Constitution d'un lexique verbal polylexical " fondamental " du français : repérage à partir de corpus et modélisation

Doriane Simonnet

Constitution d'un lexique verbal polylexical « fondamental » du français : repérage à partir de corpus et modélisation

Doriane Simonnet

Sous la direction d'Agnès Tutin

Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles - LIDLEM

UFR LLASIC
Département Informatique intégrée en Langues, Lettres et Langage (I3L)

Mémoire de master 2 mention Sciences du langage - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2016-2017

Constitution d'un lexique verbal polylexical « fondamental » du français : repérage à partir de corpus et modélisation

Doriane Simonnet

Sous la direction d'Agnès Tutin

Laboratoire de Linguistique et Didactique des Langues Etrangères et Maternelles - LIDLEM

UFR LLASIC
Département Informatique intégrée en Langues, Lettres et Langage (I3L)

Mémoire de master 2 mention Sciences du langage - 20 crédits

Parcours : Industries de la Langue

Année universitaire 2016-2017

Remerciements

Je tiens avant tout à remercier Agnès Tutin de m'avoir donné l'opportunité de réaliser ce mémoire, de m'avoir patiemment suivie, écoutée et guidée. Merci également à Olivier Kraif pour sa collaboration et ses conseils.

Un grand merci aussi à mes collègues, Ali, Anne-Laure, Judith, Louise, Pauline et William d'avoir su créer le climat d'entre-aide, d'échange et de soutien moral mutuel dans lequel ce stage, et plus généralement ces deux années de Master, se sont déroulés.

Merci à mes proches, tout particulièrement Antoine, ma mère Christine, Philippe et Amandine pour leur soutien, leurs nombreux encouragements et leur patience sans relâche à mon égard dans les moments difficiles.

UNIVERSITÉ Grenoble Alpes

DÉCLARATION

1. Ce travail est le fruit d'un travail personnel et constitue un document original.
2. Je sais que prétendre être l'auteur d'un travail écrit par une autre personne est une pratique sévèrement sanctionnée par la loi.
3. Personne d'autre que moi n'a le droit de faire valoir ce travail, en totalité ou en partie, comme le sien.
4. Les propos repris mot à mot à d'autres auteurs figurent entre guillemets (citations).
5. Les écrits sur lesquels je m'appuie dans ce mémoire sont systématiquement référencés selon un système de renvoi bibliographique clair et précis.

NOM : Simonnet

PRENOM : Doriane

DATE : 27/06/2017

SIGNATURE : 

Sommaire

INTRODUCTION 8

Partie I - LA NOTION DE LEXIQUE FONDAMENTAL 10

CHAPITRE 1. DEFINITION 11

1. TRAITS GENERAUX 11
2. INTERETS 11
3. PROBLEMATIQUES INHERENTES A LA CREATION DES LISTES DE LEXIQUE FONDAMENTAL 12

CHAPITRE 2. HISTORIQUE 12

1. TRAVAUX PRECURSEURS 12
2. LE FRANÇAIS FONDAMENTAL DE GOUGENHEIM 13
3. LISTES DE FREQUENCE RECENTES 14

CHAPITRE 3. CRITERES DEFINISSANT L'ASPECT « FONDAMENTAL » DU LEXIQUE 17

1. FREQUENCE 17
2. DISPERSION 17
3. DISPONIBILITE 18

Partie II - LE DOMAINE DE LA PHRASEOLOGIE 19

CHAPITRE 4. CARACTERISATION DES EXPRESSIONS POLYLEXICALES 20

1. TRAITS GENERAUX 20
2. APPROCHES PHRASEOLOGIQUE ET STATISTIQUE 21
3. PROPRIETES CARACTERISANTES ET DISCRIMINANTES 22
4. CLASSIFICATIONS 25
5. LA PHRASEOLOGIE ET LE TRAITEMENT AUTOMATIQUE DES LANGUES 25

CHAPITRE 5. VERS UN LEXIQUE POLYLEXICAL VERBAL FONDAMENTAL 26

1. SPECIFICITES DES EXPRESSIONS POLYLEXICALES VERBALES 26
2. EXPRESSIONS POLYLEXICALES FONDAMENTALES ? 27

Partie III - POSSIBILITES D'EXTRACTION DES UNITES PHRASEOLOGIQUES 32

CHAPITRE 6. PRINCIPES GENERAUX ET MESURES D'ASSOCIATIONS 33

1. PRINCIPES GENERAUX DES METHODES D'EXTRACTION 33
2. MESURES D'ASSOCIATION 33

CHAPITRE 7. APPROCHES ET METHODES 34

1. APPROCHE CORPUS-BASED VS. APPROCHE CORPUS-DRIVEN 34
2. METHODES STANDARDS UTILISEES EN MATIERE D'EXTRACTION 35

Partie IV - REPERAGE DES EXPRESSIONS POLYLEXICALES DANS LE CORPUS 37

CHAPITRE 8. RESSOURCES ET DONNEES 38

1. DESCRIPTION DU CORPUS 38
2. DESCRIPTION ET EVALUATION DE LA LISTE INITIALE 42

CHAPITRE 9. PROCESSUS D'EXTRACTION 43

1. PRINCIPES GENERAUX DE LA METHODE EMPLOYEE 43
2. DEFINITION DES PATRONS CATEGORIELS 44
3. SCRIPT D'EXTRACTION 46

CHAPITRE 10. TESTS ET EVALUATIONS DE LA METHODE APPLIQUEE 49

1. EVALUATION D'UNE PREMIERE VERSION DU SCRIPT D'EXTRACTION ET RECALIBRAGE 49
2. COMPARAISON DE LA METHODE AVEC CELLE D'UNE EXTRACTION PAR ALR 56

Partie V - CONSTITUTION DE LA LISTE 62

CHAPITRE 11. SELECTION DES ITEMS 63

1. FILTRAGE DE LA LISTE SUR LE CRITERE DE LA VALIDITE DES EXPRESSIONS CANDIDATES 63
2. REDIMENSIONNEMENT DES MESURES STATISTIQUES 69
3. FILTRAGE DE LA LISTE SUR LE CRITERE DE L'ASPECT FONDAMENTAL DES EXPRESSIONS – METHODE ADOPTEE ET EVALUATIONS 71

CHAPITRE 12. ESTIMATION DE LA CAPACITE DE LA METHODE A PRODUIRE UNE LISTE D'EXPRESSIONS FONDAMENTALES 72

1. EVALUATION DE L'IMPACT DE LA SOUS-REPRESENTATION DE L'ORAL SPONTANE 73
2. EVALUATION DES SEUILS UTILISES 74
3. RELATIVISATION DE LA VALEUR INFORMATIVE DE LA MESURE DE DISPERSION UTILISEE. 76

Partie VI - DESCRIPTION DES EXPRESSIONS ET MODELISATION 79

CHAPITRE 13. ANNOTATIONS ET EXTRACTIONS DE CERTAINES CARACTERISTIQUES DES EXPRESSIONS POLYLEXICALES FONDAMENTALES 80

1. CARACTERISTIQUES ANNOTEES 80
2. CARACTERISTIQUES EXTRAITES 84

CHAPITRE 14. EXTRACTIONS COMPAREES DE LA SOUS-CATEGORISATION 88

1. REPERES THEORIQUES ET RESSOURCES EXISTANTES 88
2. EXTRACTION PAR METHODES COMPAREES 90
3. SELECTION DES INFORMATIONS DE SOUS-CATEGORISATION PERTINENTES 95
4. EXTRACTION DES POSSIBILITES DE CLITICISATION DES PRONOMS REGIS 96

CHAPITRE 15. MODELISATION 97

1. MODELISATION DE LA SOUS-CATEGORISATION A PARTIR DU MODELE DU LEFFF ET DE DICOVALENCE ET OBSERVATIONS PERMISES 97
2. PROPOSITION DE MODELISATION DE L'INTEGRALITE DES INFORMATIONS EXTRAITES 102

CONCLUSION ET PERSPECTIVES 106

INTRODUCTION

Le travail que nous proposons dans cette étude consiste en la réalisation d'un lexique « *fondamental* » d'expressions polylexicales verbales, à partir d'un repérage sur corpus. Partant de la constatation qu'une telle ressource est à ce jour inexistante pour le français, cette démarche nous semble intéressante à plus d'un titre. Un tel lexique pourrait en effet se révéler très utile, non seulement du point de vue du traitement automatique des langues, mais également pour satisfaire des besoins dans les domaines de la didactique des langues et de la psycholinguistique.

Mais, le but d'une telle entreprise ne réside pas seulement dans les possibilités qu'offre la ressource produite en termes de réutilisations postérieures. D'un point de vue théorique, elle permettra également de répondre à certaines questions linguistiques et méthodologiques. Les processus de création d'un tel lexique se déclinent en effet en une gamme de différentes possibilités théoriques et techniques, dû si bien aux aspects linguistiques de l'objet concerné qu'à la diversité des outils, techniques et approches que propose le traitement automatique des langues dans l'exploitation des corpus. Ainsi, le procédé que nous utiliserons devra être élaboré avec soin, de manière à être efficace et cohérent avec les attentes linguistiques et applicatives en matière de résultats. Sa mise en place entrainera donc en amont la nécessité d'une pré-réflexion méthodologique aiguisée. De la même façon, les approches linguistiques théoriques que nous serons amenée à prendre seront décisives. Elles devront donc être soumises à tout un questionnement qu'il s'agira d'affronter.

En outre, nous nous attendons à voir émerger de nos résultats des données permettant de répondre à des interrogations en matière d'étude de la phraséologie, ce qui constitue également un des intérêts majeurs de notre étude. Entre autres, il sera sans doute possible, d'établir une évaluation quantitative de la part de la polylexicalité dans la langue, réputée importante dans la littérature. De plus, une description de certaines caractéristiques syntaxico-sémantiques des expressions polylexicales pourra être proposée.

Les deux aspects qui définissent l'objet que nous souhaitons voir émerger, c'est-à-dire, les caractéristiques « *fondamentale* » et polylexicale, relèvent du domaine complexe de la linguistique qu'est l'étude du lexique. Galisson (1976, p.7) estime que ce dernier est « *rebelle à la structuration[, s]a nature et ses dimensions ne le prédisposent pas […] aux mises en forme systématiques.* ». Il s'agira donc d'une entreprise complexe et ce, en premier lieu, de par l'essence-même de son objet. Au-delà de cette difficulté, s'ajouteront celles plus générales de l'ambiguïté de la langue, particulièrement présente d'ailleurs dans les unités polylexicales, et de la complexité de la syntaxe, qui posent inexorablement problème dans le traitement automatique des langues.

