Actes en ligne de DoSciLa 2013 *La langue en contexte*
Paris Diderot, CLILLAC-ARP, 5 avril 2013

Pour un dictionnaire spécialisé orienté vers la mise en discours

Aleksandra LICZNER

Université Lyon 2, ED 484 3LA, Centre de Recherche en Terminologie et Traduction
86, rue Pasteur - 69365 LYON CEDEX 07, FRANCE
Tél. : +33(0)4 78 69 72 13 – Fax : +33(0)4 72 72 09 46
Courriel : Aleksandra.Liczner @univ-lyon2.fr

ABSTRACT

This article is part of a preparatory phase of a project with multilingual terminology database (French-Spanish-Polish) within the Internet law that is called DiTerm. It is a specialized dictionary created from a collection of legal, academic and popular texts. It is also designed as an aid in translation which aims to describe practices observed in the language of a given specialty. We first present the stages of the terminology analysis that allowed extracting a large number of the information of a different kind such as linguistic, cognitive, communicative, social and cultural. In the second part of the article, we will focus on presenting a consignment template of various lexical-semantic relationships (derivational, actantial, circumstantial, collocational), inspired by the model of lexical functions (LF) developed by Mel'čuk and his co-workers.

RÉSUMÉ

Le présent article s'inscrit dans la phase préparatoire d'un projet de base de données terminographiques multilingue (français – espagnol – polonais) du domaine du droit de l'Internet, baptisée *DITerm*. Il s'agit d'un dictionnaire spécialisé créée à partir d'un corpus de textes juridiques, universitaires et de vulgarisation et conçu comme une aide à la traduction (et à la rédaction technique) dont l'ambition est de décrire les usages observés dans la langue de spécialité donnée. Nous présentons d'abord les étapes de l'analyse terminologique qui a permis d'extraire un grand nombre d'informations linguistiques, cognitives, communicationnelles, socioculturelles. Dans la deuxième partie de l'article, nous nous concentrons sur la présentation d'une méthode de consignation de différentes relations lexico-sémantiques (relations dérivationnelles, actantielles, circonstancielles, collocationnelles), inspirée du modèle des fonctions lexicales (FL) développé par Mel'cuk et ses collaborateurs.

1. INTRODUCTION

Les ressources terminologiques conventionnelles destinées aux traducteurs ne répondent pas à tous les besoins de leurs utilisateurs, notamment à celui de l'autonomie discursive. En effet, la plupart des dictionnaires spécialisés ou glossaires auxquels ont recours les traducteurs dans leur travail quotidien, ne fournissent pas suffisamment d'informations sur le fonctionnement des termes dans leur univers discursif. Tout particulièrement, on constate l'absence :

- de données concernant les relations sémantiques ou liens conceptuels que les termes entretiennent avec d'autres termes ou unités lexicales
- de renseignements sur les combinaisons lexicales typiques dans lesquelles les termes se trouvent.

Or il s'agit des informations indispensables permettant de produire un texte non seulement cohérent au niveau terminologique mais aussi « correct » (c'est-à-dire « lisible » et « riche »), sur le plan stylistique.

Comment peut-on donc rendre les dictionnaires destinés à la traduction spécialisée plus performants, plus utiles et plus proches de la réalité ?

Ce problème a déjà été soulevé par de nombreux terminologues et terminographes (entre autres Marie-Claude L'Homme, Meyer, Heid et Freibott, Dancette, Cohen, Kocurek, Pavel, Mathieu-Colas), qui considèrent que les ressources terminolographiques modernes devraient se fixer comme objectif de faire une description globale de la langue de spécialité. Ainsi, il est nécessaire de chercher à décrire toutes les propriétés linguistiques de chaque terme (ses propriétés syntaxiques, lexico-sémantiques, pragmatiques, combinatoires) et cela sans négliger la dimension conceptuelle.

Le présent article s'inscrit dans la phase préparatoire d'un projet de base de données terminographiques multilingue (français – espagnol – polonais) du domaine du droit de l'Internet, baptisée *DITerm*. Son objectif principal est de proposer un modèle de description complète des unités terminologiques qui rende compte aussi bien de la dimension cognitive des termes (leur place dans la structure conceptuelle) que des dimensions linguistique et communicative [Sag90].

Il s'agit d'un dictionnaire spécialisé créée à partir d'un corpus de textes juridiques, universitaires et de vulgarisation et conçu comme une aide à la traduction (et à

la rédaction technique) dont l'ambition est de rendre compte des usages observés dans la langue de spécialité donnée.

