnée de navigation, donnée de localisation*	**Hyponymie**
Mult (*donnée à caractère personnel*)	*base de données, fichier de données*	**Méronymie**
contenu illicite 638 occurrences **STRUCTURE ACTANCIELLE : *contenu illicite* X de Y utilisé par Z**		
FONCTION LEXICALE	**VALEUR**	**TYPE DE RELATION**
Syn (*contenu illicite*)	*contenu illégal, contenu préjudiciable*	**Synonymie**
Anti (*contenu illicite*)	*contenu légal, contenu autorisé, contenu licite*	**Antonyme**
Gener (*contenu illicite*)	*contenu, contenu numérique, contenu mis en ligne, donnée numérique, information*	**Hyperonymie**
Spec (*contenu illicite*) *FL proposée par Grimes	*contenu à caractère pornographique, contenu à caractère pédophile, contenu raciste, contenu à caractère violent, contenu contrefaisant, contenu diffamatoire*	**Hyponymie**

téléchargement 1282 occurrences **STRUCTURE ACTANCIELLE : *téléchargement* réalisé par X de Y à partir de Z vers W**		
FONCTION LEXICALE	**VALEUR**	**TYPE DE RELATION**
Syn (*téléchargement*)	*téléchargement descendant, download*	**Synonymie**
Conv (*téléchargement*)	*mise à disposition, upload*	**Conversivité**
Contr (*téléchargement*)	*streaming, lecture en flux continu,*	**Co-hyponymie**
Gener (*téléchargement*)	*service, technique*	**Hyperonymie**
Spec* (*téléchargement*) *FL proposée par Grimes	*téléchargement légal, téléchargement illégal*	**Hyponymie**
communication au public en ligne 800 occurrences **STRUCTURE ACTANCIELLE : *communication au public en ligne* de Y réalisé par X**		
FONCTION LEXICALE	**VALEUR**	**TYPE DE RELATION**
Syn (*communication au public en ligne*)	*acte de communication au public en ligne*	**Synonymie**
Contr (*communication au public en ligne*)	*correspondance privée, communication audiovisuelle*	**Co-hyponymie**
Gener (*communication au public en ligne*)	*communication électronique, communication au public par voie électronique*	**Hyperonymie**
Spec* (*communication au public en ligne*) *FL proposée par Grimes	*communication au public en ligne d'œuvres protégés, mise à disposition du public d'œuvres protégées, mise à disposition du public non autorisée d'œuvres ou d'objets protégés,*	**Hyponymie**
service de communication au public en ligne 733 occurrences **STRUCTURE ACTANCIELLE : *service de communication au public en ligne* X proposé par Y à Z**		
FONCTION LEXICALE	**VALEUR**	**TYPE DE RELATION**

Spec* (*service de communication au public en ligne*) *FL proposée par Grimes	*site web, plate-forme d'échange de pair-à-pair, plateforme d'hébergement, site de téléchargement, blog, forum de discussion, site de streaming de musique, moteur de recherche, vidéo à la demande*	**Hyponymie**
Contr (*service de communication au public en ligne*)	*service de médias audiovisuels, service de médias audiovisuels à la demande, service de courrier électronique*	**Co-hyponymie**

Tableau 37. Exemples de relations paradigmatiques modélisées au moyen des FL

8.2.2 Encodage des relations actancielles

Quant aux relations actancielles, leur encodage au moyen des fonctions lexicales paraît naturel et évident à condition que la structure actancielle des termes décrits soit correctement identifiée. Rappelons que nous avons déjà abondamment traité cette problématique à la section précédente. Nous nous limiterons donc ici à la présentation de quelques exemples de modélisation de ce type de liens à l'aide des FL.

Comme nous pouvons le constater en analysant les tableaux ci-dessous, les fonctions lexicales **S_1**, **S_2**, **S_3,** … permettent de consigner de façon systématique les noms typiques d'agent, de patient, de destinataire, etc., c'est-à-dire, les noms des participants de la situation évoquée par l'unité terminologique à sens prédicatif. Cependant, nous voudrions attirer l'attention sur le fait que les dérivés sémantiques nominaux actanciels dégagés au moyen des FL **S_1**, **S_2**, **S_3,** etc. correspondent aux participants d'une situation linguistique, reflétée dans la définition propositionnelle du mot-clé et non pas aux participants d'une situation réelle. Ainsi, dans le cas de *donnée à caractère personnel* (donnée à caractère personnel X sur Y utilisée par Z), l'application des FL permet d'obtenir les valeurs correspondant à trois participants obligatoires: les données personnelles (actant incorporé), la personne que concernent ces données et la personne qui traite les données en question. Or, en réalité (nous entendons par là

la réalité juridique), la situation évoquée par le terme implique encore d'autres protagonistes. En effet, à part la *personne concernée* et le *responsable du traitement*, termes correspondant aux actants (voir le Tableau 38), le cadre juridique prévoit d'autres participants tels que: *destinataire d'un traitement de données à caractère personnel* (« *toute personne habilitée à recevoir communication de ces données autre que la personne concernée, le responsable du traitement* » [CORPUS_DONNEES]), ou *autorité de contrôle*, c'est-à-dire une autorité chargée de la protection des données à caractère personnel comme la *CNIL*. Cependant, il est nécessaire de souligner qu'en faisant référence à ces protagonistes, nous situons notre analyse à un niveau conceptuel, ce qui complique la description de ces différents rôles au moyen des fonctions lexicales. En effet, du point de vue strictement linguistique, les sémantèmes (destinataire) ou (autorité de contrôle) doivent être considérés comme des participants extérieurs.

donnée à caractère personnel 3913 occurrences

STRUCTURE ACTANCIELLE : *donnée à caractère personnel* X sur Y utilisée par Z

FONCTION LEXICALE	VALEUR	TYPE DE RELATION
$\mathbf{S_2}$ (*donnée à caractère personnel*)	*personne concernée, public concerné, utilisateur, internaute, abonné*	**Actant $_2$**
$\mathbf{S_3}$ (*donnée à caractère personnel*)	*responsable du traitement, sous-traitant, pirate informatique, fournisseur de services de communications électroniques accessibles au public, fournisseur d'hébergement*	**Actant$_3$**

téléchargement 1282 occurrences

STRUCTURE ACTANCIELLE : *téléchargement* de X réalisé par Y à partir de Z vers W

FONCTION LEXICALE	VALEUR	TYPE DE RELATION
$\mathbf{S_1}$ (*téléchargement*)	*internaute, utilisateur, abonné, pirate, personne*	**Actant$_1$**
$\mathbf{S_2}$ (*téléchargement*)	*article de la presse en ligne, œuvre protégés, œuvre musicale, fichier musical, musique, phonogramme, film, fichier, image, logiciel, copie d'un programme, logiciel de jeux, jeux vidéo, vidéo, contenu musical,*	**Actant$_2$**

	contenu audiovisuel, contenu illicite	
S_3 (*téléchargement*)	*site Internet, Internet, site de partage de vidéo, serveur, réseau de communication au public en ligne, réseau P2P, réseau peer-to-peer*	**$Actant_3$**
S_4 (*téléchargement*)	*support informatique, ordinateur, disque dur*	**$Actant_4$**
consentement préalable 114 occurrences **STRUCTURE ACTANCIELLE : consentement préalable de X donnée à Y pour une action** *Z*		
FONCTION LEXICALE	**VALEUR**	**TYPE DE RELATION**
S_1 (*consentement préalable*)	*personne concernée, public concerné, utilisateur, internaute, abonné, personne prospectée, destinataire de la communication commerciale, consommateur*	**$Actant_1$**
S_2 (*consentement préalable*)	*responsable de traitement, sous-traitant*	**$Actant_2$**
S_3 (*consentement préalable*)	*traitement des données personnelles, prospection commerciale, prospection directe, diffusion de la publicité comportementale, installation des cookies*	**$Actant_3$**
hébergeur 479 occurrences **STRUCTURE ACTANCIELLE : hébergeur X de Y sur W pour le compte de Z**		
FONCTION LEXICALE	**VALEUR**	**TYPE DE RELATION**
S_2 (*fournisseur d'hébergement*)	*blog, site Internet, vidéo, plate-forme de partage de vidéo, site de commerce en ligne, contenu illicite, contenu, données, informations, forum de discussion, fichier, page personnelle,*	**$Actant_2$**
S_3 (*fournisseur d'hébergement*)	*destinataire du service, hébergé, client, fournisseur de contenu, éditeur du contenu, internaute, utilisateur, utilisateur du service de communication au public en ligne*	**$Actant_3$**
S_4 (*fournisseur d'hébergement*)	*serveur, serveur dédié, machine, site, plate-forme*	**$Actant_4$**

Tableau 38. Exemples de relations actancielles modélisées au moyen des FL.

8.2.3 Encodage des relations syntagmatiques

Dans les pages qui suivent, nous nous pencherons sur la question de l'encodage des relations syntagmatiques observées dans le corpus *DITerm* et plus particulièrement sur la modélisation des collocations verbales. Comme le souligne Polguère (2003 : 117), la modélisation d'un phénomène linguistique tel que la collocation est une représentation qui permet d'expliquer, anticiper ou reproduire ce phénomène. L'auteur (*ibid.*) postule, par ailleurs, que c'est grâce à la notion Sens-Texte de fonction lexicale que le problème de la description des collocations a trouvé une solution adéquate et satisfaisante. Pour notre part, nous proposons de démontrer l'efficacité descriptive des FL dans un projet terminographique.

Comme nous l'avons vu dans le premier chapitre de ce travail, une grande partie des relations collocationnelles verbales peuvent être encodées au moyen d'un nombre restreint de fonctions lexicales syntagmatiques standard telles que : **Oper$_i$**, **Func$_i$**, **Labor$_{ij}$**, **Real$_i$**, **Fact$_i$** et **Labreal$_{ij}$**. Les FL en question se présentent de façon naturelle par triplets. Comme le soulignent Mel'čuk *et al.* (1995 : 138), la trinité de ces FL *« […], reflète le fait fondamental de la syntaxe des langues naturelles qui ne distinguent que TROIS rôles syntaxiques majeurs, soit trois types d'actants syntaxiques de surface : le Sujet Grammatical, le Complément d'Objet Direct et les Compléments d'Objet Indirect […].* »

Le premier triplet est formé par les FL **Oper$_i$**, **Func$_i$** et **Labor$_{ij}$** qui formalisent la notion de *verbe support*. Comme le souligne Maurice Gross (1998), la notion de verbe support[77] a été introduite, d'abord sous le nom de « verbe opérateur », par Z. S. Harris en vue de traiter syntaxiquement les relations de nominalisation. Il s'agit des verbes sémantiquement vides (ou traditionnellement considérés comme tels) qui remplissent une fonction uniquement syntaxique. En effet, les verbes supports servent à verbaliser les noms prédicatifs, c'est-à-dire à construire des phrases avec ce type d'unités à sens prédicatif. Ainsi, il est possible de mettre en relation deux phrases synonymes dont l'une contient le couple : verbe support et nom prédicatif et l'autre un verbe ordinaire, dit distributionnel (Gross M., *ibid.*). Les exemples ci-dessous, tirés du corpus *DITerm*, montrent cette relation d'équivalence. Soulignons par

[77] Il convient de rappeler le rôle de Maurice Gross (1981) dans la définition de la notion des verbes supports ainsi que de souligner l'importance des travaux menés au LADL (entre autres Gross G. 1989) concernant la description de ce phénomène.

ailleurs, qu'il est possible de décrire ce type de relation d'équivalence à l'aide des règles lexico-syntaxiques de paraphrasage (Milićević 2003, 2007).

consentir au traitement des données à caractère personnel = donner son consentement au traitement des données à caractère personnel

« Eu égard à ce qui précède, les formulaires d'inscription en ligne que les personnes physiques doivent remplir pour s'identifier et ***consentir au traitement de données les concernant*** *seront réputés satisfaire à l'exigence de consentement explicite pour autant que toutes les autres conditions soient satisfaites. Ainsi, pour ouvrir un dossier médical personnel en ligne, les patients peuvent* ***donner leur consentement au traitement*** *en communiquant leurs coordonnées et en cochant une case spécifique pour marquer leur accord.* » [CORPUS_DONNEES]

héberger un contenu = assurer l'hébergement d'un contenu

« *Deux dispositions de la LCEN doivent être envisagées à ce stade : le destinataire du service ne doit pas être sous le contrôle de* ***l'intermédiaire qui assure l'hébergement des contenus*** *(a), le service doit être proposé à un public et le message transporté ne doit pas être qualifié de correspondance privée (b).* »[RDLI_2008] *« Ce régime de responsabilité bénéficiant aux « personnes physiques ou morales qui assurent, même à titre gratuit, pour mise à disposition du public par des services de communication au public en ligne, le stockage de signaux, d'écrits, d'images, de sons ou de messages de toute nature fournis par des destinataires de ces services »(5), rien ne s'oppose, a priori, à ce qu'il trouve à s'appliquer à* ***tout prestataire qui héberge du contenu*** *en provenance d'internautes* » [RDLI_2011]

Figure 34. Phrases synonymes – règle d'équivalence

Ainsi, les fonctions lexicales $\mathbf{Oper_i}$, $\mathbf{Func_i}$ et $\mathbf{Labor_{ij}}$ permettent d'encoder différents types de construction à verbe support et ne se distinguent que syntaxiquement, selon que leur mot clé – le terme vedette – est leur Sujet, leur CO^{dir} ou bien leur CO^{indir}. La fonction $\mathbf{Oper_i}$ prend le mot clé comme son CO^{dir}, $\mathbf{Func_i}$ comme son Sujet et $\mathbf{Labor_{ij}}$ comme son CO^{indir}. Dans le Tableau 39, nous présentons les valeurs de ces fonctions obtenues pour trois mots-clés : *consentement préalable, obligation générale de surveillance, téléchargement illégal*. Nous voudrions également attirer l'attention sur le fait que l'indice actanciel de chaque FL est

déterminé par le rôle des actants syntaxiques profonds (ASynP) des mots-clés : l'indice $_1$ renvoie à l'actant SyntP I (= X) du mot-clé, l'indice $_2$ à l'actant SyntP II (= Y) et l'indice $_3$ à l'actant SyntP III (= Z), etc. (Mel'čuk *et al. ibid.* : 139). Comme nous pouvons le constater en analysant les exemples ci-dessous, l'utilisation des indices actanciels de type 12 21 permet, entre autres, d'encoder des verbes supports converses (voir Gross G. 1989).

Soulignons au passage qu'en modélisant les phénomènes de la combinatoire lexicale au moyen des fonctions lexicales, nous nous sommes inspirée des ressources développées dans le cadre de la Lexicologie Explicative et Combinatoires telles que le *LAF* (Mel'čuk et Polguère, 2007), le *DiCo*[78] (Mel'čuk et Polguère) et le *DiCoInfo* (L'Homme). Nous avons par ailleurs repris la formalisation proposée par Mel'čuk et Polguère. Ainsi, nous utilisons le *tilde* « ~ » pour référer au mot-clé et la formule « [ART ~] » indique que les collocatifs (qui correspondent aux valeurs retournées par les FL) prennent la base (mot-clé) comme complément d'objet direct, avec présence obligatoire d'un déterminant.

MOT CLÉ : consentement préalable **STRUCTURE ACTANCIELLE : consentement préalable de X donnée à Y pour une action** *Z*	
FONCTION LEXICALE SUPPORT	**VALEUR**
Oper$_{12}$ (consentement préalable)	*donner* [ART ~], *accorder* [ART ~], *exprimer* [ART ~], *manifester* [ART ~]
Essayer_de Oper$_2$ (*consentement préalable*)	*demander* [ART ~], *solliciter* [ART ~]
Oper$_{21}$ (*consentement préalable*)	*obtenir* [ART ~], *recueillir* [ART ~]
Oper$_{31}$ (*consentement préalable*)	*requérir* [ART ~]
Func$_3$ (*consentement préalable*)	*concerne* [N]
MOT CLÉ : obligation générale de surveillance **STRUCTURE ACTANCIELLE : obligation générale de surveillance de Z imposée par X à Y**	

[78]Une version compilée du *DiCo* est accessible en ligne via l'interface *DiCouèbe* consultable à l'adresse suivante http://olst.ling.umontreal.ca/dicouebe/index.php.

FONCTION LEXICALE SUPPORT	VALEUR
Func$_2$ (*obligation générale de surveillance*)	*incomber* [à N = Y], *peser* [sur N =Y]
Oper$_{12}$ (*obligation générale de surveillance*)	*imposer* [ART ~ à N=Y]
Oper$_2$ (*obligation générale de surveillance*)	*avoir* [ART ~], *être tenu* [à ART ~]
Labor$_{12}$ (*obligation générale de surveillance*)	*soumettre* [N à ART ~]
MOT CLÉ : téléchargement illégal **STRUCTURE ACTANCIELLE : téléchargement illégal par la personne X de contenu Y réalisé à partir de Z vers W**	
FONCTION LEXICALE SUPPORT	**VALEUR**
Oper$_1$ (*téléchargement illégal*)	*procéder* [à ART ~], *réaliser* [ART ~],
Oper$_1^{usual}$ (*téléchargement illégal*)	*pratiquer* [ART ~], *s'adonner* [à ART ~ *]*
Oper$_3^{usual}$ (*téléchargement illégal*)	*servir* [à ART ~], *proposer* [ART ~]
LiquFunc$_0$ (*téléchargement illégal*)	*bloquer [ART ~]*

Tableau 39. Exemples de relations collocationnelles modélisées au moyen des FL supports.

Remarquons que parmi les FL utilisées ci-dessus, il y en a trois dont l'aspect ne correspond pas exactement à la forme régulière. Il s'agit notamment des FL **Oper$_1^{usual}$, Oper$_3^{usual}$** et **Essayer_de Oper$_2$**. Selon la typologie proposée par Jousse (2010 : 86), les deux premières doivent être considérées comme des fonctions standard à caractère spécial car elles comportent un pointeur vers les composantes sémantiques de la définition du mot-clé. Comme le rappelle l'auteure (*ibid.*), certaines FL contiennent des indications sur la quantité (**quant**), la durée (**temps**), le caractère ponctuel (**actual**) ou habituel (**usual**) d'un événement dénoté par une unité lexicale et son dérivé sémantique ou collocatif. Ainsi, dans notre exemple (Tableau 39), la notation **usual** permet donc d'indiquer que l'opération de *téléchargement illégal* est quelque chose d'habituel. Il convient de souligner que les FL standard avec pointeurs vers la définition sont révélatrices du lien profond existant entre les définitions des unités lexicales (dans notre cas – terminologiques) et les relations de dérivation sémantique ou de collocation qui leur sont associées (*ibid.*). En ce qui concerne la

FL complexe **Essayer_de Oper$_2$**, elle contient un élément (**Essayer_de)** défini par Polguère (2007 : 52) comme *FL localement standard*[79]. En effet, il s'agit des relations dont le caractère universel reste à prouver mais qui peuvent être encodées d'une façon régulière dans le contexte d'une langue naturelle donnée. Ainsi, étant donné que les FL telles que **Essayer_de**[80] ou **De_nouveau** (utilisée plus loin) sont très fréquentes en français, elles seront considérées comme des FL localement standard de la langue française.

Le deuxième triplet comprend les FL **Real$_i$**, **Fact$_i$** et **Labreal$_{ij}$** qui, comme nous l'avons déjà vu dans la première partie de ce travail, « expriment *grosso modo le sens (réaliser les « « objectifs » inhérents de la chose [désignée par le mot clé]) »* (Mel'čuk *et al. ibid. :* 141). Les fonctions en question permettent donc de décrire les réalisations typiques correspondant aux mots-clés. Comme c'est le cas pour les FL supports, les FL de réalisation ne se distinguent que syntaxiquement : **Real$_i$** prend donc son mot clé en tant que CO^{dir}, **Fact$_i$** prend son mot clé en tant que Sujet et **Labreal$_{ij}$** prend son mot clé en tant que CO^{indir}. Les indices actanciels se déterminent de la même façon que pour les FL supports (voir le Tableau 40). Soulignons également qu'il est parfois nécessaire d'indiquer le degré de réalisation, c'est-à-dire l'ordre dans lequel les activités liées au mot-clé sont normalement effectuées. Les différentes étapes peuvent ainsi être signalées par les chiffres romains (voir le Tableau 42).

MOT CLÉ : contenu illicite **STRUCTURE ACTANCIELLE : contenu illicite de X communiqué par Y à Z**	
FONCTION LEXICALE	**VALEUR**
Fact$_0$ (*contenu illicite*)	*circuler* [ART ~]
Real$_1$ (*contenu illicite*)	*diffuser* [ART ~], *afficher* [ART ~]
Real$_2$ (*contenu illicite*)	*mettre à disposition* [ART ~], *fournir* [ART ~], *poster* [ART ~], *publier* [ART ~], *déposer* [ART ~],
Real$_3$ (*contenu illicite*)	*consulter* [ART ~], *visionner* [ART ~]

[79] « *A* local standard LF *of* ***Lang*** *is an LF which has been proven to fulfil the* ***Broadness*** *and* ***Diversity conditions*** *for standardness only for the specific natural language* ***Lang****. In order to distinguish them clearly from true standard LFs, local standard LFs are named with formulas based on* ***Lang*** *rather than Latin.* » (Polguère 2007 : 52).

[80] Il convient de souligner que les FL en question ont été proposées par Polguère (2007).

	télécharger [ART ~]
MOT CLÉ : hébergeur **STRUCTURE ACTANCIELLE : hébergeur X de Y sur W pour le compte de Z**	
FONCTION LEXICALE	**VALEUR**
Fact$_2$ *(hébergeur*)	*stocker [N = Y], accueillir [N = Y], héberger [N = Y], assurer l'hébergement [de N = Y], assurer le stockage [de N = Y], fournir un service d'hébergement [à N = Z],*
ContFact$_{23}$ (*hébergeur*)	*maintenir [N = Y] en ligne*
FinFact$_2$ (*hébergeur*)	*retirer [N = Y], supprimer [N = Y]*

Tableau 40. Exemples de relations collocationnelles modélisées au moyen des FL de réalisation.

Les fonctions lexicales présentées ci-dessus se combinent fréquemment avec deux autres sous-ensembles de FL, notamment les verbes phasiques et les verbes causatifs. Les premières : **Incep**, **Cont** et **Fin** ont un caractère aspectuel et expriment les trois phases différentes d'un état ou d'un évènement, à savoir le début, la continuation et la fin. L'autre sous-ensemble est formé par les FL **Caus**, **Liqu** et **Perm** qui expriment les trois types de causation d'un état ou d'un évènement (Mel'čuk 2003, 23-24). Comme nous pouvons le constater dans les exemples qui suivent (voir le Tableau 41), l'application de ces fonctions lexicales standard (qui s'expriment obligatoirement en combinaison avec d'autres FL pour former des fonctions complexes), à des noms prédicatifs ou à des quasi-prédicats permet d'obtenir des résultats très intéressants, riches et variés.

donnée à caractère personnel 3913 occurrences **STRUCTURE ACTANCIELLE : donnée à caractère personnel X sur Y utilisée par Z**		
FONCTION LEXICALE	**VALEUR**	**TYPE DE RELATION**
Real$_3$	*traiter* [ART ~], *utiliser* [ART ~], *exploiter* [ART ~],	**[Z] utiliser ~**
AntiVerReal$_3$	*violer* [ART ~]	**[Z non autorisé] utiliser ~**
NonPermAntiVerReal$_3$	*protéger* [ART ~]	**[Qqn] faire en sorte d'empêcher que [Z non autorisé] utilise ~**
IncepReal$_3$	*accéder à* [ART ~]	**[Z] commencer à**

		utiliser ~
$Fact_3$	*circuler [parmi N = Z]*	**~ est utilisée par Z**
$CausFact_3$	*communiquer* [ART ~ à N=Z] *transmettre* [ART ~ à N=Z]	**[Qqn] communiquer ~ à Z**
$CausAntiVerFact_3$	*divulguer* [ART ~ à N=Z]	**[Qqn] communiquer ~ à Z non autorisé**
$Liqu_1Func_0$	*effacer* [ART ~], *supprimer* [ART ~], *détruire* [ART ~]	**[Qqn] supprimer ~**
Non $Liqu_1Func_0$	*conserver* [ART ~]	**[Qqn] ne pas supprimer ~**
CausDegrad	*endommager* [ART ~], *altérer* [ART ~]	**[Qqn] rendre non valide ~**
$NonPermFact_0$	*bloquer* [ART ~]	**[Qqn] empêcher que ~ soit utilisée**
$Caus_1Func_0$	*collecter* [ART ~], *recueillir* [ART ~]	**[Qqn] créer ~**
$CausOper_1$'	*rectifier* [ART ~], *modifier* [ART ~]	**[Qqn] remplacer X ~ par X'**
nom de domaine 2329 occurrences **STRUCTURE ACTANCIELLE : nom de domaine X de Y utilisé par Z**		
FONCTION LEXICALE	***VALEUR***	**TYPE DE RELATION**
$Caus_3Oper_3$	*choisir* [ART ~], *créer* [ART ~], *déposer* [ART ~], *réserver* [ART ~], *enregistrer* [ART ~] *acquérir* [ART ~] *acheter* [ART ~]	**[Qqn] se causer d'avoir ~**
$Caus_3Oper_3$'	*céder* [ART ~ à N=Z']	
$Oper_3$	*posséder* [ART ~],	**avoir ~**
$ContReal_3$	*garder* [ART ~], *conserver* [ART ~],	**[Z] continue à utiliser ~**
$Real_3$	*exploiter* [ART ~], *utiliser* [ART ~],	**[Z] utiliser ~**
$FinFunct_0$	*Expirer*	**[X] cesser d'exister**
$Liqu_1Func_0$	*supprimer* [ART ~] *effacer* [ART ~],	**[Qqn] supprimer ~**
$CausNonAble_1Fact_0$	*bloquer* [ART ~], *geler* [ART ~],	**[Qqn] empêcher que ~ soit utilisée**
$AntiVerReal_3$	*pirater* [ART ~], *usurper* [ART ~],	**[Z] non autorisé utilise ~**

Tableau 41. Collocations verbales des termes *données à caractère personnel* et *nom de domaine* encodées au moyen des FL.

Cependant, il nous paraît nécessaire d'attirer l'attention sur quelques difficultés liées à l'encodage des relations collocationnelles des unités terminologiques. Tout d'abord, il convient de souligner que le choix d'une fonction adéquate ainsi que l'identification et la numérotation des indices actanciels ne sont pas évidents. En effet, ceci est dû au fait qu'il n'est pas toujours facile d'identifier la structure actancielle d'un nom (or la plupart des termes extrait du corpus *DITerm* sont considérés soit comme des noms prédicatifs soit comme des quasi-prédicats). Comme le remarquent Mel'čuk *et al.* (*ibid.* : 139), notre intuition linguistique privilégie avant tout l'analyse actancielle des verbes. Ainsi, afin de dégager les ASynP des mots-clés, nous avons été amenée soit à trouver des équivalents verbaux des noms prédicatifs, soit à identifier la situation impliquant les termes qualifiés de quasi-prédicats.

Deuxièmement, l'appareil des FL standard simples ne semble pas suffisant pour décrire toute la richesse des relations que les termes du domaine du droit de l'Internet entretiennent avec leurs collocatifs verbaux. En effet, pour rendre compte de certaines relations, nous avons été obligée d'avoir recours à des fonctions complexes[81] qui se présentent sous forme de chaînes de symboles de FL simples constituantes. Ceci amène parfois à des formules alambiquées et difficilement déchiffrables. Or l'objectif de notre recherche est de proposer un modèle de description clair et basé sur l'intuition. Signalons également que le recours à des FL standard complexes s'est avéré, dans certains cas, insuffisant. Il s'agit notamment des relations marquées techniquement et spécifiques au domaine du droit de l'Internet comme : *transférer des données à caractère personnel, archiver des données à caractère personnel, enregistrer des données à caractère personnel, sauvegarder des données à caractère personnel*. La solution pour rendre compte de ces nuances sémantiques serait de chercher à les décrire au moyen d'autres variétés de FL comme des FL semi-standard ou non-standard.

Nous souhaiterions également attirer l'attention sur le fait que dans certains cas, l'application des FL à des termes du domaine du droit de l'Internet ne retourne pas une seule valeur mais toute une série de valeurs. Ainsi, la fonction lexicale **$\mathbf{Caus_3Oper_3}$** appliquée au mot-clé *nom de domaine* permet de dégager jusqu'à 7 collocatifs : *choisir, créer, déposer, réserver, enregistrer, acquérir*. Or, il s'agit là de verbes sémantiquement pleins qui devraient

[81] Rappelons qu'on appelle *fonction lexicale complexe* : « *un enchaînement de FL simples syntaxiquement liées, cet enchaînement ayant une valeur globale qui exprime, de façon indécomposable, le sens de l'enchaînement entier* » (Mel'čuk 2003 : 27).

posséder leur propre description, surtout dans le cadre d'un projet terminographique qui se fixe comme objectif de proposer des distinctions sémantiques fines. Ce problème ne concerne pas uniquement les fonctions syntagmatiques mais aussi, comme nous l'avons pu constater plus haut, les fonctions paradigmatiques et actancielles. Prenons l'exemple du terme *consentement préalable* pour lequel nous avons obtenu une série des dérivés actanciels qui se distinguent entre eux par de nombreux traits sémantiques. Là aussi, la solution serait de recourir à des fonctions semi-standard ou non-standard.

MOT CLÉ : téléchargement illégal **STRUCTURE ACTANCIELLE : téléchargement illégal par la personne X de contenu Y réalisé à partir de Z vers W**	
FONCTION LEXICALE SUPPORT	**VALEUR**
Oper$_1$ (*téléchargement illégal*)	*procéder* [à ART ~], *réaliser* [ART ~],
Oper$_1^{usual}$ (*téléchargement illégal*)	*pratiquer* [ART ~], *s'adonner* [à ART ~ *]*
Oper$_3^{usual}$ (*téléchargement illégal*)	*servir* [à ART ~], *proposer* [ART ~]
LiquFunc$_0$ (*téléchargement illégal*)	*bloquer [ART ~]*
NonPermFunc$_0$ (téléchargement illégal)*[82]	*empêcher* [ART ~], *prévenir* [ART ~]
Essayer de NonPermFunc$_0$ (*téléchargement illégal*)	*lutter* [contre à ART ~], *agir* [contre ART ~], *s'attaquer* [à ART ~]
Real@[83] I (*téléchargement illégal*))*	*surveiller* [ART ~], *traquer* [ART ~]
Real@ II (*téléchargement illégal*)	*détecter* [ART ~]
Real@ III (*téléchargement illégal*)	*sanctionner* [ART ~]

Tableau 42. Relations syntagmatiques du terme *téléchargement illégal* encodées au moyen des FL.

Finalement, nous tenons à aborder une question essentielle. Il s'agit d'une difficulté méthodologique majeure rencontrée lors de l'encodage des relations collocationnelles dans le domaine du droit de l'Internet. En effet, en proposant la modélisation de liens syntagmatiques au moyen des FL, nous avons constaté qu'un certain nombre de syntagmes extraits du corpus *DITerm* sont de haute fréquence non pour des raisons linguistiques, mais conceptuelles. Les

[82] * indique qu'il s'agit plutôt d'une relation d'ordre conceptuel.

[83] @ renvoie à un participant non actanciel (un participant extérieur).

liens ainsi repérés ne relèvent pas de la phraséologie mais de la combinatoire libre de forte compatibilité conceptuelle. Se pose alors la question de savoir s'il est possible, dans tels cas, d'avoir recours aux FL ? Reprenons l'exemple de *téléchargement illégal* déjà présenté dans le Tableau 39 (voir le Tableau 42 ci-dessus).

Rappelons que les FL sont révélatrices du lien profond qui existe entre les définitions sémantiques des unités lexicales (terminologiques dans notre cas) et les relations de collocation qui leur sont associées. Nous admettons donc que pour encoder certaines relations syntagmatiques relevant plutôt de l'ordre concetpuel que linguistique, il serait possible de s'apppuyer sur des descriptions des termes conçues délibérément afin d'expliciter des liens existant entre les deux éléments. Ainsi, si ⁽téléchargement illégale⁾ est défini comme une activité illégale en ligne est les verbes *lutter, agir, empêcher, dissuader, traquer, détecter, sanctionner, disuader* comme des actions menées contre les activités illégales en ligne, il est éventuellemnt possible d'encoder ces liens concetpuels forts au moyen des FL. Cependant, une telle démarche pourrait conduire à dénaturer le système des FL, qui exige une distinction nette entre les phénomènes tels que combinatoire libre et combinatoire restreinte, liens de dérivations sémantiques et liens conceptuels forts. Pour illustrer ce qui vient d'être dit, prenons l'exemple du terme *responsable du traitement*.

En effet, alors que la modélisation des collocatifs verbaux de réalisation liés aux quasi-prédicats dénotant des artéfacts ne pose pas de difficultés particulières :

Real$_3$ (*nom de domaine*) = *exploiter* [ART ~], *utiliser* [ART ~]
Real$_2$ (*spam*) = *envoyer* [ART ~]
Real$_3$ (*spam*) = *recevoir* [ART ~]
Real$_2$ (*contenu illicite*) = *publier* [ART ~]
Real$_2$ (*contenu illicite*) = *consulter* [ART ~]
Real$_3$ (*donnée à caractère personnel*) = *traiter* [ART ~]*, exploiter* [ART ~] ;

l'identification des verbes renvoyant à des actions typiques dans la situation où le mot-clé correspond à un agent (un acteur de l'Internet), s'avère plus problématique. Ceci est dû au fait que le rôle de différents acteurs de l'Internet se définit, en grande partie, en fonction d'éléments extralinguistiques (par exemple, par rapport à leur statut juridique). On rencontre, bien évidement, des cas évidents comme :

Fact$_2$ (*hébergeur*) = *héberger* [ART ~], *stocker* [ART ~]
Fact$_2$ (*internaute*) = *consulter* [ART ~]*, visiter* [ART ~]
Fact$_3$ (*internaute*) = *naviguer* [sur ART ~]
Fact$_2$ (*titulaire d'accès à Internet*) = *utiliser* [ART ~]

Cependant, dans d'autres situations, il est beaucoup plus difficile d'établir ce type de relation. Ainsi, on peut se demander quels sont « les objectifs inhérents » à la fonction de *responsable du traitement* (personne X responsable du traitement des données à caractère personnel Y). Est-ce qu'il *veille à la sécurité de…, garantit la confidentialité de…, assure la protection de…, fait respecter les obligations relatives aux…, organise, gère*… les données à caractère personnel? L'examen du corpus *DITerm* a permis de recenser un nombre important de cooccurrents verbaux dénotant des actions typiques réalisées par le responsable du traitement. Nous proposons (voir le Tableau 43), de les organiser en utilisant la FL **Fact** et en faisant appel aux chiffres romains afin d'indiquer le degré de réalisation des objectifs liés à ce mot-clé. Nous avouons toute de même que cette description ne découle pas de l'analyse sémantique du terme (et plus précisément de sa définition propositionnelle) mais s'appuie sur des éléments extralinguistiques, ce qui est contraire aux principes de la LEC. En effet, les actions recensées et encodées ici à l'aide de la FL **Fact** ainsi que leur ordre correspondent à la réalité juridique, c'est-à-dire aux obligations imposées par la loi. L'utilisation des FL y est donc méthodologiquement infondée* car sert à décrire les liens purement conceptuels.

MOT CLÉ : responsable du traitement **STRUCTURE ACTANCIELLE : personne X responsable du traitement de données à caractère personnel Y**	
FONCTION LEXICALE	**VALEUR**
Fact$_2$ (*responsable du traitement*) *	*veiller à la conformité* [de N = Y], *assurer la sécurité* [de N=Y], *garantir le respect des règles de protection* [de N = Y],
Fact$_{2\ I}$ (*responsable du traitement*) *	*déterminer les finalités* [de N = Y], *déterminer les moyens* [de N = Y], *fixer la durée* [de N = Y]
Fact$_{2\ II}$ (*responsable du traitement*) *	*informer la personne concernée* [de N =Y],
Fact$_{2\ III}$ (*responsable du traitement*) *	*demander le consentement de la personne concernée* [pour N=Y],
Fact$_{2\ IV}$ (*responsable du traitement*) *	*accomplir les formalités préalables* [à N =

	Y], *déclarer* [N = Y], *effectuer la demande d'autorisation préalable à la mise en œuvre* [de N = Y]
Fact$_{2\ V}$ (*responsable du traitement*) *	*notifier la violation* [de N=Y]

Tableau 43. Encodage de cooccurrents verbaux du terme *responsable du traitement*

8.2.4 Encodage des relations circonstancielles

Passons maintenant à la modélisation des relations circonstancielles. Comme nous l'avons vu plus haut, les circonstants sont des unités qui font référence à des circonstances dans lesquelles se déroule la situation représentée par le prédicat (comme le lieu, le temps, le résultat, le moyen, la manière, l'instrument) sans toutefois contribuer à leur sens. En présentant le concept de la prédicativité, nous avons également remarqué que la distinction entre l'actant sémantique et le circonstant n'est pas toujours facile à réaliser dans les faits. Pour illustrer ce problème, Mel'čuk (2004 : 30) propose d'établir une hiérarchie des circonstants en fonction de ce qu'il appelle *caractère circonstanciel* :

Adverbes > Temps (Durée) > Lieu > Manière > Cause > Résultat > Moyen > Instrument

Circonstants ----------→ *Actants*

Ainsi, à l'extrême gauche du continuum se trouvent les circonstants prototypiques. En revanche, plus on se déplace vers la droite du continuum, plus les circonstant perdent de leur caractère circonstanciel en se rapprochant des actants. D'après Mel'čuk (*ibid.*), l'instrument a beaucoup plus de chance de jouer le rôle d'un actant que par exemple la cause. Le Tableau 44 présente quelques exemples de modélisation des relations circonstancielles existant entre les unités terminologiques dans le domaine du droit de l'Internet. Comme nous pourrons le constater, il est tout à fait possible de capturer ce type de liens à l'aide des FL standard.

téléchargement 1282 occurrences **STRUCTURE ACTANCIELLE : *téléchargement illégal* réalisé par X de Y à partir de Z vers Z**		
FONCTION LEXICALE	**VALEUR**	**TYPE DE RELATION**

$\mathbf{S_{instr}}$ (*téléchargement illégal*)	*logiciel de partage, logiciel d'échange de pair à pair*	**Instrument**
$\mathbf{S_{med}}$ (*téléchargement illégal*)	*P2P, peer-to-peer, pair à pair,*	**Moyen**
$\mathbf{S_{loc}}$ (*téléchargement illégal*)	*Internet, réseau de communications électroniques,*	**Lieu**
$\mathbf{S_{res}}$ (*téléchargement illégal*)	*contenu contrefaisant*	**Résultat**

Tableau 44. Exemples de relations circonstancielles modélisées au moyen des FL.