Ce travail est tout de même, nous le pensons, nécessaire ; bien que les expressions polylexicales soient omniprésentes dans la langue, il semblerait que des données quant à leur fréquence soient absentes dans la littérature (Heid, 2008). Notre travail constituerait alors une des premières tentatives d'association de la dimension fréquentielle, inhérente à la notion de lexique « *fondamental* », à un inventaire descriptif des unités polylexicales.

Une étude réalisée par Benigno (2012) a déjà entamé une large réflexion autour de la notion de « *collocations fondamentales* », en se basant, elle aussi, sur une extraction à partir d'un corpus. Cependant, ce travail s'est limité aux collocations, et ce, à partir de dix mots pivots. Nous entendons, pour notre part, extraire un nombre très largement supérieur d'expressions de plusieurs types, c'est-à-dire en ne nous limitant pas aux collocations.

La première partie de ce mémoire proposera une description de la notion de lexique fondamental. La deuxième abordera quant à elle la question du domaine de la phraséologie, nous permettant ainsi d'exposer les problématiques liées à la caractéristique polylexicale du lexique que nous souhaitons créer. Nous ferons ensuite, dans la Partie III, un bref inventaire des possibilités qui s'offrent à nous en matière d'extraction des expressions polylexicales.

Après avoir ainsi caractérisé notre objet d'étude et recensé les alternatives proposées par le traitement automatique des langues pour son extraction à partir de corpus, nous expliquerons, dans la quatrième partie, la méthode que nous avons choisi d'utiliser pour le repérage des expressions. La partie V portera ensuite sur la constitution de notre lexique à partir des expressions repérées ; ces dernières devront être filtrées sur la base de différents critères afin de créer une liste qui corresponde au mieux à nos attentes. Différentes caractéristiques de chaque expression qui aura été retenue comme fondamentale seront alors décrites, et le résultat de ces descriptions devra être structuré par une modélisation qu'il s'agira d'établir. Nous dédierons donc la dernière partie de ce mémoire à la présentation de ce travail de description et de modélisation.

Partie I

-

LA NOTION DE LEXIQUE FONDAMENTAL

Chapitre 1. Définition

La notion de lexique fondamental s'impose comme un des éléments principaux de l'objet de notre travail. Il nous semble donc nécessaire de la définir à travers les différentes conceptions et réalisations de travaux organisées autour d'elle.

Nous commencerons, dans ce chapitre, par donner une définition générale du lexique fondamental. Puis nous présenterons les intérêts qu'il comporte. Enfin, nous ferons un bref état des lieux des problématiques inhérentes à sa création.

1. Traits généraux

Le « *vocabulaire fondamental* » est « *le noyau lexical d'une langue, le vocabulaire dont chaque locuteur dispose pour ses actes communicatifs, élémentaires et quotidiens* », « *un ensemble de mots réputés essentiels pour penser ou exprimer les concepts de base dans une langue* » (Benigno, 2012, pp.26;18). La réalisation de lexiques fondamentaux consiste, dans les grandes lignes, en la création d'une liste fréquentielle de mots, qui peuvent être sélectionnés selon différents critères que nous expliciterons.

Pour le français, l'ouvrage de référence est *Le Français Elémentaire* (1954) ou *Français fondamental* (1959), de Gougenheim, dont la liste de fréquence fut élaborée à partir de grandes enquêtes sur la langue orale.

2. Intérêts

La nécessité de construire de tels lexiques émerge initialement de besoins en enseignement des langues. Il s'agissait de produire une liste de mots essentiels qui devaient être connus des apprenants. Ils peuvent cependant servir dans plusieurs autres domaines comme le traitement automatique des langues (TAL) ou encore la psycholinguistique (voir par exemple Bonin, Chalard, Méot, & Fayol (2001)). Un lien existe d'ailleurs entre les résultats permis par ces vocabulaires en psycholinguistique acquisitionniste et en didactique ; Grossmann (2011) souligne le caractère quelque peu désuet de l'utilisation de ces listes en enseignement des langues, mais prend position quant à la nécessité de son maintien à condition d'une certaine remise à neuf. Il argumente en effet qu'elles permettent d'ajuster l'apprentissage sur l'acquisition ; il a été démontré que les mots les plus fréquents sont acquis en premier (Tomasello, (2003), cité par Grossmann), et que leur activation est plus facile dans le lexique mental[1]. Or, les mots de ces « *vocabulaires fondamentaux* », comme nous allons le voir, sont en grande partie choisis pour leur fréquence élevée dans la langue.

[1] Nous ne développerons pas ici les détails de ces concepts ; voir, par exemple, Levelt (1999)

3. Problématiques inhérentes à la création des listes de lexique fondamental

Benigno (2012, p.27), précise que plusieurs types de listes peuvent être construits :

- Des listes de mots qui se retrouvent dans tous les contextes (très fréquents)
- Des listes de mots essentiels (mots très fréquents et mots peu fréquents mais importants en termes de conceptualisation de base)
- Des listes de mots adaptées à un but précis.

Quoi qu'il en soit, établir un tel inventaire qui soit à la fois idéal et stable est impossible, car le noyau lexical subit des variations individuelles liées, d'une part, aux expériences de chacun dans son acquisition du vocabulaire et, d'autre part, aux situations de communication (*Ibid.*). On ne peut donc prétendre ni à l'exhaustivité de la liste, ni à une couverture totale des contextes et besoins langagiers (*Ibid.*).

De plus, il faut préciser qu'une telle liste sera certainement soumise à certain un vieillissement au fil des années, dû au caractère évolutif de la langue (Galisson, 1976). A titre d'exemple, nous trouvons dans la liste de Gougenheim les mots *dispensaire*, *plume* (pour l'écriture), *phono*, *livre* (unité de mesure), ou encore *bicyclette*, dont l'usage s'est largement affaibli depuis la publication de l'ouvrage.

Ces quelques constatations dénotent toute la difficulté de constitution d'un vocabulaire fondamental et soulignent son inexorable imperfection. Ceci n'a pas empêché à de nombreuses entreprises de réalisation de voir le jour, comme nous allons le montrer dans le chapitre suivant.

Chapitre 2. Historique

Comme nous l'avons souligné, les motivations premières de ces listes sont d'ordre pédagogique et leur utilité a été démontrée dans de nombreuses études liées à ce domaine. Nous nous concentrerons sur celles qui ont été réalisées pour le français. Il faut cependant noter que des démarches similaires ont été entreprises pour d'autres langues depuis les années 20 du siècle dernier[2].

1. Travaux précurseurs

La première initiative souvent citée est celle de l'abbé de l'Epée (XVIII[e] siècle), réalisée dans le cadre d'une entreprise de constitution de la LSF[3] ; il établit, dans ce but, une liste de 5 400 mots qu'il estimait indispensables pour l'éducation des enfants sourds-muets (recensé notamment dans Galisson (1976)).

[2] Pour un inventaire non exhaustif mais relativement fourni sur l'anglais et l'italien, voir Benigno (2012, p. 20-22).

[3] Langue des Signes Française

Plus proches de notre époque, des travaux visant à construire des listes de fréquence ont été entrepris dès les années 1920. On peut notamment citer *A French Word Book*, publié par V.A.C. Henmon en 1924, qui consistait en une liste de fréquence d'environ 9 000 lexies constituée à partir de 400 000 mots issus de textes littéraires (recensé dans Benigno (2012, p.19)). Un autre exemple est celui de *French Word Book* de Vander Beke, publié en 1929. Il contient plus de 6 000 éléments extraits à partir des quelques 1,1 millions de mots issus de textes de différents genres. (*Ibid.*). Bien qu'il inclue dans sa liste des mots littéraires et/ou archaïques, ce dernier travail a inspiré les auteurs du *Français fondamental* (*Ibid.*), œuvre de référence qui nous allons immédiatement présenter.

2. Le Français fondamental de Gougenheim

L'historique de réalisation et de publications de l'œuvre que nous présentons, ainsi que la liste de ses caractéristiques sont largement inspirés de ceux présentés dans Benigno (2012).

Le Français fondamental est né d'une initiative de l'U.N.E.S.C.O. qui entendait « *diffuser les grandes langues de civilisation* » (*Ibid.,* p.23). Une commission spéciale basée à L'Ecole Normale Supérieure de Saint-Cloud et dirigée par Gougenheim fut alors chargée de sa réalisation. Il a initialement été publié sous le titre *Français Elémentaire* en 1954 par le Centre National de Documentation Pédagogique du Ministère de l'Education Nationale. Une deuxième édition est publiée en 1956, intitulée *L'élaboration du français élémentaire. Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Il devient ensuite *Dictionnaire fondamental* en 1958, puis *Français Fondamental 1er degré* et *Français Fondamental 2e degré* en 1959. Enfin, une dernière édition, le *Dictionnaire fondamental,* paraît en 1971. Elle est publiée, comme son nom l'indique, sous la forme d'un dictionnaire. On notera que dans cette dernière édition, des expressions et proverbes ont été ajoutés.

Quelques caractéristiques notables du *Français Elémentaire* relevées par trois auteurs sont les suivantes :

- La langue dénotée par la liste ne se veut pas différente du français normal. (Benigno, 2012, p.5)
- L'œuvre s'adresse à un public large mais reste avant tout un outil destiné aux professeurs. (*Ibid.*)
- Dans l'approche adoptée pour la constitution de la liste, les auteurs ont préféré la qualité à la quantité (Galisson, 1976)
- La liste est constituée de 1 138 mots et a été établie à partir de la transcription d'enregistrements de 163 conversations, qui représentent un total de 321 000 occurrences pour 8 000 mots différents. Quelques-uns de ces enregistrements ont été réalisés à l'insu des locuteurs, mais cela n'était pas le cas pour la grande majorité du corpus (Gougenheim, 1964). De plus, les origines régionales, l'âge

et les niveaux socioculturels des locuteurs ne suivaient pas une distribution tout à fait équilibrée, mais ils étaient, somme toute, assez variés (*Ibid.*). Ce choix d'utiliser des enregistrements était tout à fait original dans le cadre d'un tel travail pour l'époque. Il provoquera une certaine déstabilisation des idées reçues quant aux représentations linguistiques préétablies concernant le fonctionnement de la langue parlée, alors peu étudiée (Cortier, 2006).