Dans les pages qui suivent, nous rendons compte de l'état d'avancement du projet et plus particulièrement de deux étapes essentielles pour sa mise en place:

- analyse terminologique du discours spécialisé lié au domaine du droit de l'Internet à partir des données textuelles trilingues (section 3)

- recherche d'une méthode de description des propriétés des unités terminologiques afin de rendre compte de leur comportement dans l'univers discursif. Le modèle proposé s'inspire du modèle des fonctions lexicales (FL) développé par Mel'čuk et ses collaborateurs (section 4).

Mais il s'impose avant cela (section 2) de décrire plus en détail les objectifs généraux du projet et son intérêt.

2. *DITERM* ET SES OBJECTIFS ?

Ainsi, pour chaque traducteur, l'idéal serait d'avoir à sa disposition un outil terminologique qui procure toutes sortes d'informations aussi bien concernant la dimension linguistique (nature linguistique des termes, leur comportement en langue, détails sur le sens, sur la combinatoire, etc.) que conceptuelle (relations entre les termes permettant la structuration des connaissances).

Le *DITerm* se fixe donc comme objectif de faire une description globale (dans la mesure du possible) de la langue de spécialité liée au domaine du droit de l'Internet sans, toutefois négliger la dimension conceptuelle.

Le *DITerm* tente avant tout de répondre aux besoins de compréhension et d'autonomie discursive des utilisateurs.

Afin d'atteindre ces objectifs, il faut mettre en œuvre deux stratégies (souvent considérés comme concurrentes [Dan05] ou bien incompatibles [Hom04], 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.

Bien évidemment, il ne s'agit pas ici de développer un projet de ressource ontologique du droit de l'Internet, (ce dernier s'inscrirait plutôt dans le cadre de l'ingénierie des connaissances juridiques).

En revanche, le *DITerm* doit permet aux traducteurs (ou aux rédacteurs techniques) de :

- trouver des descriptions des nuances de sens facilitant la compréhension des notions ;
- trouver, pour chacun des termes, l'ensemble des autres termes ou unités lexicales partageant avec le terme 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.

Une attention particulière est donc portée au traitement des phénomènes liés à la combinatoire lexicale aussi bien paradigmatique (l'analyse des dérivations sémantiques) que syntagmatique (collocations ou les co-occurrences restreintes).

Selon Mejri [Mer11], un discours spécialisé se définit tout d'abord par sa phraséologie. Comme le remarque L'Homme [Hom04], 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, pour atteindre le degré nécessaire de lisibilité dans la langue cible, le traducteur doit savoir combiner les termes à d'autres unités lexicales propres à un discours de spécialité donnée.

Le *DITerm* se veut une contribution à l'enrichissement et à l'amélioration des ressources terminographiques.

3. TERMES EN CONTEXTE – ANALYSE DU CORPUS

Afin d'atteindre ces objectifs, c'est-à-dire, proposer un modèle de dictionnaire thématique orienté vers la mise en discours, il est nécessaire de commencer par une analyse terminologique des données textuelles.

Le texte est à la base de tout travail terminologique, il est une valeur confirmée. On ne peut plus nier les influences de la linguistique de corpus ni les acquis de la terminologie textuelle :

« (...) la terminologie doit " venir " des textes pour mieux y " retourner " » [Bou90]

D'après Cabré [Cab07], nous considérons le texte spécialisé comme la production linguistique qui se manifeste dans le cadre de la communication professionnelle et dont la finalité est professionnelle. Le texte de spécialité est un habitat privilégié des termes. Pour pouvoir décrire les usages d'une langue de spécialité, il est donc indispensable de constituer (et ensuite d'analyser) un corpus composé de textes de spécialité du domaine en question, un corpus équilibré permettant d'extraire des données représentatives. La sélection rigoureuse des textes garantie la qualité de la recherche terminologique. Afin d'assurer la valeur de cette analyse, il est nécessaire de prendre en compte aussi bien les paramètres linguistiques (contexte linguistique, co-texte, voisinage local) qu'extralinguistiques (contexte, situation d'action

langagière), comme les interlocuteurs et leurs connaissances sur le domaine, conditions de dialogue, buts poursuivis, tâches à résoudre.