Soulignons qu'en essayant de systématiser les relations circonstancielles classiques que les termes du domaine en question entretiennent avec d'autres unités terminologiques ou lexicales, nous avons remarqué qu'une grande partie de ces liens correspondent aux fonctions **Sinstr** et **Smed**. Rappelons que les deux FL permettent de dégager les dérivés sémantiques représentant respectivement des instruments ou des moyens. En effet, l'analyse du corpus *DITerm* nous a permis d'extraire un nombre important de termes génériques qui renvoient aux noms d'instruments ou à ceux de moyens. Il s'agit des termes tels que : *technologie* (1940 occurrences), *technique* (5701 occurrences), *système* (5093 occurrences), *instrument* 896 (occurrences), *outil* (1275 occurrences), *moyen* (5104 occurrences), *mécanisme* (1203 occurrences), *mesure* (8563 occurrences), *dispositif* (2477 occurrences). Leur haute fréquence d'emploi dans le corpus *DITerm* peut être considérée comme un indice de la présence de ce type de liens dans le domaine du droit de l'Internet. De plus, nous avons identifié un certain nombre d'expressions telles que : *au moyen de* (245 occurrences), *via* (649 occurrences), *à l'aide de* (145 occurrence), etc., témoignant aussi de l'importance des relations de type *action / instrument* ou *action / moyen* dans le domaine. Nous avons donc décidé de les examiner de plus près. Pour ce faire, nous avons adopté le principe selon lequel la relation action / instrument peut être repérée dans le corpus à l'aide de marqueurs linguistiques tels que *avec, à l'aide de, au moyen de, grâce à, permet/permettent de* tandis que la relation action / moyen peut être repérée grâce aux expressions *par, par le moyen de, via*. Le tableau ci-dessous (Tableau 45) en donne un aperçu :

MOT CLÉ et STRUCTURE ACTANCIELLE	**FONCTION LEXICALE**
service de communication au public en ligne 733 occurrences ***service de communication au public en***	$\mathbf{S_{med}}$ *(service de communication au public en ligne) = procédé de communication*

***ligne* X proposé par Y à Z**	*électronique, Internet*
hameçonnage 36 occurrences ***hameçonnage* pratiqué par X**	$\mathbf{S_{med}}$ *(hameçonnage) = courrier électronique, site web falsifié*
traçage 54 occurrences *traçage* par X de Y	$\mathbf{S_{instr}}$ *(traçage) = cookie, témoin de connexion*
donnée comportementale 25 occurrences ***donnée comportementale* X sur Y utilisée par Z**	$\mathbf{S_{instr}}$ $\mathbf{Caus_1Func_0}$ *(donnée comportementale) = cookies, témoin de connexion*
communication commerciale non sollicitée 63 occurrences ***communication commerciale non sollicitée* envoyée par X à Y**	$\mathbf{S_{med}}$ *(communication commerciale non sollicité) = service de courrier électronique* $\mathbf{S_{instr}LiquFact_3}$ (*communication commerciale non sollicité) = logiciel anti-spam*
contenu illicite 638 occurrences ***contenu* illicite de X communiqué par Y à Z**	$\mathbf{S_{med}Real_2}$ *(contenu illicite) = service de communication au public en ligne* $\mathbf{S_{instr}Real_3}$ *(contenu illicite) = logiciel d'échange de pair-à-pair* $\mathbf{S_{med}Rela_3}$ *(contenu illicite) = streaming, flux continu, lecture continue* $\mathbf{S_{instr}CausNonAble_1Fact_3}$ *(contenu illicite) = outil de contrôle parental* $\mathbf{MultS_{instr}CausNonAble_1Fact_3}$ *(contenu illicite) = mesures techniques de protection, MTP*

Tableau 45. Exemples de relations circonstancielles de type Instrument ou Moyen.

Remarquons que l'encodage des dérivés circonstanciels se complique si l'on a affaire à des quasi-prédicats (c'est notamment le cas des termes *donnée comportementale* et *contenu illicite*). En effet, nous partons de l'hypothèse selon laquelle la structure syntaxique profonde associée aux fonctions **Sinstr** et **Smed** indique que ces fonctions se rapportent toujours au premier actant de l'événement dénoté par l'unité prédicative. Par conséquent, pour décrire le lien qui relie une entité engagée dans une situation donnée et l'instrument mis en œuvre, nous

devons faire appel aux fonctions syntagmatiques. (voir les tableaux 40, 41). Par exemple, si la relation collocationnelle *visionner un contenu illicite* correspond à la formule :

Rela$_3$ *(contenu illicite) = visionner* [ART ~] *;*

des termes dénotant le (moyen utilisé pour visionner un contenu illicite) peuvent être dégagés à l'aide de la fonction suivante :

S$_{med}$Rela$_3$ *(contenu illicite) = streaming, flux continu, lecture continue*.

Quant au terme *donnée comportementale*, son lien avec les termes dénotant (outil permettant de collecter des données comportementales) peut être dégagé au moyen de la règle suivante :

S$_{instr}$ Caus$_1$Func$_0$ *(donnée comportementale) = cookie, témoin de connexion,*

où

Caus$_1$Func$_0$ *(donnée comportementale)* = collecter [ART ~], recueillir [ART ~].

En effet, grâce à ce type de fonctions complexes, il est possible de mettre en évidence un nombre important d'unités reliées au terme vedette. Cependant, en dégageant les dérivés sémantiques circonstanciels, il faut être extrêmement prudent et ne pas oublier que la dérivation sémantique est une relation fondée sur une parenté de sens et non pas sur des propriétés extralinguistiques. Rappelons que selon Mel'čuk et Polguère (2007 : 18), une lexie L_2 est dite sémantiquement dérivée d'une lexie L_1 si et seulement si les trois conditions suivantes sont remplies :

- les définitions de L_1 et L_2 possèdent des composantes de sens communes (dans le cas le plus typique, L_2 se définit en termes de L_1)

- la relation sémantique entre L_1 et L_2 est récurrente dans la langue (en l'occurrence, dans le corpus)

- la relation entre L_1 et L_2 s'exprime éventuellement, mais pas nécessairement, de façon morphologique dans la langue.

Il est donc impossible d'encoder au moyen des fonctions lexicales les relations qui relèvent de l'ordre conceptuel, comme par exemple *données à caractère personnel* et *mécanisme opt-in, données à caractère personnel* et *droit d'opposition* ou bien *téléchargement illégal* et *procédure de réponse graduée, dispositif Hadopi, mécanisme de sanction et d'avertissement, courriel d'avertissement, courrier recommandé.*

En essayant de modéliser les relations circonstancielles, nous avons également constaté que les FL standard ne permettent pas de distinguer les différents types d'instruments et de moyens mis en œuvre. En effet, l'analyse du corpus *DITerm* a relevé la présence de deux sortes de circonstants. D'un côté, nous avons recensé un grand nombre de dispositifs, mesures, procédures juridiques, de l'autre, il s'agit des outils techniques et des technologies propres à l'informatique. Nous considérons que dans le cadre de cette étude, il serait nécessaire de trouver une méthode permettant de décrire ces différences. Le recours à des FL non-standard semble une solution adéquate à condition de respecter les critères évoqués ci-dessus.

8.3 Bilan de l'application des FL au projet *DITerm*

Pour conclure ce chapitre, nous voudrions dresser un bilan rapide de l'application des FL à notre projet terminographique. Comme nous avons pu le constater, le modèle des FL (quoique jugé parfois trop aride), constitue un outil très puissant permettant de décrire de façon uniforme une multitude de relations lexicales qu'un terme du domaine de l'Internet entretient avec d'autres unités et ceci aussi bien sur l'axe paradigmatique que sur l'axe syntagmatique. En effet, l'outil des FL permet de mettre en évidence le lien conceptuel profond qui existe entre dérivations sémantiques (synonymie, antonymie, nom d'actants ou de circonstants) et collocations. Comme le souligne Polguère (2003 : 11) : « *Ces deux types de phénomènes relèvent de liens lexicalement contrôlés et leur standardisation fait appel aux mêmes universaux linguistiques*. » De plus, le formalisme des FL assure rigueur et

systématicité dans la description des relations lexico-sémantiques tout en permettant de rendre cette description plus compacte.

Cependant, il convient de souligner que notre proposition d'encodage se base uniquement sur les FL dites standard[84], ce qui a ses conséquences méthodologiques. Rappelons que pour être considérées comme standard, les FL doivent répondre aux principes d'universalité, de cardinalité et de diversité (Polguère 2007) évoqués dans le premier chapitre de notre travail. Cependant, comme le soulignent Mel'čuk *et al.* (1995 : 150), les FL standard ne couvrent pas totalement l'immense ensemble des cooccurrences lexicales restreintes[85]. Cette remarque est d'autant plus valable s'agissant d'un travail terminographique où un grand nombre des relations sont caractéristiques d'un domaine donné et ne répondent pas à l'encodage formel. C'est notamment le cas de notre projet. En effet, nous avons constaté que les FL standard, même y ajoutant les FL complexes, les FL localement standard et les FL standard avec pointeurs vers la définition, ne permettent pas de couvrir l'ensemble des dérivations sémantiques et des collocations recensées dans le corpus *DITerm*. Ainsi, pour décrire les relations dont le sens est spécifique et donc non généralisable, la solution serait de recourir à des FL semi-standard (FLSS) et des FL non standard (FLNS). Rappelons que les FL semi-standard sont des formules dérivées des FL standard, c'est-à-dire qu'elles sont constituées d'une FL standard et d'un élément en français venant ajouter une composante de sens non prise en charge par la FL standard. (Jousse 2010 : 76). Ainsi, la relation collocationnelle *enregistrer des données à caractère personnel* pourrait être décrite de la manière suivante :

$Caus_1Func_0$ sur un support de stockage (*donnée à caractère personnel*) = enregistrer [ART ~].

Comme nous pouvons le constater, l'élément en français apporte une précision sémantique qui permet de modifier les valeurs par rapport à la fonction standard : **$Caus_1Func_0$** (*donnée à*

[84] Traditionnellement, l'ensemble des FL standard comprend une liste de 56 FL standard simples (Mel'čuk *et al.* 1995 : 129), les FL complexes correspondant à des enchaînements de FL standard simples ainsi que des configurations de FL, c'est-à-dire des suites de FL simples qui ne sont pas syntaxiquement liées entre elles (idem 149). Pour les besoins de notre étude, nous avons inclus dans ce groupe les FL localement standard (Polguère 2007 : 52) et les FL standard avec pointeurs vers la définition (Jousse 2010 : 86).

[85] D'après Mel'čuk *et al.* (1995 : 127-128) et Polguère (2007 :48), une FL standard doit s'appliquer à un nombre important d'arguments (= mots-clés) et elle doit retourner un nombre élevé d'éléments dans sa valeur. De plus, elle ne doit pas être limitée aux unités d'un seul champ sémantique.

caractère personnel) = collecter [ART ~], recueillir [ART ~]. Quant aux FL non standard, elles sont encodées par des gloses formulées dans la langue de description :

Conserver ~ quelque part = stocker [ART ~]
Envoyer ~ à quelqu'un[86] = transférer [ART ~]

Comme le soulignent Mel'čuk *et al.* (*ibid.*), ces FL ne sont pas prévisibles et ne peuvent pas être dégagées de façon méthodique. En effet, les FL semi-standard et FL non standard forment un ensemble très hétérogène et irrégulier. De plus, comme le remarque Jousse (2010 : 103) [87], il n'existe pas de véritables règles de formation de l'encodage des FLSS et FLNS. Étant donné ce caractère hétérogène et l'absence de normes d'écriture des fonctions irrégulières, la modélisation des relations au moyen du formalisme des FL ne permet pas de préserver leur systématicité. En effet, les aspects formels des différentes variétés de FL (FL standard, semi-standard et non standard) divergent de manière significative ce qui contribue à former trois ensembles disjoints de FL. Leur présence dans une même fiche terminographique rendrait la description désordonnée et parfois illisible. Or un des avantages indéniables de l'encodage standard est sa rigueur, sa prévisibilité et sa systématicité. Néanmoins, si l'on tient à proposer une description complète des unités terminologiques au moyen des FL, il est impossible de se baser uniquement sur les FL standard.

En effet, le recours à d'autres variétés de FL, notamment aux FL semi-standard et aux FL non standard, est étroitement lié à la notion de granularité (Kahane et Polguère 2001). Ainsi, quand on cherche à décrire les relations des unités lexicales (terminologiques dans notre cas), on doit décider d'un degré de précision de cette description. Si on opte pour une granularité large, on aura tendance à regrouper sous un même encodage un grand nombre de liens. Par contre, le choix d'une granularité moins restreinte amènera à l'attribution d'encodages distincts pour représenter les différents liens étudiés. Comme le souligne Polguère (2007: 56).: « *The notion of standardness interacts closely with the notion of granularity in the formal language of LF. The more granularity you use, the more risks you*

[86] En proposant ces FL semi-standard et non-standard, nous nous sommes inspirée de l'encodage proposé dans *DiCoInfo*.
[87] Il convient de mentionner le travail de Jousse (2010) dont l'objectif est la standardisation et la normalisation des FL irrégulières.

have to need to handle non standard LFs. » Si on décide de faire appel uniquement aux FL standard, on choisit un encodage moins précis. Pour illustrer cette situation, revenons aux exemples du Tableau 41. Nous y voyons que la fonction lexicale **Caus$_3$Oper$_3$** appliquée au mot-clé *nom de domaine* permet de dégager 7 collocatifs: *choisir, créer, déposer, réserver, enregistrer, acquérir, acheter*. Ces verbes, sémantiquement pleins, partagent certains traits sémantiques importants, mais se distinguent entre eux par d'autres traits plus ou moins nombreux. Pour faire ressortir ces différences, on a donc la possibilité de faire appel à des FL semi-standard qui permettent d'apporter des éléments de sens supplémentaires qui ne sont pas pris en charge par la FL standard :

Moyennant une somme d'argent Caus$_3$Oper$_3$ (*nom de domaine)* = acheter [ART ~]

De manière formelle Caus$_3$Oper$_3$ (*nom de domaine)* = déposer [ART ~], enregistrer [ART ~]

Ainsi, la décision quant au degré de précision de la description réalisée au moyen des FL (bien évi*bid.*ent là où il est possible d'attribuer une FL standard) revient au terminographe. Et même si, dans le cas d'un projet terminologique, le choix d'une plus grande granularité nous semble naturel, il faut savoir que l'opération de scission des FL standard en plusieurs FL semi-standard ou non standard a une grande incidence sur la clarté de la description. Comme le remarque Jousse (2010 : 152), le très fin degré de granularité choisi pour l'identification des relations lexicales conduit à une démultiplication des formules de FLNS et FLSS et complexifie la recherche d'analogies entre liens lexicaux.

En résumé, nous considérons que l'arsenal des FL standard ne permet pas de fournir une description complète et détaillée des relations entre les unités terminologiques du domaine du droit de l'Internet. De l'autre côté, la possibilité de recourir aux FL semi-standard et non standard qu'offre le modèle mel'čukien ne constitue pas, du point de vue terminographique, une solution entièrement satisfaisante, ceci à cause du caractère hétérogène et irrégulier de ses différentes variantes de FL. Néanmoins, nous soutenons que malgré les difficultés liées aux aspects formels, le modèle des fonctions lexicales est un outil exceptionnel apportant une réponse adéquate au problème, très complexe, de la modélisation des relations lexicales entre les termes.

Cependant, il ne faut pas oublier que l'encodage réalisé au moyen des FL s'appuie fortement sur la décomposition du sens. En effet, comme nous l'avons vu, les FL sont révélatrices du lien profond qui existe entre les définitions sémantiques des unités lexicales (terminologiques dans notre cas) et les relations de dérivation sémantique ou de collocation qui leur sont associées. Les FL, de par leur nature (elles sont toutes ancrées dans le contenu sémantique des unités), ne permettent donc pas de systématiser les liens conceptuels. Or, comme nous l'avons vu plus haut, la modélisation des relations entre les unités terminologiques appartenant au domaine du droit de l'Internet demande parfois une analyse en termes de référents ce qui pose un problème d'incompatibilité du point de vue méthodologique. En effet, le recours aux FL exige une distinction nette entre les phénomènes tels que combinatoire libre et combinatoire restreinte, liens de dérivations sémantiques et liens conceptuels forts. Nous avons démontré que dans certains cas, il serait possible de contourner ce problème méthodologique en créant délibérément des définitions sémantiques permettant de tisser des liens entre des cooccurrents de forte compatibilité conceptuelle. Néanmoins, cette démarche s'avère délicate lorsque l'on est face à des liens purement conceptuels, de parenté sémantique très faible. Une autre solution serait de proposer un système de liens dérivés de relations sémantiques permettant de décrire les relations d'ordre conceptuel. Comme nous l'avons vu dans la partie théorique de ce travail (voir la section 2.4), c'est la solution adoptée par Dancette dans la version électronique du DAD (2006). Rappelons qu'il s'agit du modèle des relations lexico-sémantiques (RLS) qui a finalement conduit l'auteure à développer un autre système de balisage des liens, système mis en œuvre dans le projet DAMT (2010) et composé de deux outils formels distincts : relations sémantique (RS) et relations associatives (RA).

Pour notre part, nous considérons que pour proposer une description complète de ces deux phénomènes de nature différente, c'est-à-dire, pour :

a) rendre compte des dérivés sémantiques et de la phraséologie des termes

et

b) refléter les affinités conceptuelles des termes ;

il faut faire appel à deux outils formels distincts qui pourraient se présentent, d'une manière uniforme, sous forme de formules explicites en français standardisé. En effet, nous pensons, tout comme Polguère (2003), que les FL doivent être appréhendées indépendamment de leur formalisation habituelle, qui ne semble pas la plus appropriée à notre projet. Dévêtir la notion de fonction lexicale de l'appareillage formel d'encodage auquel elle est habituellement associée permet d'y voir un outil conceptuel visant à rendre compte des phénomènes linguistiques d'une façon plus naturelle et intuitive. De l'autre côté, nous pensons que la représentation des liens conceptuels peut également se baser sur une sorte de paraphrasage. Nous y reviendrons dans la partie suivante de ce travail consacrée à la présentation de notre modèle de description des unités terminologiques du domaine du droit de l'Internet. Mais avant cela, nous devons aborder la question de la systématisation des relations conceptuelles entre les termes, question qui ne peut pas être prise en charge par le modèle des FL.

Chapitre 9. Tentative de systématisation des relations conceptuelles entre les termes

Rappelons que l'objectif principal de cette étude est de proposer un modèle de description complète des unités terminologiques du domaine du droit de l'Internet, un modèle qui reflète aussi bien la dimension linguistique des termes (leur nature linguistique, les relations lexico-sémantiques qu'ils entretiennent avec d'autres unités, les combinaisons lexicales typiques dans lesquelles ils se trouvent), que leur dimension cognitive (c'est-à-dire la place des termes dans la structure conceptuelle du domaine et leurs liens avec d'autres termes). Comme nous l'avons souligné à maintes reprises, nous envisageons les termes comme des unités lexicales pourvues de sens spécialisés et considérons qu'il faut les appréhender avant tout dans leur fonctionnement linguistique. Néanmoins, nous estimons qu'en proposant une description globale des termes, il est impossible de faire abstraction de leurs propriétés cognitives. Nous rejoignons ainsi l'idée de Cabré[88] (2007: 98-101) selon laquelle les unités terminologiques doivent être considérées comme des unités polyédriques, c'est-à-dire des entités à dimensions multiples. D'après Cabré, l'accès à ces dernières peut se faire par des points d'entrée (ou des *portes*, pour reprendre sa terminologie) différents : la linguistique, les sciences cognitives et les sciences de la communication. Comme l'ajoute l'auteure (*ibid.* : 99), chaque *porte* d'entrée exige une théorie adaptée et chaque théorie doit être cohérente avec les théories adaptées aux autres portes d'entrée. En effet, ce caractère multidimensionnel du terme soulève toute une série de questions d'ordre méthodologique. Comme le souligne L'Homme (2005 : 1121), les modèles théoriques et descriptifs de la terminologie proposent des solutions différentes à ces nombreuses questions. Le terme y est présenté tantôt comme un moyen d'expression de concepts dont le contenu s'ancre dans la structuration des connaissances d'un domaine spécialisé (optique conceptuelle), tantôt comme une unité lexicale dont la description mène à une structuration lexicale dudit domaine (optique lexico-sémantique). D'après l'auteur (*ibid.*), le terminologue doit faire un choix

[88] Il s'agit d'un modèle dit *des portes* développé par Cabré (1999, 2003, 2007) dans le cadre de sa Théorie communicative de la terminologie, une conception théorique permettant de réconcilier différents points de vue sur le terme.

parmi les options offertes et assumer ses conséquences méthodologiques. Elle rend ainsi attentif à l'incompatibilité méthodologique de ces deux approches.

En voulant proposer une méthode de description terminologique permettant de rendre compte du double statut des termes, nous nous demandons tout de même dans quelle mesure il est possible de rapprocher ces deux démarches. Bien évidement, nous ne prétendons pas à la création d'une ontologie du domaine du droit de l'Internet. Nous considérons, tout comme L'Homme (2005 : 1123) ou Cabré (2007 : 82), que la tâche de structuration conceptuelle doit être confiée à des spécialistes des domaines et non pas aux linguistes, dont le rôle est la description du comportement des termes en langue. Cependant, à l'instar des projets proposant d'intégrer la dimension cognitive dans les modèles descriptifs des langues de spécialité tels que les travaux de Dancette (*DAD, DAMT*) [89], Schmidt (*Kicktionary*) ou Faber (2011) ou Faber *et al.* (2005) (*EcoLexicon*), nous voudrions trouver une méthode permettant de saisir et de systématiser les liens conceptuels entre les termes. En effet, nous considérons que la structure lexicale d'une langue de spécialité reflète en quelque sorte la structure conceptuelle du domaine en question. Nous nous interrogeons en même temps sur la possibilité de mener cette analyse tout en restant dans l'optique lexico-sémantique. Comme nous l'avons vu plus haut, les fonctions lexicales issues de la Théorie Sens-Texte apportent une solution adéquate quant à la description de la nature linguistique des termes. En revanche, le modèle mel'čukien n'est pas adaptée à la formalisation des relations intervenant sur le plan conceptuel. Nous devons donc réfléchir à la manière d'intégrer et d'organiser des données d'ordre conceptuel sans toutefois renoncer à la description basée sur des éléments linguistiques. Dans les sections suivantes, nous allons donc essayer de systématiser les relations conceptuelles observées dans le corpus *DITerm*. Nous terminerons par la présentation des résultats de notre travail, dont le but est de trouver un modèle hybride de description des unités terminologiques du droit de l'Internet permettant de rendre compte aussi bien des relations lexico-sémantiques que des liens conceptuels.

[89] Nous avons consacré une partie du premier chapitre à la description de ces projets.

9.1 La qualification juridique comme porte d'accès au schéma du domaine du droit de l'Internet

L'analyse des unités terminologiques au moyen de la linguistique doit se faire à partir des textes (Cabré 2007 : 99). Notre tentative de systématisation des liens conceptuels entretenus par les termes du domaine du droit de l'Internet sera donc réalisée à la suite des observations faites sur l'ensemble des données linguistiques provenant du corpus *DITerm*. Nous considérons que les relations ainsi dégagées constitueront, au moins en partie, un reflet de la structure conceptuelle du domaine en nous permettant une description basée plutôt sur des éléments linguistiques qu'extralinguistiques. Néanmoins, nous sommes consciente que cette tâche demandera un certain recul sur la réalité linguistique. En effet, comme le souligne L'Homme (2005 : 1122), pour structurer le champ conceptuel d'un domaine donné, le terminologue doit s'aligner sur des modes d'appréhension des connaissances définies par les spécialistes. Nous espérons tout de même que notre analyse restera conciliable avec la démarche linguistique que nous prônons.

Ainsi, comme nous l'avons signalé dans la partie précédente de ce travail, l'étude des données extraites du corpus *DITerm* nous a permis de relever un nombre important de termes génériques qui peuvent servir de point d'entrée à la structure conceptuelle du droit de l'Internet. Il s'agit surtout des mots clés du vocabulaire juridique de base reconnus comme étant porteurs des notions fondamentales du droit. À cette liste s'ajoute aussi un certain nombre de mots génériques appartenant à la langue générale. En effet, la présence importante de termes génériques et abstraits dans notre corpus est liée au caractère général et impersonnel de la loi. Bien qu'elle soit considérée comme un instrument de traitement de la réalité, la loi représente des principes universellement valables sans tenir compte des circonstances, des situations et des individus concrets. C'est pour répondre à cette exigence de généralité et d'impersonnalité de la loi que le langage juridique fait appel à un grand nombre de termes génériques. Les termes en question correspondent aux catégories conceptuelles abstraites et sont communs à toutes les branches du droit. Ce sont des outils de classification juridique par excellence.

En effet, pour que la loi s'applique, un passage de l'abstrait au concret (ou autrement dit du général au singulier) s'impose. On parle alors de l'opération de spécification qui en règle générale se résume à la démarche de qualification juridique. Précisons que la

qualification juridique est un mécanisme intellectuel qui a vocation à faire entrer un élément de fait dans une catégorie juridique. Comme le remarquent Bourcier et Fernàndez-Barrera (2011 : 129), la qualification permet de relier le monde des faits au système conceptuel du droit en assurant un continuum entre les catégories factuelles plus concrètes et les catégories juridiques conceptuelles plus abstraites. Il s'agit là d'un passage obligatoire afin de rendre le droit opératif face à des évènements du monde. Selon les auteurs (*ibid.*), il est donc possible de faire un rapprochement entre les théories juridiques de la classification et la construction de systèmes conceptuels. En effet, dans une perspective terminologique et ontologique, la qualification juridique peut être conçue comme un niveau intermédiaire de conceptualisation qui permet le passage du niveau terminologique (du niveau des données textuelles) au niveau notionnel. Bourcier et Fernàndez-Barrera (*ibid.*) proposent d'utiliser ce modèle dans l'élaboration d'ontologies juridiques, notamment dans la construction d'une ressource sémantique dans le domaine de la régulation du bruit. Afin de réaliser leur projet, les auteurs ont fait appel à la *middle-out strategy*[90] (Breuket *et al.* 2002 : 21). Cette méthodologie se base sur la bidirectionnalité de l'analyse où l'on part simultanément d'un ensemble de concepts très abstraits et d'un ensemble de termes réalisés dans des corpus du domaine. L'ontologie noyau sélectionnée pour le niveau *top* de la ressource est l'ontologie de concepts juridiques fondamentaux de LIKIF-Core (Breuker *et al.* 2002), qui contient 15 modules réutilisables séparément. La méthode de modélisation termino-ontologique proposée par Bourcier et Fernàndez-Barrera (2011) consiste en l'appariement des termes extraits des corpus aux concepts noyaux de LIKIF-Core. Le passage des données terminologiques au modèle conceptuel ne se fait pas directement. En effet, comme l'expliquent les auteurs (*ibid.* : 136), la transformation d'un terme en classe ontologique est réalisée par l'intermédiaire d'une classe ontologique intermédiaire, résultat du travail de qualification juridique. Cette classe intermédiaire est représentée par un terme général et abstrait. La Figure 35 présente le modèle de transformation d'un terme en classe ontologique à l'aide d'une classe intermédiaire proposé par Bourcier et Fernàndez-Barrera (*ibid.* : 137).

La méthode de modélisation termino-ontologique présentée ci-dessus semble être une piste intéressante. À l'instar de Bourcier et Fernàndez-Barrera, nous avons donc décidé de tirer profit du mécanisme de qualification juridique. Cependant, compte tenu des objectifs de

[90] Comme le soulignent Bourcier et Fernàndez-Barrera, cette méthodologie s'oppose aux approches exclusivement *top-down* (basée sur les théories conceptuelles d'un domaine) et *bottom-up* (partant des évidences textuelles fournies par le corpus).

notre travail, qui visent à refléter aussi bien la dimension cognitive des termes que leur dimension linguistique (et ceci du point de vue d'un linguiste et non pas de celui d'un expert du domaine), nous avons été amenée à apporter quelques adaptations méthodologiques.

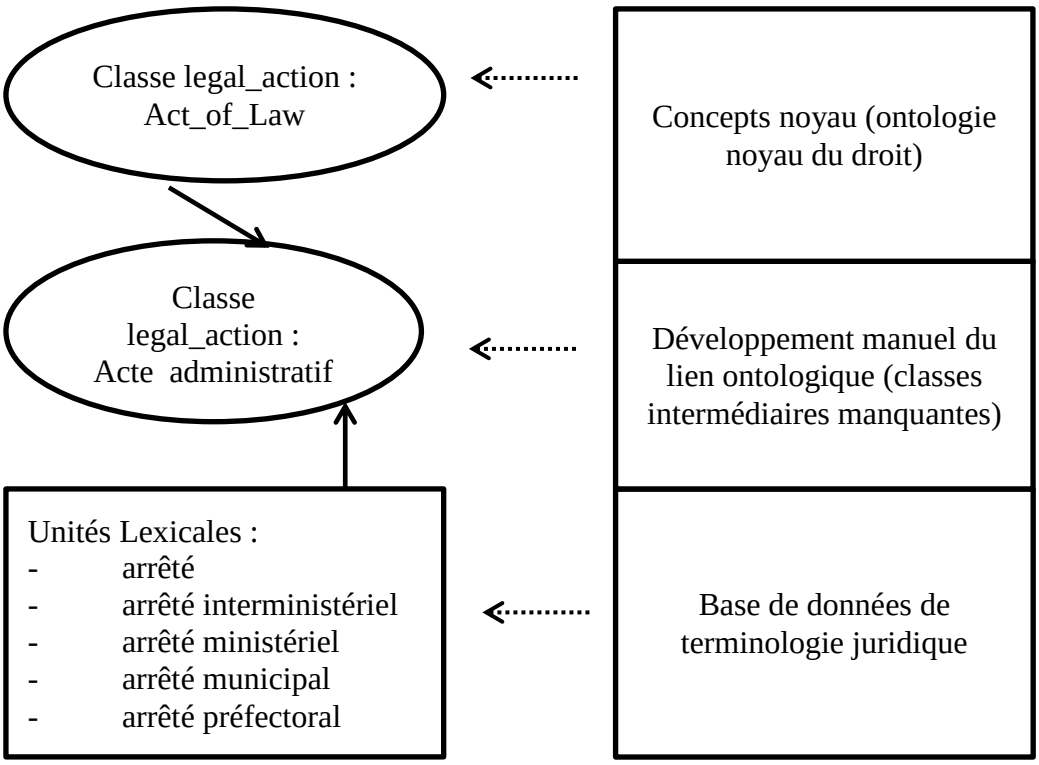


Figure 35. Méthodologie *middle-out* pour la construction de la ressource sémantique - ex. acte_administratif (Bourcier et Fernàndez-Barrera (2011: 137)

En effet, au lieu de tenter l'appariement des termes extraits du corpus aux classes ontologiques préexistantes en créant des classes ontologiques intermédiaires, nous avons décidé de chercher des traces de l'opération de qualification juridique directement dans les textes. Nous avons donc opté pour une démarche qui part des données textuelles et qui permet d'identifier, dans le corpus *DITerm,* des classes de termes abstraits. L'analyse de ces derniers permet, par la suite, de connecter les termes concrets propres au domaine du droit de l'Internet à des catégories abstraites appelés *cadres* (ou *frames*) *sémantico-conceptues*, définis à la suite des observations menées sur l'ensemble des textes.

Rappelons que nous nous sommes déjà intéressée, quoique de manière indirecte, à la qualification juridique en considérant que l'analyse des termes généraux nous permettra d'accéder à des termes concrets spécifiques du droit de l'Internet. En effet, comme nous l'avons vu dans la partie précédente de ce travail, certains termes génériques peuvent entrer dans la composition des unités terminologiques de niveau inférieur en permettant ainsi de dégager des séries de termes complexes. De plus, nous avons démontré que l'étude du contexte dans lequel se trouvent les unités considérées comme étant génériques peut conduire à l'identification des termes simples ou complexes sémantiquement apparentés aux termes génériques, mais ne partageant pas avec ces derniers de composante formelle. L'étude du contexte des termes juridiques nous a aussi permis d'accéder à des données d'ordre conceptuel. Maintenant, nous avançons l'hypothèse selon laquelle les termes génériques, en l'occurrence le vocabulaire juridique de base (4) ainsi que certains mots génériques appartenant à la langue générale (9), correspondent à des catégories conceptuelles plus abstraites ayant le rôle de relier des unités terminologiques (relevées auparavant) aux classes ontologiques du domaine du droit. Afin de mettre en évidence la relation existant entre les données linguistiques et conceptuelles, nous avons été amenée à procéder par étapes. La Figure 36 schématise notre démarche.

Ainsi, nous avons commencé notre analyse en organisant les termes génériques[91] extraits du corpus *DITerm* en ensembles sémantiques (ou classes sémantiques – voir l'Homme 2004 : 212), que nous avons appelés *familles sémantiques*. Rappelons que Cornu (2005 : 198) parle de communautés de voisinage. Selon l'auteur, dans le vocabulaire juridique, certains termes ont des sens voisins, il est donc possible de les rapprocher. Lors de cette étape, réalisée selon la méthode manuelle, nous avons donc identifié 17 familles sémantiques, c'est-à-dire 17 ensembles de termes entre lesquels existe une parenté de sens. Remarquons, qu'il serait possible de dégager d'autres familles sémantiques, ce qui contribuerait certainement à l'enrichissement de notre description. Néanmoins, nous avons choisi de restreindre notre analyse aux unités qui affichent une fréquence importante en considérant qu'il s'agit des termes qui correspondent aux concepts principaux du domaine du droit de l'Internet.

Deuxièmement, nous avons procédé à l'appariement des termes spécifiques (qui ont été repérés à partir des termes génériques lors de l'étape de l'extraction des candidats-

[91] Il s'agit surtout des données appartenant aux catégories (4) et (9) (voir les sections 6.3.2.2 et 7.1.2).

termes[92]) aux ensembles sémantiques identifiés précédemment. Le rapprochement des termes génériques et spécifiques ainsi que l'analyse de leur environnement contextuel contenant de nombreuses informations de type encyclopédique a permis d'apporter d'importantes précisions d'ordre conceptuel. Nous avons constaté, par exemple, que le terme *activité* renvoie aussi bien à des services offerts via Internet aux utilisateurs du réseau (*activité de fourniture d'accès, activité d'hébergement*) qu'aux activités menées par ces derniers à l'aide d'une connexion Internet (*téléchargement*). Quant aux acteurs de l'Internet, nous en avons distingué deux principales catégories : prestataires de services et utilisateurs de services. En ce qui concerne les différents types d'*instruments* ou d'*outils*, nous avons remarqué que l'on pourrait les diviser en trois groupes : instruments techniques, procédures juridiques et dispositifs technico- juridiques.

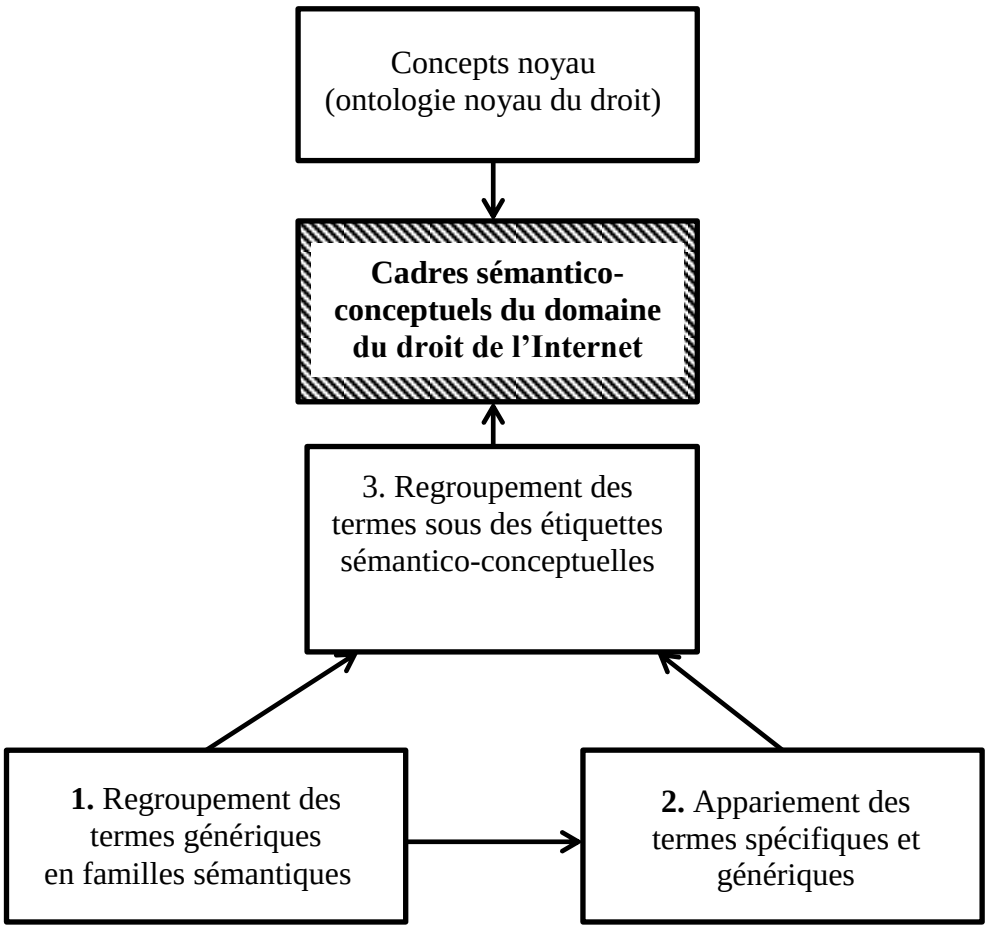


Figure 36. Méthode de structuration des données conceptuelles adoptée dans le projet *DITerm*

[92] Voir à ce propos le chapitre 6.

L'analyse de l'environnement contextuel du terme générique *droit* nous a permis de dégager trois classes de termes spécifiques, notamment : a) termes qui désignent une faculté légalement reconnue à quelqu'un par une autorité publique d'agir de telle ou telle façon ou de jouir de tel ou tel avantage, b) termes qui dénotent les personnes qui possèdent juridiquement un droit, c) termes qui renvoient à une règlementation. Quant aux termes qui renvoient à des mesures légales, nous avons constaté qu'il était possible de les diviser en deux groupes : textes législtatifs et règles. Nous avons également remarqué que dans certains cas, l'étude du contexte dans lequel se trouvent les termes génériques ne permet pas d'accéder directement à des termes spécifiques mais conduit plutôt à la découverte d'informations d'ordre conceptuel. En effet, l'étude des termes spécifiques ainsi que l'analyse contextuelle nous a permis d'identifier un plus grand nombre de catégories abstraites.