- Dans la dernière édition (*Dictionnaire fondamental* de 1971), le nombre des mots s'élève à 3 500 (Benigno, 2012).
- La liste comporte des mots fréquents et des mots disponibles[4]. Les seconds ont été sélectionnés sur la base d'une enquête réalisée dans plusieurs écoles, pour laquelle on demandait aux élèves quels mots étaient considérés comme utiles (Gougenheim, 1964).

Cet ouvrage a été très critiqué, notamment par Marcel Cohen qui, en 1955 dans *Français Elémentaire ? Non*, considère ce travail comme imparfait et résultant en une langue « *imprécise et réduite* » (cité dans Benigno (2012, p.26)). De plus, dans une visée d'exploitation didactique, Galisson (1976) soulève le problème de la forme sous laquelle le vocabulaire est présenté : il s'agit d'une liste triée par ordre alphabétique. Or, l'auteur estime que cette forme rend très difficile le choix du vocabulaire par rapport à une situation énonciative donnée, et le didacticien/l'enseignant qui voudrait construire un texte à partir de cette liste devra la parcourir entièrement et noter, au fur et à mesure, les mots qui pourraient convenir.

Il n'en reste pas moins que ce travail reste l'œuvre de référence en termes de « *vocabulaire fondamental* », outil didactique dont l'utilisation s'est peu à peu essoufflée dans l'enseignement des langues, mais dont l'utilité, dans ce domaine comme dans d'autres, reste d'actualité (Grossmann, 2011). Nous proposons, dans le point suivant, deux exemples de listes de fréquence récentes afin d'obtenir un aperçu la production et des buts de ces dernières à l'heure actuelle.

3. Listes de fréquence récentes

La première des listes que nous souhaitons présenter a été constituée dans le cadre du projet Lexique par New, Pallier, Ferrand, & Matos (2001). Elle est libre d'accès et comporte 130 000 entrées.

Elle a été construite à partir : 1) de textes publiés entre 1950 et 2000 issus du corpus Frantext[5] de l'ATLIF[6], 2) de pages web contenant les mots formes extraits sur Frantext . Ces

[4] voir point 3. du Chapitre 3

[5] http://www.frantext.fr/

[6] Laboratoire d'Analyse et Traitement Informatique de la Langue Française – CNRS et Université de Lorraine (http://www.atilf.fr/)

pages web ont été sélectionnées par la soumission des mots formes extraits du corpus au moteur de recherche FastSearch[7]. L'utilisation de ce dernier permet de mesurer la proportion du nombre de pages[8] où lesdits mots apparaissent et permet, donc, d'obtenir une indication quant à leur dispersion[9]. La liste ainsi créée inclut les formes fléchies des mots et comporte de nombreuses indications fréquentielles.

La Figure 1 est une capture d'écran du lexique obtenu, téléchargé au format Excel à partir de la plateforme sur laquelle il est hébergé (toutes les colonnes du tableau n'ont pas été capturées).

1	1_ortho	2_phon	3_lemme	4_cgram	5_genre	6_nombre	7_freqlemfil	8_freqlemliv	9_freqfilms2	10_freqlivre	11_infover
142146	étrusques	etRysk	étrusque	ADJ		p	0,33	0,54	0,19	0,41	
142147	étrusques	etRysk	étrusque	NOM		p	0,12	0,07	0,1	0,07	
142148	étréci	etResi	étrécir	VER	m	s	0	0,27	0	0,07	par:pas;
142149	étrécie	etResi	étrécir	VER	f	s	0	0,27	0	0,07	par:pas;
142150	étrécir	etResiR	étrécir	VER			0	0,27	0	0,07	inf;
142151	étrécisseme	etResis°m@	étrécisseme	NOM	m	s	0	0,07	0	0,07	
142152	étrécit	etResi	étrécir	VER			0	0,27	0	0,07	ind:pas:3s;
142153	étrésillon	etRezij§	étrésillon	NOM	m	s	0	0,07	0	0,07	
142154	étude	etyd	étude	NOM	f	s	44,22	63,04	11,35	19,66	
142155	études	etyd	étude	NOM	f	p	44,22	63,04	32,87	43,38	
142156	étudia	etydja	étudier	VER			70,9	30,41	0,21	1,15	ind:pas:3s;
142157	étudiai	etydjE	étudier	VER			70,9	30,41	0,02	0,34	ind:pas:1s;
142158	étudiaient	etydjE	étudier	VER			70,9	30,41	0,12	0,47	ind:imp:3p;
142159	étudiais	etydjE	étudier	VER			70,9	30,41	1,23	0,81	ind:imp:1s;ir

Figure 1: Extrait de la liste de fréquence du projet Lexique telle que téléchargeable à partir de la plateforme http://www.lexique.org

Cette liste de fréquence est aujourd'hui utilisée pour de nombreux projets dans des domaines diversifiés[10] comme la lexicographie, la didactique des langues, la psycholinguistique ou encore la phonologie[11].

Nous notons la faible variété des genres textuels exploités lors de l'élaboration de ce projet. En effet, le corpus utilisé est constitué principalement de romans et, dans une moindre mesure, de poésie, essais et écrits scientifique. La question de la juste représentativité de la langue se pose donc ici ; étant donné que la création de cette liste vise à ce qu'elle soit utilisée dans divers champs de recherche, nous pensons qu'il aurait peut-être été préférable que les sources utilisées eussent été de natures plus variées. Ceci aurait en effet permis d'obtenir des données fréquentielles plus représentatives de la langue en général.

[7] www.alltheweb.com ; moteur de recherche qui a depuis été racheté par Yahoo!

[8] sur les 14,27 millions répertoriées par FastSearch

[9] Voir point 2. du Chapitre 3

[10] Pour une liste détaillée de ces projets, consulter : http://www.lexique.org/utilisations.php

[11] Des informations phonologiques sont disponibles pour chaque mots dans les tableaux (transcription phonétique, structure syllabique, etc.)

Une seconde entreprise notable est la base de données lexicale MANULEX (Lété, Sprenger-Charolles, & Colé, 2004). Elle propose une liste de fréquence de 48 000 mots formes et 23 000 lemmes. Ces derniers ont été extraits à partir de 54 manuels scolaires de lecture utilisés à plusieurs niveaux de l'école primaire. Ce travail vise donc davantage à répertorier le type de lexique que rencontrent les élèves que de constituer une liste de fréquence sur le français en général (Grossmann, 2011, à propos de MANULEX). Comme pour le projet Lexique, la notion de dispersion est prise en compte dans les statistiques.

MANULEX s'adresse principalement aux chercheurs en psycholinguistique acquisitionniste travaillant sur l'écrit, ainsi qu'aux enseignants[12]. La Figure 2 représente un extrait de cette base de données telle que téléchargeable sur le site (encore une fois, toutes les colonnes n'ont pas été capturées). Les abréviations en têtes de colonnes signifient: le nombre de lettres (NLET), la catégorie morphosyntaxique (SYNT), les mesures de fréquence brute (F), de dispersion (D), de fréquence d'usage pour 1 million de mots (U) et de Standard Frequency Index[13] (SFI) par niveau scolaire des manuels d'origine.

LEMMES	NLET	SYNT	CP *F*	CP *D*	CP *U*	CP *SFI*	CE1 *F*	CE1 *D*	CE1 *U*	CE1 *SFI*
agenda	6	NC								
agénor	6	NP								
agenouillé	10	ADJ								
agenouiller	11	VER	1	0,00	0,50	37,03	2	0,27	1,88	42,74
agent	5	NC	14	0,54	46,80	56,70	15	0,75	33,14	55,20
agf	3	NP	3	0,00	1,57	41,95				
agglomération	13	NC					1	0,00	0,07	28,63
agglomérer	10	VER								
agglutinant	11	ADJ								
agglutiner	10	VER								

Figure 2 : Extrait de la base de données lexicale Manulex telle que téléchargeable à partir de la plateforme http://www.manulex.org/

A partir de ces deux exemples, nous pouvons établir trois constats quant aux listes de fréquence récentes. Le premier est celui du fait que leur forme informatisée permet des traitements faciles et, ainsi, de plus larges exploitations que dans le cas les listes précédemment évoquées. Le deuxième est que le nombre de mots traités est plus important que celui des listes plus anciennes (cela étant d'ailleurs aussi dû à l'informatisation des traitements). Enfin, ces listes sont « purement » des listes de fréquence. Elles consistent en effet en l'extraction de mots et au nombre de leurs occurrences dans des textes ; aucun filtre ne semble être appliqué quant à l'inclusion ou non de ces mots dans les listes (i.e. on ne choisit pas de ne retenir que les mots les plus fréquents). On s'éloigne donc de la conception de « *vocabulaire fondamental* » tel que définie par Gougenheim. Il nous semble par conséquent opportun de revenir sur la notion même de cet aspect « *fondamental* » en détaillant les critères qui peuvent être employés pour le caractériser.

[12] Voir : http://www.manulex.org/

[13] Indice de fréquence calculé par transformation logarithmique à partir de la fréquence brute.

Chapitre 3. Critères définissant l'aspect « fondamental » du lexique

Plusieurs critères permettent de définir l'aspect fondamental d'un mot. Selon Carter (1998), les mots fondamentaux doivent apparaitre comme génériques, neutres et non marqués (par exemple, ils ne doivent pas être marqués culturellement ou appartenir à un type de discours particulier). Il préconise d'utiliser un ensemble de tests faisant appel à des locuteurs natifs afin de déterminer si les mots envisagés comme « *fondamentaux* » entrent dans ces critères. Il précise également que ces tests ne peuvent se substituer aux mesures statistiques et que, une catégorisation aux limites strictes étant impossible, résonner en termes d'échelle semble plus correct.

Benigno (2012, pp.31-34) distingue trois critères qui semblent faire relativement consensus dans la littérature : la fréquence, la dispersion et la disponibilité.

1. Fréquence

La fréquence est communément utilisée dans tout travail de réalisation d'une liste fondamentale. Selon Gougenheim (1964, p.31) elle est un « *critère objectif, permettant de déterminer scientifiquement les mots les plus* usuels ». Il faut cependant préciser que les fréquences des mots dépendent fortement des données textuelles utilisées pour leurs calculs ; l'auteur prend pour exemple le mot *roi* (p. 138) et explique que sa fréquence variera en fonction :

- de la nature du texte : on aura une fréquence importante, par exemple, dans les ouvrages d'histoire
- du pays de provenance des textes : ceux qui ont sont sous un système monarchique produiront des écrits qui sont susceptibles de contenir davantage le mot *roi*
- des circonstances : par exemple, le décès d'un roi engendrera une fréquence plus importante du mot dans les textes produits à la suite de cet événement

Par conséquent, bien que nécessaire, la fréquence n'est pas considérée comme suffisante ni par cet auteur ni par de nombreux autres.