Cette étape d'analyse terminologique correspond à trois tâches :

- constitution du corpus
- extraction des unités terminologiques
- observation du fonctionnement des termes choisis dans le contexte (comportement, usages, combinatoire)

3.1. Constitution du corpus

Le droit de l'Internet ne peut pas être considéré comme un nouveau droit à part entière. C'est un ensemble des règles de droit applicables aux activités qui mettent en œuvre l'Internet. Il s'agit d'une matière extrêmement vaste et transversale qui 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. Les normes qui le structurent sont tirées du droit commercial, droit d'auteur, droit civile, notamment du droit des contrats, des libertés publiques. Il était donc très difficile de délimiter le champ de recherche et de définir les sources à partir desquelles récolter les textes. Il faut souligner qu'il n'existe aucun dictionnaire consacré au droit de l'Internet ; la nomenclature est donc à créer de toutes pièces.

Finalement, nous avons retenu 5 sous-domaines :

- l'Internet et les données personnelles,
- l'Internet et les contrats (e-commerce),
- l'Internet et la propriété intellectuelle,
- l'Internet et la responsabilité délictuelle des acteurs,
- l'Internet et la sécurité.

Le corpus trilingue (espagnol, français, polonais) est composé de textes de différents niveaux de spécialité provenant de sites officiels (communautaires, gouvernementaux), de revues juridiques en ligne, de portails consacrés aux aspects juridiques d'Internet. L'ensemble du corpus est divisé en 2 blocs :

- d'un côté, 3 corpus parallèles en espagnol, français et polonais composés de textes communautaires (directives, règlements, rapports, décisions, avis, arrêts, textes de vulgarisation) ;
- de l'autre côté, 3 corpus comparables en espagnol, français et polonais qui possèdent des caractéristiques communes et sont composés de textes juridiques nationaux, de commentaires de spécialistes (avocats, universitaires), dossiers, articles spécialisés, articles de vulgarisation.

Le corpus comprend environs 2 500 000 occurrences pour chaque langue.

3.2. Extraction des unités terminologiques

Le corpus a été analysé à l'aide de deux outils :

- l'extraction automatique des candidats termes (termes simples et termes complexes nominaux) a été réalisée à l'aide du logiciel TermoStat, développé à l'OLST de l'Université de Montréal par Patrick Drouin. Ce logiciel d'acquisition automatique de termes s'appuie sur une approche contrastive, c'est-à-dire qu'il utilise une méthode de mise en opposition de corpus spécialisés (notre corpus du domaine du droit de l'Internet) et non spécialisés (corpus de référence non technique constitué d'articles de journaux ou de textes communautaires). Cette technique a permis de faire émerger les unités dont la fréquence dans le corpus du droit de l'Internet est proportionnellement beaucoup plus élevée que dans le corpus non spécialisé. D'après L'Homme [Hom04], nous considérons que la spécificité est un indice fort du statut terminologique des unités.

- le logiciel NooJ, développé par Max Silberztein de l'Université de Franche Comté qui m'a permis d'appliquer au corpus un ensemble des expressions régulières afin d'extraire des mots ou des ensembles des mots correspondant aux patrons lexico-syntaxiques préalablement identifiés. Cette analyse a enrichi nos résultats.

Etant donné que les deux tiers des termes du vocabulaire juridique constituent des mots composés [Cor05], il est nécessaire d'attirer l'attention sur la difficulté de distinguer les séquences à potentiel terminologique des séquences à potentiel collocationnel [Man03], d'autant plus que la reconnaissance de ces deux phénomènes est fondée (à ce stade de l'analyse) sur la fréquence.

3.3. Observation des usages - environnement linguistique des termes

Une fois les candidats termes sélectionnés, nous avons utilisé chaque terme comme nœud [Pea98], afin d'observer le comportement de ces unités terminologiques dans leur univers discursif et de relier les termes à ses cooccurrents en observant les contextes dans lesquels ils apparaissent.

Comme le rappelle Rute Costa [Cos05] en citant Rastier (1998), on peut opposer deux conceptions du contexte linguistique : contexte comme zone d'extension, relativement au signe ; ou de restriction, relativement au texte. Dans le cadre de notre étude terminologique, nous nous intéressons au contexte linguistique (co-texte) entendu comme une zone d'extension qui permet d'élargir l'information conceptuelle, sémantique et syntactique de l'unité de signification (unité terminologique) à travers son analyse. Il est primordial d'isoler et d'observer le contexte (co-texte) qui entoure une unité terminologique car celui-ci fournit des éléments nécessaires pour mieux comprendre le fonctionnement du terme.

