Le discours dictionnairique : analyse systématique des structures sémantiques

Margareta Kastberg Sjöblom[1], Max Reinert[2]

[1]ATST, Centre Jacques Petit, EA3187, Université de Franche-Comté, UFR SLHS, 30, rue Mégevand, 25 030 Besançon cedex, margareta.kastberg@univ-fcomte.fr

[2]Laboratoire Printemps, CNRS, UMR 8085, Université de Versailles-St-Quentin-en-Yveline, 78 047 Guyancourt-cedex, max.reinert@wanadoo.fr

Abstract

The purpose of this paper is to enter into a deeper analyse of the study presented at the JADT conference 2006 – a dictionary corpus and the result of the exploration of lexicostatistical methods in order to analyse the nomenclature and the examples that appear in the bilingual dictionary. The continued study is more precisely focuses on the semantic aspects and the associations of themes of a corpus extracted from the totality of the textual contents of the French articles of a bilingual dictionary, *Norstedts stora fransk-svenska ordbok*, the Large French-Swedish Dictionary (1998). The object is to automatically extract the bound structures, in order to evolve the semantic constellations and associations of themes in this particular corpus. The vocabulary analysis proposed by the *Alceste* software proceeds by a hierarchical downward classification and allows a systematic classification of the vocabulary. The semantic classification is here related to the idea of posture more than the real lexical content and it makes it possible to isolate the lexical structures from the dictionary corpus, structures that seem to perfectly fall into specific semantic categories, and also to pave the way for a detailed description of the language diffused by the bilingual dictionary.

Résumé

Nous avons présenté lors des JADT 2006 un travail sur un corpus dictionnairique et le résultat de l'analyse lexicométrique « traditionnelle » de celui-ci en vue d'étudier la nomenclature et le choix de mots qui figurent dans le dictionnaire bilingue. Cet article, qui propose un prolongement de cette étude, s'intéresse plus précisément à l'aspect sémantique et aux associations thématiques d'un corpus dictionnairique extrait de la totalité du contenu textuel des articles français d'un dictionnaire bilingue, *Norstedts stora fransk-svenska ordbok, le Grand Dictionnaire français-suédois (*1998). L'objectif est d'extraire les structures signifiantes les plus fortes, afin de dégager les constellations sémantiques et les associations thématiques contenues dans ce corpus particulier. La méthode de l'analyse planifiée que propose le logiciel *Alceste,* qui procède par la classification descendante hiérarchique et permet un classement systématique du vocabulaire. Le classement sémantique, qui s'inspire de l'idée de posture plus que du véritable contenu lexical, permet de dégager du corpus dictionnairique des structures qui se regroupent parfaitement en classes sémantiques et permettent une description détaillée de la langue qui est diffusée par ce dictionnaire bilingue.

Mots-clés : lexicographie, dictionnairique, classification automatique, sémantique, discours.

1. Introduction

Nous avons présenté lors des JADT 2006 un travail sur un corpus dictionnairique et le résultat de l'analyse lexicométrique « traditionnelle » de celui-ci en vue d'étudier la nomenclature et le choix de mots qui figurent dans le dictionnaire bilingue. En tant que lexicographes nous sommes en effet constamment confrontés au choix, souvent assez personnel, des entrées à faire figurer dans un dictionnaire et des exemples à fournir pour exemplifier leur usage. Ce

travail de rédaction aboutit à un corpus dictionnairique marqué par ses auteurs, et socialement et culturellement très imprégné de ce choix. Toutefois, pour le grand public le texte dictionnairique, avec sa nomenclature et ses exemples, est perçu comme un discours neutre, ayant gardé encore aujourd'hui un caractère quasi normatif. La tendance à considérer le dictionnaire comme un ouvrage « unique », de vérité absolue, donne aussi à ce discours un statut particulier. Dans le cas du dictionnaire bilingue il faut ajouter que le dictionnaire constitue aussi une vitrine d'un pays et d'une culture.

Or, on constate très facilement qu'avec les multiples rééditions des anciens dictionnaires, même les dictionnaires qui semblent récents véhiculent souvent une image assez désuète de la langue cible, et diffusent une image linguistique et culturelle bien loin des réalités. Dans le monde anglo-saxon, germanique et scandinave, où l'anglais domine largement en tant que langue de communication, l'apprentissage du français est de plus en plus considéré comme une activité de luxe et on a souvent l'impression que les maisons d'édition entretiennent cette image d'une langue très noble, très « chic », pour ainsi attirer une sorte d'élite culturelle des francophones. Nous trouvons encore dans le dictionnaire français-suédois, pour exemplifier l'usage du substantif *sortie*, des exemples comme *à la sortie du sermon, faire sa sortie du port* et *sortie de bal,* qui n'incarnent plus le français contemporain, et qui illustrent bien l'image que l'on veut entretenir au détriment d'exemples utiles à l'utilisateur contemporain.

Pour avoir une idée plus précise sur la langue que diffuse le dictionnaire nous avons analysé le vocabulaire phrastique d'un dictionnaire bilingue de façon systématique avec le recours des techniques lexicométriques. Nous proposons ici un approfondissement de l'étude faite il y a deux ans qui s'appuyait sur le logiciel *Hyperbase*. Le présent travail s'intéresse plus précisément à l'aspect sémantique du dictionnaire et aux associations thématiques. Après une description de ce corpus assez particulier, et de sa récolte, nous exposerons les résultats de l'analyse sémantique et de la division en mondes lexicaux stabilisés proposée par l'application de la technique que propose le logiciel *Alceste*. En effet, le corpus phraséologique d'un dictionnaire, c'est-à-dire les phrases et les exemples à l'intérieur du dictionnaire, est bien une forme de corpus clos et il pourrait même être considéré comme un genre de discours qui s'adapterait, après un traitement informatique adéquat, à ce genre d'analyse qui permet de prendre en considération simultanément la totalité du corpus.