2. Dispersion

Le critère statistique de la dispersion permet de ne pas retenir comme fondamentaux des mots qui seraient très fréquents mais seulement présents dans certains types de textes, et qui n'apparaissent donc que dans des contextes limités. Une distribution très inégale de la fréquence d'un même mot dans les textes peut donc permettre de lui refuser le statut de mot fondamental (Benigno, 2012, p.34).

3. Disponibilité

Le terme de disponibilité est utilisé initialement par Gougenheim. Les mots « *disponibles* » ne sont pas nécessairement fréquents mais ils sont rattachés à des situations et thèmes de conversation quotidiens ; ils sont donc caractérisés par une importante « *utilité communicative* » (Benigno, 2012, p. 32). Il s'agit généralement de mots « *concrets* », « *usuels* » et « *utiles* » (Gougenheim, 1964, pp. 138;145). Ils « *font partie du vocabulaire actif : ils sont compris et utilisés. Ils servent moins souvent que les mots fréquents, mais ils expriment des notions si familières, qu'ils demeurent constamment à notre disposition.* » (Galisson, 1976, p.16).

Gougenheim, (1964, pp.138-139) donne l'exemple du mot *fourchette* ; bien que nous utilisions cet objet quotidiennement, nous ne prononçons son nom que rarement, sans pour autant pouvoir passer plusieurs jours ou semaines sans le faire. Il précise que vouloir identifier les mots disponibles ne revient pas à savoir les repérer (car ils sont extrêmement nombreux), mais plutôt à estimer le « *degré de disponibilité* » d'un mot donné. Il définit ce degré comme « *la présence plus ou moins immédiate de ces mots dans notre mémoire* » (p. 152).

Comme nous l'avons vu, ces trois critères ne sont pas toujours employés dans la création de vocabulaires fondamentaux. Cela est surtout vrai pour la disponibilité. Il nous a tout de même paru utile de les lister afin de cerner un peu mieux la complexité et les procédés de création qui caractérisent la notion de lexique « fondamental ».

Ayant défini, dans cette première partie, la notion de vocabulaire fondamental, il convient à présent de présenter les caractéristiques du deuxième aspect de notre objet d'étude, à savoir la nature polylexicale du lexique que nous proposons de construire.

Partie II

-

LE DOMAINE DE LA PHRASEOLOGIE

Chapitre 4. Caractérisation des expressions polylexicales

Les expressions polylexicales sont un phénomène particulier et très intéressant pour la linguistique. Elles posent en outre de nombreux problèmes en TAL. Nous proposons dans ce chapitre une définition de ces objets linguistiques, à travers les traits généraux qui les caractérisent, les approches qui ont été utilisées pour les décrire, l'inventaire de leurs propriétés, les classifications qui les répertorient et enfin leur mise en perspective avec le TAL.

1. Traits généraux

Le domaine de la phraséologie renvoie à l'étude des expressions polylexicales, c'est-à-dire aux séquences composées de plusieurs mots graphiques formant des unités (*pomme de terre*, *mettre la charrue avant les bœufs*, *faire allusion*, *à bientôt*, etc.). Il peut être défini comme « *l'étude des structures, des sens et de l'utilisation des combinaisons de mots*»[14] (Cowie, 1988, cité dans Granger, Paquot *et al.*, 2008).

Dès le début du XX^{ème} siècle, la question des expressions polylexicales est soulevée par Charles Bally (1951)[15], qui en fait une véritable discipline. L'auteur parle alors d'« *unités lexicologiques* » ou « *unités phraséologiques* », qu'il caractérise comme des groupes de mots qui, par force de répétition dans l'usage, finissent par acquérir un « *caractère* usuel » et à former des « *unités indissolubles* » (p.66).

Les objets linguistiques considérés dans ce domaine seront indifféremment désignés, dans le cadre de ce mémoire, sous les termes « *unités phraséologiques* », « *unités polylexicales* » ou « *expressions polylexicales* », bien que des nuances existent entre ces différents termes (Tutin, 2010b). Ils constituent en outre un groupe très hétérogène :

> *« there is no unified phenomenon to describe, but rather a complex of features that interact in various, often untidy ways and represent a broad continuum between non-compositional (or idiomatic) and compositional groups of words. » (Moon, 1998, p.6)*

Il peut en effet s'agir de combinaisons très différentes les unes des autres, puisque leurs caractéristiques formelles, syntaxiques et sémantiques divergent largement entre les sous-classes de combinaisons qu'englobe la notion d'*expressions polylexicales*. D'autre part, aucun système de sous-classification ne fait l'unanimité dans la littérature.

[14] Notre traduction

[15] La date donnée ici correspond à un celle d'une réédition posthume de l'ouvrage, la première datant de 1909

Ces unités phraséologiques représentent un phénomène omniprésent dans la langue ; Jackendoff (1995) estime en effet que le lexique accessible aux locuteurs dans les situations de la vie quotidienne comporte au moins autant d'expressions polylexicales que de lexies simples. Elles sont encore plus nombreuses dans les langages spécialisés (Heid, 2008). Cependant, une quantification empirique précise de ce phénomène dans la langue générale vient à manquer dans la littérature (*Ibid.*).

Les caractéristiques que nous venons de donner sont relativement générales, et nous avons brièvement montré que le phénomène est assez complexe et diversifié. Il est, en outre, abordé différemment par deux approches que l'on oppose traditionnellement.

2. Approches phraséologique et statistique

Les travaux en matière de phraséologie se sont inscrits dans deux courants principaux que nous allons détailler ci-après. Il s'agit de l'approche phraséologique et de l'approche statistique.

Nous pouvons noter l'existence d'une troisième approche, la tradition anthropologique. Principalement soutenue par Veronika Teliya, elle vise à « *étudier différentes manières et formes d'interaction entre la culture et le langage résultant en la formation de phraséologismes comme incarnations et transmetteurs de l'information culturelle de générations en générations* »[16] (Zykova, 2013). Cette approche restant assez marginale et peu exploitée, nous n'en développerons pas les caractéristiques. Il nous a tout de même semblé que son existence méritait d'être mentionnée.

2.1. Approche phraséologique

Fondée sur la tradition classique russe, l'approche phraséologique, se base sur l'idée d'un continuum allant des unités les plus figées aux plus transparentes[17]. Elle s'est développée entre les années 40 et 60 du siècle dernier. La préoccupation principale des linguistes adhérant à cette approche est d'établir des critères linguistiques pour distinguer les différents types d'unités phraséologiques et, plus spécifiquement, de distinguer les unités transparentes[18] et variables des combinaisons libres (i.e. qui ne sont pas phraséologiques) (Granger et al., 2008). Cette tradition sera suivie par de nombreux auteurs dont Cowie et Mel'čuk.

On doit principalement à cette approche le fait d'avoir instauré la phraséologie comme une discipline à part entière, mais aussi d'avoir su créer une terminologie et un

[16] Notre traduction

[17] Voir point 3. du présent chapitre

[18] *Ibid.*

inventaire des critères nécessaires à la catégorisation et à l'analyse des unités polylexicales (*Ibid*.).

2.2. Approche statistique

Fondée sur la tradition anglo-saxonne, l'approche statistique a pour représentants principaux J.R. Firth, à qui l'on doit la célèbre phrase « *you shall know a word by the company it keeps* » (Firth, 1957), M. Halliday et J. Sinclair. Cette tradition accorde une importance toute particulière au corpus, considéré comme la source de repérage des expressions polylexicales à partir de critères statistiques et non linguistiques (contrairement à la tradition phraséologique) (Granger et al., 2008).

Les unités polylexicales, telles qu'envisagées par ce courant, englobent si bien des termes associés sur le plan paradigmatique, comme *médecin* et *hôpital*, que des associations syntagmatiques comme *argument de poids* (Grossmann & Tutin, 2002). Les premières sont considérées comme périphériques ou en dehors du domaine de la phraséologie par l'approche phraséologique, mais font bel et bien partie intégrante de la classe d'objets étudiés par l'approche statistique (Granger et al., 2008).

Ce courant est notamment critiqué par Gaatone (1997, p.168, cité dans Granger *et al*., 2008), qui s'oppose à la tendance « *qui consiste à exagérer l'importance des expressions figées et à poser que "tout est phraséologique"* ».

Les approches phraséologique et statistique divergent donc en termes de définition du concept adopté. De plus, comme le souligne Benigno (2012), elles ont chacune leurs limites. En effet, la première inclut le risque de passer à côté d'associations assez libres mais représentatives de l'usage linguistique, tandis que la deuxième peut restreindre le repérage d'associations peu fréquentes mais pertinentes. Il est en outre intéressant de noter que Granger et al. (2008) préconisent un rapprochement entre les deux approches dans le but de faire avancer le domaine de la phraséologie.

Quoi qu'il en soit, les auteurs qui traitent de ce domaine utilisent certaines propriétés des expressions polylexicales afin de les étudier. Il nous a paru utile de dresser dans le point suivant une liste non-exhaustive de ces propriétés.

3. Propriétés caractérisantes et discriminantes

Afin de décrire les unités polylexicales, de les classer et de les distinguer des associations libres, les spécialistes examinent leurs propriétés linguistiques. Nous proposons ici un petit inventaire de celles qui sont utilisées à cet escient. Cette étape vise à permettre une meilleure clarté des termes utilisés par la suite, et nous nous limiterons donc aux notions

dont nous nous servirons dans le cadre de cette étude. Nous distinguerons trois groupes de propriétés: formelles, sémantiques et syntaxiques.

3.1. Propriétés formelles

La **polylexicalité** est considérée comme la condition nécessaire à la caractérisation des objets observés comme étant des unités phraséologiques. Elle consiste en le fait qu'ils soient composés de plusieurs « *mots* » (Granger et al., 2008). Mel'čuk (2013, p.1) parle d'énoncés « *multilexémiques* » qu'il définit comme « *une configuration de deux ou plus lexèmes syntaxiquement liés* ». Il pose cet aspect comme une des deux conditions initiales nécessaires à la prise en considération d'un objet linguistique comme unité polylexicale (ou « *phrasème* » selon sa terminologie). Il est intéressant de noter que si cette propriété semble triviale et évidente, la difficile définition de ce qu'est un « *mot* » vient la complexifier. Ceci est d'autant plus significatif que l'instabilité graphique de certaines expressions fait qu'elles peuvent être transcrites sous plusieurs formes, dont certaines sont composées de plusieurs « *mots* » et d'autres non ; Granger *et al.* (2008, p.8) donnent pour exemple les différentes graphies anglaises *good will*, *good-will* et *goodwill*.