Ensuite, chaque ensemble a reçu une étiquette sémantico-conceptuelle qui rend compte des caractéristiques communes de ses éléments. En effet, nous parlons de l'étiquette sémantico-conceptuelle car d'un côté, cette dernière est attribuée à la suite d'une analyse sémantique et rassemble les termes génériques apparentés par le sens. De l'autre côté, l'appariement des termes spécifiques aux familles de termes génériques se fait sur la base des éléments extralinguistiques liés à l'opération de qualification juridique dont les traces sont omniprésentes dans notre corpus. Le tableau ci-dessous (Tableau 46) présente les résultats de notre analyse (les chiffres placés à droite des termes génériques se rapportent à leur fréquence d'apparition dans le corpus *DITerm*)

termes génériques regroupés en familles sémantiques	**exemples de termes spécifiques repérés à partir des termes génériques [93] et ÉTIQUETTE SÉMANTICO-CONCEPTUELLE**
acte 3608 action 3056 activité 5333 agissement 279 fait 1702 opération 1784	{ACTION ou ACTIVITÉ} menée en ligne
	pratique de la communication comportementale, pratique des échanges de fichiers musicaux sur Internet, fait de visionnage, opération de traitement des données à caractère personnel, acte de téléchargement,

[93] Comme nous l'avons vu dans le chapitre 6, pour repérer les termes hyponymes, nous avons eu recours à différents patron lexico-syntaxiques tels que : (terme hyperonyme sing.|terme hyperonyme pl.)(de|des|du|de la)<N><A>*de*<N>*<A>* ; (terme hyperonyme sing.|terme hyperonyme pl.)<A>*

pratique 3159	{ACTION ou ACTIVITÉ ILLICITES} menées en ligne
	actes de parasitisme, acte de contrefaçon sur Internet, actes de piratage, fait de mise en ligne d'un contenu illicite,
	{SERVICE proposé EN LIGNE}
	activité de communications électroniques, activité de commerce électronique activité de courtage aux enchères par voie électronique, activité de fourniture de liens commerciaux, activité de moteur de recherche activité de référencement activité de stockage de contenus, activité de fourniture d'accès à Internet activité du WEB 2.0
service 18 931 prestation 1009	{SERVICE proposé EN LIGNE}
	service de communication au public en ligne, service de la société de l'information, service de courrier électronique, service de partage de vidéo, service de référencement, services de musique en ligne, prestation d'hébergement, prestation de connexion à haut débit
acteur 1227 fournisseur 1864 éditeur 1758 prestataire 1854 intermédiaire 969	{ACTEUR DE L'INTERNET – PRESTATAIRE}
	fournisseur de moteur de recherche, fournisseur de réseau publicitaire, fournisseur d'accès Internet, fournisseur d'hébergement, fournisseur de contenu, fournisseur de liens sponsorisés, prestataire du Web 2.0., prestataire intermédiaire de services Internet, prestataire de stockage,
acteur 1227 auteur 6228 usagers 214 utilisateur 5696	{ACTEUR DE L'INTERNET – UTILISATEUR}
	utilisateur du réseau, utilisateur du site utilisateur du service
donnée 21 826 information 11 923 contenu 8764 œuvre 6949	{OBJET mis EN LIGNE}
	donnée à caractère personnel, données relatives au trafic,

document 2121	base de données, contenu numérique, contenu audiovisuel, contenu multimédia, œuvre protégée
via 644 technologie 1940 technique 5701 système 5093 instrument 896 outil 1275 moyen 5104 mécanisme 1203 mesure 8563 dispositif 2477 système 5093	{TECHNOLOGIE ou MACHINE}
	outil de recherche, outil de suggestion de mots-clés, système « adwords », système anti-virus, système d'échange P2P, système de nommage Internet, technologie de reconnaissance faciale, technique des hyperliens, technique du DNS, technique du streaming, technologie P2P
	{DISPOSITIF JURIDICO-TECHNIQUE}
	mécanisme de « opt-in », mécanisme de « opt-out », mesure de blocage du site, mesure de filtrage généralisé, mesures de protection technique, instrument de filtrage, outil de contrôle parental, système d'empreintes numériques,
	{PROCÉDURE}
	mécanisme d'avertissement et de sanction, dispositif de réponse graduée, mécanisme de notification,
charte 753 clauses 587 code 2113 directive 609 disposition 3810 droit 16738 loi 9293 norme 1325 principe 4612 recommandations 486 régime 1974	{TEXTE LÉGISLATIF}
	la loi dite Hadopi 1, la loi dite LCEN, Loi du 6 janvier 1978, règles du commerce électronique, règles du droit d'auteur, Loi Informatique, Fichiers et Liberté, directive 2002/58/CE, directive « vie prive et communications électroniques », recommandations de la Cnil
	{RÈGLE}

règlement 3069 réglementation 970 règle 2911 régulation 702	règles de protection des données à caractère personnel, règles d'entreprises contraignantes, principe de loyauté règles de nommage règle du consentement préalable
procédure / procédures 5571 formalité / formalités 232	{PROCÉDURE}
	procédure de notification et de retrait, procédure alternative de résolution de litiges, procédure d'enregistrement par étape, procédure du « double clic », formalités préalables auprès de la Cnil
conforme/conformes, 613 légal(es) 1390 légalement 221 légitime(s) 1253 licéité 201 licence 1902 licite(s) 510	{CONFORME À LA LOI}
	offre légale, site légal, streaming légal, utilisation licite d'une œuvre, services licites de musique en ligne, contenu licite, *les données doivent être collectées et traitées de manière loyale et licite*[94]*,* téléchargement licite,
agrément 181 autorisation 1768 autorisé(es) 1521 permettre 255 permis 588	{AUTORISATION}
	autorisation préalable de la Cnil, autorisation unique, autorisation de reproduction, autorisation d'exploitation autorisation des ayants droit, *la prospection directe par courrier électronique est autorisée si […],* *ce référencement est permis*
illégal (es/aux) 529 illégalement 87 illicite (s) 3191 illicitement 51	{ACTIVITÉ ou ACTION ILLICITES}
	mise en ligne illicite, mise à disposition illicite d'œuvres ou d'objets protégés par un droit de propriété littéraire et artistique, téléchargement illégal
	{OBJET ILLICITE EN LIGNE}

[94] Les contextes comportant des termes génériques et apportant de précisions d'ordre contextuel sont mis en italique.

	site illicite, contenu illicite
interdiction 723 interdire 688 interdit (e) 573 non autorisé (es)	{ACTIVITÉ ou ACTION ILLICITES}
	usage non autorisé d'un nom de domaine, échange non autorisé de fichiers musicaux, reproduction ou diffusion non autorisée de programme, vidéogramme ou phonogramme, accès non autorisé à un système de traitement automatisé de données, mise à disposition non autorisée au public de phonogrammes sur Internet, échanges non autorisés sur les réseaux P2P,
	{PERSONNE QUI REALISE UNE ACTIVITÉ ou ACTION ILLICITES}
	personne non autorisée
obligation 6101 obligatoire 326 responsabilité 5415 devoir/doit/doivent 9678	{OBLIGATION}
	obligation de surveillance générale, obligation de conservation des données d'identification, obligation de contrôle a priori, obligation de filtrage, obligation de notification des violations de données à caractère personnel, obligation de prompt retrait d'un contenu illicite, obligation de sécurisation de l'accès Internet
avantage 608 droit / droits 29 096 habilité 171 pouvoir 1848	{DROIT}
	droit à l'oubli numérique, droit à la confidentialité des données à caractère personnel, droit à la copie privée, droit à la protection des données à caractère personnel, droit de suppression, droit de rectification,
	{AYANT DROIT}
	ayant droits 793 titulaire de droits 1005 titulaire des droits d'auteur 1587 titulaire du nom du domaine 54
abus 407 abusif/abusive 209 attaque 326 atteinte 3962 crimes 199 délictuel(le) 211 délit 3962 déloyal(e) 715 distorsion 107 erreur 252	{ACTIVITÉ ou ACTION ILLICITES}
	violation de données à caractère personnel, enregistrement abusif de noms de domaines utilisation abusive ou frauduleuse des données à caractère personnel, *conservation de données relatives au trafic porte atteinte au droit fondamental à la confidentialité des communications,* contenu portant atteinte à un droit d'auteur ou à un droit

faute 996 fraude 271 illégal(es/aux) 529 illégalement 87 illicite(s) 3191 illicitement 51 infraction 2249 manquement 505 violation 1498	voisin, *usurpation d'identité numérique constitue un délit,* *internautes qui chargeraient ou téléchargeraient des contenus en infraction avec les droits d'auteur*
	{INFRACTION}
	conservation de données relatives au trafic porte atteinte au droit fondamental à la confidentialité des communications, contenu portant atteinte à un droit d'auteur ou à un droit voisin, *usurpation d'identité numérique constitue un délit,* *internautes qui chargeraient ou téléchargeraient des contenus en infraction avec les droits d'auteur*
contrôle 3468 contrôler 324 défense 723 lutte contre 1044 lutter contre 352 prévenir 503 prévention 668 protection 10 112 protégé(es) 645 protéger 863 veiller à 1468	{ACTIVITÉ DE CONTRÔLE ET DE PROTECTION}
	protection des données à caractère personnel, lutte contre le téléchargement illégal, lutte contre la contrefaçon numérique, lutte contre la cybercriminalité, lutte contre les contenus illicites, lutte contre le piratage en ligne, lutte contre le spam
	{AUTORITÉ}
autorité 6803 organisme 1314 organisation 1538 association 1045 commission 8117	autorité nationale chargée de la protection des données, autorité de contrôle nationale, Autorité de régulation des communications électroniques et des postes, HADOPI, Haute Autorité pour la diffusion des œuvres et la protection des droits sur Internet, Commission nationale de l'informatique et des libertés, CNIL

Tableau 46. Appariement des termes spécifiques aux catégories notionnelles abstraites.

9.1.1 Structuration du schéma du domaine du droit de l'Internet

En catégorisant les unités terminologiques par ensembles sémantiques et en analysant leur environnement contextuel afin de leur attirbuer des étiquettes sémantico-conceptuelles, nous avons constaté qu'il était possible de diviser les données en deux grands blocs thématiques : le vocabulaire de la norme et le vocabulaire de l'action. Ceci correspond partiellement à l'idée de Cornu (2005 : 128), selon laquelle le lexique des termes fondamentaux, pris dans son ensemble, se divise en trois grandes masses, chacune au service d'une finalité primordiale du droit. L'auteur distingue donc le langage de l'établissement du droit (le vocabulaire de la norme, de la règle, de l'investiture ou bien celui du pacte), le langage de l'action (le vocabulaire qui désigne des faits et des actes auxquels est attaché un effet de droit) et le langage du dénouement (le langage de la demande en justice, celui de l'instance, de la preuve, du jugement, de l'exécution). Étant donné que ce dernier ne revêt pas de caractère particulier d'une branche du droit à l'autre et par conséquent n'a aucun impact sur la constitution de la terminologie spécifique au droit de l'Internet, nous n'allons pas le prendre en considération en travaillant sur le schéma du domaine.

Notre analyse a donc fait apparaître deux univers qui composent le champ sémantico-conceptuel du droit de l'Internet: celui de la norme et celui de l'action. Ceci ne fait que confirmer l'idée selon laquelle le droit de l'Internet peut être considéré comme le produit d'une combinatoire de concepts à partir de deux domaines de connaissances, à savoir le droit et les nouvelles technologies de l'information et de la communication. De plus, nous avons constaté que les étiquettes apposées sur les familles de termes (qui, comme nous l'avons vu plus haut, constituent un résultat indirect de l'opération de qualification juridique), correspondent à des catégories conceptuelles abstraites et peuvent être utilisées dans la représentation de la structure notionnelle du domaine. Nous avons donc décidé de trouver un modèle de représentation de connaissances de référence qui pourrait s'adapter facilement à notre démarche, la démarche qui : « *invite à élaborer un type spécifique de champ notionnel, un champ notionnel répondant aux exigences du droit et à sa logique, aux exigences qui sont liées aux marques propres de la pensée juridique.* » (Cornu 200 :193).

Dans ce dessein, nous nous sommes intéressée à l'ontologie de concepts juridiques fondamentaux de LKIF-Core[95] mise en place par Breuker *et al.* 2007 dans le cadre du projet

[95] The Legal Knowledge Interchange Format (LKIF), http://www.estrellaproject.org/lkif-core/

européen ESTRELLA et exploitée par Bourcier et Fernàndez-Barrera (2011) pour la création de la ressource sémantique dans le domaine de la régulation du bruit (voir plus haut). Rappelons que l'ontologie noyau LKIF-Core contient 15 modules réutilisables séparément : *top, place, mereology, time, process, role, action, expression, legal-action, legal-role, norm, modification, rule*. Parmi ces derniers, il y en a cinq qui ont particulièrement attiré notre attention, notamment *action, role, norm, legal-action* et *legal-role*. Le module *action* représente les actions en général, sans rapport avec des thèmes et des situations précises. Il contient des concepts de base tels que : *agent, action, artefact.* L'*agent* renvoie aussi bien à un individu qu'à un groupe (une organisation), et il est doté d'un certain nombre de compétences qui peuvent être exprimées par les *rôles* qui lui sont assignés L'*artefact* est un objet conçu par l'agent dans un but précis et il peut avoir une certaine fonction (Breuker *et al.* 2007 : 29-30). Comme nous pouvons le constater, le module *action* est étroitement lié au module *role* qui définit un ensemble de concepts permettant de catégoriser les objets, les individus et les groupes respectivement en fonction de leur usage ou de leur statut (*ibid.* : 30-31). Le module suivant, *norm*, englobe des classes de concepts considérés comme étant centraux dans l'ontologie juridique, notamment : *norme, obligation, interdiction, autorisation, droits, pouvoirs, violation*. La *norme* exprime, d'un côté une valeur déontique en ce sens qu'elle qualifie l'acceptabilité de certaines actions ou choses. Dun autre côté, elle a un aspect directif car elle engage l'agent à agir selon certaines règles (*ibid.* : 33-37). En ce qui concerne les deux derniers modules, à savoir *legal-action* et *legal-role*, ils constituent l'extension des modules *action* et *role* comprenant des concepts relatifs aux *action, agent, artefact* ou *role* dans le contexte juridique. La Figure 37 présente un fragment de LKIF-Core contenant les modules (accompagnés de certaines classes de concepts) retenus pour notre étude.

Comme nous l'avons mentionné plus haut, chaque module de l'ontologie LIKF-Core comprend un certain nombre de classes de concepts très abstraits. Nous avons donc décidé de les mettre en parallèle avec les catégories dégagées suite à l'analyse des données extraites du corpus *DITerm*. La comparaison a montré des similitudes importantes. Ainsi, le module *action* de LIKF-Core correspond à l'activité liée à l'Internet, et ses concepts de base tels que *agent, action, artefact* renvoient respectivement aux acteurs de l'Internet (utilisateurs et prestataires de services), aux actions et activités menées sur Internet (services offerts via Internet ou activités des internautes) et aux technologies ou objets mis en ligne (qui correspond respectivement aux étiquettes sémantico-conceptuelles suivantes : {ACTEUR DE L'INTERNET : UTILISATEUR}, {ACTEUR DE L'INTERNET : PRESTATAIRE},

{SERVICE EN LIGNE}, {SERVICE EN LIGNE}, {ACTIVITÉ ou ACTION}, {TECHNOLOGIE ou MACHINE}, {OBJET EN LIGNE}. Quant au module *norme*, il englobe les classes de concepts qui ont été identifiés par nous comme des concepts clés du domaine du droit et correspondent aux étiquettes sémantico-conceptuelles telles que : {RÈGLE}, {OBLIGATION}, {DROIT}, {CONFORME À LA LOI}, {AUTORISATION}, {INFRACTION}. Les modules *legal action* et *legal role* renvoient à leur tour aux catégories telles que : {PROCÉDURE}, {AUTORITÉ}, {ACTIVITÉ DE CONTRÔLE ET DE PROTECTION}.

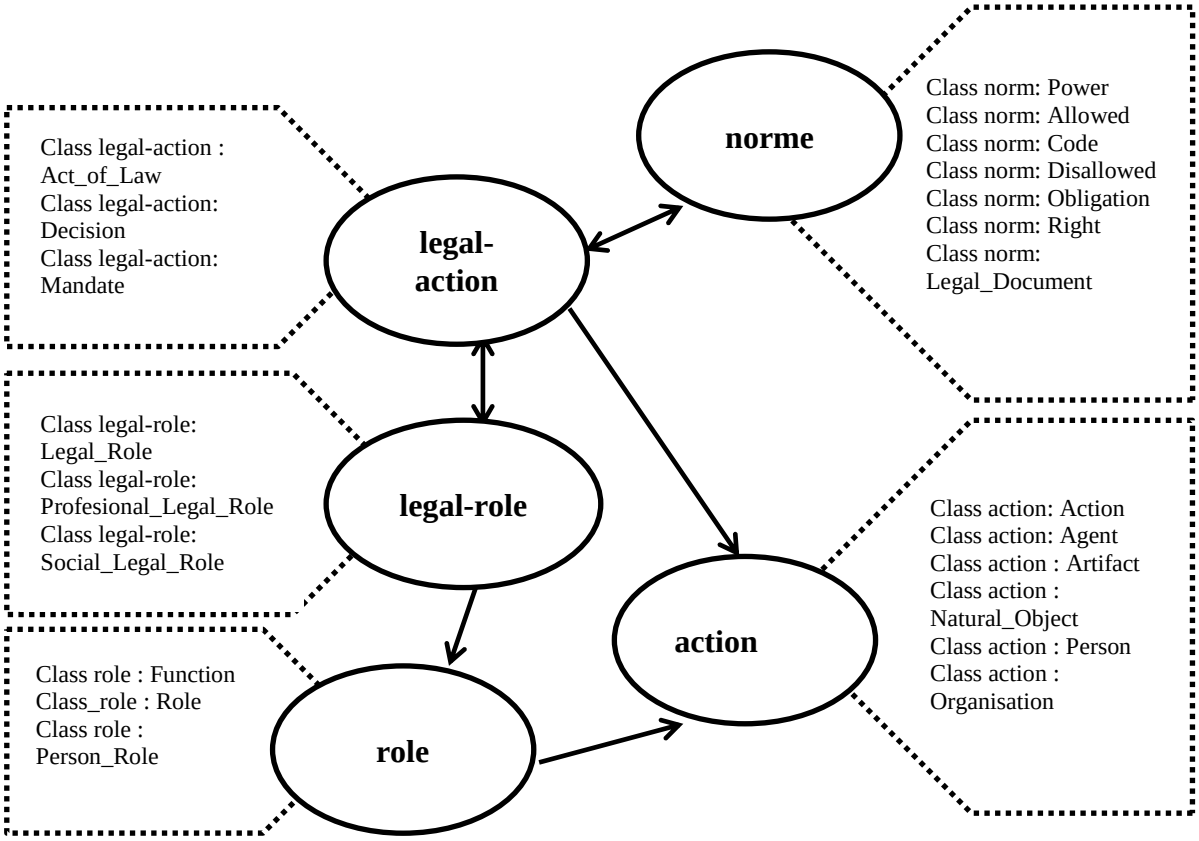


Figure 37. 5 modules de l'ontologie noyau LKIF-Core (*ibid.* : 48) réutilisés dans le projet *DITerm*

Comme nous pouvons le constater, le rapprochement entre les éléments de l'ontologie LIKF-Core et les catégories sémantico-conceptuelles dégagées à la suite des observations faites sur le corpus *DITerm* permet de connecter les données linguistiques à un modèle de représentation de connaissances de référence. Et même si cette démarche demande un certain recul par rapport à la réalité linguistique, elle fait ressortir la structure du domaine du droit de

l'Internet et facilite la systématisation des relations entre les termes. La Figure 38 représente le schéma conceptuel tel qu'il se dégage de notre analyse.

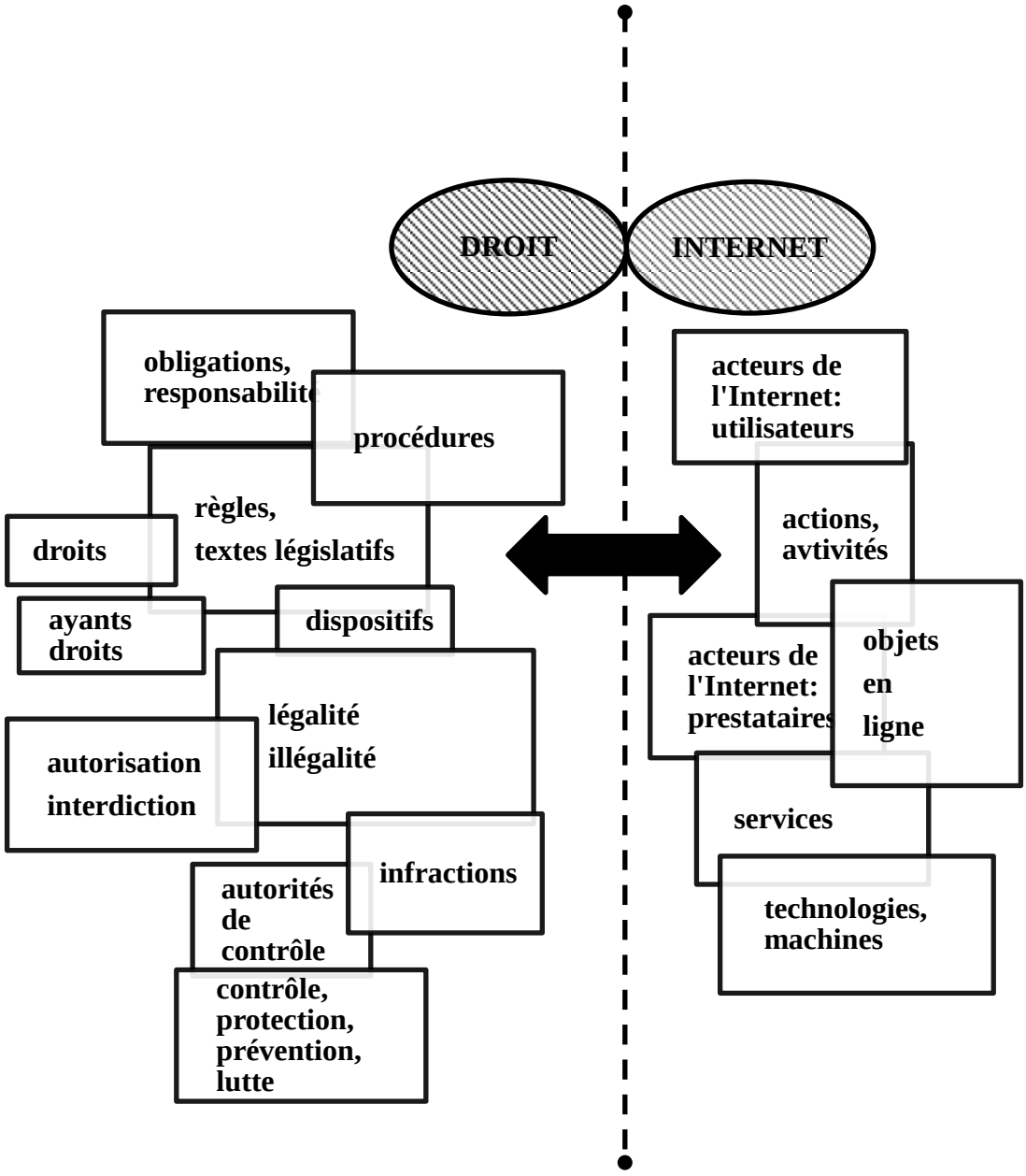


Figure 38. Schéma conceptuel du domaine du droit de l'Internet

9.2 Description des relations conceptuelles des termes au moyen de *cadres*

Le schéma présenté ci-dessus résulte donc du rapprochement entre les 5 modules choisis de l'ontologie LIKF-Core et l'ensemble des catégories sémantico-conceptuelles identifiées à la suite de l'appariement des termes spécifiques aux familles de termes génériques. Le schéma représente deux champs notionnels qui entrent en interaction en tissant de nombreux liens entre les différents éléments de la structure. Chaque élément de cette structure doit être considéré comme un cadre (ou *frame*) sémantico-conceptuel. Nous nous référons ici à la théorie des cadres (*frames* en anglais) entendue au sens large. En effet, comme le souligne Martin (2007 : 411), il existe au moins deux courants dont l'approche est centrée sur la notion de *cadre* (*frame*, *schéma*, *scénario*). Le premier, issu des travaux de Fillmore (voir Fillmore 1982, Fillmore et Baker 2009) exploite la dimension linguistique du concept, le deuxième, attribué à Minsky (1975) est axé sur les aspects purement cognitifs. Comme le soulignent Fillmore et Baker (2009 : 313-320), il est nécessaire de faire la distinction entre les *frames* conceptuels à la Minsky permettant de structurer des informations relatives à une situation donnée indépendamment de la manière dont elles se manifestent à travers la langue et les *frames* sémantiques développés par Fillmore qui relèvent de notre connaissance de la langue et permettent d'associer des formes linguistiques à des structures conceptuelles.

9.2.1 Le terme considéré comme une unité à charge conceptuelle évoquant un *cadre sémantico-conceptuel*

Selon Minsky (1975 : 212) :

« *A frame is a data-structure for representing a stereotyped situation, like being in a certain kind of living room, or going to a child's birthday party. Attached to each frame are several kinds of information. Some of this information is about how to use the frame. Some is about what one can expect to happen next. Some is about what to do if these expectations are not confirmed*. ».

Un *frame* à la Minsky peut donc se définir comme un ensemble de connaissances, de croyances et d'expériences relatives à une situation donnée. Autrement dit, il s'agit d'une structure de données représentant une situation stéréotypée. Chaque *frame* contient plusieurs types d'informations qui permettent de placer cette situation dans un contexte. Comme le souligne Martin en citant les propos de Minsky (1977 : 355 dans Martin 2007 : 412), un cadre conceptuel correspond à une série de questions que l'on doit poser à propos d'une situation hypothétique pour que cette dernière puisse être pleinement comprise. Certaines informations associées à un cadre constituent des éléments fondamentaux et restent inchangées, d'autres, stockées dans les *terminals* (Minsky 197 :212) doivent être précisées ou modifiées selon la situation. En effet, un *frame* peut être décrit comme une matrice divisée en plusieurs compartiments (*slots*) et destinée à accueillir différents types de données (*fillers).* Les *slots* représentent ici les relations conceptuelles générales tandis que les *fillers* renvoient à des données précises permettant de concrétiser et instancier les catégories abstraites. Les cadres sont organisés entre eux dans des systèmes (*frame-systems*) permettant de formuler une idée complète ou corriger les attentes par rapport à une situation donnée. Pour Fillmore et Baker (2009 : 314) :

« *Frames, in this sense, play an important role in how people perceive, remember, and reason about their experiences, how they form assumptions about the background and possible concomitants of those experiences, and even how one's own life experiences can or should be enacted.* ».

Les cadres conceptuels de Minsky permettent donc de structurer la connaissance, d'identifier les relations qu'entretiennent entre eux les différents ensembles de concepts mais aussi de diriger la perception. Comme le souligne Dancette (2011b : 287), la représentation d'un concept au moyen d'un cadre « à la Minsky » donne une idée de la nature de l'information que l'on cherche à « encadrer ». De plus, il convient de souligner que ce type de *frames* rend compte de la tendance qu'ont les concepts (et indirectement les termes) à se réunir dans l'esprit du locuteur pour former ce que Cornu (2005 : 202) appelle des groupes d'intervention. En effet, le recours à un concept correspondant à un cadre de connaissance active toute une série d'associations permettant d'accéder à un ensemble de termes liés directement ou indirectement à ce concept. Cette situation peut être visualisée par des graphes que Dancette (2011b : 287) propose de désigner sous l'appellation de *carte conceptuelle* et qui illustrent cet

enchaînement d'idées. Remarquons que le fait qu'un concept en suggère un autre et un terme en évoque un autre est un phénomène qui doit absolument être pris en compte dans un modèle de description des unités terminologiques. La théorie des cadres de Minsky offre une méthode efficace pour représenter ces relations.

Compte tenu de tous les éléments mentionnés plus haut, nous considérons que le système de *frames* à la Minsky constitue un outil intéressant de représentation de connaissances en terminologie. Comme nous l'avons vu précédemment, les groupes de termes extraits du corpus *DITerm* et rassemblés sous les étiquettes sémantico-conceptuelles correspondent à des catégories conceptuelles abstraites qui pourraient, selon nous, être représentées au moyen des cadres. En effet, chacune de ces catégories évoque une situation donnée qui devraient être décrite en prenant en compte un contexte particulier et les relations qui y sont liées. La question se pose pourtant de savoir si un tel model conceptuel peut s'adapter à un projet qui est mené dans une perspective linguistique. En cherchant une réponse à ce problème d'ordre méthodologique, nous nous sommes intéressée au projet lexicographique de Martin (2007, 2008), dans lequel l'auteur propose d'adapter les *frames* de Minsky à la description du vocabulaire. En s'inspirant de la structure *slot-filler* de la théorie des cadres (voir plus haut), Martin met en place un système de cadres sémantico-conceptuels (*conceptual semantic frames*) destinés à représenter le sens mais aussi la charge conceptuelle des unités lexicales. Comme le souligne l'auteur (2007 : 412), les cadres sémantico-conceptuels ne font pas la distinction entre les informations linguistiques et encyclopédiques. L'objectif est de refléter la connaissance subjective et stéréotypée des locuteurs telle qu'elle ressort de leur usage de la langue. Les cadres sémantico-conceptuels offrent donc un modèle définitionnel des unités lexicales qui est basé sur la description de leur relations avec d'autres unités (considérés comme des concepts lexicalisés).

À l'instar de Martin (*ibid.*), nous proposons donc de considérer les catégories abstraites dégagées à la suite de l'analyse des données textuelles comme des cadres sémantico-conceptuels permettant de rassembler les informations aussi bien linguistiques qu'encyclopédiques. Rappelons que ces catégories constituent les principaux éléments de la structure conceptuelle du domaine du droit de l'Internet (voir la Figure 38). L'approche basée sur les *frames* permettra donc de rendre compte des relations que ces éléments entretiennent avec d'autres catégories de la structure du domaine et par conséquent guidera la description des termes appartenant à ces classes conceptuelles. Prenons l'exemple de la catégorie des

acteurs de l'internet : prestataires. Nous admettons (mais à cette étape, il s'agit juste d'une hypothèse), que la classe regroupant les acteurs de l'Internet peut être reliée à d'autres classes constituant le schéma du domaine du droit de l'Internet par le biais d'un certain nombre de relations. La Figure 39 représente cette situation.

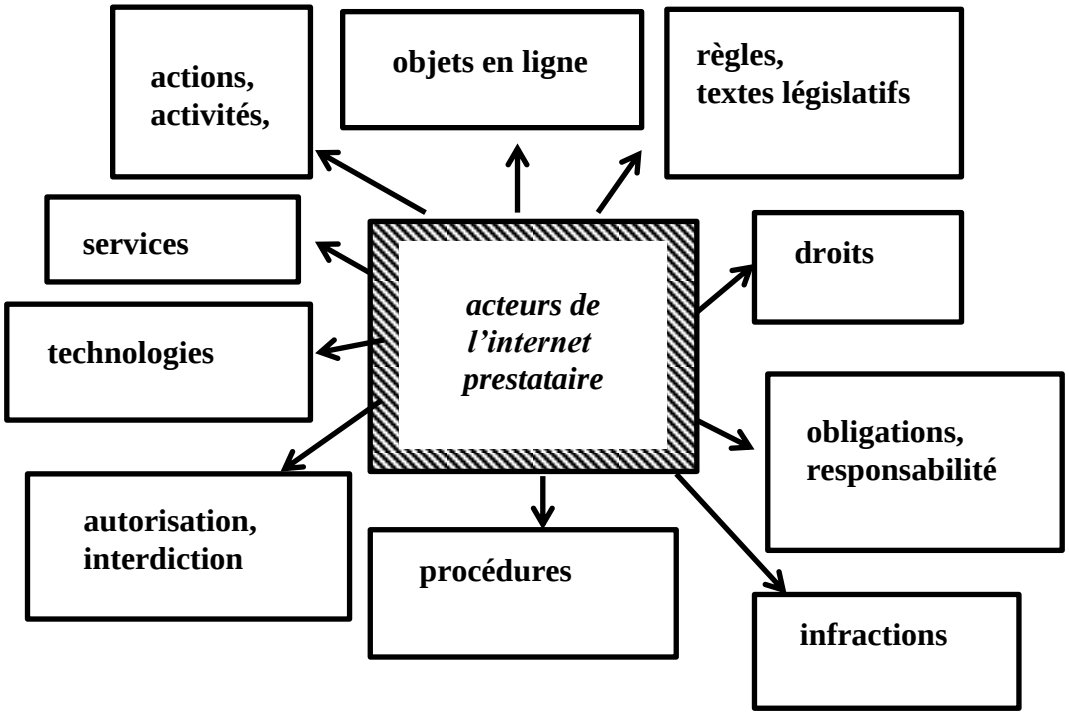


Figure 39. Catégorie *acteurs de l'Internet : prestataires* et ses relations avec d'autres catégories

.

Comme le montre la figure ci-dessus (rappelons que, pour l'instant, il ne s'agit que de relations hypothétiques, données à titre d'exemple), nous supposons que chaque acteur de l'Internet se caractérise par une activité qui lui est propre, que pour exercer cette activité, il fait appel à une technologie, et qu'au titre de son activité, il est soumis à une réglementation, à des procédures, relève d'un régime de responsabilité, est tenu à un certain nombre d'obligations, etc. Nous admettons donc qu'il est possible d'assimiler une catégorie conceptuelle abstraite à un cadre sémantico-conceptuel qui se caractérise par un certain nombre de relations avec d'autres cadres sémantico-conceptuels. La Figure 40 représente notre démarche (notons que la liste complète des cadres sémantico-conceptuels utilisé dans notre projet est fournie à l'Annexe IV).

Catégories conceptuelle abstraite : **acteurs de l'Internet : prestataires →**	Cadre sémantico-conceptuel correspondant : {ACTEUR DE L'INTERNET : PRESTATAIRE}
Catégories conceptuelles abstraites liées à la catégorie : **acteur de l'Internet : prestataires →**	Cadres sémantico-conceptuels liés : {ACTIVITÉ ou ACTION} {SERVICE INTERNET} {OBJET EN LIGNE} {RÈGLE} {AUTORITÉ JURIDIQUE} {TEXTE LÉGISLATIF} {PROCÉDURE} {DISPOSITIF TECHNICO-JURIDIQUE} {OBLIGATION} {ACTIVITÉ ou ACTION ILLICITES} {OBJET ILLICITE EN LIGNE}

Figure 40. Passage d'une catégorie conceptuelle abstraite vers un cadre sémantico-conceptuel.

Ainsi, pour décrire un terme appartenant à cette catégorie, il faut le situer dans un contexte défini par le cadre sémantico-conceptuel correspondant et prendre en compte toutes les relations qui sont propres à ce cadre. Si nous souhaitons définir un terme correspondant au cadre sémantico-conceptuel. {ACTEUR DE L'INTERNET : PRESTATAIRES} et décrire son fonctionnement dans les textes du droit de l'Internet (qui constituent un reflet de l'univers conceptuel du domaine), nous devons prendre en compte un certain nombre d'éléments qui caractérisent la situation évoquée par ce cadre. Cependant, compte tenu du caractère de notre travail, les relations propres aux cadres sémantico-conceptuels retenus ne doivent pas être définies *a priori*, par rapport à un modèle de représentation de connaissance déjà existant, mais identifiées à partir des données textuelles. En effet, nous nous intéressons à la dimension conceptuelle des termes pour pouvoir expliquer leur comportement dans les textes et non pas pour les rattacher à la structure conceptuelle du domaine. Il convient pourtant de souligner que le repérage des relations permettant de caractériser des cadres sémantico-conceptuels propres au domaine du droit de l'Internet n'est pas une tâche facile. Nous considérons tout de

même qu'un examen approfondi de l'environnement contextuel des termes reliés à ces cadres permettra de dégager des liens spécifiques à chaque situation et de les généraliser par la suite.

9.2.2 Annotation de l'environnement contextuel des termes au moyen des *cadres sémantiques* à la Fillmore comme moyen de structuration des relations conceptuelles au sein du domaine du droit de l'Internet

À cet effet, nous nous sommes tournée vers un modèle d'annotation sémantique inspiré de *FrameNet* (Baker 2009, Fillmore *et al.* 2003), projet lexicographique développé à l'Université de Berkeley qui constitue une application de la sémantique des cadres de Fillmore (Fillmore 1982, Fillmore et Baker 2009, Fontenelle 2009). Rappelons que les cadres sémantiques, contrairement aux *frames* de Minsky, sont ancrés dans la réalité linguistique. En effet, selon Fillmore, le sens des mots peut être décrit par rapport à ce qu'il a appelé *semantic frames,* à savoir des représentations schématiques des structures conceptuelles correspondant aux expériences, pratiques, croyances des membres d'une communauté linguistique donnée :

« *The central idea of Frame Semantics is that word meanings must be described in relation to semantic frames – schematic representations of the conceptual structures and patterns of beliefs, practices, institutions, images, etc. that provide a foundation for meaningful interaction in a given speech community.* »

(Fillmore *et al.* 2003:235).

Fillmore postule qu'on ne peut comprendre la signification de bien des mots, de manière optimale, qu'en tenant compte du contexte événementiel ou situationnel dans lequel ils s'inscrivent (Baker 2009 : 32). Les mots sont ainsi interprétés par le biais des cadres conceptuels évoqués dans l'esprit des interlocuteurs.

Comme le souligne Fontenelle (2009 : 163) le *frame* de Fillmore se définit donc comme une sorte de scénario ou un schéma qui sous-tend l'utilisation d'un item lexical ainsi que sa compréhension. En effet, la sémantique des cadres part du principe selon lequel le lexique est construit sur la base de « connaissances d'arrière-plan » (background knowledge) (L'Homme 2015 : 32) :

« *[…] Frame Semantics is the study of how linguistic forms evoke or activate frame knowledge, and how the frames thus activated can be integrated into an understanding of the passages that contain these forms.* »

(Fillmore et Baker 2009 : 317).