Nous nous intéresserons ici, au titre d'exemple d'application possible, aux phrases extraites du grand dictionnaire français-suédois, *Norstedts stora svensk-franska ordbok* (1998)[1].

2. Corpus dictionnairique : articles et exemples dans le dictionnaire – échantillon de langue

La récolte d'un corpus dictionnairique n'est pas aisée. Premièrement, le dictionnaire est presque toujours d'un produit commercial, appartenant à une maison d'édition. Etant donné le coût important d'élaboration d'un dictionnaire et la concurrence rude entre les différents éditeurs, ces corpus sont pratiquement inaccessibles pour la recherche publique[2]. Deuxièmement, la particularité de sa présentation, sa typographie et son métalangage font aussi que le dictionnaire est extrêmement difficile à scanner et à océriser, et l'accès au corpus

[1] *Norstedts stora fransk-svenska ordbok*, le Grand Dictionnaire français-suédois (1998) Stockholm, Norstedts, (74.000 mots et phrases selon l'éditeur).

[2] Ce n'est pas seulement le cas des éditeurs privés, le *TLFi* n'a libre accès que partiellement et depuis peu de temps et *Frantext* n'est pas accessible dans son intégralité.

dictionnairique dans sa globalité, par les différentes versions de dictionnaires sous forme électronique ou de cédéroms, est quasiment impossible, à cause des systèmes de protection. Il faut donc, comme nous l'avons fait, se battre pour obtenir un corpus d'un éditeur, comportant la nomenclature et le contenu des différents articles sous format numérisé[3].

Que faut-il ensuite considérer comme étant le corpus dictionnairique ? Il s'agit d'un côté de la nomenclature, les entrées ou les vedettes, et de l'autre du contenu de chaque article. Dans le cadre d'une étude comme la nôtre c'est le deuxième aspect qui nous intéresse, celui de la langue et de la phraséologie interne, représentées dans l'article dictionnairique.

L'article dictionnairique est une définition large qui inclut - en dehors de l'entrée dictionnairique (la vedette) et des indications métadictionnairiques - toutes sortes d'indications sous forme d'échantillons de langue, contenant une forme de la vedette, afin de donner des indications sur la construction phraséologique, les collocations et les expressions idiomatiques.

Les séquences sont souvent construites sous la forme d'un exemple, suivi d'une indication sur la signification. Dans le dictionnaire unilingue il s'agit souvent d'une paraphrase, dans le dictionnaire bilingue d'une traduction. Al-Kasimi (1977 : 88) utilise le terme de « illustrative examples » et définit l'exemple comme étant « any phrase or sentence that illustrates the use of the item defined or translated ».

Le dictionnaire fournit en effet une indication de construction, de signification ou une expression idiomatique, et l'exemple donné suit très souvent directement cette séquence. Il est intéressant de noter que généralement un échantillon paraphrasé ou traduit ne peut pas être produit « mot à mot » à partir de l'information donnée auparavant, et que ce type d'échantillon dans un dictionnaire sert d'explication souvent nécessaire, comme dans l'exemple suivant : **hund,** chien *där ligger (det) en ~ hund begraven il y a anguille sous roche*. (traduction littérale : *là, il y a un chien enterré*).

Toutefois, l'exemple ne se limite pas à confirmer une information donnée, il sert très souvent à l'amplification, ou même à indiquer une divergence de l'usage. Un échantillon de langue paraphrasé ou traduit peut avoir une position assez indépendante par rapport à l'entrée dictionnairique à laquelle il s'adresse : **kupa 1.** cloche**, ~** *handen «* mettre sa main en coupe ».

On est donc à l'intérieur des exemples dans une microstructure qui se révèle assez indépendante de l'article dans son ensemble. C'est l'ensemble du contenu - phrases, syntagmes ou fragments de phrases - qui constitue notre corpus dictionnairique, le corpus informatisé de l'inventaire français du dictionnaire *Norstedts stora svensk-franska ordbok*.

Notre corpus est constitué par les phrases et les syntagmes qui constituent la partie française des articles du dictionnaire *Norstedts stora svensk-franska ordbok* ; il englobe selon la lecture du corpus par *Alceste* 111 504 occurrences et 16 626 formes distinctes reparties sur les 26 lettres de l'alphabet, choisies ici comme les jalons des différents sous-corpus, ou des lignes étoilées qui introduisent chaque unité de contexte initiale (U.C.I.), pour employer la terminologie d'*Alceste*. La fréquence moyenne par forme est de 7 et le nombre d'hapax s'élève à 8320, des chiffres habituels dans l'analyse de corpus.

[3] Ce n'est pas seulement le cas des éditeurs privés, le *TLFi* n'a libre accès que depuis peu de temps et *Frantext* n'est pas accessible dans son intégralité.

3. Associations thématiques

3.1. Divisions sémantiques

Nous avons évoqué dans l'introduction notre très fort sentiment de subjectivité et de préférence personnelle quant à la langue que diffuse le dictionnaire bilingue. Dans des études antérieures sur le vocabulaire d'un autre dictionnaire bilingue[4] nous avons pu constater le manque de certaines classes sémantiques et l'hypertrophie de certains autres certains domaines, notamment un manque total de termes techniques et informatiques et curieusement une forte présence des domaines de l'équitation et de l'escrime (Kastberg : 2006).