Le **caractère binaire** désigne quant à lui le fait qu'une expression soit composée de deux éléments. Chez Mel'čuk, ces deux éléments sont majoritairement deux lexies, tandis que Grossmann & Tutin (2002) préfèrent les envisager comme constituants (qui peuvent donc consister en plusieurs mots), afin de pouvoir inclure des expressions syntagmatiques, comme *fort comme un turc*. Cette caractéristique, nous y reviendrons, ne s'applique pas nécessairement à toutes les expressions polylexicales.

3.2. Propriétés sémantiques

Le recours à la sémantique est essentiel pour effectuer une distinction entre les différents types d'associations lexicales. De nombreuses propriétés sont utilisées pour cela. Nous présentons ici les plus significatives d'entre elles.

La première est la **non-compositionnalité**. Elle désigne le fait que la somme des sens des composants d'une association ne permette pas de déduire le sens de cette association. Par exemple, le sens de l'expression *prendre un râteau* ne peut être déduit par addition des sens de *prendre* et de *râteau*. Mel'čuk (2013) précise que cette notion ne doit pas être confondue avec celle de transparence/opacité. La **transparence** désigne le fait qu'une expression puisse être comprise par un locuteur sans qu'il ne la connaisse. Les associations qualifiées de transparentes ont donc la caractéristique de porter un sens qui soit déductible, comme dans l'expression *célibataire endurci* (exemple emprunté à Grossmann & Tutin, (2002)). L'**opacité**, par opposition, se réfère aux cas où une telle déduction est impossible, comme, par exemple, dans le cas de *poser un lapin* dont le sens semble difficilement déductible, hors contexte, par un locuteur non natif.

La **dissymétrie des composants** (terme emprunté à Grossmann & Tutin (2002)) désigne le fait que, dans le cas de les collocations[19], un des éléments de la séquence conserve son sens habituel.

Enfin, la dernière notion que nous abordons ici est celle de **figement**. Nous la présentons dans cette partie, mais notons que certains de ses aspects sont syntaxiques plus que sémantiques. Cette propriété peut être considérée comme l'une des plus importantes dans le travail de description et de classification des unités phraséologiques. Elle est utilisée pour 1) distinguer les expressions polylexicales des associations libres, 2) établir des classifications entre les différentes expressions polylexicales retenues. Gross (1996) établit une liste des critères nécessaires à la description de ce figement ; il évoque notamment l'opacité sémantique, le blocage des propriétés syntaxiques transformationnelles ou encore le blocage des paradigmes synonymiques (i.e. l'impossibilité de remplacer un mot par un de ses synonymes).

3.3. Propriétés syntaxiques

La **flexibilité syntaxique** désigne le fait que les alternances syntaxiques (passivation, relativisation, etc.) soient possibles ou non pour chaque expression examinée**.** Ainsi, l'expression *mordre la poussière* ne peut être passivisée (**la poussière a été mordue*) (exemple emprunté à Tutin (2010b)) ; cela est en revanche possible pour *tourner la page*[20] (*la page est tournée*)[21].

La **non-insertion de modifieurs** est une caractéristique de certaines expressions qui refusent des modifieurs, comme dans l'expression *mariage blanc/*mariage très blanc*.

Enfin, l'**aspect atypique des constructions** désigne la particularité qu'ont certaines unités polylexicales d'avoir une structure syntaxique qui s'éloigne de la langue standard comme, par exemple, l'absence de déterminant dans les constructions à verbes support[22] comme *avoir faim* (Grossmann & Tutin, 2002).

Les différentes caractéristiques que nous venons de présenter ont été utilisées entre autres pour établir des classifications. Nous allons maintenant faire un bref point sur ces dernières.

---

[19] *Ibid.*

[20] Sens figuré (*passer à autre chose*)

[21] Par exemple *: « Le peuple voudrait croire **que cette page est** maintenant définitivement **tournée**.* », extrait par nous-même à partir de l'outil Lexicoscope (Kraif, 2016) sur le corpus frWack (Baroni, Bernardini, Ferraresi, & Picci, 2010)

[22] Voir 1.2. du Chapitre 5

4. Classifications

Les classifications sont nombreuses dans la littérature. Elles peuvent donc être considérées comme une richesse, mais ont l'inconvénient de nuire quelque peu à la communication entre les linguistes et de créer un sentiment de flou dans le domaine ; les terminologies divergent et se recoupent parfois en qualifiant du même terme deux éléments différents (Granger et al., 2008). Les différences entre ces classifications sont largement dues aux différents critères utilisés dans la distinction des catégories (*Ibid.*). En outre, elles sont habituellement établies selon des buts précis : lexicologie/lexicographie, pédagogie, psycholinguistique ou encore TAL (*Ibid.*).

Malgré les nombreuses divergences terminologiques et taxonomiques qui divisent les auteurs, quelques grandes classes sont généralement présentes dans chacune de leurs travaux :

- Le point de départ de chacune de ces classifications est la distinction entre les associations libres et les expressions polylexicales, qui reçoivent des dénominations différentes selon l'auteur ; on parlera notamment de « *phrasèmes* » chez Mel'čuk (1995;1998;2013).
- Les expressions à fonction pragmatique (*how do you do ?, de rien*) forment souvent une catégorie particulière ; on trouve, par exemple, les « *formules* » chez Cowie (1998;1988) et les « *pragmatèmes* » chez Mel'čuk (1995;1998;2013).
- Une autre catégorie récurrente est celle des proverbes et des locutions.
- Les phrases idiomatiques non-compositionnelles comme *poser un lapin* en français ou *spill the bean* pour l'anglais forment généralement une catégorie à part (voir par exemple Cowie, 1998; Lo Cascio, 2000; Mel'čuk, 2013).
- Enfin, les collocations constituent la dernière grande catégorie commune à de nombreuses classifications (par exemple, Nattinger & DeCarrico, 1992; Lo Cascio, 2000; Grossmann & Tutin, 2002; Mel'čuk, 2013) ; elles sont communément caractérisées par une structure base/collocatif, dans la laquelle la base impose une restriction de sélection lexicale sur le collocatif. Par exemple, dans l'expression *rendre des comptes*, la base (*comptes*) impose le choix du collocatif verbal (*rendre*).

Au-delà du fait de donner lieu à des divergences théoriques, la complexité du phénomène polylexical a des répercussions pratiques, notamment dans le TAL.

5. La phraséologie et le traitement automatique des langues

Selon la formule restée célèbre, Sag, *et al.* (2002) qualifient les unités phraséologiques de « *pain in the neck for NLP* ». L'implémentation de la phraséologie au

TAL n'est pourtant pas récente : dès les années 1960, le logiciel de traduction automatique SYSTRAN contenait un lexique d'expressions polylexicales (fait rapporté par Heid (2008)).

La phraséologie ne constitue pas un sous domaine ni une certaine approche du TAL ; elle est plutôt un élément qui tend à être utilisé par les différentes branches et applications de la discipline. Heid (2008, p.338) explique :

> *« This is because phraseological units need to be identified, described, classified, represented and manipulated with respect to all levels of linguistic description, such as morphology and morphosyntax, syntax, semantics and, last but not least, with respect to contrastivity between languages. »*

Si son emploi est varié, c'est que la prise en compte de la phraséologie se révèle d'une grande utilité ; l'auteur sus-cité va jusqu'à affirmer que la prise en compte des phénomènes phraséologiques est la condition *sine qua non* à toute réalisation d'application TAL à grande échelle. Il précise également que deux orientations principales se dessinent quant à l'utilisation de la phraséologie en TAL : la modélisation informatique de phénomènes linguistiques et la réalisation d'applications diverses aux besoins variés (parsing syntaxique, génération automatique de texte, etc.). En outre, il souligne le fait que, dans le domaine du TAL, les problèmes de classification prennent tout leur sens ; les jeux d'étiquettes de catégories en fonction des modèles adoptés et des besoins des applications sont variés.

A la suite de cette description de ce que sont les expressions polylexicales, nous allons à présent tenter de dresser un portrait de l'objet plus précis que nous souhaitons traiter dans le cadre de ce travail.

Chapitre 5. Vers un lexique polylexical verbal fondamental

Les unités phraséologiques que nous souhaitons inclure dans le cadre de notre étude auront deux caractéristiques principales ; elles devront être verbales et *fondamentales*. Si les spécificités des expressions ayant la première caractéristique ont maintes fois été traitées par le passé, l'objet formé par des expressions qui auraient la caractéristique d'être fondamentales a rarement été théorisé.

1. Spécificités des expressions polylexicales verbales

Nous allons traiter dans notre travail d'expressions polylexicales verbales, i.e. contenant un verbe. Derrière cette définition générale, il convient de rappeler que se cachent

des types d'expressions différentes (quelle que soit la classification utilisée), figées et non-figées, et des phénomènes variés inhérents aux expressions.

1.1. Phénomènes d'alternance

Au vu des éléments que nous avons jusqu'à présent décrits à propos des unités phraséologiques, une constatation nous semble évidente : des phénomènes d'alternances syntaxiques comme la passivation ou la relativisation vont parfois altérer la forme des constructions dans lesquelles ces expressions sont incluses et, potentiellement, modifier l'ordre linéaire d'apparition de ses composants. Par exemple, dans *prendre une décision/une décision a été prise*, les deux composants n'apparaissent pas dans le même ordre. Cette constatation est d'autant plus intéressante et problématique qu'elle ne peut pas s'appliquer à toutes les unités verbales.

Ce phénomène est typique des expressions polylexicales verbales et méritait d'être souligné.

1.2. Collocations à verbes support

Les collocations à verbe support, terme que l'on doit à Maurice Gross (1981), sont des constructions particulières. Elles sont considérées dans la littérature comme des structures singulières. Comme l'indique la terminologie du phénomène, les verbes inclus dans ces constructions ont pour fonction de servir de « *support* » syntaxique au nom qui va, lui, exprimer le prédicat sémantique (par exemple, Alonso Ramos, 1999). Par exemple, dans la séquence *prendre une décision*, le verbe sert uniquement à fournir des informations sur le mode, le temps, l'aspect et la personne, tandis que le nom exprime le sens de l'expression (exemple emprunté à Benigno, 2012).