Chaque cadre peut être caractérisé par un certain nombre de protagonistes (la notion de *protagoniste* doit être prise au sens large du terme car elle recouvre des éléments qui ne sont pas nécessairement humains), appelés *éléments du cadre*. Comme le souligne Fontenelle (2009 : 163), la sémantique des cadres se donne pour but de consigner la façon dont la langue relie les éléments du cadre aux constituants syntaxiques qui dépendent de façon syntagmatique des items lexicaux qui évoquent ce cadre. Selon l'auteur (*ibid.*), l'objectif final est de décrire les constellations possibles gravitant autour de ces items lexicaux.

9.2.2.1 Panorama des ressources basées sur la théorie des cadres

Le projet *FrameNet* est une ressource lexicale qui propose de décrire le vocabulaire anglais selon les principes de la théorie de Fillmore. En effet, c'est une représentation fidèle des concepts de la sémantique des cadres à tel point que les deux sont souvent confondus. À ce jour, la base de données *FrameNet* [96], accessible sur l'Internet, contient plus de 10 000 unités lexicales et environs 1 000 cadres illustrés par quelque 200 000 exemples annotés manuellement. Chaque acception d'un mot est donc associée à un cadre particulier et chaque cadre regroupe plusieurs unités lexicales. Les *frames* sont décrits au moyen de scénarios conceptuels comprenant un certain nombre d'éléments (*Frame Elements FE*) qui peuvent correspondre à des entités, des processus, des propriétés, etc. Les différents éléments des cadres sont étiquetés, ce qui permet de savoir comment ils sont réalisés dans le langage courant. Il convient de souligner que *FrameNet* fait une distinction entre éléments obligatoires (*core Frame Elements*) et facultatifs (*non-core Frame Elements*). Comme le souligne Baker (2009 : 33), les éléments sont regroupés dans la partie noyau (*core)* lorsqu'ils sont inhérents à la définition du cadre et ont tendance à apparaître dans des positions syntaxiques fondamentales ; les éléments du cadre de la partie périphérique (*non-core*) ne sont pas spécifiques à ce cadre ; ils sont communs à toutes sortes d'autres cadres. Les phrases annotées

[96] La ressource *FrameNet* (ainsi que la documentation concernant le projet) est disponible à l'adresse suivante : https://FrameNet.icsi.berkeley.edu/fndrupal/.

qui viennent compléter la description des *frames* permettent de connecter le niveau de description conceptuel à des réalisations linguistiques véritables. Ces annotations font apparaître plusieurs façons d'instancier les éléments du cadre. À titre d'illustration, nous proposons de reproduire, ci-dessous, le cadre *Prohibiting*.

Prohibiting

Definition:

In this frame a State_of_affairs is prohibited by a Principle. Raising constructions are common in this frame. In this frame the Principle which prohibits the State_of_affairs is not an agent who denies permission to a specific individual or group of individuals, and thus differs from the Authority in the Deny_permission frame.

Guns were PROHIBITED in the airport.

Code 1425 BANS large trucks in the tunnel.

FEs:

Core:

Principle [Prin] **Semantic Type:** Artifact	A moral, legal, or social norm which rules a State_of_affairs to be inadmissible behavior or an outcome of behavior. The Establishment Clause of the First Amendment plainly PROHIBITS the establishment of a national religion by Congress.
State_of_affairs []	

Non-Core :

Circumstances [cir]	The Circumstances are the conditions under which the State_of_affairs is prohibited.
Explanation [Exp]	The reason for which the State_of_affairs is prohibited.
Place [Place] **Semantic Type:** Locative_relation	The Place is where the prohibiting occurs.
Time [Time] **Semantic Type:** Time	The Time is when the prohibiting occurs.

Lexicla Units: *ban.n, ban.v, bar.v, forbid.v, outlaw.v, prohibit.v, prohibition.n, proscribe.v*

Figure 41. Frame Prohibiting extrait de la base lexicale FrameNet (consulté le 08.04.2016)

Comme le souligne L'Homme (2015 : 30), le modèle de Fillmore, bien que conçu à l'origine pour la langue de manière générale, revêt des attraits évidents du point de vue du lexique spécialisé, dont celui de rendre possible une connexion entre des descriptions linguistiques (basées sur un corpus) et un niveau de représentation plus abstraite, celui du cadre, se voulant une modélisation conceptuelle d'une situation évoquée par des ensembles d'unités lexicales. En effet, ces dernières années, la sémantique des cadres ainsi que la méthodologie développée dans le cadre du projet *FrameNet* ont intéressé un certain nombre de terminologues qui ont exploité le modèle pour élaborer leurs ressources terminologiques. Ces travaux devraient faire l'objet d'une étude approfondie, menée à part, que nous ne pouvons pas offrir ici. Nous voulons pourtant mentionner brièvement quelques travaux qui ont particulièrement attiré notre attention. L'application la plus aboutie de la sémantique des cadres à un domaine spécialisé qui devient en quelque sorte un modèle de référence a été proposée par Schmidt (2009). L'auteur a conçu une ressource lexicale multilingue (allemand, anglais, français), portant sur le domaine du football, appelée *Kicktionary*[97]. Ce dictionnaire électronique contient environ 2 000 unités organisées dans une structure hiérarchisée comprenant 16 scènes représentant les différents aspects du jeu. Chaque scène inclut plusieurs cadres qui comprennent un certain nombre d'éléments. Par exemple, la scène **Goal** renvoie à 10 cadres : Award_Goal (11 LUs), Celebrate_Goal (5 LUs), Concede_Goal (7 LUs), Convert_Chance (8 LUs), Goal (85 LUs), Multiple_Goals (8 LUs), Overcome_Goalkeeper (7 LUs), Own_Goal (4 LUs), Prepare_Goal (6 LUs), Score_Goal (7 LUs). Le cadre Goal, qui fait partie de la scène **Goal**, contient, quant à lui, 14 éléments qui contribuent à sa définition, à savoir : SCORER, SCORER_TEAM, SHOT, RESULTING_SCORE, SOURCE, PREPARING_EVENT, PREVIOUS_SCORE, PART_OF_BODY, MOVING_BALL, TARGET, GOAL, CONCEDING_TEAM, BALL, PATH. Il convient de souligner que Schmidt, en proposant l'organisation en scènes et cadres, se réfère aux premiers travaux de

[97] *Kicktionary* est accessible sur l'Internet à l'adresse suivante : http://www.kicktionary.de/index.html.

Fillmore où ce dernier établit une distinction nette entre *scene*, une entité purement conceptuelle et *frame*, structure de nature plutôt linguistique : « *Whereas a scene is defined in terms of pieces of abstract (and possibly non-linguistic) knowledge, the notion of a frame is concerned with the properties of concrete linguistic means of expressing this kind of knowledge.* » (Schmidt 2009 : 103)

JuriDiCo[98] (Pimentel 2011, 2012) est un autre projet basé sur la sémantique des cadres qui paraît intéressant de point de vue de notre étude. En effet, c'est une base de données terminologique portant sur le domaine du droit. A l'heure actuelle, *JuriDiCo* recense une centaine de verbes juridiques en anglais, portugais et français. Les verbes sont décrits au moyen de 35 cadres propres aux procédures de jugement comme par exemple : **Verdict, Crime, Contesting, Investigate, Appellate, Law applicability, Issues**[99], etc. Dans son projet, Pimentel propose de définir les éléments des cadres à partir de la structure actancielle des verbes étudiés en supposant qu'il est possible de mettre en parallèle les actants et les participants obligatoires (*Core Elements*) des cadres évoqués par les verbes.

Le troisième projet que nous voudrions présenter brièvement ici est le *Framed DiCoEnviro*[100] (L'Homme *et al.* 2014, L'Homme 2015), une versions « en frames » du *DiCoEnviro*, décrit en détail dans le premier chapitre de ce travail. Rappelons que le *DiCoEnviro* n'a pas été conçu à l'origine dans la perspective de rendre compte des cadres sémantiques. Les auteurs ont toutefois utilisé les descriptions de la version originale (notamment la description de la structure actancielle des termes et les annotations contextuelles) pour pouvoir identifier des cadres sémantiques correspondant au domaine de l'environnement. Il convient également de souligner que la méthodologie adoptée dans le *Framed DiCoEnviro* diffère légèrement de celle proposée dans *FrameNet* (L'Homme 2015 : 33). En effet, contrairement à *FrameNet* où les étiquettes utilisées pour distinguer les éléments des cadres sont définies en fonction d'un cadre spécifique, *Framed DiCoInfo* propose un ensemble d'étiquettes pouvant s'appliquer à un grand nombre de termes. Selon nous, ceci est une solution intéressante contribuant à la structuration sémantico-conceptuelle du domaine de l'environnement.

[98] *JuriDico* est accessible sur l'Internet à l'adresse suivante : http://olst.ling.umontreal.ca/cgi-bin/juridico/search.cgi.

[99] Une partie des cadres utilisés dans *JuriDico* s'inspirent des *frames* proposés dans FameNet

[100] *Framed DiCoEnviro* est accessible sur l'Internet à l'adresse suivante : http://olst.ling.umontreal.ca/dicoenviro/framed/index.php.

9.2.2.2 Proposition de description de l'environnement contextuel des termes du *DITerm* au moyen des cadres sémantiques à la Fillmore

Revenons maintenant à notre projet. En analysant les cadres sémantiques décrits dans *FrameNet* et *JuriDiCo,* nous nous sommes aperçue qu'un certain nombre de *frames* pourraient être exploités dans le cadre de notre travail. Nous voudrions cependant attirer l'attention sur une question fondamentale. Contrairement aux projets évoqués plus haut, projets qui font appel aux cadres sémantiques afin de décrire les unités qui évoquent ces cadres, nous proposons d'utiliser ces derniers afin de mettre en évidence les relations qui caractérisent les termes que nous avons retenus. En effet, l'analyse de l'environnement contextuel de ces termes a montré que la plupart d'entre eux correspondent à des éléments des cadres évoqués par d'autres items. Cependant, pour comprendre notre démarche, il faut bien distinguer deux notions auxquelles nous avons décidé de faire appel, à savoir les cadres sémantico-conceptuels et les cadres sémantiques. En effet, selon nous, il s'agit de deux types de *frames* qui renvoient à deux niveaux de conceptualisation différents. Comme nous l'avons souligné plus haut, chaque terme de notre base de données terminographique peut être associé à une catégorie plus abstraite représentée au moyen d'un cadre sémantico-conceptuel auquel nous avons associé une étiquette sémantico-conceptuelle.. Pour nous, un cadre sémantico-conceptuel est une structure conceptuelle (dégagée sur la base des critères aussi bien sémantiques que conceptuels) qui permet de rattacher un terme donné à un schéma notionnel du domaine. Nous considérons que l'organisation des termes en cadres sémantico-conceptuels facilite la description des liens qui relient ces termes à d'autres termes. En effet, nous partons de l'hypothèse que chaque cadre sémantico-conceptuel est lié à un certain nombre d'autres cadres sémantico-conceptuels, c'est-à-dire, il se caractérise par un ensemble défini de relations. Ces dernières, bien qu'établies en fonction des éléments extralinguistiques (juridiques), se manifestent, d'une manière ou d'une autre, dans le discours. Quant aux cadres sémantiques, nous les appréhendons dans leur dimension linguistique, comme des structures schématiques permettant de connecter des réalisations linguistiques à un niveau de description conceptuel. Nous supposons que l'analyse contextuelle des unités terminologiques menée en termes des cadres sémantiques permet de refléter les relations propres aux cadres sémantico-conceptuels auxquelles appartiennent ces unités.

En effet, nous avons remarqué que les termes (qui sont associés à des structures de nature conceptuelle et cognitive que nous appelons cadres sémantico-conceptuels et qui correspondent à un niveau de conceptualisation plus abstrait) se réalisent dans le discours dans des combinaisons de formes linguistiques propres au langage juridique qui peuvent être analysées au moyen de structures schématiques appelées cadres sémantiques. L'analyse de l'environnement contextuel des termes a montré que ces derniers font partie des constellations gravitant autour de différentes unités évoquant des cadres typiques du domaine du droit. Il s'agit notamment :

a) des unités correspondant à des verbes juridiques: *prescrire, commettre, sanctionner, imposer, obliger, prévoir, autoriser, interdire, respecter, devoir, soumettre, établir,* etc (les verbes en question appartiennent à la catégorie 7 de notre classification des termes– voir la section 7.1)

Ex. *Le* **fournisseur d'hébergement [terme]** ***est*** *donc* ***soumis*** **[unité évoquant un cadre sémantique]** *à une procédure de notification et de retrait de contenu illicite qu'il va devoir respecter pour éluder sa responsabilité.* [RDLI_2009]
ou
Le fournisseur d'hébergement est donc ***soumis*** **[unité évoquant un cadre sémantique]** *à une procédure de notification et de retrait de* **contenu illicite [terme]** *qu'il va devoir respecter pour éluder sa responsabilité.* [RDLI_2009]

b) des unités correspondant à des noms prédicatifs considérés comme étant porteurs des notions fondamentales du droit tels que : *responsabilité, obligation, droit, injonction, illégal, conforme*, etc. (catégorie 4 de notre classification – voir la section 7.1).

Ex. *Le Parlement européen a, en effet, adopté, le 24 novembre dernier, la nouvelle directive-cadre sur les télécommunications (« paquet télécoms ») procédant à la révision de la directive précitée du 12 juillet 2002 qui établit notamment* ***l'obligation*** **[unité évoquant un cadre sémantique]** *d'obtenir le* **consentement préalable [terme]** *des utilisateurs avant l'installation de « cookies » sur leur ordinateur* [RDLI_2010]

Comme nous l'avons souligné précédemment (voir la section 7.1.2), les unités en question n'ont pas été retenues comme termes spécifiques du domaine du droit de l'Internet car elles sont communes à toutes les branches du droit. Cependant, ces unités revêtent pour nous un

intérêt particulier car elles permettent d'identifier des relations qui relient un terme donné à d'autres termes du droit de l'Internet en contribuant ainsi à la structuration de la connaissance dans notre domaine de spécialité. Remarquons entre parenthèses (voir les exemples cités plus haut), que les termes que nous decrivons peuvent se trouver dans un contexte immédiat d'une unité évoquant un cadre sémantique (« **fournisseur d'hébergement [terme]** ***est*** *donc* ***soumis*** **[unité évoquant un cadre sémantique] »**), ou bien dans un contexte plus éloigné (***soumis*** **[unité évoquant un cadre sémantique]** *à une procédure de notification et de retrait de* **contenu illicite**) ce qui rend l'analyse plus complexe.

Nous avons décidé par la suite de rapprocher les deux types d'unités évoquant des cadres sémantiques typiques du langage juridique et repérées dans l'entourage des termes qui nous intéressent avec des données provenant de *FrameNet* et de *JuriDiCo*. Le Tableau 47 présenté ci-dessous illustre notre démarche.

unités évoquant des « cadres juridiques » extraits du corpus *DITerm*	[CADRES SEMANTIQUES] FRAMENET *+ unités lexicales associées*	[CADRES SEMANTIQUES] JURI*DICO* *+ termes associés*
autoriser, autorisation, permettre, permission, interdiction, interdire	**Permitting** In this frame a State_of_affairs is permitted by a Principle. Raising constructions are common in this frame. In this frame the Principle which sanctions the State_of_affairs is not an agent who grants permission to a specific individual or group of individuals, and thus differs from the Grantor in the Grant_permission frame. Termes anglais: allow.v, entitle.v, permit.n, permit.v, sanction.v **Prohibiting** In this frame a State_of_affairs is prohibited by a Principle. Raising constructions are common in this frame. In this frame the Principle which prohibits the State_of_affairs	**Authorization** A judicial decision or other source of law or an authority (law) allows or does not allow a certain person (Protagonist) to engage in a certain kind of behavior (Act). Termes: authorize, permit, preclude, prohibit

	is not an agent who denies permission to a specific individual or group of individuals, and thus differs from the Authority in the Deny_permission frame. Unités lexicales: ban.n, ban.v, bar.v, forbid.v, outlaw.v, prohibit.v, prohibition.n, proscribe.v	
illégal (es/aux) *illégalement* *illicite* *illicitement* *conforme/conformes,* *légal(es)* *légalement* *légitime(s)* *licéité* *licence* *licite(s)*	**Legality** Words in this frame describe the status of an Action with respect to a Code of laws or rules. An Object may also be in violation or compliance of the Code by virtue of its existence, location or possession. Unités lexicales: *criminal.a, fair.a, illegal.a, illicit.a, lawful.a, legal.a, legitimate.a, licit.a, permissible.a, prohibited.a, unlawful.a, wrong.a, wrongful.a, wrongly.adv*	
imposer, *obliger,* *faire obligation,* *faire peser une obligation,* *faire injonction,* *mettre à la charge de qqn une obligation*	**Imposing_obligtion** A Duty is imposed on a Responsible_party according to a Principle which regulates how the Responsible_partyshould respond to a Situation. The Situation may be expressed metonymically by reference to an Obligator, whose action invokes the Principle. It is only rarely the case that the Principle and the Situation/Obligator are both expressed overtly Unités lexicales: *bind.v, charge.n, charge.v, commit.v, obligate.v, oblige.v, pledge.v, require.v*	[Order] A Judge (or Judgment or Law) imposes an order or obligation (Duty) on somebody (Protagonist). Termes: *commit, impose, order, require*

Tableau 47. Comparaison de données provenant du corpus *DITerm* avec les cadres répertoriés dans FrameNet et Juri*DiCo* (consultés le 15.01.2016)

La comparaison des données répertoriées dans *FrameNet* et *JuriDiCo* avec celles extraites du corpus *DITerm* nous a permis de dégager un certain nombre de *frames* pouvant être exploités dans le cadre de notre projet (une liste de tous les cadres sémantiques utilisés dans ce travail est fournie à l'Annexe V). Néanmoins, il convient de souligner que les critères sémantiques pris en compte au moment de la définition de ces cadres dans *DITerm* étaient beaucoup moins stricts que ceux adoptés dans *FrameNet* ou *JuriDiCo*. Nous allons essayer de l'expliquer à travers l'exemple du cadre **OBLIGATION**. En effet, l'analyse des *frames* relatifs à la situation d'obligation renvoie, dans FrameNet, à trois cadres différents, soit : **Being_obligated**, **Being_obligatory** et **Imposing_obligtion**. Chaque cadre décrit cette situation dans une autre perspective permettant de dégager trois structures syntaxico-sémantiques et de différencier les sens des unités associées à ces *frames*. La Figure 41 reproduit les définitions des trois cadres ainsi que des phrases annotées qui font apparaître plusieurs façons d'instancier ces cadres :

Being_obligated :
Under some Condition, usually left implicit, a Responsible_party is required to perform some Duty. If they do not perform the Duty, there may be some undesirable Consequence, which may or may not be stated overtly.
Ex.
It **is** my DUTY to fight any attack on the Brotherhood .

This country has the RESPONSIBILITY to support its citizens' right to express themselves.

I am OBLIGATED to pay or they'll double the fine.

Being_obligatory :

Under some Condition, usually left implicit, a Duty needs to be fulfilled by a Responsible_party. If the Duty is not performed, there may be some undesirable Consequence for the Responsible_party, which may or may not be stated overtly.
Compare this frame to the Being_obligated frame.
It would BEHOOVE us to comprehend and personalize what Jesus had to tell his disciples

Ex.
Pressed too much at the OBLIGATORY BBQ. INI

When I was in HS, marching band *was* MANDATORY in order to participate in any of the other ensembles INI

Should insurance *be* COMPULSORY?

Imposing_obligation

A Duty is imposed on a Responsible_party according to a Principle which regulates how the Responsible_partyshould respond to a Situation. The Situation may be expressed metonymically by reference to an Obligator, whose action invokes the Principle. It is only rarely the case that the Principle and the Situation/Obligator are both expressed overtly.

Ex.
They escaped total Soviet invasion and occupation only by entering into a separate agreement thatOBLIGATED them to military action against the retreating German armies.

The lease agreements BOUND them to make rent payments to Homeowners Rescue.

It was also discovered that with out her knowledge, he had COMMITTED her to a new TV series and he had already taken an advance on the money.

The Generality's invitation to give a conference on the theme OBLIGATED me to study Gaudí's work even more.

Figure 42. : Cadres Being_obligated, Being_obligatory et Imposing_obligtion, extraits de la base lexicale FrameNet (consulté le 15.01.2016)

Cependant, contrairement à *FrameNet* et *JuriDiCo*, le recours aux *frames* sémantiques, dans le cas de notre travail, n'a pas pour but de fournir une description des unités qui évoquent ces cadres car, comme nous l'avons démontré plus haut, il s'agit des termes génériques communs à toutes les branches du droit et non pas spécifiquement au droit de l'Internet. En effet, nous faisons appel au modèle des cadres sémantiques pour pouvoir systématiser l'annotation des contextes dans lesquels apparaissent les termes spécifiques au domaine du droit de l'Internet et en même temps mettre en évidence les relations que ces derniers entretiennent avec d'autres termes du domaine. Nous avons donc décidé de considérer un cadre sémantique comme une structure qui englobe différents aspects de la même situation. Ainsi, le *frame* **OBLIGATION** utilisé dans notre projet correspond à quatre formules d'annotation :

OBLIGATION

Un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou un {ACTEUR DE L'INTERNET : UTILISATEUR} est / n'est pas obligé de réaliser une {ACTIVITÉ ou ACTION} concernant un {OBJET EN LIGNE}, une {PROCEDURE} ou un {DISPOSITIF TECHNOCO-JURIDIQUE} dans certaines circonstances.

OU

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE LÉGISLATIF} oblige un {ACTEUR

DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} à réaliser une {ACTIVITÉ ou ACTION} concernant un {OBJET EN LIGNE}, une {PROCEDURE}, un {DISPOSITIF TECHNOCO-JURIDIQUE dans certaines circonstances.

OU

L'obligation d'une {ACTIVITÉ ou ACTION} un {OBJET EN LIGNE}, une {PROCEDURE}, un {DISPOSITIF TECHNICO-JURIDIQUE appartient à un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou un {ACTEUR DE L'INTERNET : UTILISATEUR}.

OU

Une {ACTIVITÉ ou ACTION}, une {PROCEDURE}, un {DISPOSITIF TECHNICO-JURIDIQUE} est une obligation d'un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou un {ACTEUR DE L'INTERNET : UTILISATEUR}.

Figure 43. Cadre sémantique OBLIGATION dans *DITerm*

Remarquons que **OBLIGATION** comprend trois éléments de cadre centraux, notamment: a) un protagoniste qui doit remplir un devoir et qui correspond, dans notre cas, aux catégories {ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR; b) un devoir à remplir qui renvoie aux catégories {ACTIVITÉ ou ACTION}, {ACTIVITÉ ou ACTION} concernant un {OBJET EN LIGNE}, {PROCEDURE}, ou {DISPOSITIF TECHNICO-JURIDIQUE}; c) une autorité juridique ou une règle qui imposent le devoir en question et qui correspondent aux catégories : {AUTORITÉ JURIDIQUE} ou {TEXTE LÉGISLATIF} L'élément **dans certaines circonstances**, quant à lui, doit être considéré comme un élément périphérique. Comme nous pouvons le constater, le cadre du domaine du droit de l'Internet a une portée nettement restreinte : les participants sont limités à ceux qui interviennent dans les questions juridiques liées aux nouvelles technologies. Ce caractère plus spécialisé des cadres facilite la mise en évidence de certains liens d'ordre sémantico-conceptuels. Remarquons que les étiquettes utilisées pour distinguer les éléments du cadre correspondent à celles attribuées précédemment aux catégories regroupant les différentes familles de termes (appelés *cadres sémantico-conceptuels*). Nous essayons par-là d'établir une connexion entre le schéma notionnel du domaine de l'Internet et les réalisations linguistiques des relations qui existent entre les termes. En effet, nous considérons que le recours au système d'étiquetage basé sur le

concept des *cadres sémantico-conceptuels* lors du processus d'annotation contextuelle au moyen des *cadres sémantiques* permet de rendre compte de la double dimension du terme (linguistique et conceptuelle) et de rapporcher deux types de démarches, notamment une approche conceptuelle et une approche lexico-sémantique.

Ainsi, l'identification des cadres sémantiques dans le cadre de notre projet a comme objectif principal de systématiser l'annotation de l'environnement contextuel des termes et de capturer ses relations d'ordre sémantico-conceptuelles. En effet, ce type d'annotation a été choisi pour nous permettre de visualiser l'interaction des termes étudiés avec d'autres unités et de définir les types de liens qui les unissent. Comme le montrent les phrases annotées ci-dessous (Figure 44), la place et le rôle des termes au sein des cadres sémantiques varient d'un contexte à l'autre. Les deux premiers exemples présentent la situation où le terme instancie un des éléments de cadre obligatoires. L'exemple suivant est un cas particulier. Il s'agit d'un terme complexe dont une des composantes: *obligation,* évoque le cadre en question et les deux autres : *surveillance générale* correspondent à un des participants obligatoires de la situation évoquée par *obligation*. Ceci est lié à la propriété de nombreux termes complexes qui ont un sens compositionnel. Quant aux deux derniers exemples, ils montrent que l'annotation au moyen des cadres sémantiques permet de rendre compte de relations indirectes et complexes. En effet, dans les deux cas de figure, les termes n'apparaissent pas dans l'environnement contextuel direct des unités évoquant le cadre. Il est pourtant possible de dégager le lien indirect qui existe entre ces items.

CADRE : OBLIGATION

Exemple d'annotation de l'environnement contextuel du terme *hébergeur*

La loi pour la confiance dans l'économie numérique du 21 juin 2004, impose à **l'hébergeur** *du site une obligation de retirer les contenus illicites présents sur un blog ou un site, sauf à engager sa responsabilité* [RDLI_2007]

La jurisprudence met, au surplus, à la charge de **l'hébergeur** ***l'obligation de surveiller des contenus*** *notifiés afin qu'ils ne réapparaissent pas.* [LEGALIS.COM]

Exemple d'annotation de l'environnement contextuel du terme *obligation de surveillance générale*

Et d'en conclure que sa responsabilité ne peut pas être engagée sur le fondement de la LCEN car elle n'a pas une **obligation de surveillance générale** *du site et a pris les*

mesures nécessaires à la suite des signalements de la société Maceo, les annonces en cause ne présentant aucun caractère manifestement illicite. [RDLI_2012]

Exemple d'annotation de l'environnement contextuel du terme *contenu illicite*

En effet, l'article 6-I-2 de la LCEN fait peser sur les fournisseurs d'hébergement une obligation de supprimer ***les contenus illicites****, en plaçant ceux-ci dans une situation délicate, puisqu'il pourrait s'agir, pour eux, d'apprécier le caractère licite ou illicite d'un contenu.* [RDLI_2008]

Exemple d'annotation de l'environnement contextuel du terme *donnée à caractère personnel*

Le responsable du traitement est tenu de veiller à la sécurité des ***données à caractère personnel*** *et de leur traitement, ainsi qu'au respect de la finalité de celui-ci.* [CORPUS_DONNEES]

Figure 44. Exemple d'annotation des contextes au moyen des cadres sémantiques (cadre OBLIGATION)

Ainsi, remarquons que l'analyse des termes insérés dans des contextes annotés au moyen des cadres sémantiques et basée sur le système d'étiquetage renvoyant aux cadres sémantico-conceptuels permet de réaliser les actions suivantes :

a) accéder à d'autres termes du domaine du droit de l'Internet. Par exemple, l'annotation des contextes dans lesquels apparaît le terme *hébergeur* nous a conduite au terme *obligation de retrait de contenus illicite.*

b) identifier les liens existant entre le terme donné et d'autres termes du domaine. L'analyse de l'environnement contextuel du terme *donnée à caractère personnel* a fait apparaître un réseau de termes interconnectés par un type défini de liens: *responsable du traitement* -> est obligé d'assurer -> *sécurité des données à caractère personnel, obligation de sécurité des données à caractère personnel* -> l'obligation du -> *responsable du traitement*

c) caractériser les liens qui définissent le cadre sémantico-conceptuel auquel est associé le terme donné.

En résumant, nous considérons que notre modèle d'annotation basée sur la théorie des cadres sémantiques ainsi que sur le système d'étiquetage des catégories abstraites auxquelles apartiennet les termes permet de rendre compte du comportement des termes dans leur

environnement contextuel ainsi que de systématiser les relations conceptuelles du droit de l'Internet et de connecter les termes à la structure notionnelle du domaine tout en restant dans une démarche linguistique.

Chapitre 10. *DITerm*, proposition de modélisation des données terminographiques du domaine du droit de l'Internet - à la recherche d'un modèle hybride

Tout au long de cette partie de notre travail, nous avons essayé de systématiser la multitude de relations que les termes du domaine du droit de l'Internet entretiennent avec d'autres unités aussi bien sur le plan linguistique que conceptuel. Pour ce faire, nous avons mis en œuvre deux stratégies (souvent considérées comme concurrentes ou bien incompatibles), notamment la description linguistique des termes basée sur leurs caractéristiques lexico-sémantiques et la structuration des liens conceptuels qui les relient au schéma notionnel du domaine. En considérant que ces deux types de relations se manifestent d'une manière ou d'une autre dans le discours, nous avons posé l'hypothèse selon laquelle il est possible d'en rendre compte tout en restant dans une démarche purement linguistique. Ainsi, dans les pages qui suivent, nous voudrions présenter notre proposition de modélisation des données terminographiques dont l'objectif est de mettre en évidence aussi bien la dimension linguistique que conceptuelle du terme.

10.1 Nomenclature

Tout d'abord, il est convient de souligner que le *DITerm* (voir l'Annexe VI) n'est pas un dictionnaire complet du domaine du droit de l'Internet. En effet, pour reprendre l'expression de Mel'čuk et Polguère (2007 : 15), il s'agit d'un échantillon de dictionnaire, un échantillon du noyau de la terminologie du domaine, c'est-à-dire des termes qui ont un poids important dans le réseau terminologique du droit de l'Internet. Nous avons choisi les termes qui sont fréquents et qui entretiennent une multitude de liens, aussi bien sur le plan linguistique que conceptuel. Bien que le *DITerm* soit un dictionnaire en construction, les fiches présentées dans le cadre de ce travail correspondent aux articles dont la rédaction est

terminée. Soulignons tout de même qu'une nomenclature cible de deux centaines de termes a été constituée et qu'elle est répertoriée à l'Annexe III. Sa description pourrait faire l'objet d'un projet postdoctoral.

10.2 Article

Chaque article correspond à une acception spécialisée qui doit être en lien avec le domaine du droit de l'Internet. En effet, les termes sont décrits par le biais de leur appartenance au domaine. Ceci permet de résoudre le problème de la polysémie due à leur caractère dynamique (voir la section 7.1). Ainsi, même si un terme donné fonctionne également en dehors du droit de l'Internet (c'est notamment le cas de *responsable du traitement*), il sera décrit par rapport aux liens qui le lient au domaine et considéré en l'occurrence comme un acteur de l'Internet.
Les articles sont découpés en plusieurs rubriques qui font l'objet des sections suivantes :

- **ENTRÉE**
- **DEFINITION DU TERME**
- **RELATIONS**

Les trois sections sont à leur tour, divisées en un certain nombre de sous-sections, chacune consacrée à un autre aspect de la description du terme.

10.3 Entrée

Chaque entrée présente le terme en français accompagné de l'information sur la partie du discours à laquelle il appartient. S'il s'agit d'un nom, l'indication de la partie du discours est suivie de la mention du genre. L'entrée comporte également les équivalents espagnols et polonais du terme.

CONSENTEMENT PRÉALABLE (n. m.)

ESPAGNOL : **CONSENTIMIENTO (n. m.) DEL INTERESADO**
POLONAIS: **ZGODA (n. f.) NA PRZETWARZANIE DANYCH OSOBOWYCH**

Figure 45. Entrée dans le *DITerm* (Annexe VI : 6)

10.4 Définition du terme

Cette section est divisée en quatre rubriques, chacune consacrée à la définition du terme mais abordée sous un angle différent :

- **FORME PROPOSITIONNELLE**
- **{CADRE SÉMANTICO-CONCEPTUEL} auquel appartient le *terme vedette***
- **DÉFINITION**
- **CONTEXTES DÉFINITOIRES annotés**

10.4.1 Forme propositionnelle

Chaque terme décrit est présenté avec sa forme propositionnelle, c'est-à-dire avec une formule représentant sa structure actancielle (la notion de *structure actancielle* a été introduite et étudiée à la section 8.1). La forme propositionnelle explicite le nombre d'actants que possède le terme. Les actants y sont représentés traditionnellement, au moyen des variables X, Y, Z, W (à l'instar des dictionnaires tels que le *DEC*, le *DiCo*, le *LAF* – voir la section 1.4.3). Leur rôle par rapport au terme décrit n'est pas précisé ni illustré au moyen d'un système d'étiquettes comme c'est le cas, par exemple, dans le *DiCoInfo* ou le *DiCoEnviro* (voir la section 2.3). En revanche, nous nous inspirons de ces deux projets en proposant une mention de l'actant typique qui constitue quant à lui une sorte d'étiquette sémantique. Afin de le distinguer du reste du texte, nous proposons de l'écrire en police non proportionnelle. L'indication de l'`actant typique` précède la variable et correspond à l'une des réalisations linguistiques de l'actant. L'actant typique est choisi en fonction des critères suivants[101] :

a) il doit s'agir de la réalisation qui sera évoquée le plus naturellement dans la définition du terme ;

[101] Nous adoptons les critères adoptés dans le cadre des projets *DiCoInfo* (*DiCoInfo* Manuel : 15) et *DiCoEnviro* (DiCoEnviro Manuel : 13).

b) il doit s'agir, en règle générale, de la réalisation rencontrée le plus fréquemment dans l'environnement du terme décrit lors de l'observation de ses occurrences ;

c) il doit s'agir d'un terme générique qui englobe les autres réalisations.

De plus, nous avons décidé d'indiquer le cadre sémantico-conceptuel auquel appartient l'actant typique. Cette information apparaît entre accolades à la suite de la variable et permet de connecter les éléments à la structure notionnelle du domaine. La présentation de la forme propositionnelle s'appuie donc sur un système de double étiquetage (sémantique et sémantico-conceptuel) auquel nous aurons recours tout au long de la présentation (voir la section 10.5.1.3).

TÉLÉCHARGEMENT ILLÉGAL (n. m.)

FORME PROPOSITIONNELLE: *téléchargement illégal* {ACTIVITÉ ou ACTION ILLICITES} d'une `œuvre` X {OBJET EN LIGNE} réalisé par un `internaute` Y {ACTEUR DE L'INTERNET : UTILISATEUR} à partir d'un `site` Z {SERVICE INTERNET} vers un `ordinateur` {TECHNOLOGIE ou MACHINE}

Figure 46. Forme propositionnelle du terme *téléchargement illégal* dans le *DITerm* (Annexe VI : 147)

10.4.2 Cadre sémantico-conceptuel auquel appartient le terme vedette

Cette rubrique précise l'appartenance du terme décrit à un cadre sémantico-conceptuel. Rappelons que selon nous, un cadre sémantico-conceptuel est une structure conceptuelle, dégagée sur la base de critères aussi bien sémantiques que conceptuels, qui permet de rattacher un terme donné à un schéma notionnel du domaine (voir la section 9.2.1). Chaque terme appartenant à un cadre sémantico-conceptuel se caractérise par un ensemble défini de relations avec d'autres cadres, identifiées sur la base des observations faites sur le corpus. La mise en évidence des relations se fait au moyen des cadres sémantiques (voir la section 9.2.2). La prise en compte de ces liens facilite la description de la dimension conceptuelle du terme.

Formellement, un cadre sémantico-conceptuel est une sorte d'étiquette correspondant à un mot ou groupe de mots, ayant une signification très générale et renvoyant à une catégorie abstraite. Les cadres apparaissent entre accolades est sont suivis de la mention des relations qui les caractérisent en les reliant à d'autres cadres du domaine. En effet, la rubrique décrivant l'appartenance du terme à un cadre sémantico-conceptuel se présente sous forme d'un tableau qui fournit deux type d'informaions :

a) informations sur les cadres sémantico-conceptuels liés au cadre sémantico-conceptuel auquel appartient le terme en question. Ces cadres sont identifiés suite à l'analyse du contexte dans lequel apparaît le terme, analyse basée sur l'annotation sémantique au moyen des cadres sémantiques (voir ci-dessous)

b) information sur les cadres sémantiques dans lesquels s'insère le terme appartenant à un cadre sémantico-conceptuel donné. Cela permet de mettre en évidence les relations que le terme appartenant à un cadre sémantico-conceptuel en question entretient avec d'autres cadres sémantico-conceptuels du domaine (voir plus haut).

{CADRE SÉMANTICO-CONCEPTUEL} auquel appartient *hébergeur*: {ACTEUR DE L'INTERNET : PRESTATAIRE}	
hébergeur est lié aux {CADRES SÉMANTICO-CONCEPTUELS} suivants :	→ {ACTIVITÉ ou ACTION} → {SERVICE EN LIGNE} → {OBJET EN LIGNE} → {AUTORITÉ JURIDIQUE} → {TEXTE LEGISLATIF} → {ACTEUR DE L'INTERNET : UTILISATEUR} → {RÈGLE} → {OBLIGATION} → {PROCÉDURE} → {INFRACTION} → {OBJET ILLICITE EN LIGNE}
hébergeur s'insère dans les **CADRES SÉMANTIQUES** suivants :	→ **RÉGLEMENTATION** → **RESPONSABILITÉ** → **OBLIGATION**

	→ **SOUMIS À UNE MESURE LÉGALE** → **RESPECT DE LA RÉGLEMENTATION**

Figure 47. Description d'un cadre sémantico-conceptuel dans le *DITerm* (Annexe VI : 100)

10.4.3 Définition

Les définitions doivent apporter des précisions d'ordre aussi bien linguistique (sémantique) qu'encyclopédique (conceptuel). En effet, chaque définition reprend et explicite les informations fournies dans les deux rubriques précédentes, à savoir : *Forme propositionnelle* et *Cadre sémantico-conceptuel*. Ainsi, tout comme dans *DiCoInfo* (voir *DiCoInfo* Manuel : 17), la définition dans *DITerm* est construite à partir de la forme propositionnelle et indique explicitement les actants en utilisant la notation des actants typiques qui constituent une sorte d'`étiquettes sémantiques`. Il s'agit d'une définition analytique, c'est-à-dire une définition par genre prochain et différences spécifiques qui doit rendre compte des principaux liens (intervenus aussi bien sur le plan linguistique que conceptuel), dégagés sur la base de l'observation du comportement du terme en contexte. Quant au genre prochain, il est l'hyperonyme du terme défini et correspond à la réalisation évoquée le plus naturellement dans la définition du terme. Ce terme hyperonyme est utilisé ainsi comme une `étiquette sémantique` au même titre que l'actant typique. Remarquons entre parenthèses qu'un terme correspondant à un genre prochain dans la définition d'une unité terminologique donnée peut jouer le rôle d'un actant typique dans la définition d'une autre unité. Cela permet de tisser un réseau de termes typiques du droit de l'Internet censés correspondre à des `étiquettes sémantiques` dont l'utilisation facilite et uniformise la description des termes et de leurs relations.