Cette étude était effectuée « à la main » et s'inspirait du classement décrit et utilisé par Eveline Martin (Martin, 1993 : 30-35), initialement proposé par O. Klapp[5] (d'après sa bibliographie et des systèmes que fournissent Bouty et Aziza (Aziza, 1978 : 205)), ce classement divisait les entrées lexicales en diverses catégories principales comme *animés*, *inanimés*, *univers-registres* et des sous-catégories telles *types*, *figures*, *végétaux*, *objets*, *sensible-naturel*, *sociopolitique*, *ludique* etc. De ce travail, nous nous souvenons surtout de la difficulté du classement et du traitement de la polysémie avec une catégorie « *autres* » qui grandissait au fur et à mesure que l'analyse avançait.
Certains analyseurs morpho-syntaxiques proposent aussi un classement sémantique qui peut se révéler intéressant. Le programme *Cordial* fait appel à un thésaurus de référence, où sont cataloguées les disciplines, les concepts et les connaissances.

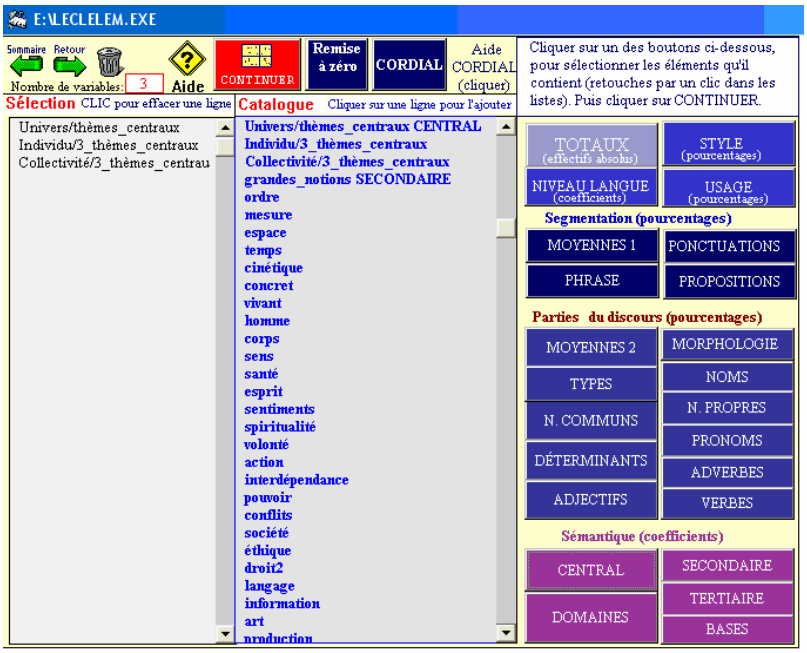


Figure 1 : analyse sémantique de Cordial

Le traitement sémantique proposé par *Cordial* vise à rendre compte des idées, des sentiments, des actions, c'est-à-dire des thèmes exprimés dans un texte. Les résultats auxquels ce classement conduit ne sont pas sans intérêt. Pourtant, on se rend vite compte des faiblesses et

[4] Vising J. (1950) : *Svensk-fransk ordbok, Dictionnaire français-suédois* Stockholm, Svenska Bokförlaget Albert Bonnier.

[5] Klapp O. : Bibliographie d'histoire littéraire française et Bulletin analytique de linguistique française, 1985 – 1990.

des incertitudes du codage sémantiques, avec une opacité quasi-totale quant aux définitions de certains concepts comme par exemple la catégorie *cinétique* proposé par le programme. Il convient aussi de souligner une certaine fragilité de l'analyse due à l'homographie et même à la polysémie, et sans doute ces étiquettes s'appliquent-elles mal à certains corpus.

Le logiciel *Alceste* procède d'une autre manière et s'inspire d'une autre philosophie de base émanant de l'idée de posture plus que du véritable contenu lexical.

Il s'agit de prendre en considération les aspects dynamiques qui divisent l'activité lexicale, tenant compte d'un côté de l'imaginaire et de l'autre du symbolique. On s'est ici inspiré de l'idée de « répétition » chez Foucault, considérant le discours comme une activité d'usage qui véhicule des mots. On ne touche qu'au mimésis, aux discriminations répétées, et il ne s'agit pas de directement diviser le vocabulaire préétabli sur le contenu lexical.

3.2. L'analyse planifiée d'Alceste

L'analyse d'Alceste fonctionne par étapes, le logiciel effectuant une méthode de classification descendante hiérarchique. Cette méthode procède par fractionnements successifs du texte. Elle repère les oppositions les plus fortes entre les mots du texte et extrait ensuite des classes d'énoncés représentatifs.

Nous avons donc préparé notre corpus dictionnairique, comprenant la totalité du contenu textuel français des articles du dictionnaire *Norstedts*, pour l'exécution de l'analyse de son vocabulaire. La première étape de l'analyse consiste en l'identification des locutions, des mots outils et en une lemmatisation des données qui permettent d'établir un dictionnaire du vocabulaire de ce corpus.

Durant cette opération les mots sont catégorisés à l'aide de clés catégorielles qui serviront par la suite à choisir les mots analysés ou bien à les rejeter ou à les mettre en supplément.

Cette opération exige évidemment le recours à un dictionnaire de référence. Dans *Alceste* elle se base sur la division formelle des 10 000 mots pleins du corpus étalon ; le calcul aboutit à une liste des catégories avec leur fréquence.