Si les deux caractéristiques de la classe des expressions verbales sont assez faciles à remarquer pour peu que l'on s'intéresse à la littérature portant sur la phraséologie, la définition de ce qu'est une expression fondamentale est bien moins aisée à établir.

2. Expressions polylexicales fondamentales ?

Les notions d'expressions polylexicales et de lexique fondamental semblent rarement avoir été unies au sein d'un même objet d'étude. Nous présenterons dans ce point une entreprise relativement ancienne qui s'inscrit dans cette optique, et une, plus récente. Cette dernière constitue, à notre connaissance, la seule tentative actuelle de définition d'un objet à la fois polylexical et fondamental.

2.1. Inventaire thématique et syntagmatique du français fondamental *: Initiative de construction de syntagmes à partir du vocabulaire fondamental du Gougenheim*

Dans *Inventaire thématique et syntagmatique du français fondamental,* publié en 1976, R. Galisson (1976) expose son travail de regroupement des mots du *Français fondamental* en « *syntagmes usuels* ». Si le terme « *syntagme* » est employé, les objets concernés sont ceux communément désignés aujourd'hui comme « *collocations* » (et ce dernier terme est d'ailleurs également utilisé par l'auteur, quoique plus rarement)[23].

Pour former ces syntagmes, Galisson n'a retenu que les « *lexèmes de désignation* », à savoir les substantifs, verbes et adjectifs. Il a tout d'abord procédé à un classement des substantifs par thèmes. Par exemple, *ferme, champ, cheval, tracteur, etc.* ont été regroupés sous l'étiquette *AGRICULTURE*. Ensuite, afin de passer de l'échelle lexématique à celle syntagmatique, l'auteur a fait défiler tous les verbes et adjectifs avant et après chaque substantif. Ici, il précise que : « *Des affinités se sont manifestées, des connexions se sont établies, des enchaînements se sont créés* » (p.13). Les critères de tri des syntagmes ainsi produits ont été, dans les mots de ce dernier, les « *lieux commun* », les « *banalités de la vie quotidienne* » et le « *sentiment linguistique* » de ses collègues (p.14) ; il juge ces critères « *raisonnables, mais approximatifs* » (p.14). Ceci constitue le principal point faible de ce travail. L'auteur le reconnaît volontiers en admettant le caractère « *trop artisanal* » de cette méthode (p.19). Les syntagmes alors construits consistent en un terme stable et un terme commutable, également qualifié de « *collocatif* » (p.19).

Notons que, comme l'indique l'auteur, les expressions figées telles que les proverbes ou dictons ont été exclues. Les raisons invoquées sont leur manque de productivité et leur appartenance à un niveau avancé de connaissance de la langue (le travail de Galisson s'inscrit dans une logique pédagogique d'enseignement des langues). De plus, les collocatifs dits « *illimités* »[24] n'ont pas été inclus à tous les syntagmes dont ils pouvaient faire partie et ce, par souci de clarté (p.19).

De plus, la volonté de l'auteur était de regrouper en syntagmes les mots du *Français fondamental*, et seulement les mots du *Français fondamental*. La conséquence de ce choix est que certains mots ont dû être exclus car aucun des autres lexèmes de la liste ne permettaient, par association, de construire de syntagme usuel.

En outre, un problème énoncé par l'auteur est la polysémanticité des mots de la liste de Gougenheim ; en effet, aucune définition n'est associée aux mots est on ne sait pas, par

[23] Cette petite confusion terminologique est due au fait que, si le terme de « *collocations* » était, à l'époque de la publication de l'ouvrage, parvenu en France depuis l'école britannique, son emploi a mis un certain temps à s'imposer dans la communauté scientifique, qui lui préférait donc le terme de « *syntagme* » malgré son ambiguïté (Hausmann & Blumenthal, 2006).

[24] Verbes comme *avoir, faire, acheter,* etc., ou adjectifs comme *petit, beau,* etc.

exemple, si *bois* fait référence à la matière, au bois de chauffage ou à une petite forêt. Or, cela a fortement compliqué sa tâche de classification thématique.

L'aboutissement de ce travail consiste en des syntagmes classés par thèmes et présentés de façon schématique (des exemples sont reproduits en Figures 3). Il constitue une démarche tout à fait notable en ce sens 1) qu'il propose, dès 1976, une définition de la collocation en termes de lexème-noyau/lexème-satellite et 2) qu'elle se base sur le vocabulaire fondamental dans une volonté de le structurer sous forme d'expressions polylexicales.

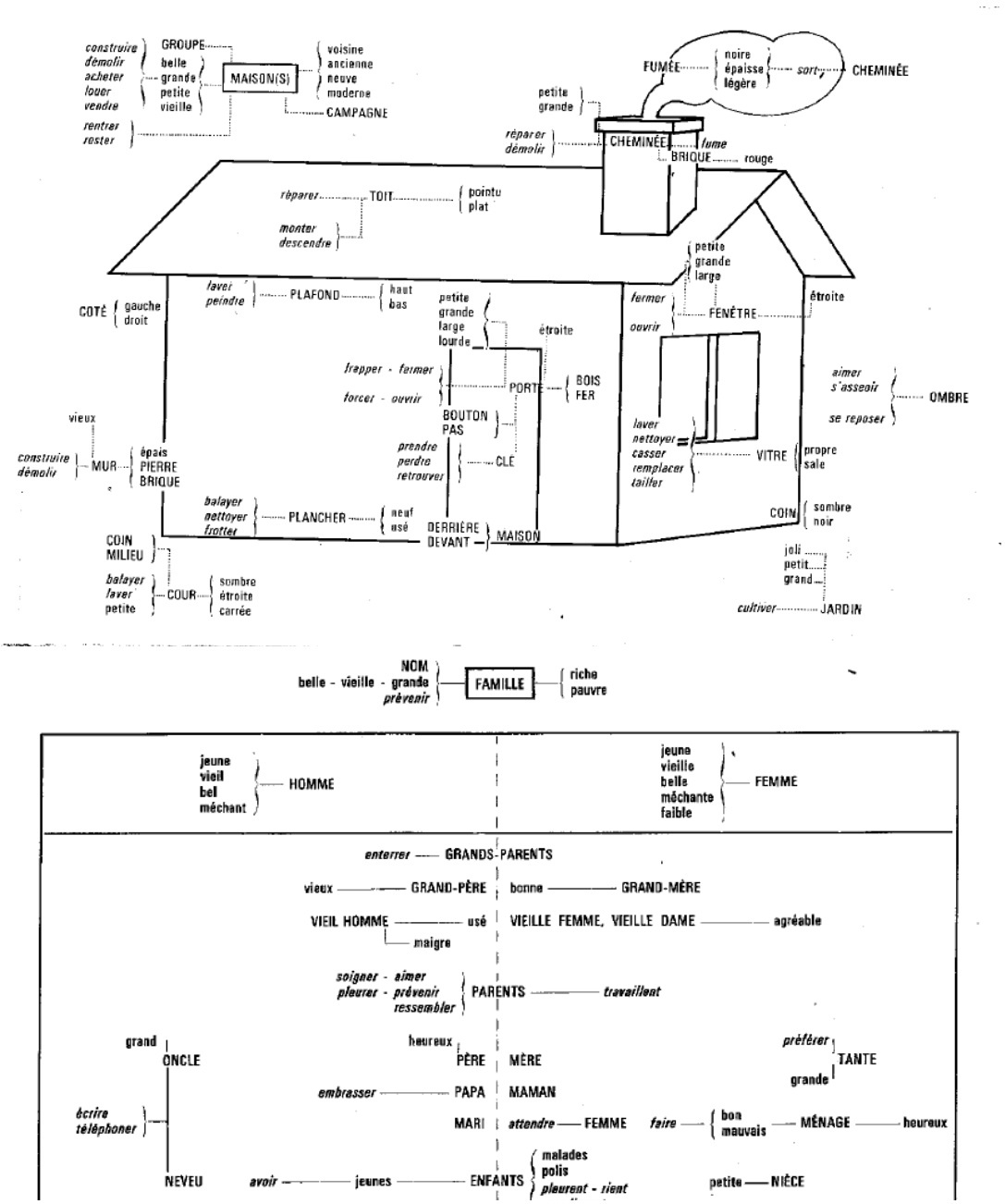


Figure 3: Exemples de syntagmes autour du thème de la maison puis de la famille construits par Galisson, tels que présentés dans l'ouvrage *Inventaire thématique et syntagmatique du français fondamental*

2.2. La notion de collocation fondamentale. Etude de corpus en vue d'une exploitation didactique, *thèse doctorale de Veronica Benigno : un travail de définition et d'extraction*

La thèse doctorale de Veronica Benigno (2012) nous semble encore plus intéressante dans le cadre de notre étude. L'auteur a en effet mené une réflexion théorique et méthodologique sur les « *collocations fondamentales* » ; il s'agissait tout d'abord de définir cette notion et d'élargir ainsi « *la notion de "vocabulaire fondamental" à la dimension syntagmatique* » (p.3). Puis, il lui a fallu développer une méthode d'extraction sur corpus pour constituer un échantillon d'associations fondamentales permettant une analyse. Cette étude, comme son titre l'indique, se base dans une optique de didactique des langues. Dans la lignée de Gougenheim, l'auteur estime que la sélection des collocations fondamentales doit s'effectuer à partir des critères de fréquence, dispersion et disponibilité.

Benigno se sert de dix mots pivots représentant, selon le *Thésaurus Larousse* de Péchoin (1991), des « *événements sociaux* » (p.87). Ces mots sont tous des substantifs et sont appelés à constituer la « *tête* » de la collocation. Ils sont répertoriés comme fondamentaux par Gougenheim (1971). Cette caractéristique est apparue essentielle à l'auteur, motivée par l'idée qu'une collocation constituée de deux mots non fondamentaux ne peut aboutir elle-même à une collocation fondamentale.

Le corpus est frWaC de Baroni et al. (2010), qui est constitué de textes issus du web, a été utilisé pour les besoins de cette étude. Sur ce dernier, les expressions candidates ont été extraites à partir des mots pivots grâce à des scripts Perl développés par Olivier Kraif (2011). Ils fournissaient, pour chaque association (mot pivot)-(mot_x) rencontrée, un ensemble de mesures statistiques de fréquences et de mesures d'association[25].