De plus, la définition contient des informations encyclopédiques que nous avons identifiées grâce à l'annotation et l'étiquetage de l'environnement contextuel des termes. En effet, ces informations sont liées à l'appartenance d'un terme à un cadre sémantico-conceptuel donné (voir la rubrique précédente) et reflètent les relations que ce dernier enretient avec d'autres cadres sémantico-conceptuels. Ainsi, chaque élément de la définition peut être défini

au moyen d'une étiquette sémantico-conceptuelle correspondant à un cadre sémantico-conceptuel auquel il appartient. Remarquons que toutes les informations contenues dans la définition sont explicitées, grâce aux contextes annotés, dans la partie consacrée à la description des relations.

HÉBERGEUR (n. m.)

FORME PROPOSITIONNELLE: *hébergeur* {ACTEUR DE L'INTERNET : PRESTATAIRE} de `données` Y {OBJET EN LIGNE} ou d'un `site` Y {SERVICE EN LIGNE} sur un `serveur` W {TECHNOLOGIE ou MACHINE} pour le compte d'un `utilisateur` Z {ACTEUR DE L'INTERNET : UTILISATEUR}

DÉFINITION: *hébergeur* est un `intermédiaire technique` {ACTEUR DE L'INTERNET : PRESTATAIRE} qui offre à l'`utilisateur` {ACTEUR DE L'INTERNET : UTILISATEUR}, un `service` permettant d'accueillir sur son serveur {ACTIVITÉ OU ACTION} un `service` de communication au public en ligne (blog, page personnelle, site) {SERVICE EN LIGNE} ou des `données` de toute nature (écrits, images, sons, vidéo, un message, annonces) {OBJET EN LIGNE}. Un `intermédiaire technique` {ACTEUR DE L'INTERNET : PRESTATAIRE} qui agit en qualité d'hébergeur bénéficie d'un régime de responsabilité qui lui est propre {RÈGLE}. En effet, il peut être tenu comme responsable des `données` stockées, uniquement s'il a connaissance de l'existence des `données` et s'il n'agit pas promptement pour retirer les contenus à caractère illicite {PROCÉDURE}. Ainsi, l'hébergeur n'est pas soumis à une obligation de surveillance générale des `données` stockées {OBLIGATION} mais il a une obligation de retirer les contenus illicites dès qu'il en prend connaissance {OBLIGATION}.

Figure 48. Exemple d'une définition dans le *DITerm* (Annexe VI : 101)

10.4.4 Contextes définitoires

La dernière rubrique de la première section comporte ce que nous appelons des *contextes définitoires,* c'est-à-dire des contextes qui contiennent des définitions plus au moins formelles du terme décrit. Ils illustrent la manière dont les termes sont définis par les experts du domaine. Rappelons que les contextes définitoires ont été extraits du corpus *DITerm* à l'aide de marqueurs lexico-syntaxiques (voir les sections 4.3.1, 6.3.2.2 ainsi que Pearson 1998 :135 – 190). Ces éléments linguistiques nommés par Pearson (*ibid.*), *connective phrases* sont identifiés graphiquement (ils apparaissent sur fond noir avec un texte en blanc).

CONTEXTES DÉFINITOIRES :
Le consentement, *entendu comme l'autorisation donnée par une personne physique de traiter les données la concernant, peut être signifié de différentes manières.* [CORPUS_DONNÉES]

Figure 49. Exemple d'un contexte définitoire dans le *DITerm* (Annexe VI : 25)

Remarquons que les contextes définitoires apparaissent tout au long de l'article afin d'apporter des précisions d'ordre encyclopédique et d'illustrer des relations taxinomiques.

10.5 Description des relations

La deuxième partie de l'article, la plus importante, est consacrée à la description des relations lexicales et conceptuelles que le terme en question entretient avec d'autres unités du vocabulaire du droit de l'Internet. On recense parmi ces dernières :

a) les termes qui forment le noyau dur, c'est-à-dire la nomenclature proprement dite du domaine du droit de l'Internet à partir de laquelle se réalise le discours spécialisé. Les termes en question sont ou seront décrits dans des articles qui leur sont (ou seront) consacrés.

b) les unités considérées comme vocabulaire de soutien ou vocabulaire annexe qui ne font pas l'objet d'articles à part mais apparaissent comme des éléments descriptifs rattachés aux termes choisis (voir la section 5.1.1.2).

10.5.1 Règles de base

10.5.1.1 Recours aux formules explicites

Afin de rendre compte des liens entretenus par les termes décrits, nous faisons appel à des formules explicites, fondées sur un métalangage compréhensible par tous, une sorte de français standardisé (Mel'čuk et Polguère 2007 : 21). Notre proposition d'encodage s'aligne d'une certaine façon sur ceux mis en œuvres dans le *DiCo* (Polguère 2003a, 2003b), le *LAF* (Mel'čuk et Polguère 2007) et le *DiCoInfo* (voir la section 2.3). Remarquons pourtant que dans le cadre de notre travail, l'encodage au moyen du français standardisé ne correspond que partiellement à la vulgarisation des fonctions lexicales (voir la section 8.3). En effet, le recours aux formules explicites nous permet de décrire, à l'aide d'un formalisme unique, aussi bien les relations lexicales que les liens conceptuels. La forme d'explication des liens que nous avons choisie présente, selon nous, de nombreux avantages. Nous proposons de les détailler ci-dessous.

a) L'encodage traditionnel des FL étant souvent jugé trop sophistiqué et abstrait, nous avons opté pour une façon plus claire et intuitive de présenter les liens lexicaux entretenus par les termes. Nous considérons que le recours aux formules explicites permet de simplifier au maximum la description terminographique en la rendant accessible à différents types de publics.

b) La représentation des relations lexicales au moyen d'un encodage vulgarisé des FL est une solution qui paraît tout à fait naturelle. Comme le souligne Polguère (2003), les FL doivent être appréhendées indépendamment de leur formalisation habituelle. En effet, d'après l'auteur, les FL ne sont pas une fiction théorique développée dans le cadre de la théorie Sens-Texte, mais correspondent bien à un fait linguistique universel, observable, dont la maîtrise est une partie essentielle de notre connaissance linguistique. En effet, les FL standard doivent être considérées comme des universaux linguistiques dénotant des liens lexicaux récurrents. Elles sont toutes associables à un contenu sémantique donné de nature très vague et qui n'est pas nécessairement un

primitif, au sens où l'entend par exemple Wierzbicka (*ibid.* : 8). Ainsi, dévêtir la notion de fonction lexicale de l'appareillage formel d'encodage auquel elle est habituellement associée permet d'y voir un outil conceptuel visant à rendre compte des phénomènes linguistiques d'une façon plus simple, brute et directe. (voir la setion 8.3). Ceci amène Polguère (*ibid.* : 16) à dire, qu'il est possible d'encoder les liens de FL en utilisant une sorte de langue « contrôlée » fondée sur les universaux sémantiques associés aux FL, dans laquelle on peut paraphraser de tels liens. Polguère (*ibid.*) met l'accent sur l'importance du lien unissant fonctions lexicales et paraphrase linguistique. Il attire également l'attention sur la nécessité de la mise au point de bonnes techniques de paraphrasage sémantique utilisant des langues contrôlées.

c) En ce qui concerne le projet *DITerm*, nous avons commencé notre travail de modélisation des liens lexicaux en faisant appel à l'appareil des FL standard (voir la section 5.2.1.). Le recours à l'encodage formel nous a permis d'analyser et de systématiser les relations entretenues par les termes et d'en saisir la véritable nature. Une fois la modélisation au moyen des FL effectuée, nous avons entrepris le paraphrasage des liens capturés à l'aide des FL en s'appuyant sur les gloses de vulgarisation proposées dans le *DiCo*, le *DiCoInfo*. Le Tableau 48, ci-dessous présente notre démarche.

Encodage au moyen des FL	**Gloses de vulgarisation**
Oper$_1$ (téléchargement illégal)= procéder [à ART ~], réaliser [ART ~],	Un internaute réalise un *téléchargement illégal*.
Oper$_1^{usual}$ (téléchargement illégal)= pratiquer [ART ~], s'adonner [à ART ~]	Un internaute réalise un *téléchargement illégal* régulièrement.
Oper$_3$ (téléchargement illégal) = servir [à ART ~], proposer [ART ~]	Un site donne la possibilité d'un *téléchargement illégal*.
LiquFunc$_0$ (téléchargement illégal) = bloquer [ART ~]	Un intermédiaire technique ou un logiciel fait en sorte que le *téléchargement illégal* ne soit plus réalisé par un internaute.

NonPermFunc$_0$ (téléchargement illégal) = empêcher [ART ~], prévenir [ART ~],	Une autorité juridique ou un intermédiaire technique fait en sorte que le *téléchargement illégal* ne commence pas à avoir lieu.

Tableau 48. Exemples de gloses de vulgarisation dans le *DITerm*.

Même si notre description des liens lexicaux s'appuie largement sur le formalisme des FL, nous avons décidé de présenter, dans les articles du *DITerm*, un seul niveau d'explication de ces relations, celui des formules explicites, rédigées en langue contrôlée.

d) Le recours aux formules en français standardisé permet aussi d'uniformiser la description au moyen des FL. Rappelons que les FL standard ne permettent pas de couvrir l'ensemble des dérivations sémantiques et des collocations recensées dans le corpus *DITerm*. Ainsi, pour décrire les relations dont le sens est spécifique (et donc non généralisable), la solution serait de recourir à des FL semi-standard (FLSS) et à des FL non standard (FLNS). Or, comme le remarque Jousse (2010 : 103) – voir la section (8.3), il n'existe pas de véritables règles de formation de l'encodage des FLSS (constituées d'une FL standard et d'un élément en français) et FLNS (encodées par des gloses formulées entièrement dans la langue française). Étant donné le caractère hétérogène de l'ensemble des FL, le recours à ce formalisme ne permet pas de préserver la systématicité de la description. Nous considérons donc qu'un recours généralisé à des formules en français standardisé permet d'éviter la présence, dans une même fiche, de trois différentes formes d'encodage.

e) Finalement, rappelons également que les FL permettent la description de deux phénomènes : les dérivations sémantiques et les collocations, relations lexicales fondées sur l'héritage de propriétés sémantiques. Or, comme nous avons pu le constater (voir les sections 8.2.3, 8.3), un grand nombre de syntagmes relevés dans notre corpus se caractérise par une forte compatibilité conceptuelle qui relève de la combinatoire libre. Ces relations qui existent entre les termes du domaine sont déterminées plutôt par rapport à la réalité juridique qu'à la réalité linguistique et pour

les décrire, on doit prendre en compte les propriétés référentielles qui renvoient à des concepts situables en dehors de la langue. (voir la section 8.3).

Nous soulignons que ce type de liens conceptuels doit absolument être pris en considération dans une description terminologique. En effet, il paraît utile de formaliser dans une base de données terminographique destinée aux traducteurs, des expressions polylexicales qui ne sont pas à strictement parler des collocations, dans la mesure où les équivalents de traduction d'un même lexème peuvent varier en fonction de leurs divers cooccurrents. Néanmoins, il est nécessaire de distinguer les deux types de phénomènes (combinatoire libre et combinatoire restreinte, liens conceptuels forts et liens de dérivation sémantique). Ainsi, pour pouvoir rendre compte des affinités conceptuelles sans dénaturer le système des FL, il faudrait créer un outil formel séparé. Craignant tout de même que cela n'alourdisse notre description, nous nous sommes tournée vers le procédé de paraphrasage linguistique, qui selon nous est adapté aussi bien à la description des relations sémantiques (voir plus haut) que conceptuelles.

10.5.1.3 Système de double étiquetage

Comme le remarque Polguère (2011 : 197), les lexies forment un ensemble bien trop vaste pour être véritablement appréhendé dans sa totalité. Confrontés au gigantisme du lexique, nous cherchons donc à le rationaliser : « *nous classons pour connaître et reconnaître.* ». D'après l'auteur (2011 : 199), il n'y a pas un unique système de classification des lexies : *« De par la nature du lexique et de par la nature de l'ensemble du système de la langue […], plusieurs classifications sont nécessaire, chacune ayant une finalité donnée*. » Quant au vocabulaire propre à un domaine de spécialité donné (qui ne constitue qu'une partie infime du lexique en général), nous supposons qu'il est plus facile de le saisir et de le décrire dans son ensemble. Cependant, en proposant un modèle de structuration des unités terminologiques, il est nécessaire de prendre en considération le fait que les termes contenus dans les textes spécialisés font référence à des concepts qui ont été classés et définis par les experts du domaine. Un système de classification des unités terminologiques doit donc se concevoir (au moins en partie) relativement au domaine des concepts. Comme le soulignent Dancette et Halimi (2007 : 553), pour comprendre et traduire un texte spécialisé, le traducteur, qui ne détient pas le savoir de l'expert, doit pouvoir accéder aux connaissances utiles (aussi bien linguistiques que conceptuelles). La prise en compte de la dimension conceptuelle du

terme lors de la mise en place d'un système de structuration des données terminologiques s'impose donc comme une solution incontournable. Soulignons qu'il existe de multiples systèmes de classification des unités terminologiques. Leur caractère dépend aussi bien d'un projet visé que d'un type de domaine de spécialité. Parmi différents modèles possibles, on en peut distinguer deux auxquels, traditionnellement, ont recours les terminographes. D'un côté, il s'agit d'un classement aristotélicien ou autrement dit ontologique, proposant une structuration des termes de façon hiérarchique, de l'autre côté, d'un classement par « champs ». Dans le cadre de notre projet, nous avons décidé de faire appel à ce dernier type de classement, notamment à un classement par « champs » qui vise à rassembler des termes possédant des composantes sémantico-conceptuelles communes.

Rappelons (voir à ce sujet le chapitre 9), que chaque terme du domaine du droit de l'Internet appartient à une catégorie abstraite, défini comme un cadre sémantico-conceptuel. Ce dernier est identifié sur la base de critères aussi bien sémantiques (l'étape d'appariement des termes aux familles sémantiques) que conceptuel (l'étape d'analyse des informations contextuelles). Comme nous l'avons vu plus haut, l'appartenance d'un terme donné à un cadre sémantico-conceptuel est signalée dans la première partie de l'article (voir la section 10.4). Son appartenance à une catégorie sémantico-conceptuelle se manifeste également par le biais des relations hyperonymies qu'il entretient avec d'autres termes (voir la rubrique ***terme vedette*** **APPARTIENT À LA CATÉGORIE** reproduite dans la Figure 50). Le nom du cadre sémantico-conceptuel devient ainsi une sorte d'étiquette sémantico-conceptuelle associée au terme (pour la liste complète des étiquettes sémantico-conceptuelle, nous renvoyons à l'Annexe IV). Remarquons que le système des étiquettes sémantico-conceptuelles dégagées à la suite de l'analyse sémantico-conceptuelle des termes extraits du corpus *DITerm* (voir les sections 9.1 et 9.2.1), est structuré autour des étiquettes nominales, car, pour reprendre les propos de Polguère (2011 : 10), « *le Nom est justement l'entité lexicale de 'nommage'* ». Les étiquettes sémantico-conceptuelles sont écrites en majuscules et apparaissent entre accolades. Il paraît nécessaire de souligner qu'un terme du droit de l'Internet peut posséder plusieurs étiquettes sémantico-conceptuelles ce qui s'explique par le caractère complexe et multidimensionnel des concepts du domaine. C'est notamment le cas du terme *consentement préalable* (voir l'exemple de la Figure 50) qui est lié à 5 cadres sémantico-conceptuels, à savoir {DROIT}, {OBLIGATION}, {PROCÉDURE}, {RÈGLE}, {AUTORISATION}.

consentement préalable **APPARTIENT À LA CATÉGORIE de**	
{AUTORISATION} Autorisation	consentement (n. m.)
{RÈGLE}	principe (n. m.) de *consentement préalable*, règle (n. f.) de *consentement préalable*,
{OBLIGATION}	exigence (n. f.) de *consentement préalable*, obligation (n. f.) de *consentement préalable*
{PROCÉDURE}	mécanisme (n. m.) de *consentement préalable*, système (n. m.) de *consentement préalable*
{DROIT}	droit (n. m.) au *consentement préalable* droit (n.m.) d'opposition

Figure 50. Étiquetage sémantico-conceptuel dans le *DITerm* (Annexe VI : 8)

Les étiquettes sémantico-conceptuelles permettent ainsi de connecter un terme donné au champ conceptuel du domaine. Il est tout de même important de souligner que nous ne faisons pas appel à l'étiquetage sémantico-conceptuel uniquement pour indiquer la place d'un terme donné dans la structure notionnelle du domaine mais aussi pour offrir une caractérisation conceptuelle des termes avec lesquelles ce dernier entretient des relations d'ordre sémantique et conceptuel. En effet, chaque terme relié d'une manière ou d'autre au terme vedette y est présenté avec une étiquette sémantico-conceptuelle correspondante. Ainsi, le recours au système d'étiquettes sémantico-conceptuelles permet de rendre compte d'un réseau de termes associés et de signaler les différents liens qu'ils les unissent. Les étiquettes sémantico-conceptuelles sont également utilisées dans l'annotation des contextes dans lesquels apparait le terme vedette. Comme nous l'avons vu plus haut (voir la section 9.2.2) et comme nous le verrons dans la section suivante (voir la section 10.5.1.3), l'annotation de l'environnement contextuel au moyen des cadres sémantiques à la Fillmore contribue à la structuration du champ notionnel du domaine. Rappelons que les termes du droit Internet gravitent autour de différentes unités évoquant des cadres sémantiques propres au langage juridique. Afin d'uniformiser l'annotation de tels contextes, nous avons décidé de faire appel au système d'étiquettes sémantico-conceptuelles permettant de caractériser les termes qui constituent des éléments de cadre centraux. Il s'est avéré que l'étiquetage sémantico-conceptuel des éléments de cadre centraux permet d'identifier et d'accéder à des liens conceptuels indirects que le terme en question entretient avec d'autres termes (voir la Figure 51).

Comme nous pouvons le constater en analysant entre autres l'exemple présenté ci-dessus (revenons à la Figure 50), dans certains cas, nous utilisons un système de double étiquetage :

{AUTORISATION} et `autorisation`
{SERVICE EN LIGNE} et `site`
{ACTEUR DE L'INTERNET : UTILISATEUR} et `internaute`

où l'étiquette notée en police non proportionnelle[102] correspond à une étiquette sémantique, concept développé dans le cadre du projet *DiCo* (voir à ce sujet Polguère 2003b et Polguère 2011). En effet, contrairement aux étiquettes sémantico-conceptuelle, une étiquette sémantique renvoie à l'univers purement linguistique. Pour citer les propos de Polguère (2011 : 2001), une étiquette sémantique ne se conçoit pas relativement au domaine des concepts, mais relativement au système de la langue, c'est-à-dire, il n'est pas nécessaire d'avoir recours à des notions autres que celles de la linguistique pour la construire. Cependant, il convient de préciser que la notion d'étiquette sémantique telle qu'elle est utilisée dans le cadre de notre projet n'a pas été théorisée ni définie de façon claire. En effet, le système des étiquettes sémantiques dans le *DITerm* s'est construit au fur et à mesure du développement du projet. Les étiquettes ont été identifiées de façon inductive, dans le cadre du processus de modélisation des données terminologiques, c'est-à-dire au moment de la description sémantique du terme vedette et de ses relations. Ainsi, elles ont été déterminées par rapport à trois types de relations : a) le terme vedette et son hyperonyme, b) le terme vedette et les actants typiques de la situation relative au terme vedette, c) le terme vedette et les participants extérieurs à la situation relative au terme vedette. En pratique, les étiquettes sémantiques correspondent à des termes typiques du droit de l'Internet, c'est-à-dire aux termes qui apparaissent le plus naturellement dans l'environnement contextuel du terme décrit et qui peuvent, en même temps, jouer le rôle d'un terme hyperonyme englobant d'autres expressions linguistiques apparentées sémantiquement. Étant donné que les étiquettes sont définies au niveau de la microstructure, le recours à l'étiquetage sémantique n'a pas pour objectif de proposer un modèle de structuration des données terminologiques. En revanche, la mise en place du système des étiquettes sémantiques doit aider dans l'extraction de régularités

[102] Nous avons emprunté cette convention d'écriture au projet DiCo (Mel'čuk et Polguère), voir (Polguère 2003b) et (Polguère 2011).

au niveau de l'information lexicale à l'intérieur de chaque article et contribuer à la construction d'un métalangage de vulgarisation uniformisé et standardisé. Pour terminer, remarquons que dans certain cas, le nom d'une étiquette sémantique et identique au nom d'une étiquette sémantico-conceptuelle comme par exemple : {DROIT} et `droit`, {OBLIGATION} et `obligation` ou {AUTORISATION} et `autorisation`. Cela est dû au fait qu'il existe des connexions entre étiquettes sémantiques et concepts structurant la cognition et que chaque étiquette sémantique est associée en quelque sorte à un concept correspondant (Polguère 2011 : 201).

.

10.5.1.3 Annotation des contextes au moyen des cadres sémantiques à la Fillmore

La description des relations que le terme vedette entretient avec d'autres termes du domaine du droit de l'Internet (il s'agit aussi bien des relations paradigmatiques, collocationnelles que des liens conceptuels), est accompagnée d'un certain nombre de contextes annotés au moyen des cadres sémantiques (comme nous l'avons vu à la section 9.2.2, la méthodologie utilisée s'inspire largement des travaux réalisés dans le cadre du projet *FrameNet*). Les contextes servent bien évidement à illustrer de quelle manière le terme vedette s'utilise concrètement dans les textes spécialisés et à fournir des renseignements supplémentaires quant à son fonctionnement linguistique. Cependant, ce n'est pas la seule raison pour laquelle nous avons fait appel à ce modèle d'annotation. En effet, nous avons remarqué que les relations décrites au moyen des formules explicites s'insèrent quant à elles dans d'autres types de relations. Prenons l'exemple de *données à caractère personnel* (voir la Figure 51). Le terme en question entretient avec le verbe *transférer* la relation que nous proposons de décrire au moyen de la formule suivante : « faire passer des *données à caractère personnel* d'un endroit ou d'une personne à un/une autre ». L'analyse l'environnement contextuel du groupe verbal *transférer des données à caractère personnel* (ou de sa nominalisation *transfert* (n. m.) *des données à caractère personnel*), a montré que les expressions en question s'insèrent, quant à elles, dans d'autres types de relations comme par exemple :

- « *[…]* ***le transfert de données personnelles*** *[…]* *est* *[…]* *prohibé.* »
- « *[…]* *il est interdit de* ***transférer les données à caractère personnel*** *[…]*. »
- « *L'article 226-22-1 du Code pénal sanctionne de cinq ans d'emprisonnement et de 300 000 euros d'amende* ***le fait de transférer les données personnelles*** »

- « *L'article 42 de la directive* <u>subordonne</u> ***les transferts des données à caractère personnel*** *[…].* »

En effet, il s'agit des relations typiquement juridiques, propres à toutes les branches du droit qui peuvent être décrites au moyen des cadres sémantiques à la Fillmore (comme nous l'avons vu à la section 9.2.2).

<table>
<tr><td colspan="3">SITUATION: Le protagoniste principal fait passer des données à caractère personnel d'un endroit ou d'une personne à un/une autre.</td></tr>
<tr><td>PROTAGONISTE PRINCIPAL</td><td>PROCÈS IMPLIQUANT données à caractère personnel</td><td>AUTRES PARTICIPANTS OU CIRCONSTANTS</td></tr>
<tr><td>{ACTEUR DE L'INTERNET : PRESTATAIRE} ou

{ACTEUR DE L'INTERNET : UTILISATEUR}

responsable du traitement

ou

{PERSONNE QUI REALISE UNE ACTIVITÉ ou ACTION ILLICITES} :

personne non autorisée</td><td>transférer des données à caractère personnel

(transfert de données à caractère personnel)

transmettre des données à caractère personnel

(transmission de données à caractère personnel)</td><td>source :
[d'un endroit] :
d'un État (n. m.) membre,
depuis l'Union (n. f.) européenne,
depuis la Communauté (n. f.),
destination :
[à un endroit] :
vers un pays (n. m.) tiers,
vers un État (n. m.) établi hors de l'Union Européenne,
vers un État (n. m.) tiers,
en dehors de l'Union (n. f.) européenne,
à l'extérieur (n. m.) de l'UE,
à l'étranger (n. m.)
moyen :
[via Internet] :
via un réseau (n. m.) de communications électroniques
manière :
[sans l'aide des moyens automatisés] :
manuellement
but :
[pour lutter contre les infractions] :

à des fins (n. f. pl.) de coopération policière et judiciaire</td></tr>
<tr><td colspan="3">AUTORISATION ou INTERDICTION

Une {ACTIVITÉ ou ACTION} concernant donnée à caractère personnel : [l'action de faire passer des données à caractère personnel d'un endroit ou d'une personne à un</td></tr>
</table>

autre] est interdite.

Le transfert de données personnelles vers un pays extérieur à l'Union européenne est en principe prohibé, à moins que la société exportatrice de données ne mette en œuvre un mécanisme juridique permettant de s'assurer du niveau de protection apporté aux données transférées. [RDLI_2012]

Selon l'article 25 de la directive n° 95/46/CE, il est interdit de transférer les données à caractère personnel en dehors de l'Union européenne vers un pays tiers qui n'a pas un niveau de protection adéquat des données personnelles. [RDLI_2011]

AUTORISATION ou INTERDICTION

Un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} est / n'est pas autorisé à réaliser une {ACTIVITÉ ou ACTION} concernant des *données à caractère* personnel [l'action de faire passer des *données à caractère personnel* d'un endroit ou d'une personne à un autre.

A contrario, un responsable de traitement présent sur le territoire de l'Union européenne peut librement transférer les données personnelles vers d'autres sociétés du groupe situées dans un État membre de l'Union européenne ou de l'Espace économique européen(6). [RDLI_2011]

AUTORISATION ou INTERDICTION

Une {ACTIVITÉ ou ACTION} concernant *donnée à caractère personnel* : [l'action de faire passer des *données à caractère personnel* d'un endroit ou d'une personne à un autre] est autorisée.

Les transferts de données à caractère personnel d'un État membre vers un pays tiers ayant un niveau de protection adéquat sont autorisés. [RDLI_2012]

PUNI PAR LA LOI

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE LEGISLATIF} punit une {ACTIVITÉ ou ACTION} : [l'action de faire passer des *données à caractère personnel* d'un endroit ou d'une personne à un autre] d'une peine.

L'article 226-22-1 du Code pénal sanctionne de cinq ans d'emprisonnement et de 300 000 euros d'amende le fait de transférer les données personnelles vers un état extérieur à l'Union européenne qui ne dispose pas d'une législation de protection équivalente à celle de l'Europe communautaire, et reconnue comme telle (ce qui n'est pas le cas de l'Île Maurice), en violation des mesures prises par la CNIL ou la Commission européenne. [RDLI_2008]

SOUMIS À UNE MESURE LÉGALE

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE LEGISLATIF} soumet une

{ACTIVITÉ ou ACTION} réalisée par l'acteur de l'Internet : [l'action de faire passer des *données à caractère personnel* d'un endroit ou d'une personne à un autre l'Internet] à une {RÈGLE}

L'article 42 de la directive subordonne les transferts vers des pays tiers pour lesquels la Commission n'a pas adopté de décision constatant un niveau de protection adéquat, à la présentation de garanties appropriées, notamment des clauses types de protection des données, des règles d'entreprise contraignantes et des clauses contractuelles. [CORPUS_DONNEES]

TERMES LIÉS : transfert (n. m.) des données à caractère personnel {ACTIVITÉ ou ACTION}, règles (n. f. pl.) d'entreprise contraignantes (*ou* règles (n. f. pl.) internes d'entreprise, *ou* BCR) {RÈGLE}, clauses (n. f. pl.) contractuelles {RÈGLE}

Figure 51. Annotation de l'environnement contextuel de *données à caractère personnel* au moyen des cadres sémantiques à la Fillmore, (Annexe VI : 79)

En effet, l'existence de ce type de liens est due au fait que le vocabulaire du droit de l'Internet est un vocabulaire technique qui acquiert, dans le contexte du droit, une charge supplémentaire de sens juridique. Il s'agit de termes liés à des événements qui produisent des effets juridiques, d'où la nécessite d'une description qui doit être réalisée à deux niveaux (technique et juridique). L'annotation au moyen des cadres sémantiques permet de connecter le terme vedette au schéma notionnel du domaine et de mettre en évidence les liens conceptuels indirects qu'il entretient avec d'autres unités. Comme nous pouvons le constater en examinant les contextes reproduits plus haut, l'expression *transférer des données à caractère personnel*, identifiée comme une relation de deux unités : *transférer* et *données à caractère personnel* (et décrite dans la première partie du tableau), s'insère dans plusieurs cadres sémantiques propres au domaine du droit, notamment dans les cadres :

SOUMIS À UNE MESURE LÉGALE,
PUNI PAR LA LOI,
AUTORISATION ou INTERDICTION[103].

Ainsi, tandis que les unités telles que : *subordonner*, *sanctionner*, *être autorisé*, *être prohibé*, *être interdit* évoquent les cadres sémantiques cités plus haut, les expressions *transférer des données à caractère personnel* ou *transfert* (n. m.) *des données à caractère personnel*

[103] La liste complète des cadres sémantiques du domaine du droit utilisés dans notre projet est fournie à l'Annexe V.

constituent un des éléments centraux des cadres. Les autres éléments des cadres sémantiques sont représentés par d'autres termes du domaine du droit de l'Internet. Remarquons que tous les termes appartiennent à des cadres sémantico-conceptuels définis antérieurement (voir la rubrique CADRE SÉMANTICO-CONCEPTUEL, le chapitre 9) et peuvent être décrits au moyen du système des étiquettes déjà introduit dans la première rubrique de l'article (voir la CADRE SÉMANTICO-CONCEPTUEL)[104]. Chaque contexte annoté est précédé d'une formule d'annotation présentant le cadre sémantique et les codes couleur : l'unité évoquant le cadre apparait en texte blanc sur fond noir, les éléments du cadre apparaissant sur un fond en couleur. Chaque élément de cadre est associé à une ou plusieurs étiquettes sémantico-conceptuelles (qui apparaissent entre accolades). Le recours au système d'étiquetage sémantico-conceptuelles permet d'uniformiser l'annotation et d'accéder à des liesn conceptuels indirects que

10.5.2 Organisation des liens paradigmatiques, syntagmatiques et conceptuels

La description des relations lexicales et conceptuelles est proposée sous forme de plusieurs tableaux. Les relations sont organisées en respectant l'ordre suivant :

10.5.2.1 Relations hiérarchiques

Le premier tableau est consacré aux relations hiérarchiques qui correspondent soit aux relations paradigmatiques (basées sur une description lexicographique en sème), soit aux relations conceptuelles (basées sur des éléments extralinguistiques) – voir la section 8.2.1. Le tableau comporte les rubriques suivantes (la présence de toutes les rubriques n'est pas obligatoire).

- rubrique **« VARIANTE ou PROCHE de** *terme vedette* » qui regroupe les synonymes, les quasi-synonymes et les sens voisins du terme décrit

- rubrique « *terme vedette* **APPARTIENT A LA CATEGORIE de** » qui d'un côté évoque l'hypéronyme du terme qui s'exprime dans une relation paradigmatique

[104] Le liste complète des cadres sémantico-conceptuels utilisés dans notre projet est fournie à l'Annexe IV.

analysée en termes de sèmes et de l'autre côté, présente des termes généraux qui renvoient à des concepts juridiques ou techniques et doivent être analysés en termes de référents. Ces derniers peuvent s'exprimer dans le corpus dans une relation syntagmatique : acte de téléchargement, statut d'hébergeur, règle de consentement préalable.

- rubrique « **DIFFÉRENT de** *terme vedette* » rassemble des co-hyponymes qui sont décrits soit en termes de référents soit en termes de composantes sémantiques.

- rubrique « **SORTE de** *terme vedette* » regroupe les hyponymes du terme qui peuvent être décrits soit en termes de référents, soit en termes de composantes sémantiques.

- rubriques « **CONTRAIRE à** » et « **CONVERSIF de** » qui décrivent les antonymes, contraires et contrastifs, par des relations basées sur l'analyse en termes de composantes linguistiques.

La plupart des relations sont illustrées par des contextes définitoires comportant une annotation relative aux marqueurs lexico-syntaxiques permettant d'identifier la relation en question (voir plus haut).

10.5.2.2 Propriétés et caractéristiques

Le deuxième tableau est consacré aux relations dont la description est basée uniquement sur des éléments extralinguistiques. Il s'agit notamment des caractéristiques et des propriétés associées au concept véhiculé par le terme et définies par rapport à la réalité juridique. Il s'agit des liens qui se manifestent dans le discours par des syntagmes de forte affinité conceptuelle comme *confidentialité des données à caractère personnel, neutralité d'hébergeur, consentement préalable explicite.* L'analyse de l'environnement contextuel de ces syntagmes montre que ces derniers apparaissent à leur tour dans d'autres types de relations que nous avons proposé de décrire de deux manières différentes, à savoir, a) à l'aide des formules explicites, b) au moyen des cadres sémantiques à la Fillmore (voir les sections 9.2.2, 10.5.1.1 et 10.5.1.3). Ceci permet de capturer des relations indirectes qui relient le

terme décrit avec d'autres unités et de préciser sa place dans le schéma notionnel du domaine L'exemple ci-dessous (Figure 52) illustre notre démarche :

<table>
<tr><td colspan="2">PROPRIÉTÉS, CARACTÉRISTIQUES de données à caractère personnel</td></tr>
<tr><td>Propriété de données à caractère personnel qui ne peuvent être communiquées qu'à une <code>personne autorisée</code>.</td><td>{RÈGLE}, {DROIT}, {OBLIGATION}

confidentialité (n. f.) des données à caractère personnel</td></tr>
<tr><td colspan="2">SITUATION : Le protagoniste principal fait en sorte que données à caractère personnel ne soient communiquées qu'à une <code>personne autorisée</code>.</td></tr>
<tr><td>PROTAGONISTE PRINCIPAL</td><td>PROCÈS</td></tr>
<tr><td>{ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR}

<code>responsable du traitement</code></td><td>garantit la confidentialité des données à caractère personnel
assure la confidentialité des données à caractère personnel
préserve la confidentialité des données à caractère personnel</td></tr>
<tr><td colspan="2">CONTEXTE DÉFINITOIRE

3. La confidentialité des données. Seules les personnes autorisées peuvent accéder aux données personnelles contenues dans un fichier. Il s'agit des destinataires explicitement désignés pour en obtenir régulièrement communication et des «tiers autorisés» ayant qualité pour les recevoir de façon ponctuelle et motivée (ex. : la police, le fisc). [CNIL]

OBLIGATION

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE LEGISLATIF} oblige un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou un {ACTEUR DE L'INTERNET : UTILISATEUR} à réaliser une {ACTIVITÉ ou ACTION} concernant la confidentialité des données à caractère personnel.

La loi a donc imposé au responsable du traitement des données personnelles, et plus largement à ses sous-traitants, une obligation de sécurité et de confidentialité, assorties de sanctions pénales. [CORPUS_DONNEES]</td></tr>
</table>

Figure 52. Description des liens sémantiques et conceptuels dans le *DITerm* (Annexe VI : 60)

10.5.2.3 Participants directement et indirectement impliqués dans les situations relatives au terme vedette

Le troisième tableau présente les participants directement impliqués dans les situations relatives au terme vedette. Autrement dit, il s'agit des actants identifiés au moment de la description de la structure actancielle des unités terminologiques (voir *Forme propositionnelle* et les sections 7.2.1 et 8.2.2). Comme dans les cas précédents, la présentation des liens tient sur deux colonnes. Ainsi, la colonne de gauche précise le type de lien. Pour ce faire, nous reprenons la mention de l'actant typique qui apparaît déjà dans la forme propositionnelle. L'actant typique devient une sorte d'`étiquette` exploitée dans la partie suivante de notre description. Quant à la colonne droite, elle contient une liste des réalisations linguistiques des actants qui ont été observées dans le corpus *DITerm*. Ces réalisations linguistiques sont associées à des cadres sémantico-conceptuels (au moyen d'une étiquette sémantico-conceptuelle correspondante) et sont regroupées en fonction de critères sémantiques, des précisions sémantiques étant mentionnées entre crochets.

Le quatrième tableau décrit les participants indirectement impliqués dans les situations relatives au terme. En effet, il s'agit des participants extérieurs, c'est-à-dire, de différents intervenants qui ne remplissent pas le rôle des actants (voir la section 7.2.1) mais entretiennent avec le terme-clé un certain nombre de relations d'ordre conceptuel.

10.5.2.4 Relations transversales

Le cinquième tableau regroupe des relations transversales, c'est-à-dire des relations conceptuelles propres au droit de l'Internet. Les relations en question permettent de rattacher le terme donné à la structure conceptuelle du domaine et reflètent son appartenance à un cadre sémantico-conceptuel (voir la section 9.2.1). En effet, il s'agit des liens indirects, c'est-à-dire des liens qui unissent des termes qui ne sont pas des cooccurrents directs. Ces liens sont identifiés grâce à l'annotation de l'environnement contextuel du terme au moyen des cadres sémantiques à la Fillmore (voire la section 9.2.2) et grâce à la mise en place du système d'étiquetage sémantico-conceptuel. Le tableau des relations transversales constitue donc un récapitulatif des relations conceptuelles capturées tout au long de la description du terme présentée dans l'article.