A3 : Calcul et gestion des clés :

TABLEAU DES CLES PRINCIPALES : CORPUS ETALON : TALavril05

CLE		INTITULE	EFFECTIF	EMPIR	PHI*100
Q	2	Aux_Avoir	138	1170	-11.194
R	2	Aux_Etre	531	2391	-14.109
U	1	VERBES_MODAUX	128	644	-7.544
V	1	VERBE_ER	4806	3471	8.403
W	1	VERBE_IRR	819	1135	-3.484
Y	1	FormeVerbaleFrequent	8228	6835	6.248
0	2	0p_Je	325	1111	-8.749
1	2	1p_1&2singulier	605	1493	-8.528
2	2	2p_1&2pluriel	207	1321	-11.372
3	2	3p_3sing&pluriel	1989	5039	-15.937
4	2	4d_LieuMouvementImpr	1848	2682	-5.973
5	2	5d_TemporaliteActeBu	1068	1783	-6.283
6	2	6d_EvaluatifIntensit	338	1436	-10.747
7	2	7c_DialogueProblemeA	3390	6381	-13.889
8	2	8c_LogiqueDistance	160	804	-8.426
9	2	9c_PrescriptifDemons	4733	6582	-8.454

A3 : TRANSFERT DES CLES TOPIQUES : DICTO -> DICB :
Corpus d'étalonnage : 111 obtenu le 5 avril 2005 (ETALavril05)
DICIN : 35262 mots
Nombre de mots communs avec DICTO : 3163
Nombre de mots absents de DICTO : 780

Nous relevons ici les catégories grammaticales principales présentes dans le corpus et ce tri permet de constater des caractéristiques très révélatrices du corpus dictionnairique.

En effet, nous nous trouvons dans le dictionnaire dans un espace essentiellement verbal, le dictionnaire est une espace d'action, ce qu'illustre ici la surreprésentation remarquable des verbes en *–er.* L'article dictionnairique exemplifie et met en valeur le mouvement, et le verbe en *–er t*raduit souvent le mouvement non intentionnel des activités d'usage qui relèvent de l'imaginaire, tandis que les verbes de la description sont déficitaires.

Il est aussi à remarquer dans ce tableau la relative absence de mots outils. En effet, le discours dictionnairique ne contient que peu d'argumentation, faisant appel aux mots outils dans d'autres types de discours, et le fait que nous trouvons les pronoms personnels parmi les mots les plus déficitaires lors d'une comparaison exogène n'a rien d'étonnant, compte tenu des nombreux exemples impersonnels qui caractérisent tout dictionnaire, ce que reflète par ailleurs aussi l'importance relative de verbes à l'infinitif. Ces deux phénomènes sont en effet très liés, étant donnée l'absence du pronom dans des constructions à l'infinitif comme *partir en voyage, appeler au secours*, etc. qui offrent des définitions détachées de tout ancrage énonciatif et non actualisées.
Après la reconnaissance des formes, l'élimination d'hapax et la réduction des désinences de conjugaison il nous reste 2 842 formes à analyser, 1 075 mots outils et 26 mots étoilés qui serviront de base pour la suite des calculs de données et l'analyse des unités de contexte élémentaires (U.C.E.).

B3 : Classification descendante hiérarchique de DONN.2

Nombre de mots retenus :	2 842
Nombre d'unites de contexte :	3 644
Poids du tableau analyse :	29 540

Il s'agit d'attribuer à chaque unité une appartenance à une classe topique, par quatre clés, qui divise le vocabulaire par rapport à la posture, entre le réel, l'imaginaire et le symbolique. Cette division en classes topiques est l'aboutissement d'une réflexion sur la matérialité de l'activité langagière et la notion de posture, une approche bien différente de celle de divisions sémantiques conventionnelles qui divisent le vocabulaire en classes catégorielles selon le contenu[6] ou bien d'une analyse onomasiologique traditionnelle. On ne trouvera pas dans cette démarche une grille de contenu ou une catégorisation formelle, mais la recherche d'un découpage permettant d'isoler des mondes lexicaux stabilisés à partir d'un calcul purement statistique des probabilités et des cooccurrences. Pour l'estimation statistique *Alceste* fait appel à un corpus d'étalonnage d'environ 26 millions de caractères réunissant des textes du XIX^{e} et du XX^{e} siècles[7].

C1 : Calcul des classes topiques à partir des clés topiques

distribution des marques dans les clés :

CLE A	9522.	32.20%
CLE B	5508.	18.63%
CLE C	6009.	20.32%
CLE D	8528.	28.84%

5618 u.c.e. classées soit 82.11050 %

Nombre de classes retenues : 4

[6] Voir M. Reinert « Mondes lexicaux stabilisés et analyse de discours », article publié dans ces actes qui expliquent cette méthode et cette approche philosophique du langage.

[7] Idem.

Les clés correspondent au classement suivant : Clé A « imaginaire », Clé B « réel intérieur », Clé C « réel extérieur » et Clé D « symbolique » ; la distribution des contextes de notre corpus est la suivante :

1 Classe Contexte A	1581. u.c.e. soit 23.11%
2 Classe Contexte B	1283. u.c.e. soit 18.75%
3 Classe Contexte C	1216. u.c.e. soit 17.77%
4 Classe Contexte D	1538. u.c.e. soit 22.48%

Il s'agit maintenant de comparer les classes obtenues et leur intersection :

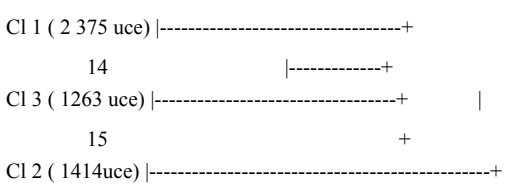
C1 : Intersection des classes RCDH1 et RCDH2