Les fréquences ont permis de dresser une première partie de la liste des candidats. Autres de ces candidats ont été repérés grâce à des mesures d'association, qui permettent d'extraire des associations peu fréquentes mais dont la force d'association est importante, (comme par exemple, *nez aquilin*) et qui sont donc considérées comme pertinentes. Cette sélection des unités polylexicales sur la base de leur pertinence (et pas simplement sur la fréquence) est considérée par l'auteur comme une façon d'inclure des collocations potentiellement *disponibles* dans sa liste. La dispersion a été prise en compte en fonction des domaines internet de provenance des unités repérées.

Les résultats obtenus ont ensuite été nettoyés, puis une sélection des unités candidates a été réalisée. Ces unités ont été soumises à des tests auprès de 90 locuteurs natifs chargés de juger du caractère fondamental des collocations proposées. L'analyse des résultats de ces tests ont : 1) confirmé que la fréquence n'était pas la seule condition au caractère

[25] Voir Chapitre 6

fondamental, 2) montré que les locuteurs tendent à considérer comme fondamentales les expressions les plus figées, même si moins fréquentes.

Les deux travaux que nous venons de présenter ont donc ce point commun de représenter une tentative d'union des deux concepts de base de notre travail. Le premier le fait dans une visée de construction d'un lexique structuré, le deuxième dans un but d'analyse et de définition d'un objet linguistique, sans dimension de construction d'un lexique large. Nous avons décidé, pour notre part, de tenter de combiner ces deux aspects dans notre travail.

En outre, les conceptions de l'objet traité divergent également selon les auteurs ; pour Galisson, les deux éléments d'une collocation doivent être fondamentaux, tandis que pour Benigno, seule la base doit l'être. Nous avons quant à nous choisi de partir de la supposition selon laquelle une expression peut-être fondamentale sans que ses composants ne le soient.

Nous avons ainsi défini notre objet d'étude. Un des objectifs de notre travail va alors consister à être capable d'extraire ses occurrences à partir d'un corpus. Pour cela, comme nous allons l'expliquer dans la partie suivante, plusieurs alternatives s'offraient à nous.

Partie III

-

POSSIBILITES D'EXTRACTION DES UNITES PHRASEOLOGIQUES

Chapitre 6. Principes généraux et mesures d'associations

Les techniques d'extraction sont variées mais reposent globalement sur certains principes généraux et, souvent, sur l'utilisation de mesures statistiques spécifiques.

1. Principes généraux des méthodes d'extraction

Les corpus utilisés pour extraire les expressions polylexicales peuvent être de différentes natures : texte brut, corpus lemmatisé et étiqueté, texte chunké ou encore corpus analysé syntaxiquement (Heid, 2008). Nombre des outils utilisés exploitent la récurrence des groupes de mots et leur fréquence de cooccurrence afin d'effectuer la tâche de repérage (*Ibid.*). De manière générale, ils ont deux tâches à effectuer (parfois réalisées par le même outil) : identifier les expressions candidates selon des critères prédéfinis et les classer selon une mesure d'association donnée (Seretan, 2011, p.31).

2. Mesures d'association

2.1. Utilité

Les mesures d'association permettent d'évaluer la force d'association qu'il existe entre les mots en neutralisant, dans une certaine mesure, les effets aléatoires (Evert, 2005 ; Heid, 2008). Leurs valeurs peuvent être comprises comme une quantification du degré de figement d'une unité polylexicale (*Ibid.*). Elles consistent en « *une formule qui calcule un score d'association à partir des informations de fréquence dans un tableau de contingence de pairs de mots* »[26] (Evert, 2005, p.75). Comme le précise Benigno (2012, p.71), elles sont notamment utilisées, dans les approches *corpus-based*[27] pour relativiser les erreurs obtenues dans le calcul des fréquences.

2.2. Fonctionnement général

De manière générale, une mesure d'association est obtenue par le procédé suivant :

- un décompte du nombre de cooccurrences de deux mots dans un corpus est effectué
- puis, ce nombre est comparé avec celui des autres contextes de chacun des deux mots ainsi qu'avec le nombre total des paires de mots du corpus dont les catégories sont similaires (par exemple, tous les couplets noms-verbes en cooccurrence dans le corpus si c'est d'un couplet de ce genre dont on veut obtenir la mesure d'association).

Ces mesures permettent l'obtention de deux résultats : le classement, i.e. le fait de placer les paires ayant le plus haut score en tête de liste, et la sélection (de par le fait qu'elles

[26] Notre traduction

[27] Voir point 1. du Chapitre 7

permettent d'établir un seuil nécessaire à l'acceptation des paires candidates) (Evert, 2005). Certaines de ces mesures se concentrent sur l'un ou l'autre de ces résultats, tandis que d'autres essayent de fournir un certain équilibre entre les deux (*Ibid.*).

S'il existe plusieurs mesures associatives, Evert & Krenn (2001, cités par Heid, 2008) estiment qu'aucune d'elles ne supplante les autres en matière d'utilité et que le choix de l'emploi de l'une ou de l'autre dépend du type d'unités polylexicales que l'on souhaite capturer. De plus, comme le souligne Seretan (2011, p.30), ces mesures sont, à certains égards, limitées ; elles ne peuvent évaluer qu'une relation binaire entre deux mots (or certaines expressions en comportent davantage), et elles permettent d'évaluer le caractère arbitraire mais non la prévisibilité des expressions. L'auteur précise également que leur utilisation n'est pas absolument nécessaire, argumentant que la simple fréquence de cooccurrence se révèle souvent tout aussi fiable (p.44).

Les processus d'extraction d'unités polylexicales, qui reposent sur les principes généraux que nous venons de décrire, se déclinent en plusieurs approches et méthodes.

Chapitre 7. Approches et méthodes

1. Approche corpus-based vs. approche corpus-driven

L'extraction d'unités phraséologiques sous-entend, cela tombe sous le sens, l'utilisation de corpus. De manière générale en linguistique computationnelle, cette dernière est abordée selon deux logiques différentes. Il s'agit de l'opposition entre approche « *corpus-based* » et approche « *corpus-driven* ».

L'approche *corpus-based* se sert du corpus comme d'un fond de données linguistiques à partir duquel on va extraire des exemplaires de phénomènes pour confirmer des hypothèses théoriques (Storjohann, 2005). Le corpus est donc considéré comme un support complémentaire (*Ibid.*). Dans cette approche, le but premier est d'analyser les usages et variations de patterns systématiques de formes ou de structures linguistiques prédéfinies (Biber, 2012).

L'approche *corpus-driven*, quant à elle, considère le corpus comme une source empirique dont les données vont permettre aux linguistes de détecter des phénomènes sans suppositions ni attentes antérieures (Storjohann, 2005, se référant à Tognini-Bonelli, 2001). Elle consiste donc à « *prendre les données présentes dans les corpus comme unique point de départ pour la formulation d'hypothèses et de conclusion*s » (Dupont, 2014, p.1). Elle va notamment permettre de constater l'existence de constructions linguistiques que les modèles théoriques omettent, et, inversement, de démontrer la rareté d'occurrence de certains mots

dans certains patterns bien que ces constructions soient possibles sur le plan théorique (Biber, 2012).

On a donc une dichotomie entre deux approches ; l'une se sert des corpus pour analyser des concepts et constructions linguistiques préétablis à travers leurs variations, l'autre prend appui sur le corpus, sans modèles prédéfinis, pour voir apparaître ces concepts et ces constructions.

Appliquées à l'extraction de la phraséologie, les approches *corpus-based* vont souvent consister en l'extraction des expressions polylexicales à partir de patterns structurels et/ou lexicaux préétablis. Nous citerons comme exemple Arnaud, Ferragne, Lewis, & Maniez (2008) qui, dans le but d'étudier la lexicalisation des séquences Adj+N en anglais, vont extraire à partir de leur corpus[28] les occurrences de ces structures commençant par 31 adjectifs différents prédéfinis. Les approches *corpus-driven* vont quant à elles extraire les unités phraséologiques selon des critères statistiques.

Ces lignes directrices donnent lieu à différentes méthodes dont nous allons à présent donner quelques exemples.

2. Méthodes standards utilisées en matière d'extraction

Le standard en matière d'extraction consiste à utiliser un corpus taggué et lemmatisé (voire, analysé syntaxiquement) à partir duquel on réalise une sélection basée sur des patterns de catégories morphosyntaxiques, suivie d'un filtrage des résultats par mesures d'association (Seretan, 2011). L'utilisation de telles méthodes peut produire en sortie des associations qui ne sont pas des collocations au sens où l'entendent de nombreux spécialistes en ce que l'on peut obtenir, par exemple, des séquences tout à fait compositionnelles et prédictibles comme *new results* (Heid, 2008). Il en résulte donc qu'un filtrage manuel s'avère souvent nécessaire si les résultats de l'extraction sont destinés à être exploités dans le cadre d'études lexicographiques. Heid estime donc que les phénomènes fréquentiels capturés par les mesures d'association sont « *nécessaires mais pas suffisants* » pour de tels travaux (*Ibid.*, p. 352).

Les déclinaisons de ce standard sont nombreuses et d'autres alternatives sont proposées. Par exemple, Seretan (2011) a extrait des collocations binaires grâce aux relations entre les mots d'un corpus analysé en dépendance. Elle a utilisé pour cela toutes les relations présentes dans le corpus, en ajoutant quelques restrictions morphosyntaxiques (par exemple, si la tête de la relation est un nom, il ne doit pas être un nom propre). Ce procédé lui a permis d'unir deux collocations en une plus complexe. Les candidats ainsi extraits ont été classés selon une mesure d'association (Log Likelihood Ratio) après avoir été regroupés en classes syntaxiques, ce qui, selon Evert & Krenn, (2001), améliore les résultats obtenus par

[28] British National Corpus

classement selon des mesures d'association. Dans ce travail, Seretan n'a pas utilisé de seuil de fréquence.

Une autre méthode, utilisée par Joseph (2013), a consisté l'extraction d'expressions polylexicales verbales par l'emploi de règles transformationnelles. Dans un premier temps, après un étiquetage et une lemmatisation du corpus, les candidats ont été extraits à l'aide de structures syntaxiques. Ont ainsi été sélectionnées les expressions polylexicales ayant comme structures : [V N]/[V DET N]/[V PREP N]/[V PREP DET N]. Dans un deuxième temps, les candidats ont été soumis à des tests de transformation morphologiques, syntagmatiques et paradigmatiques implémentées à l'aide d'expressions régulières. Les candidats ayant réussi ces tests ont ensuite été filtrés par seuils de fréquence et de figement.