10.5.2.5 Situations relatives au terme vedette

Les tableaux suivants décrivent différentes relations qui interviennent sur l'axe syntagmatique. Il s'agit aussi bien des collocations à proprement parler que des syntagmes de haute fréquence dans notre corpus qui reflètent des liens de forte compatibilité conceptuelle. Les deux types de relations sont présentés selon le même schéma, notamment au moyen d'une formule en français standardisé suivie des réalisations linguistiques possibles. S'il s'agit d'une relation collocationnelle, la formule qui l'explicite correspond à une glose de vulgarisation de FL. En revanche, si c'est un lien basé sur une forte compatibilité conceptuelle, nous faisons appel à la paraphrase entendue au sens plus large. Ainsi, chaque relation est présentée comme une situation composée des éléments suivants :

- PROTAGONISTE PRINCIPAL (qui correspond au sujet grammatical),
- PROCÈS (qui englobe les éléments de base du syntagme verbal, c'est-à-dire le verbe et son COD),
- AUTRES PARTICIPANTS (s'ils jouent le rôle d'actant) ou CIRCONSTANTS (s'ils correspondent à des circonstants – voir la section 8.2.4).

Le terme vedette, en italique, apparait : a) soit comme le protagoniste principal impliqué dans un procès (il occupe donc la fonction de sujet grammatical) ; b) soit comme un élément du procès (il occupe ainsi la place de complément d'objet direct). La deuxième partie du tableau présente les réalisations linguistiques possibles de chaque élément de la situation (observées dans le corpus *DITerm*). En effet, la plupart des éléments impliqués dans la situation décrivant une relation donnée correspondent à des participants définis dans les rubriques précédentes, notamment aux actants et aux participants extérieurs. Pour apporter plus de clarté, nous proposons donc de reprendre la mention des actants typiques et des participants extérieurs typiques introduits plus haut. Comme nous l'avons vu plus haut, ces termes typiques ouent le rôle des étiquettes sémantiques. Dans le cas où seule une partie des réalisations linguistiques regroupées sous l'étiquette de l'actant typique a été repérée dans une situation donnée (dans le corpus *DITerm*), nous apportons des précisions sémantiques qui apparaissent entre crochets. Quant aux circonstants, ils sont précédés d'une mention, mise en gras, relative au type de relation circonstancielle évoquée (ex. **manière**, **moyen**, **instrument**, **lieu**) et de précisions concernant le sens d'un circonstant donné, mises entre crochets. Nous rapportons les relations circonstancielles les

plus caractéristiques pour une situation donnée. Pour terminer, soulignons que tous les éléments impliquées dans une situation donnée sont associés à des étiquettes sémantico-conceptuelles ce qui permet de les connecter à la structure notionnelle du domaine du droit de l'Internet et en même temps d'uniformiser la description des relations. Les Figures 52 et 53 reproduisent deux exemples de situations décrites dans le *DITerm*.

<table>
<tr><td colspan="3">SITUATIONS RELATIVES à contenu illicite</td></tr>
<tr><td colspan="3">SITUATION : Le protagoniste principal communique un contenu illicite sur Internet.</td></tr>
<tr><td>PROTAGONISTE PRINCIPAL</td><td>PROCÈS IMPLIQUANT contenu illicite</td><td>CIRCONSTANTS</td></tr>
<tr><td>internaute

{ACTEUR DE L'INTERNET}

[personne qui utilise l'Internet]
[personne qui crée le contenu illicite en ligne]
[personne qui fournit le contenu illicite en ligne]
[personne qui intervient sur le contenu illicite en ligne]</td><td>met en ligne un contenu illicite
poste un contenu illicite
publie un contenu illicite
dépose un contenu illicite
met à disposition du public un contenu illicite
fournit un contenu illicite</td><td>manière :

[en grande quantité]

massivement (adv.)</td></tr>
<tr><td colspan="3">INFRACTION

Un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} commet une {INFRACTION} quand il réalise une {ACTIVITÉ ou ACTION} : [l'action de communiquer un contenu illicite sur Internet].

L'internaute qui met en ligne des contenus illicites comme des œuvres (vidéo, musique etc...), protégées par le droit d'auteur, sans autorisation de l'auteur, que ce soit en téléchargement ou streaming commet un délit de contrefaçon de droit d'auteur. [RDLI_2007]

RESPONSABILITÉ

{ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} n'est pas/est responsable d'une {INFRACTION} quand il</td></tr>
</table>

réalise une {ACTIVITÉ ou ACTION} concernant *contenu illicite* : [l'action de communiquer un *contenu illicite*].

L'éditeur d'un site internet et l'auteur qui mettent en ligne un contenu litigieux sur ce site sont responsables des atteintes au droit à l'image, aux droits d'auteur et aux droits voisins du droit d'auteur. [RDLI_2008]

PRÉJUDICE

Une {ACTIVITÉ ou ACTION} concernant *contenu illicite*: [l'action de communiquer un *contenu illicite*]nuit à {AYANT-DROIT}, à un {DROIT} de quelqu'un

a) Le non-retrait d'informations ou la mise en ligne de contenus illicites qui portent atteinte aux droits de tiers et sont signalés aux prestataires de services nous ramène à un schéma classique : mise en œuvre de leur responsabilité civile pour le préjudice subi par la personne ou encore condamnations pour injure, diffamation, atteinte à la vie privée. [RDLI_2012]

TERMES LIÉS : droit (n. m.) d'auteur {DROIT}, droits voisins du droit d'auteur {DROIT}

Figure 53. Exemple d'annotation des contextes au moyen des cadres sémantiques (Annexe VI : 33)

Conclusion

Le but de la présente recherche était de proposer une méthode de description complète des unités terminologiques du domaine du droit de l'Internet, devant servir de base à la conception d'un dictionnaire spécialisé destiné aux traducteurs dont la langue de travail est le français. Il s'agit d'un modèle hybride permettant de rendre compte à la fois de la dimension linguistique et conceptuelle des termes et plus précisément, des relations lexico-sémantiques et des liens conceptuels que ces derniers entretiennent avec d'autres unités terminologiques et lexicales appartenant au vocabulaire du domaine. Rappelons que le projet s'est construit sur la base de la constatation que les ressources terminographiques traditionnelles ne fournissent pas suffisamment d'information sur le fonctionnement du terme dans l'univers discursif et conceptuel du domaine. Tout au long de notre étude, nous avons donc essayé de répondre à la question posée en introduction :

Comment peut-on rendre les dictionnaires destinés à la traduction spécialisée plus performants et plus utiles aux traducteurs ?

Pour ce faire, nous avons décidé d'étudier les unités terminologiques sous différents angles.

Le terme en tant qu'attribut du texte de spécialité : constitution du corpus *DITerm* et extraction des candidats-termes

Comme nous l'avons montré dans la première partie de la thèse, consacrée à la présentation des fondements théoriques de notre recherche, le texte doit être à la base de tout projet terminographique, il est une valeur confirmée. On ne peut plus nier les influences de la linguistique de corpus (voir le chapitre 3) ni les acquis de la terminologie textuelle et computationnelle (voir le chapitre 4) Comme le remarque Bourigault (1999 : 30) : « *(...) la terminologie doit venir des textes pour mieux y retourner* ». En effet, l'unité terminologique n'est plus considérée comme une unité que l'on observe à l'état isolé mais comme un élément d'un ensemble qui doit être analysé en contexte. La constitution et l'analyse d'un corpus textuel équilibré et représentatif du domaine du droit de l'Internet se sont donc imposées

comme des étapes essentielles et incontournables de notre démarche. Cependant, avant de passer à l'élaboration du corpus *DITerm* (c'est ainsi que nous avons baptisé notre projet), il fallait tout d'abord délimiter deux concepts de base, à savoir *corpus spécialisé* et *texte de spécialité*. En effet, considéré comme un habitat privilégié des termes, un texte de spécialité ne se caractérise pas seulement par une haute densité terminologique mais aussi par des constructions linguistiques spécifiques, des regroupements syntagmatiques privilégiés, des moyens d'expression conventionnels qui structurent le discours spécialisé et facilitent la communication entre experts. Néanmoins, nous avons constaté qu'il était difficile d'établir le caractère spécialisé d'une langue/discours/texte en termes uniquement linguistiques. Ainsi, afin de déterminer les critères permettant d'identifier un texte spécialisé du domaine du droit de l'Internet, nous nous sommes basée sur des éléments extralinguistiques tels que: domaine, sujet, type d'interlocuteurs, fonction d'usage spécialisé, circonstances particulières de la situation de communication spécialisée, éléments qui selon nous sont plus opérationnels que les critères linguistiques. Nous avons également porté notre attention sur la spécificité des textes (discours) juridiques dont le rôle principal n'est pas la transmission de l'information ni de la connaissance mais plutôt l'établissement ou l'explication des règles.

Une fois la notion de *spécialité* cernée, nous sommes passée à la définition des critères sur lesquelles allaient reposer la sélection des documents faisant partie de notre corpus. Rappelons que le choix rigoureux des textes garantit la qualité de la recherche terminologique. Les critères tels que l'appartenance au domaine du droit de l'Internet par le biais du sujet, la taille et les dates de publication, le niveau de spécialisation, les variétés de genres et les variétés de situation de communication ont été déterminés en adéquation avec les objectifs visés et dans le but de garantir la représentativité et l'équilibre du corpus *DITerm*. À l'occasion, nous avons évoqué un certain nombre de difficultés liées à la construction d'un corpus du domaine du droit de l'Internet telles que le caractère pluridisciplinaire du domaine, l'hétérogénéité thématique, la profusion des textes et la pluralité des sources.

L'extraction des candidats-termes a été réalisée à l'aide de deux outils, à savoir *TermoStat* et *NooJ*. Soulignons que le repérage de termes du domaine du droit de l'Internet s'est avéré une tâche complexe et laborieuse, et ce malgré l'apport incontestable des outils informatiques. Pour tirer le meilleur parti de ces derniers, nous avons décidé de combiner trois

types d'indices permettant de reconnaître les termes dans une masse au préalable indifférenciée d'unités lexicales :

- indices quantitatifs qui s'appuient sur les méthodes statistiques et concernent la fréquence et la répartition d'un phénomène dans le corpus
- indices linguistiques qui correspondent à des indices formels, distributionnels et lexico-sémantiques
- extralinguistiques qui reflètent l'appartenance des termes au domaine en question

Rappelons que la linguistique de corpus utilise deux types d'approches (Leech 1992, 2005, Sinclair 1991, 2004) : l'approche déductive (dite *corpus-based*) qui fait appel aux données textuelles afin de confirmer des hypothèses théoriques et l'approche inductive (dite *corpus-driven*) qui explore les données sans *a priori* dans le but de repérer et définir des phénomènes linguistiques directement à partir des textes. Nous n'avons pas voulu opter pour une approche plutôt que pour une autre et nous avons préféré combiner les deux méthodes, sachant que chacune apporte des avantages et des inconvénients. Bien évidement, tout au long de notre analyse, nous avons essayé de rester au plus près des phénomènes qui apparaissent dans le corpus. Pourtant, nous ne nous sommes pas interdit d'utiliser des connaissances *a priori* sur la langue. Nous avons donc divisé les indices en deux groupes : les indices guidés par les données et les indices guidés par les hypothèses. Quant à ces derniers, nous avons constaté que le statut terminologique d'une unité se lie fortement à des particularités des contextes discursifs où l'unité s'insère (Krieger 2002). Ainsi, nous avons décidé de tirer avantage de phénomènes linguistique propres au discours juridique (à savoir : la condensation nominale, le recours à la nominalisation, la présence de syntagmes nominaux aux modes de formation ultracomplexes), afin d'accéder aux termes du domaine du droit de l'Internet. Pour ce faire, nous avons mis en œuvre des techniques consistant en le repérage des fonctionnements considérés comme réguliers. Nous avons défini un certain nombre de patrons morphosyntaxiques sous forme de règles que nous avons projetées sur le corpus afin de localiser les séquences correspondantes. Soulignons qu'outre un nombre important de candidats-termes, un examen minutieux des résultats nous a permis d'esquisser un portrait thématique du corpus et de prendre connaissance, par le biais des unités extraites, des notions les plus importantes du domaine.

Remarquons cependant que pour pouvoir valider le statut terminologique des candidats-termes et prendre la décision finale quant à la nature des unités à retenir, nous avons été amenée à faire appel à des indices extralinguistiques tels que l'appartenance au domaine et l'adéquation au projet visé. En étudiant les données extraites du corpus *DITerm*, nous avons dû prendre en compte trois questions, notamment : l'aspect pluridisciplinaire du droit de l'Internet, la nature déontique du langage juridique et l'absence de frontières rigides entre la langue générale et la langue spécialisée. Ainsi, l'analyse des candidats-termes nous a permis de dégager dix catégories différentes d'unités qui entrent dans la composition du vocabulaire du droit de l'Internet et qui représentent une sorte de nomenclature de travail. En effet, d'un côté, nous avons dégagé une liste des unités terminologiques constituant le noyau de la terminologie du domaine et destinées à faire l'objet des articles du dictionnaire. De l'autre côté, nous avons retenu un grand nombre d'unités caractérisées par différents niveaux de *spécialité* et correspondant à ce que nous avons appelé le *vocabulaire annexe*. Même s'il ne s'agit pas là des termes *stricto sensu*, ces unités contribuent à la structuration du domaine du droit de l'Internet aussi bien au niveau lexical que conceptuel.

Le terme en tant qu'*unité terminologique étendue* – analyse de l'environnement contextuel des termes extraits du corpus *DITerm*

Une fois la nomenclature du domaine sélectionnée, nous avons utilisé chaque terme comme nœud (Sinclair 1991, Pearson 1998) afin d'étudier le comportement de ces unités terminologiques dans leur univers discursif et de relier les termes à leurs cooccurrents en observant les contextes dans lesquels ils apparaissent. Notre méthodologie d'exploration a consisté en le repérage manuel d'unités significatives apparaissant dans les contextes (ou plutôt co-textes) associés à des termes clés. Sous le nom *unité significative*, nous avons désigné toutes sortes de données textuelles récurrentes fournissant des informations sur le sens et l'usage des termes ainsi que des indices sur l'organisation conceptuelle du domaine. En proposant d'isoler les segments répétés comportant l'unité nodale choisie, nous avons voulu renouer avec la tradition contextualiste et plus particulièrement avec les travaux de Sinclair (1991, 2004, voir aussi Tognini-Bonelli 2001), pour qui seules les unités prises dans leur contexte ont du sens. Comme nous l'avons montré dans la partie théorique de notre travail (voir le chapitre 3), selon les contextualistes, il est impossible de proposer une description linguistique d'une unité isolée, extraite de son environnement discursif, car ce qu'elle représente et signifie est déterminé par les éléments qui l'entourent. Rappelons que

d'après Sinclair (voir la section 3.4.2), l'unité lexicale doit être définie comme une *unité lexicale étendue* qui s'articule autour d'un noyau mais s'étend à des unités proches liées entre elles à différents degrés et sélectionnées en fonction de critères d'affinité. Le sens y est donc représenté comme un résultat de combinaisons de plusieurs éléments linguistiques en contexte.

Pour ce qui est de notre étude, nous avons constaté que l'unité terminologique du domaine du droit de l'Internet peut être considérée comme une *unité terminologiques étendue* dont le sens spécialisé ainsi que ses caractéristiques d'ordre conceptuel peuvent être déterminés en fonction d'éléments qui se manifestent d'une manière ou d'une autre dans son environnement contextuel. Nous avons également présumé que l'organisation de ces éléments sur l'axe syntagmatique reflète un système de relations abstrait et virtuel qui présente un double caractère aussi bien linguistique que conceptuel (nous nous inspirons du concept purement linguistique de *système* introduit par Halliday (voir la section 3.4.1.3) mais nous l'employons dans un autre sens, plus large et plus abstrait). La cooccurrence linéaire doit donc être interprétée comme une réalisation des choix faits au niveau du système et son analyse doit permettre d'accéder à ce système. Ainsi, l'analyse des concordances des termes (réalisée à l'aide du logiciel *NooJ*) nous a permis de réaliser les actions suivantes :

- extraire les cooccurrents les plus fréquents des termes (relevant aussi bien de la combinatoire restreinte que de la combinatoire libre)
- établir la structure actancielle des termes
- dégager les relations circonstancielles
- extraire les relations paradigmatiques que le terme entretient avec d'autres unités
- repérer les relations conceptuelles spécifiques au domaine du droit de l'Internet par le biais des termes génériques identifiés préalablement
- repérer des informations encyclopédiques sur le domaine de spécialité et extraire des définitions

Le terme en tant qu'*unité terminologique à sens prédicatif* – modélisation des propriétés linguistiques des termes du droit de l'Internet

En effet, nous avons constaté que les unités lexicales ou terminologiques gravitant autour des termes clés du domaine entretiennent avec ces derniers de nombreuses relations de différentes natures permettant de refléter la structure (aussi bien lexicale que conceptuelle) du domaine. Face à cette multitude de données extraites du corpus, nous nous sommes donc posé une série de questions d'ordre méthodologique. Comment systématiser les données ? Comment expliciter cette variété des liens caractérisant les termes ? Quel formalisme adopter pour décrire toute la richesse des informations ? En partant de l'hypothèse que le terme est une unité lexicale à sens spécialisé, nous avons admis qu'il était possible de l'analyser et de le décrire en s'appuyant sur des modèles empruntés à la sémantique lexicale. Nous nous sommes donc intéressée au modèle des fonctions lexicales développé par Mel'čuk et ses collaborateurs, Alain Polguère et André Clas, dans la cadre de la Lexicologie Explicative et Combinatoire (LEC) (Mel'čuk *et al.* 1995), qui constitue une composante d'une théorie plus générale, à savoir la Théorie Sens-Texte (TST) (Mel'čuk 1997). En effet, comme nous avons pu le voir dans la partie théorique de ce travail (voir le premier chapitre), l'approche mel'čukienne offre une méthode de description globale de l'unité lexicale. L'originalité des FL est de proposer un modèle fonctionnel unique qui permet de rendre compte de façon uniforme de deux types de phénomènes : les dérivations sémantiques et les collocations. Les FL mettent ainsi en lumière une multitude de relations qu'une lexie entretient avec d'autres unités aussi bien sur l'axe paradigmatique que sur l'axe syntagmatique. Cependant, en faisant appel à cet outil, il est nécessaire de prendre en compte le fait que les relations décrites au moyen des fonctions lexicales sont toutes ancrées dans le contenu sémantique de la lexie. En effet, la description lexicale réalisée à l'aide des FL s'appuie fortement sur la décomposition du sens, et il faut savoir que dans ce processus, l'identification de la nature prédicative d'une lexie donnée et la définition de sa structure actancielle constituent une étape essentielle. Soulignons tout de même que dans notre cas, l'identification de la structure actancielle s'est avérée une tâche délicate. En effet, la plupart des termes extraits du corpus *DITerm* ont été considérés soit comme des noms prédicatifs soit comme des quasi-prédicats (voir les sections 7.2.2 et 8.1) ; or il n'est pas toujours facile de définir la structure actancielle d'un nom, surtout d'un nom à sens compositionnel.

De plus, nous avons constaté qu'il n'était pas non plus évident de faire la distinction entre les participants obligatoires, optionnels et circonstants et de prendre une décision adéquate quant au nombre d'actants devant apparaître dans la forme propositionnelle du terme vedette. Rappelons que pour être considéré comme un actant, un sémantème doit remplir deux

conditions : il doit, d'une part correspondre à un participant de la situation dénoté par ⁽S⁾ et, d'autre part, il doit pouvoir s'exprimer auprès de l'expression lexicale de ⁽S⁾ dans la phrase (Polguère 2012 : 5). Remarquons également que dans la TST (Mel'čuk 2004 : 10), une situation dénotée par l'unité prédicative est une situation linguistique, à savoir une situation telle qu'elle est reflétée par la langue et exprimée dans les emplois de l'unité à sens prédicatif. La structure actancielle ne constitue pas une représentation schématique d'une situation réelle. Or, dans un projet terminologique (comme le nôtre), on est constamment tenté de mettre en évidence d'autres types de participants qui relèvent plutôt de l'ordre conceptuel.
L'étape suivante a consisté en l'encodage des relations repérées dans le corpus au moyen des FL. Nous nous sommes intéressée en particulier :

a) aux relations paradigmatiques car il est relativement facile d'établir une correspondance entre ces dernières et les liens taxinomiques, représentant le groupe de relations qui ont retenu la plus grande attention en terminologie
b) aux collocations verbales et surtout aux verbes supports et aux verbes de réalisation car ils permettent, respectivement, d'exprimer le sens des unités terminologiques à sens prédicatif et de décrire les réalisations typiques correspondant à ces dernières.
c) aux relations actancielles qui, mis à part leur rôle dans la définition du sens des unités terminologiques prédicatives, constituent un point d'entrée au schéma conceptuel du domaine mettant en évidence un certain nombre de participants impliqués dans les situations évoquées par les termes
d) aux relations circonstancielles car elles constituent une part importante des relations extraites du corpus *DITerm* et, tout comme les relations actancielles, elles représentent un point d'entrée au schéma conceptuel du domaine

L'application des FL à notre projet terminographique a montré que malgré leur aspect jugé parfois trop aride et abstrait, les fonctions lexicales constituent un outil exceptionnel apportant une réponse adéquate au problème, très complexe, de la modélisation des relations lexicales entre les termes. En effet, il s'agit d'un modèle très puissant permettant de décrire de façon uniforme une multitude de relations qu'un terme du domaine de l'Internet entretient avec d'autres unités et ceci aussi bien sur l'axe paradigmatique que sur l'axe syntagmatique. De plus, le formalisme des FL assure rigueur et systématicité dans la description tout en permettant de la rendre plus compacte. En effet, le recours au formalisme des FL nous a

permis de saisir la véritable nature des liens lexico-sémantiques qui relient les termes à d'autres unités appartenant au vocabulaire du droit de l'Internet.

Néanmoins, nous avons constaté que les FL standard ne permettent pas de couvrir l'ensemble des dérivations sémantiques et des collocations recensées dans le corpus *DITerm*. Ainsi, pour décrire les relations dont le sens est spécifique, le modèle offre la possibilité de faire appel à des FL semi-standard et à des FL non standard. Cependant, cette solution ne nous a pas paru satisfaisante car le recours simultané aux trois types de fonctions (FL standard, semi-standard et non standard) rendrait la description peu systématique. En outre, les FL, de par leur nature (comme nous l'avons souligné plus haut, elles sont toutes ancrées dans le contenu sémantique des unités), ne permettent pas de systématiser les liens d'ordre conceptuel. Or, nous avons remarqué que la modélisation des relations entre les unités terminologiques appartenant au domaine du droit de l'Internet demande parfois une analyse en termes de référents et non pas en termes de constituants linguistiques. Nous avons montré que dans certains cas, il serait possible de contourner ce problème méthodologique en créant délibérément des définitions sémantiques permettant de tisser et d'expliciter des liens qui ne relèvent pas directement de la combinatoire restreinte. Cependant, cette démarche pourrait conduire à dénaturer le système des FL, qui exige une distinction nette entre les phénomènes tels que combinatoire libre et combinatoire restreinte, liens de dérivations sémantiques et liens conceptuels forts. Nous avons donc décidé de faire appel à deux outils formels distincts qui pourraient se présentent, d'une manière uniforme, sous forme de formules explicites en français standardisé

Le terme en tant qu'élément d'un *cadre sémantico-conceptuel* – description des relations conceptuelles

Ainsi, le modèle mel'čukien n'étant pas adaptée à la formalisation des relations intervenant sur le plan conceptuel, nous avons été amenée à trouver une autre façon de systématiser les données d'ordre cognitif, ce qui a par ailleurs demandé un certain recul par rapport à la réalité linguistique. Nous avons tout de même tenu à rester au plus près des observations faites sur l'ensemble des données linguistiques provenant du corpus *DITerm*. Pour ce faire, nous avons décidé de tirer parti de l'opération de qualification juridique dont les résultats se manifestent dans le discours par une présence importante de termes génériques reconnus comme étant porteurs des notions fondamentales du droit. Il est apparu que la prise

en compte du mécanisme de qualification juridique, conçue comme un niveau intermédiaire de conceptualisation, permet de réaliser trois actions :

a) accéder à des termes concrets spécifiques du droit de l'Internet
b) relier les termes du domaine de l'Internet à des catégories conceptuelles plus abstraites, définies comme des cadres sémantico-conceptuels
c) dégager le schéma conceptuel du domaine du droit de l'Internet par le rapprochement des catégories abstraites identifiées suite à l'analyse des données extraites du corpus *DITerm* avec les éléments de l'ontologie de concepts juridiques fondamentaux de LIKIF-Core (Breuker *et al.* 2007)

En effet, en faisant référence à la théorie des cadres (Minsky 1975, Martin 2007, 2008), nous avons constaté que les termes du droit de l'Internet et plus particulièrement les relations d'ordre conceptuel qu'ils entretiennent avec d'autres termes du domaine pourraient être définis par le biais des cadres sémantico-conceptuels auxquelles ils sont rattachés. Nous avons remarqué que les cadres en question évoquaient des situations qui pourraient être décrites en prenant en compte des contextes particuliers et les relations qui y sont liées. Par conséquent, chaque terme, en raison de son appartenance à un cadre sémantico-conceptuel donné, se caractériserait par un ensemble défini de relations d'ordre conceptuel. Ces dernières, bien qu'établies en fonction des éléments extralinguistiques (juridiques), se manifestent, d'une manière ou d'une autre, dans le discours. Il s'est pourtant avéré que la mise en évidence de ces relations n'est pas une tâche facile.

Voulant rester dans une approche résolument linguistique, nous nous sommes donc tournée vers un modèle d'annotation sémantique inspiré de la sémantique des cadres (voir Fillmore 1982, Fillmore et Baker 2009) et plus précisément du projet *FrameNet* (Baker 2009, Fillmore *et al.* 2003). Ainsi, nous avons proposé d'introduire dans notre analyse deux types de *frames*, correspondant à deux niveaux de conceptualisation différents. En effet, tandis que le *cadre sémantico-conceptuel* renvoie à une structure conceptuelle permettant de rattacher un terme donné à un schéma notionnel du domaine, le *cadre sémantique* doit être appréhendé comme une structure schématique ancrée dans la réalité linguistique permettant de saisir les relations du terme par le biais de leur fonctionnement dans l'univers discursif. Nous avons constaté que les termes, associés (sur le plan conceptuel) à des cadres sémantico-conceptuels, se réalisent dans le discours (c'est-à-dire, sur le plan linguistique), dans des combinaisons de formes linguistiques typiques du domaine du droit qui peuvent être analysées au moyen des

cadres sémantiques « juridiques ». Soulignons tout de même qu'en règle générale et contrairement à d'autres projets basés sur la sémantique des cadres (voir la section 9.2.2.1), les termes décrits n'évoquent pas les cadres sémantiques choisis pour la description. En revanche, ils font partie des constellations gravitant autour des unités évoquant ces cadres qui, quant à elles, appartiennent au vocabulaire juridique général et ne font pas l'objet de notre description. Bref, nous avons proposé de faire appel à l'annotation au moyen des cadres sémantiques, car elle permet selon nous de visualiser l'interaction des termes étudiés avec d'autres unités et de définir les types de liens qui les unissent.

Ainsi, tout au long de cette étude, nous avons essayé de démontrer que le terme, unité à dimensions multiples, peut être considéré sous différents angles, notamment :

- comme un *attribut* du texte de spécialité.
- comme une *unité terminologique étendue* dont le sens spécialisé ainsi que les caractéristiques d'ordre conceptuel peuvent être déterminés en fonction des éléments qui se manifestent d'une manière ou d'une autre dans son environnement contextuel.
- comme une *unité terminologique à sens prédicatif* qui, de par sa nature linguistique, entretient de nombreuses relations lexico-sémantiques avec d'autres unités.
- comme un élément d'un *cadre sémantico-conceptuel* qui le relie au schéma notionnel du domaine.

Nous avons également essayé de démontrer que, quelle que soit la perspective, il est possible d'analyser la nature du terme ainsi que le caractère de ses relations tout en restant dans une démarche linguistique. Cette démarche nous a conduite à proposer un modèle de description hybride des unités terminologiques du domaine du droit de l'Internet qui fournit aux traducteurs une variété d'expressions précises, de combinaisons adéquates et de formulations appropriées à ce domaine de spécialité ainsi que des indices de nature conceptuelle permettant le traitement des informations contenues dans les textes à traduire. Afin de rendre compte des liens entretenus par les termes, nous avons fait appel à des formules explicites fondées sur une sorte de métalangage qui présente l'avantage d'être adapté aussi bien à la description des relations sémantiques que conceptuelles.

En effet, le modèle *DITerm* propose de décrire le terme comme le centre d'une constellation d'où partent de nombreux termes coordonnés[105]. Tout comme Lundquist (Lundquist 1998, cité dans Dancette et Halimi, 2007 : 533), nous considérons que le texte de spécialité se construit autour du concept clé qui appelle les idées associées. Le recours à ce concept active toute une série d'associations menant à un ensemble de termes liés directement ou indirectement au concept ainsi qu'à un ensemble de formules linguistiques permettant d'exprimer ces termes en contexte. Notre modèle de description des données terminologiques tente donc de rassembler et d'organiser ces éléments afin de guider le traducteur dans sa découverte du domaine.

Comme nous l'avons annoncé en introduction, le modèle de description des unités terminologiques du droit de l'Internet proposé dans le cadre de cette étude doit servir de base à la conception d'un dictionnaire spécialisé du domaine. À l'heure actuelle, il s'agit donc d'un échantillon de dictionnaire, d'un projet en construction qui, comme nous l'espérons, pourra faire l'objet d'une recherche post-doctorale. Une nomenclature cible, dont la description correspondrait à la deuxième étape de développement de notre base de données terminographique, a déjà été constituée. Bien évidement, nous n'excluons pas la possibilité de l'élargir à d'autres termes du domaine. Cela demanderait pourtant une mise à jour de notre corpus, action qui semble tout à fait naturelle dans le cadre d'un domaine en plein essor.

Nous considérons également que le travail de modélisation mené sur un nombre plus important de termes permettrait de dégager plus de régularités et d'affiner l'organisation des données contenues dans les articles du dictionnaire. Nous pensons tout particulièrement au développement du système de double étiquetage sémantique et sémantico-conceptuel et à la hiérarchisation des relations collocationnelles et conceptuelles présentées dans notre modèle sous forme de *situations*. En s'inspirant de la modélisation des collocatifs verbaux proposée dans le *DiCoInfo* (L'Homme 2009), nous supposons qu'il serait possible de regrouper les différentes *situations* dans des ensembles reflétant une certaine organisation au niveau conceptuel.

Nous envisageons aussi de situer notre futur projet dans une perspective multilingue (espagnol – français – polonais) et de proposer (outre les équivalents en espagnol et en

[105] Nous renouons ici avec l'idée des groupes d'intervention de Cornu (2005 : 202)

polonais qui apparaissent actuellement dans les entrées), la description des relations dans ces deux langues. Cela nécessitera la constitution d'un corpus trilingue qui à notre avis, devrait être divisé en deux blocs : d'un côté, un ensemble de sous-corpus parallèles en espagnol, français et polonais composés de textes communautaires ; de l'autre côté, un ensemble de sous-corpus comparables dans ces trois langues, composés de textes juridiques nationaux.

Finalement, nous considérons que notre projet, dont l'objectif principal est de refléter tout un réseau de relations, est inadapté à une version papier et devrait absolument faire l'objet d'une informatisation. Le format d'un dictionnaire en ligne et les possibilités qu'offre le numérique (listes déroulantes avec des liens hypertexte) allégeraient considérablement la présentation des données contenues dans les fiches et en faciliteraient la consultation.

Bibliographie

AARTS (Jan), 1991, Intuition-based and Observation-based Grammars. In K. Aijmer et B. Altenberg (eds.): *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* Londres/New York : Routledge, pp.44-62.

AUSSENAC-GILLES (Nathalie), CONDAMINES (Anne) et SZULMAN (Sylvie), 2002, Prise en compte de l'application dans la constitution de produits terminologiques. In *Actes des 2e assises nationales du GDR i3 Information, Interaction, Intelligence*. Nancy : Cépaduès, pp. 289–303.

AUSSENAC-GILLES (Nathalie), SÉGUÉLA (Patrick), 2000, Les relations sémantiques : du linguistique au formel. In *Cahiers de Grammaire 25, « Sémantique et corpus »*, pp. 175-198

BACHIMONT (Bruno), 2000, Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In J. Charlet, M. Zacklad, G. Kassel & D. Bourigault (éds.), *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, Chapitre 19, Paris : Eyrolles.

BAKER (Collin F.), 2009, La sémantique des cadres et le projet FRAMENET: une approche différente de la notion de « valence ». In *Langages*, 2009/4 (n° 176), pp. 32-49.

BALLY (Charles), 1951, *Traité de syntaxique française*, (troisième édition), Genève/Paris Éditions KLINCKSIECK.

BÉJOINT (Henri), 2007, Nouvelle lexicographie et nouvelles terminologies. In L'Homme, Marie-Claude et Vandaele Sylvie (dir.), *Lexicographie et terminologie : compatibilité des modèles et des méthodes.* Presses de l'Université d'Ottawa, pp. 29-78.

BÉJOINT (Henri), 2009, Lexicographie et linguistique : quelques réflexions sur le domaine anglais. In *Lexique*, n°19, pp.117-158.

BÉJOINT (Henri) et MANIEZ (François) (dir.), 2005, *De la mesure dans les termes. Hommage à Philippe Thoiron*. Coll. « Travaux du C.R.T.T », Lyon : Presses universitaires de Lyon, 456 p.

BÉJOINT (Henri) et THOIRON (Philippe) 2000, Le sens des termes. In Henri Béjoint et Philippe Thoiron (dir.), *Le sens en terminologie*, Coll. « Travaux du C.R.T.T », Lyon : Presses universitaires de Lyon, pp. 5-19.

BÉJOINT (Henri) et THOIRON (Philippe), 2010, La terminologie, une question de termes ? In *Meta : journal des traducteurs / Meta : Translators' Journal*, vol 55, n°1, Montréal : Les Presses de l'Université de Montréal, pp. 105-118.

BENSON (M.), 1989, The Structure of the Collocational Dictionary, in *International Journal of Lexicography*, vol 2 (n°1), Oxford University Press

BINON (Jean), BERTELS (Ann), VAN DYCK (Jan), VERLINDE (Serge), 2000 *Dictionnaire d'apprentissage du français des affaires*, Paris : Didier.

BINON (Jean), SELVA (Thierry), VERLINDE (Serge), 2006, Corpus, collocations et dictionnaires d'apprentissage. In *Langue française*, vol. 150, n° 150, pp. 84-98.

BOCQUET (Claude), 2008, *La traduction juridique ; fondement et méthode*. Collection Traducto, Bruxelles : De Boeck.

BOURCIER (Danièle) et FERNÀNDEZ-BARRERA (Meritxell), 2011, Relier les niveaux terminologique et conceptuel dans le domaine juridique : hypothèse sur la méthodologie *middle-out*. In *Actes de la conférence Terminologie & Ontologie : Théories et Applications Toth 2011*. Annecy, pp. 129-144.

BOURIGAULT (Didier) et JACQUEMIN (Christian), 2000, Construction de ressources terminologiques. In Jean-Marie Pierrel (ed.), *Ingénierie des langues*, Paris : Hermès, pp.215-233.

BOURIGAULT (Didier) et SLODZIAN (Monique), 1999, Pour une terminologie textuelle. In *Terminologies Nouvelles*, n°19, pp. 29-33.

BOURIGAULT (Didier), LAME G., 2002, Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit. In *Revue Traitement automatique des langues*, n° 47:1, Paris : Hermès

BOWKER (Lynne), 1998, Exploitation de corpus pour la recherche terminologique ponctuelle. In Humbley J. (dir.), *Terminologies nouvelles*, n° 18, juin 1998, pp. 22-27

BOWKER (Lynne) et L'HOMME (Marie-Claude), 2004, Ingrid Meyer, Terminologist (1957-2004). In *Terminology*, n° 12, pp. 183-188.

BOWKER (Lynne), PEARSON (Jennifer), 2002, *Working with Specialized Language - A practical guide to using corpora*, London: Routledge.

BREUKER (Joost) *et al.* 2007. *ESTRELLA. Deliverable 1.4. Ontology of Basic Legal Concepts (LKIF-Core)*, consulté le 15 septembre à l'adresse suivante : http://www.estrellaproject.org/doc/D1.4-OWLOntology-of-Basic-Legal-Concepts.pdf

CABRÉ (Maria Teresa), 1998, *La terminologie : théorie, méthode et application.* Traduit du catalan, adapté et mis à jour par Monique C. Cormier et John Humbley, Ottawa : Presses de l'Université d'Ottawa / Paris : Armand Colin, p.295

CABRÉ (Maria Teresa), 2000, Terminologie et linguistique : la théorie des portes. In Diki-Kidiri (M.), dir. *Terminologie et diversité culturelle, Terminologies nouvelles,* juin 2000, n° 21, p. 10-15.

CABRÉ (Maria Teresa), 2000, *Sur la représentation mentale des concepts : bases pour une tentative de modélisation*. In Henri Béjoint et Philippe Thoiron (dir.), *Le sens en terminologie*, Coll. « Travaux du C.R.T.T », Lyon : Presses universitaires de Lyon, pp. 20-39.

CABRÉ (Maria Teresa), 2007, La terminologie, une discipline en évolution : le passé, le présent et quelques éléments prospectifs. In L'Homme, Marie-Claude et Vandaele Sylvie (dir.), *Lexicographie et terminologie : compatibilité des modèles et des méthodes.* Ottawa : Presses de l'Université d'Ottawa, pp. 79-109.

CABRÉ (Maria Teresa), 2008, Constituer un corpus de textes de spécialité. In *Cahiers du CIEL 2007-2008*, consulté le 25.08.2015 à l'adresse suivante : < http://www.eila.univ-paris-diderot.fr/recherche/clillac/ciel/cahiers/2007-2008>

ČERMÁKOVÁ (Anna) et TEUBERT (Wolfgang), 2007, *Corpus Linguistics. A Short Introduction.* London/New York: Continuum International Publishing Group Ltd.

CHOMSKY (Noam), 1965, *Aspects of the Theory of Syntax*. Cambridge/Massachusetts: THE MIT Press.

CHOMSKY (Noam), 1986, *Knowledge of Language: Its Nature, Origin, and Use*. New York : Praeger Publishers.

COHEN, (Betty), 1986, réed. 2011, *Lexique de cooccurrents - Bourse et conjoncture économique*, Montréal, Linguatech.

CONDAMINES (Anne), 1994, Terminologie et représentation des connaissances. In *Didaskalia*, n° 5, pp. 35-51.

CONDAMINES (Anne), 1999, Approche sémasiologique pour la constitution de Bases de Connaissances Terminologiques. In V. Delavigne, M. Bouveret (dir.), *Sémantique des termes* spécialisés, Rouen : Publication de l'Université de Rouen, pp.101 - 118

CONDAMINES (Anne), 2005, Linguistique de corpus et terminologie. In *Langages*, n° 157, pp. 36-46.