Nombre minimum d'u.c.e. par classe :	456
DONN.1 Nombre de mots par u.c. :	8
Nombre d'u.c. :	3 985
DONN.2 Nombre de mots par u.c. :	9
Nombre d'u.c. :	3 644
5 052 u.c.e classées sur 6 842 soit	73.84 %
Nombre d'u.c.e. distribuées :	6 647

Distribution des u.c.e. entre les deux partitions :

	RCDH1 *	RCDH2		
	Classe *	1	2	3
	poids *	3 084	1 748	1 815
1	3 079 *	2 375	265	439
2	1 818 *	291	1 414	113
3	1 750 *	418	69	1 263

Classification Descendante Hiérarchique



Cette opération nous permet de constater que trois catégories occupent 75% (5 052 u.c.e. classées, soit 73.83806%) de la surface les clés A, B et D, la catégorie C étant fortement sous-représentée dans le discours du dictionnaire bilingue. En effet, le dictionnaire bilingue est déficitaire dans cette catégorie, à la différence d'un dictionnaire unilingue qui contient de l'information d'ordre encyclopédique faisant notamment appel aux noms propres. C'est probablement une des différences structurelles fondamentales entre un dictionnaire bilingue et un dictionnaire monolingue, surtout de grande ampleur comme *le Larousse* ou *le Robert.*

4. Profil du vocabulaire dictionnairique

La plus grande classe de notre corpus, la classe 1 qui définit le « contexte B », occupe à elle seule 47% des U.C.E. retenus dans l'analyse. Avant d'observer les différents items lexicaux du tableau, notons la troisième ligne de la fin du tableau qui nous indique que 659 des 978 items lexicaux se trouvent dans la clé B, ce qui est une valeur exceptionnelle, qui traduit une cohérence quasi parfaite du calcul statistique[8] et des valeurs très significatives.

[8] Pour plus de détails sur ce calcul voir Reinert M. *Notice d'Alceste*.

C2 : Profil des classes/formes réduites

Classe n° 1 : nombre d'uce. : 2 375. soit : 47.01%

Effectifs		Pourc	Chi2	Identification	246.	412.	59.71	29.03 o	B7 pas
131.	162.	80.86	77.00 a	BG bien	85.	119.	71.43	29.17 o	B9 cela
106.	124.	85.48	75.53 a	DD fait+	140.	216.	64.81	28.71 o	B7 pour
72.	84.	85.71	51.37 a	BG mal	147.	230.	63.91	27.63 o	C9 ca
72.	86.	83.72	47.33 a	BB temps	67.	92.	72.83	25.07 o	A1 tu
333.	550.	60.55	45.38 a	BB avoir	91.	138.	65.94	20.41 o	B7 tout
51.	57.	89.47	41.73 a	DB dire+	158.	264.	59.85	18.43 o	C3 son
544.	971.	56.02	39.20 a	BY faire	465.	868.	53.57	18.11 o	A9 quelqu'un
58.	72.	80.56	32.99 a	BC mauvais+	235.	414.	56.76	17.22 o	D9 quelque chose
62.	79.	78.48	31.91 a	AA jour+	42.	57.	73.68	16.47 o	D3 on
42.	51.	82.35	25.83 a	AY va	34.	44.	77.27	16.32 o	B7 peu
373.	667.	55.92	24.50 a	DD être	170.	291.	58.42	16.13 o	D9 c'est
134.	211.	63.51	24.05 a	BY prendre	14.	14.	100.00	15.82 o	3 il m'
29.	32.	90.62	24.59 a	CC ans	33.	43.	76.74	15.39 o	D5 ou
64.	89.	71.91	22.55 a	CC bonne+	59.	87.	67.82	15.38 o	B6 plus
40.	50.	80.00	22.06 a	BA coeur+	28.	35.	80.00	15.40 o	B7 si
61.	85.	71.76	21.27 a	CY aller	56.	83.	67.47	14.18 o	B2 vous
26.	29.	89.66	21.29 a	CC parole+	99.	160.	61.88	14.65 o	A3 ses
32.	38.	84.21	21.27 a	CE dernier+	59.	89.	66.29	13.52 o	B1 me
37.	46.	80.43	20.82 a	U pouvoir	44.	64.	68.75	12.30 o	B0 je ne
23.	25.	92.00	20.41 a	AA vieu+	14.	15.	93.33	12.96 o	B7 moins
172.	264.	65.15	36.80 b	DR est	24.	31.	77.42	11.58 o	D2 nous
100.	146.	68.49	27.85 b	a y	48.	72.	66.67	11.33 o	B7 rien
51.	69.	73.91	20.32 b	n deux	22.	28.	78.57	11.26 o	D9 ici
60.	86.	69.77	18.19 b	i bon	41.	61.	67.21	10.12 o	B3 elle
517.	990.	52.22	13.42 b	a en	19.	24.	79.17	10.01 o	C3 il se
26.	33.	78.79	13.46 b	n trois	19.	24.	79.17	10.01 o	B7 mieux
27.	36.	75.00	11.40 b	BQ ai	71.	114.	62.28	10.92 o	D9 ce
204.	288.	70.83	69.58 o	D9 que	21.	27.	77.78	10.31 o	D9 ce que
225.	340.	66.18	53.75 o	C3 il	659.	978.	67.38	202.04 s	*Cle_B
279.	441.	63.27	51.25 o	D7 ne	60.	90.	66.67	14.21 s	*H
60.	73.	82.19	36.80 o	B0 je	38.	55.	69.09	10.88 s	*Q

Les mots qui se trouvent à l'intérieur de cette classe traduisent le réel et l'affect. Nous trouvons en tête de liste des items comme *bien*, *mal*, *mauvais*, *bonne*, *mieux*, etc., des mots qui véhiculent des « valeurs » : *un homme très bien*, *être bien avec*, *des gens biens*, *en tout bien tout honneur*, *grand bien vous fasse*, *un livre mal accueilli*, *être mal compris*, *avoir mal au cœur*, etc.