L'analyse des concordances des termes (réalisée à l'aide

du logiciel Nooj) nous a permis d'observer le comportement de ces unités terminologiques, leur univers discursif et de :

- établir la structure actantielle des termes, identifier les éléments obligatoires ;
- repérer des contextes illustrant les cooccurrents les plus fréquents des termes et observer leur comportement linguistique ;
- repérer des informations sur le domaine de spécialité, extraire des définitions ;
- extraire des relations logico-sémantiques que le terme entretient avec d'autres termes, des relations spécifiques au domaine du droit de l'Internet.

4. A LA RECHERCHE D'UNE MÉTHODE DE DESCRIPTION

L'analyse du corpus a permis de repérer les cooccurrents les plus fréquents des termes ainsi que d'extraire une liste d'autres termes ou d'autres unités lexicales qui partagent une relation sémantique ou un lien conceptuel avec les termes vedette.

A titre d'exemple, nous proposons d'étudier le cas du terme *donnée à caractère personnel*. L'analyse des concordances a démontré que le terme en question entretient différentes relations lexico-sémantiques (relations sémantiques fondamentales comme synonymie, hyponymie, hyperonymie, co-hyponymie, relations actantielles, circonstancielles, collocationnelles) avec d'autres termes ou d'autres unités lexicales.

Les questions qui se posent sont : comment expliciter cette multitude de relations dans une base de données terminographiques, quel formalisme adopter pour décrire toute la richesse des informations extraites du corpus, comment systématiser les données.

4.1. Modèle des fonctions lexicales

L'enjeu principal de cette recherche est de trouver une méthode de description et de systématisation des caractéristiques lexico-sémantiques des unités terminologiques ainsi que des phénomènes phraséologiques observées dans le corpus, c'est-à-dire propres à la langue de spécialité donnée.

Ainsi, nous nous sommes intéressés au modèle des fonctions lexicales (développé par [Mel95] et ses collaborateurs Alain Polguère et André Clas dans la cadre de la Lexicologie Explicative et Combinatoire), qui offre des méthodes de description formelles, exhaustives et systématiques du lexique d'une langue.

Il faut souligner que la Lexicologie Explicative et Combinatoire (LEC) constitue une composante d'une théorie plus générale, la Théorie Sens-Texte (TST).

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.

Le modèle a déjà séduit un grand nombre de terminographes comme L'Homme, Cohen, Lainé, Dancette, Jousse, Mortchev – Bouveret. L'adaptation du modèle des fonctions lexicales à la terminologie a fait l'objet de travaux conduits par Frawley 1988, L'Homme (2002, 2004), Dancette (2003), Jousse &Mortchev (2003), Mortchev – Bouveret [Mor07]. Le projet *DITerm* cherche à s'inscrire dans cette mouvance et s'inspire de ces travaux.

L'approche mel'čukienne offre un modèle de description globale et rigoureuse de l'unité terminologique.

La table 1 présent les relations lexico-sémantiques du terme *donnée à caractère personnel* ainsi que ses cooccurrences systématisées à l'aide des FL.

En effet, 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'un terme entretient avec d'autres termes ou d'autres unités lexicales :

- sur l'axe paradigmatique, elles permettent de décrire les relations fondamentales en terminologie comme synonymie, conversion, antonymie, les relations taxonomiques (hyperonymie, hyponymie, co-hyponymie), les relations partitives (méronymie, holonymie) comme partie-tout, élément-ensemble ; phase-processus, les phénomènes de la dérivation syntaxique ainsi que les relations actancielles et circonstancielles – dérivés sémantiques ;
- sur l'axe syntagmatique, elles permettent de dégager, pour une unité terminologique donnée, des cooccurrents lexicalement contraintes.

L'attribution des FL se fait à la suite d'une observation d'un nombre élevé d'occurrences en corpus spécialisé, ce qui permet de refléter le fonctionnement linguistique réel des termes et de leurs cooccurrents. Par conséquent, ce modèle de description fournit aux traducteurs une variété d'expressions précises, de combinaisons adéquates et de formulations appropriées au domaine de spécialité donné.

Finalement, la mise en relation des termes du même champ (possible grâce aux FL), permet de rendre compte de la structure conceptuelle et sémantique du domaine, ce qui guide le traducteur dans son approche d'un nouveau domaine.