CONDAMINES (Anne) et DEHAUT (Nathalie), 2011, Mise en œuvre des méthodes de la linguistique de corpus pour étudier les termes en situation d'innovation disciplinaire : le cas

de l'exobiologie. In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 56, n° 2, pp. 266-283.

CONDAMINES (Anne), REBEYROLLE (Josette), 2000, Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In J. Charlet, M. Zacklad, G. Kassel, D. Bourigault (eds.) *Ingénierie des connaissances*, Paris : Eyrolles, pp. 225-241.

CORI (Marcel) et DAVID (Sophie), 2008, Les corpus fondent-ils une nouvelle linguistique ? In *Langages*, n° 171, pp. 111-129.

CORNU (Gérard), 2005 (3ème édition), *Linguistique juridique*, Paris : Ed. Montchrestein

CRUSE (David A.), 1986, *Lexical Semantics*, Cambridge : Cambridge University Press.

DANCETTE (Jeanne), 2003, Sens, cohérence et réseaux conceptuels. *In TTR – Traduction, Terminologie et Rédaction*, XVI/I, pp. 141-159

DANCETTE (Jeanne), 2006, Les relations lexico-sémantiques dans un dictionnaire spécialisé. In Thomas Szende (dir.), *Le français dans les dictionnaires bilingues*. Paris et Genève, Champion / Slatkine, pp. 144-156.

DANCETTE (Jeanne), 2007, Semantic Relations in the Field of Retailing. In *Terminology* n° 13 (2), pp. 201-223.

DANCETTE (Jeanne), 2011a, Un dictionnaire encyclopédique plurilingue sur thésaurus. In Marc Van Campenhoudt, Teresa Lino et Rute Costa (dir.) *Passeurs de mots, passeurs d'espoir : lexicologie, terminologie et traductologie face au défi de la diversité.* Actes des 8e Journées scientifiques du Réseau de chercheurs Lexicologie, terminologie, traduction, (Lisbonne, 15-17 octobre 2009), pp. 161-176.

DANCETTE (Jeanne), 2011b, L'intégration des relations sémantiques dans les dictionnaires spécialisés multilingues : du corpus ciblé à l'organisation des connaissances. In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 56, n° 2, pp. 284-300.

DANCETTE (Jeanne) et HALIMI (Sonia), 2005, La représentation des connaissances et son apport à l'étude du processus de traduction. In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 50, n° 2, pp. 548-559.

DANCETTE (Jeanne) et L'HOMME (Marie-Claude), 2001, Modélisation des relations sémantiques dans un dictionnaire spécialisé bilingue. In *L'Éloge de la différence : La voix de l'autre, Actes des Sixièmes Journées scientifiques de l'AUF*, Beyrouth (Liban), pp. 385-400.

DANCETTE (Jeanne) et L'HOMME (Marie-Claude), 2002, The gate to knowledge in a multilingual specialized dictionary: Using lexical functions for taxomonic and partitive relations. In Proceedings Euralex 2002, Copenhagen: University of Copenhagen, pp. 597-606.

DANCETTE (Jeanne) et L'HOMME (Marie-Claude), 2004, *Building Specialized Dictionaries Using Lexical Functions.* In Rita Temmerman et Uus Knops, (dir.) Linguistic Antverpiensia 3, pp. 113-131.

DANCETTE (Jeanne) et RÉTHORÉ (Christophe), 1997, Le dictionnaire bilingue (Anglais-Français) de la distribution : entre dictionnaire de langue et encyclopédie. *In Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 42, n° 2, pp. 229-243.

DANCETTE (Jeanne) et RÉTHORÉ (Christophe), 2000 et 2006 pour la version électronique, *Dictionnaire analytique de la distribution. Analytical Dictionary of Retailing.* Montréal : Les Presses de l'Université de Montréal

DEPECKER (Loïc), 2000, Le signe entre signifié et concept. In Henri Béjoint et Philippe Thoiron (dir.), *Le sens en terminologie*, Coll. « Travaux du C.R.T.T », Lyon : Presses universitaires de Lyon, pp. 66- 126.

DEPECKER (Loïc), 2002, *Entre signe et concept. Eléments de terminologie générale.* Paris, Presses Sorbonne Nouvelle, 200 p.

DEPECKER (Loïc) et ROCHE (Christophe), 2007, Entre idée et concept : vers l'ontologie. In *Langages*, n° 168, pp. 106-114.

DESPRÈS (Sylvie), SZULMAN (Sylvie), 2008, Réseau terminologique versus Ontologie. In *Actes de la conférence Terminologie & Ontologie : Théories et Applications Toth 2008*. Annecy, pp.17-34, consulté le 21 juin 2015 à l'adresse suivante : https://hal.archives-ouvertes.fr/hal-00423525

DESPRÈS (Sylvie), SZULMAN (Sylvie), 2005 Construction d'une ontologie du droit communautaire. In *IC - 16èmes Journées francophones d'Ingénierie des Connaissances, May 2005, Nice, France*, Presses universitaires de Grenoble, pp.85-96, 2005. <hal-01025426>

DROUIN (Patrick), 2003, Term extraction using non-technical corpora as a point of leverage. In *Terminology* 9-1, pp. 99-115.

DROUIN (Patrick), 2004, Spécificités lexicales et acquisition de la terminologie. In *Le poids des mots, Actes des 7es Journées Internationales d'Analyse statistique des Données Textuelles JADT 2004*, Presses Universitaires de Louvain, pp. 345-352

DROUIN (Patrick), LANGLAIS (Philippe), 2006, Évaluation du potentiel terminologique de candidats termes. In *Actes des 8es Journées internationales d'Analyse statistiques des Données Textuelles JADT 2006*, disponible sur le site *Lexicometrica* : http://lexicometrica.univ-paris3.fr/jadt/jadt2006/tocJADT2006.htm, pp. 389- 400.

FABER (Pamela), 2011, The dynamics of specialized knowledge representation. Simulational reconstruction or the perception – action interface. In *Terminology*, 17:1, p. 9-29.

FABER (Pamela) et SANCHEZ (Maribel-T), 2001, Codifying conceptual information in descriptive terminology management. In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 46, n° 1, pp. 192-204.

FABER (Pamela) et SANCHEZ (Maribel-T), 2005, Framing Terminology : a Process-Oriented Approach. In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 50, n° 4, consulté à l'adresse suivante : http://id.erudit.org/iderudit/019916ar

FELBER (Helmut), 1987, *Manuel de terminologie*, Unesco et Infoterm.

FILLMORE (Charles), 1982, Frame Semantics. In The Linguistic Society of Korea (éds.) *Linguistics in the Morning Calm*, Seoul, Hanshin Publishing Company, pp. 111-137.

FILLMORE (Charles J.) et BAKER (Collin), 2009, A Frames Approach to Semantic Analysis. In Bernd Heine et Heiko Narrog (éd.), *The Oxford Handbook of Linguistic Analysis*, Oxford: Oxford University Press, 313-339.

FILLMORE (Charles J.), JOHNSON (Christopher R.) et PETRUCK (Miriam R.L.), 2003, Background to FRAMENET. In *International Journal of Lexicography*, n° 16 (3), pp. 235-250.

FONTENELLE (Thierry), 2009, Sémantique des cadres et lexicographie. In *Lexique* 19, P.U.S. pp. 161-179.

FRANCIS (Gill), 1993, A Corpus-Driven Approach to Grammar - Principles, Methods and Examples. In M. Baker, G. Francis et E. Tognini-Bonelli (eds.): *Text and Technology: In honour of John Sinclair*. Amsterdam / Philadelphia: John Benjamins Publishing Company, pp.137-156.

FRANCIS (Nelson), 1992, Language corpora B.C. In J. Svartvik (eds.): *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1992*. Berlin: De Gruyter, pp. 17-34.

FRAWLEY (William), 1988, New Forms of Specialized Dictionaries, in *International Journal of Lexicography*, n° 3 vol. 1, autumn 1988, pp. 189-213.

GAUDIN (François), 2003, *Socioterminologie: Une approche sociolinguistique de la terminologie*, Bruxelles : Duculot.

GAUDIN (François) et BOUVERET (Myriam), 1996, Biolex, pistes de description sémantique. In *Lexicomatique et dictionnairique*, Actes des IVèmes journées scientifiques du réseau thématique AUPELF-UREF "Lexicologie, Terminologie et Traduction", Lyon 28-29 septembre 1995, éd. Actualité scientifique AUPELF-UREF/FLA, pp. 349-357.

GÉMAR (Jean-Claude), 1991, Terminologie, langue et discours juridique. Sens et signification du langage du droit. In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 36, n° 1, pp. 275-283.

GENTILHOMME (Yves), 1995, Contribution à une réflexion sur les locutions mathématiques. In Martins-Baltar (éds.), *La locution en discours, Cahiers du français contemporain 2*, Paris : Didier Erudilion, 1995, pp. 197-242.

GRABAR (Natalia) et HAMON (Thierry), 2004, Les relations dans les terminologies structurées : de la théorie à la pratique. In *Revue d'intelligence artificielle*, Vol. 18 (1), pp. 57-85

GRIMES (Joseph E.), 1990, Inverse Lexical Functions. In J. Steele (éds.): *Meaning-Text Theory: Linguistics, Lexicography and Implications*, Ottawa: Ottawa University Press, pp. 350 – 364.

GROSS (Gaston), 1989, *Les constructions converses du français*, Coll. Langage et Cultures n° 22, Genève / Paris : Droz,

GROSS (Gaston), 1996 : *Les expressions figées en français (noms composé et autres locutions)*, Paris/Gap : Editions Ophrys

GROSS (Maurice), 1981, Les bases empiriques de la notion de prédicat sémantique. In: *Langages*, vol. 15, n°63, pp. 7-52.

GROSS (Maurice), 1998, La fonction sémantique des verbes supports. In *Travaux de Linguistique*, De Boeck & Larcier, Duculot, 37 (1), pp.25-46. <hal-00621387>

HABERT (Benoît), 2000, Des corpus représentatifs : de quoi, pour quoi, comment ? In *Cahiers de l'Université de Perpignan* n°31, p. 11-58.

HABERT (Benoît), NAZARENKO (Adeline), SALEM (André), 1997, *Les linguistiques de corpus.* U Linguistique. Paris: Armand Colin/ Mosson,

HALLIDAY (Michael Alexander Kirkwood) & HASAN (Ruqaiya), 1976, *Cohesion in English*, London: Longman Group Ltd

HALLIDAY (Michael Alexander Kirkwood) rev. par MATTHIESSEN (Christian M.I.M.), 2014, *Halliday's Introduction to Functional Grammar*. Londres / New York: Routledge.

HALLIDAY (Michael Alexander Kirkwood), 1966 [2002], Lexis as a Linguistic Level. In Halliday M.A.K. (Jonathan Webster eds.): *On Grammar. Collected Works of MAK Halliday, Vol 1*, London/New York: Continuum, pp.158 – 172.

HALLIDAY (Michael Alexander Kirkwood), 1985 [2003], Systemic Background. In Halliday M.A.K. (Jonathan Webster eds.): *On Language and Linguistics. Collected Works of MAK Halliday, Vol. 3*, London/New York: Continuum, pp.185-198.

HALLIDAY (Michael Alexander Kirkwood), 1991, Corpus studies and probabilistic grammar. In K. Aijmer et B. Altenberg (eds.): *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* Londres /New York: Routledge, pp.30-43.

HALLIDAY (Michael Alexander Kirkwood), 1992, Language as System and Language as Instance: the Corpus as a Theoretical Construct. In J. Svartvik (eds.): *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1992*. Berlin: De Gruyter, pp. 61-77.

HAMON (Thierry) et NAZARENKO (Adeline), 2002, Structuration de terminologie: quels outils pour quelles pratiques ? In *TAL* volume 43, pp. 7-18.

HAUSMANN (Franz Josef), 1979, Un dictionnaire des collocations est-il possible ? In *Travaux de linguistique et de littérature, n°* 17 (1), pp. 187-195

HEID (Ulrich) & FREIBOTT (Gerhard), 1991: Collocations dans une base de données terminologique et lexicale. In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 36, n° 1, pp. 77-91.

HJELMSLEV (Louis), 1984 *Le Langage*, Paris : Les Éditions de Minuit, 191p.

HUMBLEY (John), 2001, Compte rendu : BÉJOINT (Henri) et THOIRON (Philippe) (dir.) 2000 : Le sens en terminologie, Lyon, coll. « Travaux du C.R.T.T », 381 p. In. *Meta : journal des traducteurs / Meta : Translators' Journal*, vol 46, n°4, Montréal : Les Presses de l'Université de Montréal, pp. 728-731.

HUMBLEY (John), 2004, La réception de l'œuvre d'Eugen Wüster dans les pays de langue française. In *Les Cahiers du C.I.E.L* 2004. pp. 33- 51

HUMBLEY (John), 2007, Vers une réception plurielle de la théorie terminologique de Wüster: une lecture commentée des avant-propos successifs du manuel Einführung in die Terminologielehre". In *Langages* n° 168, pp. 82-91.

HUNSTON (Susan), FRANCIS (Gill), 2000, *Pattern Grammar: A Corpus-Driven Approach to the Lexical Grammar of English*. Amsterdam / Philadelphia: John Benjamins Publishing Company,

JACQUES (Marie-Paule), 2005, Pourquoi une linguistique de corpus ? In G. Williams (dir.): *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes, pp. 23-30.

JAKOBSON (Roman), 1963 [2003], *Deux aspects du langage et deux types d'aphasie*, in *Essais de linguistique générale*, t. I, Paris : Les Éditions de Minuit, 43-67

JOUSSE (Anne-Laure), 2010, *Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales.* Thèse de doctorat, Université de Montréal et Université Paris Diderot (Paris 7)

JOUSSE (Anne-Laure) et MORTCHEV-BOUVERET (Myriam), 2003, Lexical functions to represent derivational relations in specialized dictionaries. In *Terminology* n° 12, vol.1, pp. 71–98.

KAHANE (Sylvain) et POLGUÈRE (Alain), 2001, Formal foundation of lexical functions. In *Actes du colloque « COLLOCATION : Computational Extraction, Analysis and Exploitation»,* Toulouse, pp. 8–15.

KOCOUREK (Rostislav), 1991a, *La langue française de la technique et de la science. Vers une linguistique de la langue savante*, Deuxième édition, Wiesbaden, Brandstetter Verlag

KOCOUREK (Rostislav), 1991b, Textes et termes. In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 36, n° 1, pp. 71-76.

KRIEGER (Maria de Graça), 2002, Terminographie juridique et spécificités textuelles. In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 47, n° 2, pp. 233-243.

L'HOMME, (Marie-Claude), 1998, Caractérisation des combinaisons lexicales spécialisées par rapport aux collocations de langue générale. In Fontenelle, T. et al. (éds.), *Proceedings EURALEX '98*, Liège, 4-8 aout 1998, Liège : Université de Liège pp. 513-522.

L'HOMME (Marie-Claude), 2000, Les enseignement d'un mot polysémique sur les modèles de la terminologie. In *Cahiers de Grammaires*, 25, *« Sémantique et Corpus »*, pp. 71-91.

L'HOMME (Marie-Claude), 2002, Fonctions lexicales pour représenter les relations sémantiques entre termes. In *TAL*, vol 43 n° 1, Hermès Science Publications, Paris, pp. 19-41.

L'HOMME (Marie-Claude), 2004, *La terminologie : principes et techniques*, Montréal : Les presses de l'Université de Montréal (Paramètres).

L'HOMME (Marie-Claude), 2005a, Sur la notion de « terme ». In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 50, n° 4, pp. 1112-1132.

L'HOMME (Marie-Claude), 2005b, Conception d'un dictionnaire fondamental de l'informatique et de l'Internet : sélection des entrées. In *Le Langage et l'homme*, vol. 40, n°2, pp. 141-158.

L'HOMME (Marie-Claude), 2007, Using Explanatory and Combinatorial Lexicology to Describe Terms. In Wanner L (éds.), *Selected Lexical and Grammatical Issues in the Meaning-Text Theory*, Studies in Language Companion, vol. 84, Amsterdam, John Benjamins Publishing Company, 167- 202.

L'HOMME, (Marie-Claude), 2008, Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. In *Traduire*, nº 217, pp. 78-103.

L'HOMME (Marie-Claude), 2009, A methodology for describing collocations in a specialized dictionary. In Nielsen, Sandro and Sven Tarp (eds.), *Lexicography in the 21st*

Century: In honour of Henning Bergenholtz, John Benjamins Publishing Company, pp. 237–256.

L'HOMME (Marie-Claude), 2010, Designing Terminological Dictionaries for Learners based on Lexical semantics: the Representation of Actants. In Pedro A. Fuertes-Olivera, *Specialised Dictionaries for Learners,* Berlin / New York: Walter de Gruyter, pp. 141-154.

L'HOMME (Marie-Claude), 2015, Découverte de cadres sémantiques dans le domaine de l'environnement : le cas de l'influence objective; In *Terminàlia* n° 12, pp. 29-40, version électronique disponible à l'adresse suivante : http://terminalia.iec.cat, consultée le 31 septembre 2015

L'HOMME (Marie-Claude) et MEYNARD (Isabelle), 1998, Le point d'accès aux combinaisons lexicales spécialisées : présentation de deux modèles informatiques. In *TTR : traduction, terminologie, rédaction*, vol. 11, n° 1, pp. 199-227.

L'HOMME (Marie-Claude) et POLGUÈRE (Alain), 2008, Mettre en bons termes les dictionnaires spécialisés et les dictionnaires de langue générale. In François Maniez (dir.) *Lexicographie et terminologie : histoire de mots. Hommage à Henri Béjoint*, Lyon : Presses de l'Université de Lyon, pp. 191-206

L'HOMME (Marie-Claude) et VANDAELE (Sylvie), 2007, Lexicographie et terminologie : disciplines sœurs ou pratiques distinctes ? In L'Homme, Marie-Claude et Vandaele Sylvie (dir.), *Lexicographie et terminologie : compatibilité des modèles et des méthodes.* Presses de l'Université d'Ottawa, pp. 1-25.

L'HOMME (Marie-Claude), LEROYER (Patrick), ROBICHAUD (Benoît), 2012, Encoding collocations in DiCoInfo: From formal to user-friendly representations. In Sylviane Granger et Magali Paquot, *Electronic Lexicography*, Oxford University Press, pp. 211-236

L'HOMME (Marie-Claude), ROBICHAUD (Benoît) et SUBIRATS (Carlos), 2014, Discovering Frames in Specialized Domains. *In Language Resources and Evaluation*. Reykjavik (Islande): LREC,, pp. 1364-1371

LEECH (Geoffrey), 1991, The state of the art in corpus linguistics. In K. Aijmer et B. Altenberg (eds.): *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Londres / New York: Routledge, pp.8-29.

LEECH (Geoffrey), 1992, Corpora and theories of linguistic performance. In J. Svartvik (eds.): *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1992*. Berlin: De Gruyter, pp. 105-126.

LEECH (Geoffrey), 2005, Adding Linguistic Annotation. In M. Wynne (eds.): *Developing Linguistic Corpora: a Guide to Good Practice*. Oxbow Books, Oxford, pp. 17-29, disponible

à l'adresse suivante : http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm, consulté le 14.02.2015.

LEGALLOIS (Dominique), 2012, La colligation : autre nom de la collocation grammaticale ou autre logique de la relation mutuelle entre syntaxe et sémantique ? In *Corpus* [En ligne], n° 11, consulté le 15 avril 2015 à l'adresse suivante : http://corpus.revues.org/2202, pp. 31-54.

LEGALLOIS (Dominique) et TUTIN (Agnès), 2013, Présentation : vers une extension du domaine de la phraséologie. In *Langages*, n° 189, pp. 3-25.

LÉON (Jacqueline), 2007, Meaning by Collocation: the Firthian filiation of corpus linguistics. In D. A. Kibbee, (éds.): *History of Linguistics 2005: Proceedings of ICHoLS X, 10th International Conference on the History of Language Sciences,* Amsterdam / Philadelphia: John Benjamins Publishing Company, pp. 404–415.

LÉON (Jacqueline), 2008, Aux sources de la « Corpus Linguistics » : Firth et la London School. *In Langages*, n° 171, p. 12-33.

LERAT (Pierre), 1995, *Les langues spécialisées*, Paris : Presses Universitaires de France

LERAT (Pierre), 2002, Vocabulaire juridique et schémas d'arguments juridiques. In *Meta : journal des traducteurs / Meta: Translators' Journal*, vol. 47, n° 2, pp. 155-162.

LERAT (Pierre), 2006, Terme et micro-contexte. Les prédications spécialisées. In Daniel Blampain, Philippe Thoiron et Marc Van Campenhoudt (éds.), *Mots, termes et contextes*, Paris : AUF, pp. 89–98.

LERAT (Pierre), 2007, La nominalisation en *–tion* dans un texte techno-administratif. In *Actes de la conférence Terminologie & Ontologie : Théories et Applications Toth 2007*. Annecy, pp.79-92, consulté le 15 septembre 2015 à l'adresse suivante : https://hal.archives-ouvertes.fr/hal 00202639,

LONGRÉE (Dominique) et MELLET (Sylvie), 2013, Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours. In *Langages*, n° 189, pp. 65-79.

LYONS (J.), 1970 : *Linguistique générale. Introduction à la linguistique théorique*, Paris, Larousse (Collection *Langue et langage*).

LYONS (John), 1978, *Eléments de sémantique,* traduit par Jacques Durand, Paris, Larousse (Collection *Langue et langage*).

MANIEZ (François), 2003, Un modèle d'extraction des collocations en langue de spécialité. In *ASp*, n° 35-36, pp. 35-48.

MANIEZ (François), 2002, The use of electronic corpora and lexical frequency data in solving translation problems. In Bengt Altenberg et Sylviane Granger (éds.), *Lexis in Contrast, Corpus-based approaches.* Amsterdam: John Benjamins, pp. 291-306.

MANIEZ (François), 2008, Using the Web and corpora as language resources for the translation of complex noun phrases in medical research articles. In *Panacea*, N° 26, pp. 162–167.

MARTIN, Willy (2007), The Lexicon is a (kind of) Frame. In Leonel Ruiz Miyares (éds.), *Actas X Simposio Internacional Comunicación Social.* Santiago de Cuba, pp. 410-418.

MARTIN, Willy (2008), A unified approach to semantic frames and collocational patterns. In Sylviane Granger et Fanny Meunier (éds.), *Phraseology: an interdisciplinary perspective.* Amsterdam : John Benjamins Publishing Company, pp. 51-65.

MATHIEU-COLAS (Michel) & LE PESANT (M. Denis), 1998 : *Introduction aux classes d'objets*, In*: Langages*, 32e année, n°131, 1998. pp. 6-33.

McENERY (Tony) et WILSON (Andrew), 1996, *Corpus Linguistics*, Edinburgh: Edinburgh University Press.

MEJRI (Salah), 2011, Phraséologie et traduction des textes spécialisés. In Mogorron Huerta P. et Gonzalez C. (éd.), *Estudios y análisis de fraseología contrastiva: lexicografía, traducción y análisis de corpus*, Université d'Alicante, p. 125-137.

MEL'ČUK (Igor), 1988, Paraphrase et lexique dans la théorie linguistique Sens-Texte. In G.G. Bès et C. Fuchs, *Lexique*, n° 6, *Lexique et paraphrase*, Lille : Presses Universitaires de Lille, pp. 13-54.

MEL'ČUK (Igor), 1997, *Vers une linguistique sens-texte*, Leçon inaugurale faite le vendredi 10 janvier 1997 à la Chaire Internationale du collège de France, Paris, Collège de France

MEL'ČUK (Igor), 2003, Collocations dans le dictionnaire. In Thomas Szende (dir.) *Les écarts culturels dans les dictionnaires bilingues*, Paris : Honoré Champion Editeur, pp. 19-64

MEL'ČUK (Igor), 2004, Actants in semantics and syntax I: actants in semantics. In *Linguistics* 42-1. Walter de Gruyter, pp. 1-66.

MEL'ČUK (Igor), 2010, *La phraséologie en langue, en dictionnaire et en TALN*, Université de Montréal, texte consulté le 18.09.2012 à l'adresse suivante : http://www.iro.umontreal.ca/~felipe/TALN2010/Xml/Papers/all/taln2010_submission_207.pdf

MEL'CUK (Igor), 2013, Tout ce que nous voulions savoir sur les phrasèmes, mais … In *Cahiers de lexicologie : Revue internationale de lexicologie et de lexicographie*. Paris,

Classiques Garnier, téléchargé le 18/09/2012, à l'adresse suivante : http://olst.ling.umontreal.ca/pdf/MelcukPhrasemes2011.pdf, pp. 129-149

MEL'CUK, I. *et coll.* (1984, 1988, 1992, 1999) : *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques I–IV,* Montréal : Les Presses de l'Université de Montréal.

MEL'ČUK (Igor) et al. (1995), *Introduction à la Lexicologie explicative et combinatoire*, Louvain-La-Neuve : Editions Duculot

MEL'CUK (Igor), POLGUERE (A), 2006, Dérivation sémantiques et collocations dans le DiCo/LAF. In *Langue française*, vol. 150, nr 2, *Collocations, corpus, dictionnaires*, pp. 66-83.

MEL'CUK (Igor), POLGUERE (A), 2007, *Lexique actif du français*, Bruxelles : Éditions de Boeck.

MEL'ČUK (Igor) et POLGUÈRE (Alain), 2008, Prédicats et quasi-prédicats sémantiques dans une perspective lexicographique. In *Lidil* [En ligne] 37, mis en ligne le 1er septembre 2009, consulté le 13 novembre 2015 à l'adresse suivante : http://lidil.revues.org/2691

MEYER (Ingrid), BOWKER (Lynne), ECK (Koren), 1992, COGNITERM: An Experiment in Building a Terminological Knowledge Base. In *EURALEX' 92 Proceeding,* pp. 159- 172

MILIĆEVIĆ (Jasmina), 2003, *Modélisation sémantique, lexicale et syntaxique de la paraphrase langagière.* Thèse de doctorat, Département de linguistique et de traduction, Université de Montréal.

MILIĆEVIĆ (Jasmina), 2007, *La paraphrase. Modélisation de la paraphrase langagière*, Peter Lang, Bern.

MINSKY (Marvin), 1975, A Framework for Representing Knowledge. In P. Winston (éds.), *The Psychology of Computer Vision*, New York, McGraw Hill, pp. 211-277.

MORTCHEV-BOUVERET (Myriam), 2006, Fonctions lexicales pour le typage de relations syntagmatiques et paradigmatiques. Une approche lexicographique du terme. In *Terminology* n° 12, vol. 2, pp. 235–259.

MORTCHEV-BOUVERET (Myriam), 2007, Modélisation des relations lexico-sémantiques dans un dictionnaire spécialisé. In L'Homme, Marie-Claude et Vandaele Sylvie, *Lexicographie et terminologie : compatibilité des modèles et des méthodes*, Ottawa : Presses de l'Université d'Ottawa, pp. 293-320.

MURPHY (M. Lynne), 2003, *Semantic Relations and the Lexicon: Antonymy, Synonymy and Other Paradigms*, Cambridge, Cambridge University Press.

OTMAN (Gabriel), 1996, *Les représentations sémantiques en terminologie*, Paris, Masson, 216 p.

OTMAN (Gabriel), 1997, Les bases de connaissances terminologiques : les banques de terminologie de seconde génération. In *Meta : journal des traducteurs / Meta: Translators'* Journal, vol. 42, n° 2, pp. 244-256.

PAVEL (Silvia) et NOLET (Diane), 2001, *Précis de terminologie*, Hull, Bureau de la traduction, consulté le 12.09.2015 à l'adresse suivante : http://www.bt-tb.tpsgc-pwgsc.gc.ca/index.php?lang=francais&cont=692

PEARSON (Jennifer), 1998, *Terms in Context*. Amsterdam: John Benjamins Publishing Company,

PETIT (Michel), 2010, Le discours spécialisé et le spécialisé du discours: repères pour l'analyse du discours en anglais de spécialité. In *E-rea*, 8/1, pp. 1-17.

PIMENTEL (Janine), 2011, Description de verbes juridiques au moyen de la sémantique des cadres, In *Actes de la conférence Terminologie & Ontologie : Théories et Applications Toth 2011*. Annecy, pp. 145-166.

PIMENTEL (Janine), 2012a, *Criteria for the Validation of Specialized Verb Equivalents: Application in Bilingual Terminography.* Thèse de doctorat, Université de Montréal.

PIMENTEL (Janine), 2012b, *JuriDiCO: Users Guide*. Université de Montréal, Observatoire de linguistique Sens-Texte (OLST), consulté le 03.05.2013

POLGUÈRE (Alain), 2003, Collocations et fonctions lexicales : pour un modèle d'apprentissage. In F. Grossmann et A. Tutin (dir.) : *Les Collocations. Analyse et traitement*, coll. « Travaux et Recherches en Linguistique Appliquée », Amsterdam : De Werelt, pp. 117-133.

POLGUÈRE (Alain), 2003b, Étiquetage sémantique des lexies dans la base de données DiCo. In *TAL*. Volume 44 – n° 2, pp. 39-68.

POLGUÈRE (Alain), 2007, Lexical Function Standardness. In L. Wanner (éds.), *Selected Topics in Meaning Text Theory, in honour of Igor Mel'čuk*, John Benjamins, pp. 43-92.

POLGUERE (Alain), 2008 (2e édition) : *Lexicologie et sémantique lexicale*, Montréal, Les presses de l'Université de Montréal (Paramètres).

POLGUÈRE (Alain), 2011, Classification sémantique des lexies fondée sur le paraphrasage. In Seong Heon Lee (dir.), *Cahiers de lexicologie 98*, « Du Lexique aux dictionnaires en passant par la grammaire. Hommages à Chai-song Hong », Paris, Classique Garnier, pp. 197-211.

POLGUÈRE (Alain), 2012, Propriétés sémantiques et combinatoires des quasi-prédicats sémantiques. In *Scolia*, Université des sciences humaines, Strasbourg, mis en ligne le 4 juin 2012, consulté le 13 novembre 2015 à l'adresse : https://hal.archives-ouvertes.fr/hal-00703878, pp. 131-152.

POLGUÈRE (Alain) et SIKORA (Dorota), 2013, Modèle lexicographique de croissance du vocabulaire fondé sur un processus aléatoire, mais systématique. In C. Garcia-Debanc, C. Masseron et C. Ronveaux, (éds.), *Enseigner le lexique*, Namur, Presses Universitaires de Namur. pp. 35–63., <hal-00875192>

POTTIER (BERNARD), 1992 : *Sémantique générale*, Paris, Presses Universitaire de France.

RASTIER (FRANÇOIS), 1987, *Sémantique interprétative*, Paris, P.U.F., 277 p.

RASTIER (François), 1995, Le terme : entre ontologie et linguistique. In *La banque des mots* Numéro spécial 7, *Terminologie et Intelligence Artificielle*, pp. 35-65.

RASTIER (François), 2005, Enjeux épistémologiques de la linguistique de corpus. In G. Williams (dir.): *La linguistique de corpus*. Rennes : Presses Universitaires de Rennes, pp. 31-45.

RASTIER (François), 2005b, Discours et texte. In *Dossier « Textes et discours », Approfondissement théoriques, Revue Texto*, consulté en ligne le 27.08.2015 à l'adresse suivante : < http://www.revue-texto.net/Reperes/Themes/Rastier_Discours.html>

REY (Alain) & CHANTEREAU (Sophie), 1993, *Dictionnaire des expressions et locutions*. Paris: Dictionnaires LE ROBERT.

ROCHE (Christophe), 2005, Terminologie et ontologie. In *Langages*, n° 157, pp. 48-62.

ROCHE (Christophe), 2007, Le terme et le concept : fondements d'une ontoterminologie. In *Actes de la première conférence Terminologie & Ontologie : Théories et Applications Toth 2007*. Annecy, pp. 1-22.

SAGER (Juan C), 1990, *Practical Course in Terminology Processing*, Amsterdam: John Benjamins Publishing Company.

SAGER (Juan C.) 2000. Pour une approche fonctionnelle de la terminologie. In H. Béjoint et P. Thoiron (dir.), *Le sens en terminologie*, Coll. « Travaux du C.R.T.T », Lyon, Presses universitaires de Lyon, pp. 40-60.

SAUSSURE (Ferdinand de), 1972, *Cours de linguistique générale*, Éditions Payot&Rivages, Paris, 520 p.

SAUSSURE (Ferdinand. De), (1967) : *Cours de linguistique générale*, fascicule 2, Wiesbaden, Otto Harrassowitz.

SCHMIDT (Thomas), 2009, The Kicktionary – a multilingual lexical resource of football language. In Hans C. Boas, (éd.). *Multilingual FrameNets in Computational Lexicography*, Berlin/New York: De Gruyter Mouton, pp. 101-132.

SÉGUÉLA (Patrick), 2001, *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. Mémoire de thèse en Informatique, Université Toulouse.

SILBERZTEIN (Max), 2003, *NooJ Manual*, téléchargeable à l'adresse suivante : www.nooj4nlp.net

SILBERZTEIN (Max), KOEVA (Svelta), MAUREL (Denis), 2005, *Formaliser les langues avec l'ordinateur. De INTEX à Nooj*, Collection « Les Cahiers de la MSH Ledoux », Presses Universitaires de Franche-Comté, Besançon.

SILBERZTEIN (Max), TUTIN (Agnès), 2005, NooJ, un outil TAL pour l'enseignement des langues. Application pour l'étude de la morphologie lexicale en FLE. In *Alsic*, Vol. 8, n° 2, pp. 123-134, consulté le 18.09.2014.

SINCLAIR (John), 1991, *Corpus, Concordance, Collocation*, Oxford University Press, Oxford.

SINCLAIR (John), 2004, *Trust the Text. Language, corpus and discourse.* London / New York: Routledge,

SINCLAIR (JOHN), 2005, Corpus and Text - Basic Principles. In M. Wynne (éds.), *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford, Oxbow Books, pp. 1-16, accessible à l'adresse suivante: http://ahds.ac.uk/linguistic-corpora/ [consulté le 14.04.2015]

SLODZIAN (Monique), 2000, *L'émergence d'une terminologie textuelle et le retour du sens*. In H. Bejoint et P. Thoiron (dir.), *Le Sens en terminologie*, Lyon : Presses universitaires de Lyon, pp. 61-85.

SOURIOUX (Jean-Louis) et LERAT (Pierre), 1975, *Le langage du droit*, PUF. Paris.

TESNIERE (Lucien), 1988, *Éléments de syntaxe structurale*, (deuxième édition revue et corrigée), Paris : Éditions Klincksieck, 674 p.

TOGNINI-BONELLI (Elena), 2001, *Corpus Linguistics at Work*, Amsterdam / Philadelphia: John Benjamins Publishing Company.

TUTIN (Agnès) et GROSSMANN (Francis), 2002, Collocations régulières et irrégulières : esquisse de typologie du phénomène collocatif. In *Revue française de linguistique appliquée* 1/2002 (Vol. VII), p. 7-25, consulté le 10.09.2012 à l'adresse : www.cairn.info/revue-francaise-de-linguistique-appliquee-2002-1-page-7.htm.

TUTIN (Agnès), 2005, Le dictionnaire de collocations est-il indispensable ? In Th. Fontenelle (éds.). *Revue Française de Linguistique Appliquée*. *Dictionnaires : nouvelles approches, nouveaux modèles*. 2005/2, Vol. X, pp. 31-48.

VAN CAMPENHOUDT (Marc), 2001, Pour une approche sémantique du terme et de ses équivalents. In *International Journal of Lexicography*, n° 14 (3) pp. 181-209.

VAN CAMPENHOUDT (Marc), 2010, Le terme : Condensation syntaxique et condensation des connaissances en langue spécialisée. In *Romanica Wratislaviensia* 57, pp. 29–46, consulté le 15 septembre à l'adresse suivante : http://www.termisti.org/romanica_w.pdf.

VANDENDORPE (Christian), 1990, Paradigme et syntagme De quelques idées vertes qui ont dormi furieusement, in *Revue québécoise de linguistique théorique et appliquée* 9, 3, pp. 169-193

WIERZBICKA (Anna), 1996, *Semantics: Primes and Universals*. Oxford, Oxford University Press.

WILLIAMS (Geoffrey), 2001, *Sur les caractéristiques de la collocation,* In *Actes de TALN*, Tours 2-5 juillet 2001, Université de Tours, pp. 9 – 16.

WILLIAMS (Geoffrey), 2005, Linguistique et corpus. Introduction. In G. Williams (dir.): *La linguistique de corpus.* Presses Universitaires de Rennes, pp. 13-18.

WILLIAMS (Geoffrey), 2006, La linguistique et le corpus: Une affaire prépositionnelle. In *Texto, revue de linguistique en ligne*, consulté en avril 2014 à l'adresse suivante : http://www.revuetexto.net/Parutions/Livres-E/Albi-2006/Williams.pdf

Ressources lexicographiques et terminologiques disponibles en ligne

Base JuriDico, projet développé par Janine Pimenel à l'OLST de l'Université de Montréal, accessible en ligne à l'adresse suivante : http://olst.ling.umontreal.ca/cgi-bin/juridico/search.cgi

DAD, Dictionnaire analytique de la distribution / Analytical Dictionary of Retailing, Montréal, Dancette (Jeanne), 2006, accessible en ligne à l'adresse suivante : < http://falbala.ling.umontreal.ca/dad/ consulté le 04.06.2014.

DAFA, Dictionnaire d'apprentissage du français des affaires, version électronique, Binon, (Jean), Verlinde (Serge),Van Dyck (Jan), Bertels, (Anne) 2000, accessible en ligne à l'adresse suivante : http://www.projetdafa.net/ consulté le 10.05.2014.

DAFLES, Dictionnaire d'Apprentissage du Français Langue Étrangère ou Seconde, Selva (Thierry), Verlinde (Serge), version en construction accessible en ligne à l'adresse suivante : http://www.kuleuven.be/dafles/miroir/acces.php?id=

DAMT, Dictionnaire analytique de la mondialisation du travail / Analytical Dictionary of Globalization and Labour /Diccionario analítico de la globalización del trabajo, Dancette (Jeanne), 2010, accessible en ligne à l'adresse suivante : http://www.crimt.org/damt.htm , consulté le 04.06.2014.

DiCoEnviro, Dictionnaire fondamental de l'environnement, projet développé à l'OLST de l'Université de Montréal, sous la direction de Marie-Claude L'Homme, accessible en ligne à l'adresse suivante : http://olst.ling.umontreal.ca/cgi-bin/dicoenviro/search_enviro.cgi consulté le 04.06.2014.