C'est aussi dans cette catégorie que nous trouvons les verbes. Ici nous trouvons les verbes d'état et les verbes modaux et la riche fréquence des pronoms témoigne également de cette réalité[1]. Des verbes comme *faire*, *avoir*, *dire*, *prendre*, *pouvoir* et *donner* en tête de liste reflètent non seulement la description mais aussi une activité incessante à l'intérieur des articles dictionnairiques et lorsqu'il s'agit d'un discours articulé ce dictionnaire semble nettement privilégier les pronoms à la première et à la deuxième personne.

L'utilisation de pronoms reflète aussi l'argumentation et c'est bien dans cette classe que nous en trouvons les signes avec la négation et les mots-outils.

[9] La corrélation dans toute analyse lexicométrique entre le verbe et le pronom est par ailleurs bien documentée (cf. M. Kastberg Sjöblom 2002 ; 339-341).

La deuxième classe de notre corpus, la classe 2 qui définit le « contexte D », occupe 27.99% des U.C.E. retenus dans l'analyse. Encore une fois la catégorisation semble presque parfaite avec 618 des 116 items appartenant à la clé D et un Chi2 de 470.54 ! (sixième ligne avant la fin) :

Effectifs		Pourc	Chi2	Identification
38.	39.	97.44	94.05a	DE social+
36.	47.	76.60	55.61a	DE public+
20.	20.	100.00	51.66a	DE civil+
19.	19.	100.00	49.07a	CC vente+
20.	21.	95.24	47.32a	DE general+
26.	33.	78.79	42.53a	DD etat+
19.	21.	90.48	40.85a	BB peine+
19.	21.	90.48	40.85a	DD acte+
15.	15.	100.00	38.71a	DD medical+
19.	22.	86.36	37.36a	CC poste+
16.	17.	94.12	37.01a	CE chef+
16.	17.	94.12	37.01a	DD produit+
14.	14.	100.00	36.12a	BE demande+
19.	23.	82.61	34.20a	CE police+
15.	16.	93.75	34.44a	DE interet+
21.	27.	77.78	33.39a	DE ecole+
13.	13.	100.00	33.53a	DE judiciaire+
20.	26.	76.92	31.05a	DE societe+
14.	15.	93.33	31.87a	CC agent+
14.	15.	93.33	31.87a	DE credit+
22.	30.	73.33	30.79a	CE service+
12.	12.	100.00	30.95a	BB conseil+
15.	17.	88.24	30.72a	DE professionnel+
12.	12.	100.00	30.95a	DD science+
12.	12.	100.00	30.95a	DE defense+
394.	1027.	38.36	68.85b	a d'
16.	22.	72.73	21.94b	m etat
7.	7.	100.00	18.03b	x dispositif
7.	8.	87.50	14.08b	x ordinateur
9.	12.	75.00	13.19b	x publicitaire
8.	10.	80.00	13.45b	x telephonique
5.	5.	100.00	12.88b	x informatique
4.	4.	100.00	10.30	m q
4.	4.	100.00	10.30b	m securite
4.	4.	100.00	10.30b	x secr
84	224.	37.50	10.52o	D7 par
618.	1166.	53.00	470.54s	*Cle_D
111.	186.	59.68	96.21s	*I
19.	26.	73.08	26.36s	*U
141.	366.	38.52	21.73s	*A
104.	280.	37.14	12.32s	*E
118.	329.	35.87	10.83s	*S

La clé D traduit la posture du symbolique. Ici, il s'agit surtout du social et de la société avec des items comme *social*, *public*, *civil*, *médical*, *police*, etc. La posture symbolique pourrait sembler assez étonnante dans le discours dictionnairique, qui est très concret. Il s'agit d'une posture partielle car dans le social il y a une grande partie du symbolique et le thème de la société dans ce contexte est celui qui est le plus proche du symbolique.

La société est un thème très important dans le dictionnaire bilingue. Un des buts principaux du dictionnaire bilingue est en effet d'« aider » et d'« orienter » un utilisateur dans une nouvelle culture et dans une société différente de la sienne. Bien utiles sont donc des exemples comme *année civile*, *droit civil*, *guerre civile*, *partie civile* etc. En effet, si le mot *civil* est le même dans les deux langues, *civil* se traduit par *civil*, en contexte l'usage est bien différent. *Année civile*, se traduit par *kalenderår* (année de calendrier), *droit civil* par *civilrätt*, *guerre civile* par *inbördeskrig (*guerre interne*)*, *partie civile* par *målsägande* (teneur de la cause), tandis que dans l'autre sens nous trouvons *civilekonom* qui se traduit en français par *diplômé en sciences économiques*, *civilkurage* par *courage civique* et *civilförvaltning* par *administration publique*.