4.2. Problème de compatibilité du modèle

L'adaptation du modèle des FL à un projet terminographique provoque une réflexion sur la compatibilité des méthodes lexicologiques et terminologiques et sur le statut de la terminologie moderne dans une perspective linguistique.

Le formalisme demande que l'on situe notre projet dans l'optique lexico-sémantique où le terme est considéré comme unité lexicale qui véhicule un sens spécialisé et non comme étiquette de concept. Par conséquent, on se dégage du plan descriptif purement conceptuel car les relations ne sont plus établies entre des concepts (comme dans les modèles ontologiques), mais entre unités lexicales. Cela peut entraîner la perte de possibles rapports conceptuels entre termes.

Table 1. Fonctions lexicales dans *DITerm* – terme : *donnée à caractère personnel*.

STRUCTURE ACTANTIELLE : ~ permettant identifier X (Patient = Personne) et utilisé par Y (Agent = Personne ou Machine)		
FONCTION LEXICALE	**RELATION**	**TERMES/MOTS reliés**
Syn	Synonymie	donnée personnelle, donnée relative aux personnes physiques
Anti	Antonyme	donnée anonyme
Gener	Hyperonymie	donnée, information
Spec (*FL proposée par Grimes)	Hyponymie/Co hyponyme	donnée sensible, donnée biométrique, adresse IP, donnée nominative, donnée de connexion, donnée de localisation
Mult	Méronymie	fichier de données
S_{loc}	Localisation	réseau de communications électroniques, Internet, site Internet
S_1	Agent	responsable du traitement, sous-traitant, pirate informatique, fournisseur de services de communications électroniques accessibles au public
S_2	Patient	personne concernée, public concerné, utilisateur, internaute, abonné
S_3	Destinataire	destinataire
Propr * (FL non standard, adapté)	Propriété	confidentialité, intégrité, disponibilité, authenticité, sécurité
$Real_1$	L'agent réalise une action typique sur….	traiter, utiliser, gérer
S_0Real_1	Non pour l'action typique	traitement de données à caractère personnel…, utilisation de, gestion de
CausPredPejor	L'agent fait de sorte que… se dégrade	détruire, écraser, supprimer
S_0CausPredPejor	Nom pour l'action de dégradation	suppression de…, destruction…, perte…, altération…

BIBLIOGRAPHIE

[Bou99] Bourigault D. & Slodzian M. (1999), « Pour une terminologie textuelle », *Terminologies nouvelles N°19*, pp.29_32

[Cab07] Cabré M.T. (2007-2008), « Constituer un corpus de textes de spécialité », Cahier du *CIEL 2007-2008*, pp. 37-56

[Cor05] Cornu G. (2005), *Linguistique juridique*, Paris, Ed. Montchrestein

[Cos06] Coste R. (2006), « Texte, Terme, Contexte », *7es Journées Scientifiques AUF-LTT, Mots, Termes et Contextes*, Bruxelles, pp. 1-8

[Dan05] Dancette J. (2005), « Les Représentations Lexico-Sémantiques (RLS), moyen de structuration des connaissances dans les domaines spécialisés », *L'organisation des connaissances : approches conceptuelles*, La Librairie des Humanités, L'Harmattant, pp. 83-96

[Hom04] L'Homme M-C. (2004), *La terminologie : principes et techniques*, Montréal, Les presses de l'Université de Montréal (Paramètres)

[Hom08] L'Homme M-C. (2008), « Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés », *Traduire* 217, pp. 78-103

[Mor07] Mortchev-Bouveret, M. (2007), « Modélisation des relations lexico-sémantique dans un dictionnaire spécialisé », *Lexicographie et terminologie : compatibilité des modèles et des*

méthodes, Presses de l'Université d'Ottawa, pp. 293-320

[Man03] Maniez F. (2003), « Un modèle d'extraction des collocations en langue de spécialité », ASp, N° 35-36, pp. 35-48

[Mej11] Mejri S. (2011), « Phraséologie et traduction des textes spécialisés», Estudios y análisis de fraseología contrastiva: lexicografía, traducción y análisis de corpus, MOGORRON

[Mel95] Mel'cuk I. et al. (1995), *Introduction à la Lexicologie explicative et combinatoire*, Louvain-La-Neuve, Editions Duculot

[Pea98] Pearson J. (1998), *Terms in Context*, Amsterdam et Philadelphia, John Benjamins

[Sag90] Sager J C. (1990), A Practical Course in Terminology Processing, Amsterdam / Philadelphia, John Benjamins