DiCoInfo, Dictionnaire fondamental de l'informatique et de l'Internet, projet développé à l'OLST de l'Université de Montréal, sous la direction de Marie-Claude L'Homme, accessible en ligne à l'adresse suivante : http://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi consulté le 04.06.2014.

DiCouèbe, Dictionnaire en ligne de combinatoire du français, le *DiCo* en ligne, projet développé à l'OLST de l'Université de Montréal, par Igor Mel'čuk et Alain Polguère , accessible en ligne à l'adresse suivante : http://olst.ling.umontreal.ca/dicouebe/

EcoLexicon, projet développé à l'Université de Grenade par le groupe Lexicon sous la direction de Pamela Faber Benítez, accessible en ligne à l'adresse suivante http://ecolexicon.ugr.es/en/index.htm

FrameNet, projet développé à l'Institut International d'Informatique de Berkeley sous la direction de Collin Baker, accessible en ligne à l'adresse suivante : https://FrameNet.icsi.berkeley.edu/fndrupal/

Kicktionary, projet développé par Thomas Schmidt dans le cadre d'une bourse postdoctorale à l'Institut International d'Informatique de Berkeley, accessible en ligne à l'adresse suivante : http://www.kicktionary.de/index.html

WordNet, projet développé au sein du Département des Sciences Cognitives de l'Université de Princeton, accessible en ligne à l'adresse suivante : https://wordnet.princeton.edu/

Liste des figures et des tableaux

Table des matières

Annexe I : Liste des documents constituant le corpus *DITerm*

SOUS-CORPUS		DOCUMENT	NOMBRE DE DOCUMENTS	NOMBRE DE MOTS
CORPUS_UE_ET_JURISPRUDENCE_UE				
CORPUS_UE				
	CORPUS_ECOMMERCE		**49**	**477 377**
		COM 2004.28		13 901
		COM 2006.120		4 365
		COM 2006.688		4 392
		COM 2007.696		5 141
		COM 2008.798		4 695
		COM 2008.572		4 334
		COM 2009.278		4 079
		COM 2010.253		4 989
		COM 2011.941		12 379
		DIRECTIVE 1999.93		5 592
		DIRECTIVE 2000.31		11 439
		DIRECTIVE 2002.19		8 945
		DIRECTIVE 2002.20		8 410
		DIRECTIVE 2002.21		12 999
		DIRECTIVE 2002.65		6 837
		DIRECTIVE 2007.64		29 553
		DIRECTIVE 2009.140		24 578
		DIRECTIVE 2010.13		17 148
		DIRECTIVE 2011.83		17 049
		DIRECTIVE 2002.22		16 740
		COM 2003.702		13 787
		RECOM 1997.489		3 263
		COM 2011.794		13 453
		COM 2011.942		9 081
		COM 2012.698		12 543
		COM 2009.557		5 349
		CNUDCI 2001		31 581
		CNUDCI 1996		29 992

		CNUDCI 2009		55 708
		AVIS 1999.C.93.06		2 376
		AVIS 2004.C.318.08		1 652
		AVIS 2012.C.229.01		3 949
		AVIS 2012.C.351.11		4 315
		AVIS 2012.C.351.16		3 013
		COM 1998.586		6 686
		COM 2004.479		3 301
		COM 2004.841		3 822
		COM 2006.739		1 832
		COM 2008.724		18 761
		COM 2009.626		3 342
		COM 2010.571		10 893
		COM 2011.793		8 279
		COM 2012.238		18 919
		COM 2012.596		6 950
		DECISION 2009.767		931
		DECISION 2010.425		2 750
		DIRECTIVE 2002.77		3 422
		REGLEMENT 2009.976		3 513
		RESOLUTION 2008.2204.INI		4 653
	CORPUS_DONNEES		**59**	**577 936**
		COM 2003.265		10 870
		COM 2007.698		16 549
		COM 2007.087		4 728
		COM 2007.228		4 775
		COM 2010.609		8 320
		COM 2012.009		6 444
		COM 2012.010		25 194
		COM 2012.011		48 962
		DECISION 2000.520		29 976
		DECISION 2001.497		3 853
		DECISION 2004.915		1 594
		DECISION 2008.597		2 904
		DECISION 2008.977		8 951
		DIRECTIVE 95.046		13 085
		DIRECTIVE 97.066		5 267
		DIRECTIVE 2002.058		9 673
		DIRECTIVE 2006.024		6 325
		DIRECTIVE 2009.136		20 431
		REGLEMENT 2001.045		11 747
		AVIS 01_2012		18 313
		AVIS 02_2012		4 998
		AVIS 03_2012		18 488

		AVIS 04_2012		5 815
		AVIS 05_2012		15 150
		AVIS 13_2011		9 756
		AVIS 15_2011		21 202
		AVIS 16_2011		5 600
		AVIS 04_2010		2 589
		AVIS 03_2009		3 014
		AVIS 6_2007		16 836
		AVIS 7_2007		10 567
		AVIS 02_2006		5 425
		AVIS 03_2006		1 144
		AVIS 8_2006		3 053
		AVIS 10_2006		13 725
		AVIS 04_2005		4 242
		AVIS 5_2005		5 128
		AVIS TRAVAIL_2005		8 717
		AVIS 2007.04		14 958
		AVIS 2008.01		14 739
		AVIS 2008.02		2 319
		AVIS 2009.01		4 552
		AVIS 2009.05		5 768
		AVIS 2010.02		14 943
		AVIS 2010.03		9 642
		AVIS 2010.05		5 225
		AVIS 2010.08		17 567
		COM 2005.438		7 053
		COM 2005.493		4 126
		AVIS 2005.C.298.01		8 557
		DISPOSITION 2005.C.308.01		4 298
		AVIS 2009.C.128.04		12 228
		COM 2009.387		3 481
		COM 2010.170		4 494
		REGLEMENT D'EXECUTION 2011.1179		3 540
		AVIS 2012.C.34.01		12 165
		DOCUMENT DE TRAVAIL SEC 2012.0073		4 571
		AVIS 2009.C.120.08		4 335
		POSITION COMMUNE 2009.016		14 310
	CORPUS_PROPRIETE INTELLECTUELLE		**37**	**196 368**
		COM 2000.199		8 550
		COM 2007.385		4 842
		COM. 2007.836		4 269
		COM 2008.466		7 852

		COM 2009.532		4 401
		COM 2011.380 ACTA		10 776
		COM 2012.372		21 273
		COCLUSION 2008.C.319.06		1 768
		DECISION CE 2002.3639		14 763
		DIRECTIVE 2004.48		6 294
		DIRECTIVE 91.250		2 930
		DIRECTIVE 96.09		6 162
		DIRECTIVE 2001.29		9 223
		DIRECTIVE 2009.24		3 014
		RAPPORT 2008.2121 M.ORTEGO		6 692
		RAPPORT 2008.2160 LAMBRINIDIS		7 084
		RECOMMENDATION 2005.737		2 275
		RECOMMENDATION 2006.585		1 869
		REGLEMENT 2002.733		3 067
		REGLEMENT 2004.874		7 192
		RESOLUTION 2007.2153		4 355
		COM 2009.303		4 148
		COM 2008.513		4 822
		AVIS 2010.C 147.01		10 918
		COM 2009.467		4 378
		AVIS 2011.C.18.19		3 137
		COM 2012.789		2 166
		DOCUMENT DE TRAVAIL SWD 2012.205		3 211
		RESOLUTION 2010.C.8.E.19		1 236
		DECISION 2007.595		1 726
		RECOMMENDATION 2011.711		3 640
		RECOMMENDATION 2009.625		2 610
		COM 2011.616		4 317
		RESOLUTION 2006.2008.INI		4 059
		COMP C.2.38.126		1 040
		AVIS 2009.C.77.16		5 013
		RESOLUTION 2009.2178		4 240
	CORPUS_SECURITE INTERNET		**32**	**190 582**
		COM 2000.890		19 446
		COM 2001.298		12 870
		COM 2006.661		2 524
		COM 2006.663		3 328
		COM 2007.267		4 658
		COM 2008.207		4 658
		COM 2008.448		4 419
		COM 2009.064		2 451

		COM 2009.149		4 959
		COM 2011.128		16 732
		COM 2011.163		6 354
		COM 2011.556		3 913
		CONVENTION SUR LA CYBERCRIMINALITE		10 535
		DECISION 1999.276		6 464
		DECISION 2002.173		10 698
		DECISION 2005.854		6 750
		DECISION 2008.1351		5 773
		DECISION CADRE 2005.222		2 877
		REGLEMENT 2004.460		8 330
		CONCLUSION 2009.C.62.05		1 776
		COM 2004.91		3 646
		COM 2005.347		5 319
		COM 2004.341		3 610
		COM 2008.106		4 120
		AVIS 2008.C.325.14		2 920
		COM 2010.517		6 762
		COM 2009.2041		9 155
		RESOLUTION 2004.0091		6 053
		RECOMMENDATION 2006.952		3 691
		AVIS 2012.C.351.15		3 864
		COM 2012.140		3 540
		AVIS 2009.C.2.02		3 590
JURISPRUDENCE UE_72 ARRETS				
		ARRETS	**72**	**433 628**
CORPUS FR			**532**	**3 088 216**
	REVUE LAMY – RDLI		**372**	**2 161 918**
		RDLI_2012_ 11 N° 78-88 / articles	68	342 862
		RDLI_2011_11 N° 67-77 / articles	78	341 986
		RDLI_2010_11 N° 56-66 / articles	70	302 735
		RDLI_2009_11 N° 45-55 / articles	86	433 983
		RDLI_2008_11 N° 34-44 / articles	79	362 085
		RDLI_2007_11 N° 23-33 / articles	77	378 267
	CORPUS_LEGALIS.COM			
		ARRETS	**126**	**457 447**
	CORPUS_LE FORUM DES DROITS SUR INTERNET			
		Recommendations	**16**	**217 599**

	CORPUS_LOIS FR		**19**	**251 252**
		Code des postes et des communications électroniques		
		Loi n°2004-575 du 21 juin 2004 pour la confiance dans l'économie numérique		
		Rapport d'information 627 sur l'application LCEN		
		Loi Informatique et libertés n°78- 17 du 6 janvier 1978,		
		LOI 2006.961 DADVSI		
		PROJET LOI HADOPI		
		LOI 2009.669 HADOPI		
		LOI 2009.1311 HADOPI 2		
		ACCORD ELISEES		
		RAPPORT HADOPI THIOLLIERE 2008		
		DECISION 2006_540		
		DECISION 2009 590		
		DECISION 2009 580		
		AVIS HADOPI GERARD 2009		
		RAPPORT OLIVIENNES		
		RAPPORT HADOPI RIESTER 2009		
		AVIS HADOPI MARLAND MILITELLO		
		AVIS HADOPI RETAILLEAU		
		RAPPORT DADVSI 2005.02		
				5 111 267

Annexe II : Syntaxe des expressions regulières *NooJ*

Symbole	Explication	Exemple
|	opérateur de disjonction, indique un choix entre deux mot ou plus	fournisseur|prestataire
espace	opérateur de concaténation, précise que les chaînes recherchées doivent être côte à côte dans les contextes (l'opérateur de concaténation a priorité sur l'opérateur de disjonction)	fournisseur de
()	les parenthèses permettent de modifier l'ordre de priorité par défaut	(fournisseur|prestataire) de
< >	les chevrons sont utilisés pour introduire les caractères spéciaux *NooJ*	<WF>, <N>, <A>
<WF>	chaîne de caractère	<WF> en ligne
^	Le symbole correspond à la position au début de la chaine de caractères recherchée (métacaractère Perl)	<WF+MP="^distribu">
$	Le symbole correspond à la fin de la chaîne de caractères recherchée (métacaractère Perl)	<WF+MP="able$">
*	L'astérisque remplace une chaîne indéfinie (de 0 à n) de caractères	<N+MP="ion$"><N><A>*(de)* <N>*<A>*
<A>, <N>, <V>,	Les symboles entre chevrons permettent de retrouver des mots appartenant à une catégorie grammaticale (adverbe, nom,	consentement <A>

<DET>, <PREP>	verbe) ou une catégorie sémantique préalablement identifiées dans un dictionnaire *NooJ*	
<UNK>	permet de retrouver des mots non reconnues (non annotées) par *NooJ*	<UNK+MP="ing$">
[a-z]	correspond à tout caractère alphabétique en minuscules dans la plage s'étendant de "a" à "z".	<UNK+MP="[a-z]">
[A-Z]	correspond à tout caractère alphabétique en majuscules dans la plage s'étendant de "a" à "z".	<UNK+MP="[A-Z]">
-	Le moins est un opérateur qui permet d'exclure une chaîne de caractère ou une catégorie précises	<N+MP="age$"-MP="^page$"-MP="^dommage$"-MP="^message$"-MP="^davantage$">
+	Le plus est un opérateur qui permet de restreindre la recherche à une chaîne de caractère ou une catégorie précises.	
MP= "…"	Permettent d'introduire une expression régulière de type Perl	voir plus haut

Annexe III : Nomenclature du droit de l'Internet dans le *DITerm*

N°	TERME (ET SES VARIANTES)	FRÉQUENCE	CADRE SÉMANTICO-CONCEPTUEL
1.	accès (accéder)*[106]	1405	{ACTIVITÉ ou ACTION ILLICITES}
2.	accès (n. m.) Internet / accès (n. m.) à l'Internet	223 / 228	{SERVICE INTERNET}
3.	anonymisation (n. f.)	88	{ACTIVITÉ ou ACTION}
4.	autorisation (n. f.) préalable	46	{AUTORISATION} {RÈGLE}
5.	ayant (n. m.) droit	793	{AYANT DROIT}
6.	base (n. f.) de données	1940	{OBJET EN LIGNE}
7.	Blocage / mesure de blocage	401 197	{DISPOSITIF TECHNICO-JURIDIQUE}
8.	blog (n.m)	590	{SERVICE INTERNET}
9.	certificat (n.m) qualifié	74	{DISPOSITIF TECHNICO-JURIDIQUE}
10.	ciblage (n.m)	112	{ACTIVITÉ ou ACTION}
11.	cloud computing / informatique en nuage	113 /249	{SERVICE INTERNET}
12.	commerce électronique / e-commerce	1261 / 123	{ACTIVITÉ ou ACTION} {SERVICE INTERNET}
13.	communication au public en ligne	672	{SERVICE INTERNET} {ACTIVITÉ ou ACTION}
14.	communication au public par	267	{SERVICE INTERNET}

[106] Les unités accompagnées d'un astérisque appartiennent au vocabulaire de soutien ayant un fort potentiel terminologique.

	voie électronique		{ACTIVITÉ ou ACTION}
15.	communication audiovisuelle	190	{SERVICE INTERNET}
16.	communication commerciale (eu)	233	{ACTIVITÉ ou ACTION}
17.	communication électronique	3729	{SERVICE INTERNET} {ACTIVITÉ ou ACTION}
18.	connexion (connecter)	946	{ACTIVITÉ ou ACTION}
19.	consentement préalable	114	{AUTORISATION} {RÈGLE}
20.	contenu [mis en ligne]	8835	{OBJET EN LIGNE}
21.	contenu illicite	638	{OBJET ILLICITE EN LIGNE}
22.	contrefaçon [numérique]	2997	{ACTIVITÉ ou ACTION ILLICITES}
23.	contrôle parental	89	{DISPOSITIF TECHNICO-JURIDIQUE}
24.	Cookie /cookie traceur /témoin de connexion	695 / 18 /50	{TECHNOLOGIE ou MACHINE}
25.	copiage* (copier)*	178	{ACTIVITÉ ou ACTION}
26.	correspondance privée [par voie électronique]	168	{ACTIVITÉ ou ACTION} {SERVICE INTERNET}
27.	courrier électronique	598	{SERVICE INTERNET}
28.	cybersquatting / cybersquattage	35 / 4	{ACTIVITÉ ou ACTION ILLICITES}
29.	déclaration normale	15	{PROCÉDURE}
30.	déclaration préalable	62	{PROCÉDURE}
31.	déclaration simplifiée	24	{PROCÉDURE}
32.	diffusion* (diffuser*)	2064	{ACTIVITÉ ou ACTION}
33.	distribution* (distribuer*)	1564	{ACTIVITÉ ou ACTION}
34.	DNS (Domain Name Système) /système des noms de domaine	70 /20	{TECHNOLOGIE ou MACHINE}
35.	donnée biométrique	200	{OBJET EN LIGNE}
36.	donnée de connexion	115	{OBJET EN LIGNE}

37.	donnée de localisation	171	{OBJET EN LIGNE}
38.	donnée de navigation, /donnée relative au comportement de navigation	69 / 15	{OBJET EN LIGNE}
39.	donnée personnelle / donnée à caractère personnel	1050 3913	{OBJET EN LIGNE}
40.	donnée relative au trafic (des communications électroniques)	144	{OBJET EN LIGNE}
41.	donnée sensible	145	{OBJET EN LIGNE}
42.	droit d'accès	272	{DROIT} {RÈGLE}
43.	droit d'effacement	47	{DROIT} {RÈGLE}
44.	droit d'interrogation	12	{DROIT} {RÈGLE}
45.	droit de rectification	37	{DROIT} {RÈGLE}
46.	droit d'opposition	133	{DROIT} {RÈGLE}
47.	durée de la conservation des données à caractère personnel	29	{PROCÉDURE} {RÈGLE}
48.	échange* (échanger*) Partage* (partager*)	/1073 /1055	{ACTIVITÉ ou ACTION}
49.	éditeur de contenus /éditeur de contenu	147	{ACTEUR DE L'INTERNET : UTILISATEUR} {ACTEUR DE L'INTERNET : PRESTATAIRE}
50.	éditeur de service de communication au public en ligne	292	{ACTEUR DE L'INTERNET : PRESTATAIRE}
51.	formalités préalables	158	{SERVICE INTERNET}
52.	forum de discussion	349	{SERVICE INTERNET}
53.	fournisseur d'informatique en nuage	32	{ACTEUR DE L'INTERNET : PRESTATAIRE}

54.	fournisseur d'accès à Internet /fournisseur d’accès à l’Internet /fournisseur d'accès /FAI /fournisseur d’accès à un réseau de communication au public en ligne	246 /7 /1224 667 /5	{ACTEUR DE L’INTERNET : PRESTATAIRE}
55.	fournisseur de contenu / fournisseur de contenus	106	{ACTEUR DE L’INTERNET : PRESTATAIRE}
56.	fournisseur de liens commerciaux / fournisseur de liens sponsorisé / prestataire de liens commerciaux	54 / 21 / 27	{ACTEUR DE L’INTERNET : PRESTATAIRE}
57.	fournisseur de moteur de recherche	29	{ACTEUR DE L’INTERNET : PRESTATAIRE}
58.	fournisseur de réseau publicitaire en ligne	35	{ACTEUR DE L’INTERNET : PRESTATAIRE}
59.	fournisseur de services de communications électroniques	145	{ACTEUR DE L’INTERNET : PRESTATAIRE}
60.	fournisseur d'hébergement /hébergeur /prestataire d'hébergement /prestataire de stockage	479 /485 /152 / 25	{ACTEUR DE L’INTERNET : PRESTATAIRE}
61.	fournisseur du service de référencement, / prestataire d’un service de référencement sur Internet	10 /23	{ACTEUR DE L’INTERNET : PRESTATAIRE}
62.	Framing	20	{ACTIVITÉ ou ACTION}
63.	hameçonnage /phishing	36 / 53	{ACTIVITÉ ou ACTION ILLICITES}
64.	hébergement	1347	{ACTIVITÉ ou ACTION}
65.	hyperlien / lien hypertexte	129	{OBJET EN LIGNE}

		117	
66.	identité numérique	116	{OBJET EN LIGNE}
67.	information préalable (au traitement des données à caractère personnel)	17	{PROCÉDURE}
68.	interconnexion des données	12	{ACTIVITÉ ou ACTION}
69.	intermédiaire technique prestataire technique	241 / 140	{ACTEUR DE L'INTERNET : PRESTATAIRE}
70.	Internaute	3737	{ACTEUR DE L'INTERNET : UTILISATEUR}
71.	IP /adresse IP	1242 / 1040	{OBJET EN LIGNE}
72.	lien commercial / lien sponsorisé / lien promotionnel	708 / 143 / 119	{OBJET EN LIGNE}
73.	lien publicitaire	62	{OBJET EN LIGNE}
74.	marquage / watermarking / tatuage numérique	51 / 9 / 20	{DISPOSITIF TECHNICO-JURIDIQUE}
75.	mesure de filtrage / filtrage	450 /1112	{DISPOSITIF TECHNICO-JURIDIQUE}
76.	mesures techniques de protection / MTP	258 /150	{DISPOSITIF TECHNICO-JURIDIQUE}
77.	mise à disposition / upload / téléchargement ascendant	1050 / 70 / 16	{ACTIVITÉ ou ACTION}
78.	mise en ligne (mettre en ligne)	1075	{ACTIVITÉ ou ACTION}
79.	mot clé	1919	{OBJET EN LIGNE}
80.	moteur de recherche	739	{SERVICE INTERNET}
81.	navigation (naviguer)	314	{ACTIVITÉ ou ACTION}
82.	nom de domaine	1783	{OBJET EN LIGNE}
83.	Nommage	92	{ACTIVITÉ ou ACTION}
84.	notification (notifier) [de violation de données]	1189	{PROCÉDURE} {ACTIVITÉ DE CONTROLE

			ET DE PROTECTION}
85.	notification (notifier) [de contenus illicites]		{PROCÉDURE} {ACTIVITÉ DE CONTROLE ET DE PROTECTION}
86.	obligation de promptitude à retirer le contenu illicite /obligation de prompt retrait	12 / 7	{OBLIGATION}
87.	obligation d'information	21	{OBLIGATION}
88.	obligation de confidentialité	36	{OBLIGATION}
89.	obligation de sécurité	41	{OBLIGATION}
90.	obligation de surveillance de l'accès Internet	144	{OBLIGATION}
91.	obligation de surveillance générale	234	{OBLIGATION}
92.	obligation de surveillance particulière	89	{OBLIGATION}
93.	œuvre [en ligne]	5241	{OBJET EN LIGNE}
94.	offre légale (en ligne)	334	{SERVICE INTERNET}
95.	opérateur de communications électroniques / opérateur de réseaux ouverts au public et de fourniture au public de services de communications électroniques	/66 /14	{ACTEUR DE L'INTERNET : PRESTATAIRE}
96.	opt-in	80	{DISPOSITIF TECHNICO-JURIDIQUE} {RÈGLE}
97.	opt-in actif	10	{DISPOSITIF TECHNICO-JURIDIQUE} {RÈGLE}
98.	opt-in passif	5	{DISPOSITIF TECHNICO-JURIDIQUE} {RÈGLE}
99.	opt-out	91	{DISPOSITIF TECHNICO-JURIDIQUE} {RÈGLE}
100.	opt-out actif	6	{DISPOSITIF TECHNICO-JURIDIQUE}

			{RÈGLE}
101.	opt-out passif	2	{DISPOSITIF TECHNICO-JURIDIQUE} {RÈGLE}
102.	outil de signalement de contenus illicite	14	{DISPOSITIF TECHNICO-JURIDIQUE} {RÈGLE}
103.	outil de suggestion de mots clés /générateur de mots clés	36 /60	{SERVICE INTERNET}
104.	paiement en ligne	2398	{SERVICE INTERNET}
105.	pair à pair / peer to peer / P2P / peer-to-peer	105 /157 /309 /227	{TECHNOLOGIE ou MACHINE}
106.	parasitisme	223	{ACTIVITÉ ou ACTION ILLICITES}
107.	parking de noms de domaine / garage de noms de domaine	57 / 5	{ACTIVITÉ ou ACTION}
108.	pharming,	10	{ACTIVITÉ ou ACTION ILLICITES}
109.	piratage (pirater)	545	{ACTIVITÉ ou ACTION ILLICITES}
110.	Pirate	44	{PERSONNE QUI REALISE UNE ACTIVITÉ ou ACTION ILLICITES}
111.	Pistage	23	{ACTIVITÉ ou ACTION}
112.	plate-forme (plates-formes) / plateforme (plateformes) [web 2.0]	1570 60	{SERVICE INTERNET}
113.	plateforme de commerce électronique / plate-forme de commerce électronique / plateforme e-commerce	35 /117 /25	{SERVICE INTERNET}
114.	plate-forme de courtage en ligne site de courtage en ligne	20 39	{SERVICE INTERNET}
115.	plate-forme de partage de vidéos	95	{SERVICE INTERNET}
116.	prestataire de service de	69	{ACTEUR DE L'INTERNET :

	certification		PRESTATAIRE}
117.	prestataire de service de confiance qualifié	36	{ACTEUR DE L'INTERNET : PRESTATAIRE}
118.	prestataire de service de paiement en ligne	189	{ACTEUR DE L'INTERNET : PRESTATAIRE}
119.	prestataire de services de la société de l'information, /fournisseur de services de la société de l'information,	37 /15	{ACTEUR DE L'INTERNET : PRESTATAIRE}
120.	prestataire de services de musique en ligne	25	{ACTEUR DE L'INTERNET : PRESTATAIRE}
121.	prestataire du web 2.0	41	{ACTEUR DE L'INTERNET : PRESTATAIRE}
122.	principe de licéité	36	{RÈGLE}
123.	principe de loyauté	39	{RÈGLE}
124.	principe de proportionnalité	15	{RÈGLE}
125.	procédure alternative de règlement des conflits / UDRP (uniform domain name dispute resolution)	17 / 75	{PROCÉDURE}
126.	procédure de notification et de retrait	18	{PROCÉDURE}
127.	procédure en trois temps	17	{PROCÉDURE}
128.	Profilage	53	{ACTIVITÉ ou ACTION}
129.	prospection en ligne	282	{ACTIVITÉ ou ACTION}
130.	protection des données à caractère personnel/des données personnelles	568	{ACTIVITÉ DE CONTROLE ET DE PROTECTION}
131.	publicité ciblée, publicité comportementale	249 /234	{ACTIVITÉ ou ACTION}
132.	publicité en ligne /publicité sur Internet	62 / 120	{ACTIVITÉ ou ACTION}
133.	rapprochement des données	25	{ACTIVITÉ ou ACTION}
134.	référencement	662	{SERVICE INTERNET}
135.	référencement payant	92	{SERVICE INTERNET}
136.	registrar / bureau d'enregistrement	18	{ORGANISME DE CONTROLE}

		/113	
137.	registre, / organisme d'enregistrement / organisme d'attribution et de gestion des noms de domaines	100 /321 /72	{ORGANISME DE CONTROLE}
138.	règles d'entreprise contraignantes /règles d'entreprise contraignantes	68	{RÈGLE}
139.	réponse graduée /riposte graduée	/ 204 /102	{PROCÉDURE} {ACTIVITÉ DE CONTROLE ET DE PROTECTION}
140.	reproduction* (reproduire*)	1678	{ACTIVITÉ ou ACTION}
141.	réseau d'échange de pair à pair	57	{SERVICE INTERNET}
142.	réseau de communications électroniques	674	{SERVICE INTERNET}
143.	réseau ouvert au public / réseau public de communications / réseau de communications électroniques accessibles au public	112 /541 /250	{SERVICE INTERNET}
144.	réseau social en ligne / site de réseau social / plateforme de réseau social en ligne /site de réseautage social	/395 /24 /29 /8	{SERVICE INTERNET}
145.	responsable du traitement	1412	{ACTEUR DE L'INTERNET : PRESTATAIRE}
146.	service de communication au public en ligne	733	{SERVICE INTERNET}
147.	service de communications électroniques	906	{SERVICE INTERNET}
148.	service de la société de l'information	252	{SERVICE INTERNET}
149.	service web 2 0	47	{SERVICE INTERNET}
150.	signalement (signaler) [de contenus illicites]	187	{ACTIVITÉ DE CONTROLE ET DE PROTECTION}
151.	signature électronique	1363	{SERVICE INTERNET}
152.	site	9818	{SERVICE INTERNET}

	/site Internet /site web	/1759 /680	
153.	site de streaming	59	{SERVICE INTERNET}
154.	société de gestion collective	236	{ORGANISME DE CONTROLE}
155.	sous-traitant	37	{ACTEUR DE L'INTERNET : PRESTATAIRE}
156.	spam / communication commerciale non sollicitée	360 / 63	{OBJET ILLICITE EN LIGNE}
157.	streaming / en lecture seule / en flux continu	176 /10 /11	{TECHNOLOGIE ou MACHINE}
158.	système d'alerte de de signalement des contenus illicite	16	{DISPOSITIF TECHNICO-JURIDIQUE}
159.	système d'empreintes digitales ou numériques / fingerprinting	311 /18	{DISPOSITIF TECHNICO-JURIDIQUE}
160.	système de reconnaissance de contenus	24	{DISPOSITIF TECHNICO-JURIDIQUE}
161.	téléchargement (télécharger)	2127	{ACTIVITÉ ou ACTION}
162.	téléchargement descendant /download (downloading)	13 / 62	{ACTIVITÉ ou ACTION}
163.	téléchargement illégal	203	{ACTIVITÉ ou ACTION ILLICITES}
164.	titulaire de l'abonnement Internet /abonné à Internet /titulaire de la connexion Internet titulaire d'accès Internet	45 /151 /16 /89	{ACTEUR DE L'INTERNET : UTILISATEUR}
165.	traçage (tracer)	67	{ACTIVITÉ ou ACTION}
166.	traitement de données à caractère personnel	1059	{ACTIVITÉ ou ACTION}
167.	transmission (transmettre)	1407	{ACTIVITÉ ou ACTION}
168.	typosquatting	20	{ACTIVITÉ ou ACTION ILLICITES}

169.	usurpation d’identité (usurper)	458 (+97)	{ACTIVITÉ ou ACTION ILLICITES}
170.	utilisateur du réseau	214	{ACTEUR DE L’INTERNET : UTILISATEUR}
171.	verrouillage	99	{DISPOSITIF TECHNICO-JURIDIQUE}
172.	violation de données	147	{ACTIVITÉ ou ACTION ILLICITES}
173.	Visionnage	74	{ACTIVITÉ ou ACTION}

Annexe IV : Liste des cadres sémantico-conceptuels dans le *DITerm*

LISTE DES CADRES SÉMANTICO-CONCEPTUELS EXPLOITÉS DANS LE PROJET *DITerm*
{ACTEUR DE L'INTERNET : UTILISATEUR}
{ACTEUR DE L'INTERNET : PRESTATAIRE}
{AYANT DROIT}
{ACTIVITÉ ou ACTION ILLICITES}
{PERSONNE QUI REALISE UNE ACTIVITÉ ou ACTION ILLICITES}
{ACTIVITÉ ou ACTION}
{ACTIVITÉ DE CONTROLE ET DE PROTECTION}
{AUTORITÉ JURIDIQUE}
{ORGANISME DE CONTROLE}
{DISPOSITIF TECHNICO-JURIDIQUE}
{DROIT}
{OBJET EN LIGNE}
{OBJET ILLICITE EN LIGNE}
{OBLIGATION}
{PROCÉDURE}
{RÈGLE}
{SERVICE INTERNET}
{TECHNOLOGIE ou MACHINE}

{TEXTE LÉGISLATIF}

{AUTORISATION}

{INTERDICTION}

{INFRACTION}

{CONFORME À LA LOI}

Annexe V : Liste des cadres sémantiques dans le *DITerm*

CADRES SÉMANTIQUES DU DOMAINE JURIDIQUE EXPLOITÉS DANS LE PROJET *DITerm*

INFRACTION

Un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} commet une {INFRACTION} quand il réalise une {ACTIVITÉ ou ACTION}

{ACTIVITÉ ou ACTION} ou une {ACTIVITÉ ou ACTION ILLICITES} constitue une {INFRACTION}.

AUTORISATION ou INTERDICTION

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE JURIDIQUE}autorise / interdit un {AYANT DROIT} ou un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} à réaliser une {ACTIVITÉ ou ACTION}

OU

Un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} ou un {AYANT DROIT} est / n'est pas autorisé à réaliser une {ACTIVITÉ ou ACTION}

OU

Une {ACTIVITÉ ou ACTION} ou un {OBJET EN LIGNE} est autorisé/ interdit dans certaines circonstances.

DROIT

Un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET :

UTILISATEUR} a/n'a pas le droit de réaliser une {ACTIVITÉ ou ACTION}.

OU

Un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} a le droit à {une {PROCÉDURE}, un {DISPOSITIF TECHNICO-JURIDIQUE}.

OBLIGATION

Un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou un{ACTEUR DE L'INTERNET : UTILISATEUR} est / n'est pas obligé de réaliser une {ACTIVITÉ ou ACTION} concernant un {OBJET EN LIGNE}, une {PROCEDURE}, un {DISPOSITIF TECHNOCO-JURIDIQUE} dans certaines circonstances.

OU

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE JURIDIQUE} oblige un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} à réaliser une {ACTIVITÉ ou ACTION} concernant un {OBJET EN LIGNE}, une {PROCEDURE}, un {DISPOSITIF TECHNOCO-JURIDIQUE dans certaines circonstances.

OU

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE JURIDIQUE} impose une obligation concernant une {ACTIVITÉ ou ACTION}, un {OBJET EN LIGNE}, une {PROCEDURE}, un {DISPOSITIF TECHNOCO-JURIDIQUE} dans certaines circonstances.

OU

L'obligation d'une {ACTIVITÉ ou ACTION} un {OBJET EN LIGNE}, une {PROCEDURE}, un {DISPOSITIF TECHNICO-JURIDIQUE appartient à un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou un {ACTEUR DE L'INTERNET : UTILISATEUR}.

OU

Une {ACTIVITÉ ou ACTION}, une {PROCEDURE}, un {DISPOSITIF TECHNICO-JURIDIQUE} est une obligation d'un {ACTEUR DE L'INTERNET : PRESTATAIRE} ou un {ACTEUR DE L'INTERNET : UTILISATEUR}.

RÉGLEMENTATION

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE JURIDIQUE} établit une {RÈGLE}, une {PROCEDURE}, un {DISPOSITIF TECHNICO-JURIDIQUE} ou une

{OBLIGATION} concernant une {ACTIVITÉ ou ACTION} ou un {OBJET EN LIGNE} qui doit être respectée dans certaines circonstances.

OU

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE JURIDIQUE} réglemente une {ACTIVITÉ ou ACTION} ou un {OBJET EN LIGNE}.

SOUMIS À UNE MESURE LÉGALE

Une {ACTIVITÉ ou ACTION} réalisée par l'acteur de l'Internet est/n'est pas soumis à une {RÈGLE}, une {OBLIGATION}, une {PROCEDURE}, un {DISPOSITIF TECHNICO-JURIDIQUE ou à un {TEXTE JURIDIQUE} dans certaines circonstances.

OU

ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} est/n'est pas soumis à une {RÈGLE}, une {OBLIGATION}, une {PROCEDURE}, un {DISPOSITIF TECHNICO-JURIDIQUE} ou à un {TEXTE JURIDIQUE} dans certaines circonstances.

OU

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE JURIDIQUE} soumet/ ne soumet pas une {ACTIVITÉ ou ACTION} réalisée par l'acteur de l'Internet à une {RÈGLE}, une {OBLIGATION}, une {PROCEDURE} ou un {DISPOSITIF TECHNICO-JURIDIQUE dans certaines circonstances.

RESPECT DE LA RÉGLEMENTATION

{ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} respecte/ne respecte pas une {RÈGLE}, une {OBLIGATION}, un {DISPOSITIF TECHNICO-JURIDIQUE}, une {PROCEDURE} ou un {DROIT} de quelqu'un.

OU

Une {RÈGLE} ou une {OBLIGATION} ou un {DISPOSITIF TECHNICO-JURIDIQUE} ou une {PROCEDURE} un {DROIT} de quelqu'un est/n'est pas respecté(e).

REQUIS PAR LA LOI

Pour être telle (telle) que prévu(e) par la loi ou pour éviter des conséquences indésirables une {ACTIVITÉ ou ACTION} ou un {OBJET EN LIGNE}, une {PROCÉDURE} ou un {DISPOSITIF TECHNICO-JURIDIQUE} doit remplir les conditions telles que requises par la loi.

Une {ACTIVITÉ ou ACTION} ou une {PROCÉDURE} ou un {DISPOSITIF TECHNICO-JURIDIQUE} est requis(e) dans certaines circonstances.

RESPONSABILITÉ

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE JURIDIQUE} rend /ne rend pas responsable {ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR} d'une {ACTIVITÉ ou ACTION}, d'une ACTIVITÉ ou ACTION ILLICITES} ou d'une {INFRACTION} dans certaines circonstance.

OU

{ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR}n'est pas/est responsable d' une {ACTIVITÉ ou ACTION}, d'un {OBJET EN LIGNE} ou d'un {SERVICE EN LIGNE} dans certaines circonstances.

OU

{ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR}n'est pas/est juridiquement responsable d'une {ACTIVITÉ ou ACTION ILLICITES} ou d'une {INFRACTION} quand il réalise une {ACTIVITÉ ou ACTION}.

OU

{ACTEUR DE L'INTERNET : PRESTATAIRE} ou {ACTEUR DE L'INTERNET : UTILISATEUR}n'est pas/est juridiquement responsable quand il réalise une {ACTIVITÉ ou ACTION}.

LÉGALITÉ

Une {ACTIVITÉ ou ACTION} ou une {ACTIVITÉ OU ACTION ILLICITES}, un {OBJET EN LIGNE} ou une {TECHNOLOGIE ou MACHINE} est / n'est pas conforme à la loi

PUNI PAR LA LOI

Une {ACTIVITÉ ou ACTION} ou une {ACTIVITÉ OU ACTION ILLICITES}, un {OBJET EN LIGNE} ou une {TECHNOLOGIE ou MACHINE} est / n'est pas puni par la loi d'une peine.

OU

Une {AUTORITÉ JURIDIQUE} ou un {TEXTE JURIDIQUE} punit une {ACTIVITÉ ou ACTION}, une {ACTIVITÉ OU ACTION ILLICITES} ou une {INFRACTION} d'une peine

PRÉJUDICE

Une {ACTIVITÉ ou ACTION}, une {ACTIVITÉ ou ACTION ILLICITES} ou un {OBJET EN LIGNE} ou une {TECHNOLOGIE ou MACHINE} ou une personne nuit à un {AYANT-DROIT}, un {DROIT} de quelqu'un, à une {ACTIVITÉ DE CONTROLE ET DE PROTECTION}, à une {ACTIVITÉ ou ACTION} ou un {OBJET EN LIGNE}

Annexe VI : *DITerm* – modèle de dictionnaire du domaine du droit de l'Internet

Les fiches du *DITerm* sont disponibles sur le support de stockage ci-joint.