La classe 3 occupe 25% et définit la clé A, 632 des 1123 occurrences y appartiennent et le Chi2 est ici 735.43, valeur toujours aussi significative :

Classe n° 3 : nombre d'uce. : 1263. soit 25.00%

Effectifs	Pourc		Chi2	Identification	41.	76.	53.95	34.48 a	F eau
31.	34.	91.18	79.95 a	AA oeil+	15.	18.	83.33	32.78 a	AV cass+er
29.	32.	90.62	73.97 a	AF nez	13.	15.	86.67	30.51 a	AF sec+
31.	36.	86.11	72.22 a	A terre	10.	10.	100.00	30.06 a	AA doux
29.	34.	85.29	66.37 a	AF rouge+	24.	38.	63.16	29.73 a	AA bois
44.	65.	67.69	64.01 a	AF pied+	17.	23.	73.91	29.48 a	AF doigt+
21.	21.	100.00	63.26 a	AF oeuf+	14.	17.	82.35	29.92 a	AA bleu+
28.	36.	77.78	53.87 a	F feu	14.	17.	82.35	29.92 a	CC creme+
31.	44.	70.45	48.91 a	DY tirer	11.	12.	91.67	28.51 a	CC fromage+
16.	16.	100.00	48.15 a	AF huile+	13.	16.	81.25	27.09 a	CY sauter
17.	19.	89.47	42.28 a	AY marcher	218.	698.	31.23	16.78 b	z les
14.	14.	100.00	42.12 a	CV tap+er	8.	11.	72.73	13.39 b	z r
31.	48.	64.58	40.50 a	AA noir+	4.	4.	100.00	12.01 b	DY invente
20.	25.	80.00	40.53 a	AF verre+	88.	207.	42.51	35.30 o	A5 comme
15.	16.	93.75	40.46 a	DD gaz	127.	390.	32.56	12.90 o	A4 dans
16.	18.	88.89	39.33 a	AF glace+	632.	1123.	56.28	753.41 s	*Cle_A
13.	13.	100.00	39.10 a	AA noeud+	109.	249.	43.78	49.24 s	*B
31.	49.	63.27	38.64 a	AF blanc+	88.	242.	36.36	17.51 s	*F
17.	20.	85.00	38.55 a	AF fer	205.	657.	31.20	15.50 s	*C
15.	17.	88.24	36.38 a	CY pate	4.	4.	100.00	12.01 s	*K
12.	12.	100.00	36.09 a	AF botte+	51.	139.	36.69	10.42 s	*G
16.	19.	84.21	35.66 a	AF bouche+					

La clé A est celle de l'imaginaire, et nous y trouvons les couleurs et les parties du corps. Les couleurs y sont presque toutes représentées avec une préférence pour le *rouge* et le *noir* qui sont très fréquemment présentes dans les déterminations diverses comme *betterave rouge*, *chou rouge*, *vin rouge* et dans des expressions idiomatiques comme *Chaperon rouge*, *le tapis rouge* ou *se fâcher tout rouge*.

Des parties du corps, l'œil et le nez sont les plus fréquentes. L'œil est aussi très présent dans des expressions assez imagées comme *je m'en bats l'œil*, *faire un clin d'œil*, *avoir le compas dans l'œil* et *l'œil du typhon*.

C'est aussi dans cette catégorie que nous trouvons la grande masse des verbes en *–er* : *tirer*, *marcher*, *casser*, *sauter* etc. qui sont souvent des verbes d'action comme nous l'avons déjà signalé, mais qui sont aussi bien représentés dans des expressions plus imaginaires comme *faire sauter la banque*, *se faire sauter la cervelle*, *sauter du coq à l'âne* et *sauter sur l'occasion.*

La répartition sémantique du vocabulaire de cette classe est aussi nette que dans les autres classes et ce classement à partir de trois clés semble parfaitement convenir à ce corpus.

En outre, ce classement nous a permis de dégager un profil morphosyntaxique du discours dictionnairique et une nette distinction de la distribution à l'intérieur des trois classes.

Distribution des marques : Tableau des chi2 signes :

						Clé B	Clé D	Clé A
		*	Classes	*		1	2	3
	Identification	*	Poids	*		23840	12674	12172
A	AdjNom_I1_A	*	2423	*		-15	-76	178
B	AdjNom_R1_B	*	1765	*		55	1	-93
C	AdjNom_R2_C	*	2582	*		-1	47	-32
D	AdjNom_S1_D	*	3391	*		-13	255	-143

E	AdjNom_S2_E	*	2762	*	-295	1294	-276
F	AdjNom_I2_F	*	4094	*	-240	-232	1112
Q	Aux_Avoir	*	96	*	8	-4	-1
R	Aux_Etre	*	378	*	49	-12	-20
U	VERBES_MODAUX	*	90	*	42	-8	-20
V	VERBE_ER	*	1010	*	-59	-3	117
W	VERBE_IRR	*	159	*	-1	0	1
Y	FormeVerbaleFrequent	*	4171	*	144	-199	0
0	0p_Je	*	234	*	47	-14	-16
1	1p_1&2singulier	*	422	*	47	-8	-24
2	2p_1&2pluriel	*	143	*	20	0	-19
3	3p_3sing&pluriel	*	1324	*	109	-52	-22
4	4d_LieuMouvementImpr	*	1086	*	0	-6	10
5	5d_TemporaliteActeBu	*	650	*	9	-18	0
6	6d_EvaluatifIntensit	*	166	*	22	-2	-14
7	7c_DialogueProblemeA	*	2026	*	106	-41	-28
8	8c_LogiqueDistance	*	31	*	3	0	-2
9	9c_PrescriptifDemons	*	2973	*	157	-46	-56

Nombre de marques distribuées : 48 686

Les différentes catégories grammaticales sont bien reparties sur les trois classes, avec les verbes en *–er* dans la troisième classe et tous les autres verbes dans la première. Remarquons ici que les pronoms se trouve largement dans la première classe et qu'ils sont largement déficitaire dans la classe qui contient les verbes en *–er,* ce qui confirme bien ce que nous avons déjà constaté sur la particularité de l'articulation verbale dans un dictionnaire bilingue. De la même façon le fait de trouver les marques de la temporalité dans la classe B confirme bien que c'est la seule classe où nous trouvons de la « narration » et de l'argumentation. Toutes les marques de l'argumentation (sauf les marques de lieu et du mouvement, liées aux verbes en *–er* de la classe 3) se trouvent effectivement dans cette catégorie B, et la deuxième classe en est totalement déficitaire.

5. Conclusion

L'objectif de cette étude n'était pas de traiter tous les aspects sous lesquels on pourrait définir le discours dictionnairique. Il s'agit ici d'une expérience qui s'intéresse à un aspect spécifique, la répartition thématique et morphosyntaxique du vocabulaire d'un dictionnaire bilingue et la mise en application d'une méthode qui jusqu'à ce jour n'a pas été appliquée sur un corpus dictionnairique.

Cette étude d'*Alceste* confirme bien notre idée sur le choix du vocabulaire dans un dictionnaire et une volonté directionnelle assez prononcée des lexicographes vers certains domaines sémantiques, des liens de prévisibilité du contenu propre du dictionnaire, sans pour autant que l'on puisse dégager des classes thématiques « traditionnelles ». L'analyse du vocabulaire et le découpage en classes topiques permettent l'identification des mondes lexicaux stabilisés de ce corpus, dégageant des caractères sémantiques propres au discours dictionnairique. En ce sens, cette analyse complète bien les analyses effectuées auparavant, en s'appuyant sur une catégorisation basée purement sur le contenu lexical. Avec cette méthode il ne s'agit pas de chercher des champs lexicaux, des champs conceptuels, des taxinomies ou bien des isotopies dans le sens traditionnel, mais au-delà de la catégorisation lexicale, par le

biais d'une analyse statistique, d'arriver à dégager des mondes lexicaux qui se révèlent bien synthétiques et caractéristiques.

Dans le cadre de cette analyse, avec l'application de la méthode qu'offre *Alceste,* il s'avère que les classes qui ressortent dans les différentes listes peuvent sembler presque parfaites, voire construites. Toutefois, les différentes classes sont bien calculées statistiquement, en appliquant une méthode mathématique systématique et rigoureuse, sans tenir compte du contenu lexical. Et si nous avons trouvé une telle différence de distinction entre ces différentes classes, c'est que la méthode statistique correspond bien à une division sémantique nette et à un certain dynamisme de ce vocabulaire.

Au-delà de cette analyse, nous aimerions encore approfondir, malgré les difficultés matérielles, l'étude et l'exploitation systématique du discours dictionnairique, notamment des dictionnaires bilingues. Ce discours, que nous considérons comme un genre à part entière, constitue une véritable vitrine linguistique vers l'étranger, reflétant cultures et idéologies, un discours qui mérite d'être étudié tant en diachronie qu'en synchronie.

Références

Al Kasimi A. M. (1977). *Linguistics and Bilingual Dictionaries*. Leiden, Brill.

Atkins S., Zampolli A. (eds.). (1994). *Computional Approches to the Lexicon*. Oxford, Oxford University Press.

Aziza C., Oliviéri C. et Sctrick R. (1978). *Dictionnaire des symboles et des thèmes littéraires*. Nathan, Paris.

Béjoint H., Thoiron P. (1996). *Les dictionnaires bilingues*. Louvain-la-Neuve, Aupelf-Uref, Editions Duculot.

Biber D., Conrad S., Reppen R. (1998). *Corpus linguistics, Investigating Language, Structure and Use*. Cambridge, Cambridge Approaches to Linguistics.

Kastberg Sjöblom M. (2006). La nomenclature française du dictionnaire français-suédois : choix ou institution ? In Szende T. (ed.) *Le français dans les dictionnaires bilingues*. Paris, Honoré Champion, Collection Etudes de lexicologie, lexicographie et dictionnairique, p.245-262.

Kastberg Sjöblom M. (2004). Les dictionnaires dans la paire français-suédois ; une approche culturelle. In A.-M. Laurian et T. Szende (eds.), *Dictionnaires bilingues et interculturalité*. Berne, Editions Peter Lang, Collection Etudes contrastives, p.183-200.

Marello C. (1987). Examples in contemporary Italian bilingual dictionaries. In A. P. Cowie (ed.), *The Dictionary and the Language Learner.* Papers from the EURALEX Seminar at the University of Leeds, 1-3 April 1985.

Martin E. (1993). *Reconnaissance de contextes thématiques dans un corpus textuel, éléments de lexico-sémantique*. Paris, Didier Erudition, Collection Etudes de sémantique lexicale, CNRS.

Norstedts stora svenk-franska och fransk-svenska ordbok, le Grand Dictionnaire français-suédois et suédois-français. (1998). Stockholm, Norstedts Ordbok.

Rastier F. (1991). *Sémantique et recherches cognitives*. Paris, PUF, Collection Formes sémiotiques.

Reinert M. (2002). *Alceste, Manuel de référence*. Université de Saint-Quentin-en Yvelines, CNRS.

Reinert M. (2008). Mondes lexicaux stabilisés et analyse statistique de discours. In *JADT2008*.

Szende T. (ed.). (2003). *Les écarts culturels dans les dictionnaires bilingues*. Paris, Honoré Champion.

Svensén B. (2004). *Handbok i lexikografi, Ordböcker och ordboksarebte i teori och praktik*. Stockholm, Norstedts Akademiska förlag.