Enfin, dans le cadre de son mémoire de Master, Corman (2012) a utilisé un corpus arboré analysé en dépendances. La première étape du traitement a été de considérer initialement tout sous-arbre d'un arbre enraciné, d'une ampleur de 2 à 6 nœuds, comme candidat potentiel. Les sous-arbres récurrents ont ensuite été identifiés et un filtrage a été effectué pour choisir les arbres à retenir. Ce filtrage a consisté en : 1) l'élimination des expressions incomplètes, 2) l'établissement d'un seuil minimum d'occurrences, 3) un recours au critère de dispersion, 4) l'utilisation de mesures d'association.

Ces différents exemples fournissent un bon aperçu des possibilités en matière d'extraction. D'autres méthodes, comme l'approche distributionnelle[29] ou l'approche multilingue[30] sont également utilisées, mais elles sont incompatibles avec la nature de notre tâche, et nous ne développerons donc pas leur fonctionnement.

Notre choix s'est porté sur une méthode basée sur des patrons catégoriels dont nous allons expliquer le fonctionnement dans la partie suivante.

[29] Approche se basant sur l'hypothèse selon laquelle les éléments lexicaux ayant des contextes similaires partagent des composants sémantiques identiques (Heid, 2008). Elle n'utilise généralement pas de contraintes morphosyntaxiques, ce qui n'est pas compatible avec notre volonté d'extraire des expressions comportant spécifiquement un verbe.

[30] Méthode qui utilise des corpus bilingue alignés

Partie IV

-

REPERAGE DES EXPRESSIONS POLYLEXICALES DANS LE CORPUS

Chapitre 8. Ressources et données

Pour notre travail nous disposions de deux ressources initiales : un corpus de genres diversifiés et une liste d'expressions polylexicales, compilée par Agnès Tutin dans le cadre du projet AIM-WEST[31], que nous désignerons dans la suite par le terme « *liste initiale* » . Une approche que nous pouvions donc envisager aurait été de projeter sur le corpus les expressions répertoriées dans la liste initiale afin d'en extraire les mesures statistiques et les propriétés morphosyntaxiques. Dans cette optique, il nous fallait tout d'abord évaluer l'utilité de cette liste dans le cadre de notre travail.

Nous décrivons, premièrement, dans ce chapitre le corpus que nous avons utilisé, puis la liste initiale et son évaluation.

1. Description du corpus

Nous avons utilisé le corpus Contraste, initialement constitué pour la thèse de Sylvain Hatier (2016). Incluant de plusieurs genres linguistiques dans l'optique de représenter la langue générale, il a été créé dans le but d'être comparé à un corpus d'écrits scientifiques.

1.1. Composition

Le corpus comporte environ 117 millions de mots et se divise en quatre parties : romans, écrits journalistiques, transcriptions d'oral et sous-titres de films.

Le **sous-corpus littéraire** est constitué de 330 romans de 85 auteurs différents et est issu du projet Emolex[32]. Comme le détaille le Tableau 1, ces romans appartiennent à différents genres littéraires et sont, pour moitié, des traductions. La moyenne du nombre de mots par roman est d'environ 118 00, mais ce nombre varie fortement, allant de 542 000 à 731.

Genres	SF	Fiction historique	Policier	Autre	TOTAL
Nombre de romans	99	16	126	89	330
Traduits	83	10	48	25	166
Non Traduits	16	6	78	64	164
Nombre de mots	12 807 361	3 216 181	14 017 613	9 078 762	39 119 917

Tableau 1 : Composition du sous-corpus littéraire

Le **sous-corpus journalistique** est quant à lui constitué d'articles de journaux quotidiens régionaux et nationaux (*Le Monde*, *Libération*, *Le Figaro*, *Ouest France* et *Sud Ouest*) parus en 2008. Il provient également du projet Emolex.

[31] http://aim-west.imag.fr/

[32] www.emolex.eu

Les **transcriptions de français parlé** proviennent de trois sources différentes :

- Le corpus ESTER[33] (Galliano et al., 2005) qui est constitué d'une centaine d'heures de transcription d'émissions de radio (France Inter, France Info, Radio France International, Radio Télévision Marocaine, France Culture et Radio Classique) transcrites manuellement. Son but premier est l'évaluation de systèmes de reconnaissance vocale et il comporte de la parole conventionnelle (séquences d'émissions préparées) et conversationnelle (débats).
- Le corpus CFPP[34] (Branca-Rosoff, Fleury, Lefeuvre, & Pires, 2009) qui est composé de transcriptions automatiques d'interviews réalisées dans plusieurs quartiers parisiens et auprès de locuteurs d'origines socio-culturelles variées.
- Le corpus TCOF[35] (Benzitoun, Fort, & Sagot, 2012) qui propose des transcriptions de d'interactions adultes/enfants et adultes/adultes s'inscrivant dans plusieurs types de discours (réunions de travail, consultations médicales, entretiens, etc.)

Comme le montre le Tableau 2, le corpus ESTER représente la partie la plus importante du sous-corpus des transcriptions, avec 1 millions de mots, tandis que CFPP et TCOF comportent respectivement 370 000 et 340 000 mots.

Fichiers	Nombre de mots
cfpp.utf8.emolex.xml	372 039
ester.utf8.emolex.xml	1 022 873
tcof.all.utf8.emolex.xml	341 244
TOTAL	1 736 156

Tableau 2 : Composition du sous-corpus de parole transcrite

Enfin, le dernier sous-corpus est constitué de **sous-titres de films** (traduits ou non) issus du Projet OPUS (Tiedemann & Nygaard, 2004). Dû à l'absence de métadonnées quant à l'origine des sous-titres et aux œuvres présentes dans le corpus, nous ne disposons pas de ces informations.

Afin d'obtenir des mesures de dispersion significatives, nous avions besoin d'un découpage en fichiers qui soit équilibré[36]. Nous avons donc effectué la segmentation

[33] http://www.technolangue.net/article60.html

[34] http://cfpp2000.univ-paris3.fr/

[35] http://www.cnrtl.fr/corpus/tcof/

[36] Le découpage en fichiers ne l'était pas initialement

présentée ci-dessous dans le Tableau 3[37]. Cette segmentation a été effectuée automatiquement et les numéros d'identifiants des phrases ont été redéfinis de manière à ce que, à partir de 0, ils se suivent de deux en deux sur l'intégralité du corpus ; pour la suite du traitement, cela va permettre de définir le bloc de provenance d'une phrase à partir de son identifiant.

	Sous-corpus	Nombre de mots
FICTION	Science fiction	12 807 361
	Policier	14 017 613
	Fiction historique et Autre	12 294 943
JOURNALISTIQUE	*Le Figaro , Libération*	12 644 073
	Le Monde + Première moitié de *Ouest France*	11 181 797
	Deuxième moitié de *Ouest France* et *Sud Est*	14 612 035
ORAL	Premier tiers des sous-titres	12 466 977
	Deuxième tiers des sous-titres	12 842 062
	Troisième tiers des sous-titres et transcriptions	14 320 375

Tableau 3 : Segmentation du corpus

1.2. Problématiques linguistiques

Deux caractéristiques du corpus peuvent susciter des interrogations quant à son utilisation pour les besoins de notre étude qui vise à extraire des phénomènes relatifs à une langue générale et standard. Premièrement, il s'agit en partie de textes traduits, et, deuxièmement, sa partie orale est majoritairement composée de sous-titres et non de parole conversationnelle spontanée.

1.2.1. Question de la traduction

Comme nous l'avons évoqué, la moitié des romans et une partie des sous-titres de films sont des traductions. Dans le cadre de notre travail, il nous a paru nécessaire de nous interroger quant au degré de représentativité de la langue que comporte ce type de textes.

En effet, de nombreux auteurs soutiennent l'idée selon laquelle un texte traduit est caractérisé par des phénomènes qui en font une production langagière d'une nature différente de celle d'un texte écrit en langue source (voir par exemple Baker *et al.* (1993), Gellerstam (1986) et Laviosa (2002), cités par Kraif (à paraître)). Certains phénomènes semblent être typiques du texte traduit, et cela concerne notamment les collocations qui sont parfois calquées sur la langue source (Duchet, Kraif, & Castillo, 2008).

Cependant, une étude textométrique réalisée sur les romans policiers de notre sous-corpus de fiction par Kraif (à paraître) a montré qu'il n'y avait pas de différence notable

[37] Dans la pratique, afin d'obtenir un traitement des données plus rapides lors de l'extraction, chaque bloc a été divisé en 60 fichiers de tailles équivalentes, mais comme nous l'expliquerons par la suite, ce sous-découpage ne sera pas pris en compte dans le calcul des mesures de dispersion

entre les romans traduits et non traduits en termes de diversité lexicale ou d'homogénéité lexico-syntaxique. De plus, d'un point de vue phraséologique, cette étude a démontré que la traduction donnait lieu à une langue plus normalisée, dû principalement à la très faible fréquence des expressions familières, argotiques ou dialectales (*Ibid*.). Ce dernier point nous porte à penser que, dans l'optique d'un travail visant à constituer un lexique fondamental, des textes traduits peuvent tout à fait être utilisés.

1.2.2. Question de l'oral

Notre corpus oral est constitué en majeure partie de sous-titres de films. Nous avons donc ici affaire à une langue qui, en plus d'être souvent traduite, est également plus standardisée que l'oral spontané de par les contraintes qu'impose l'exercice du sous-titrage (Hatim & Mason, 2005, pp.65-66). L'oral *stricto sensu* est donc sous-représenté. De plus, le corpus ESTER comporte des enregistrements d'émissions de radio, et le CFPP contient des transcriptions d'entretiens qui ne relèvent pas de situations de discours réellement spontanées. Cette caractéristique est donc problématique dans l'optique de notre travail et il s'agira de mesurer son impact sur les résultats[38].

1.3. Traitements et forme

Pour les besoins des travaux d'Hatier (2016), le corpus a été annoté (catégories syntaxiques, lemmes, traits morphosyntaxiques) et analysé en dépendance par XIP (Aït-Mokhtar, Chanod, & Roux, 2002). Comme le rappelle Hatier (p.92), XIP réalise tout d'abord une phase de segmentation, étiquetage morphologique et catégorisation en partie du discours. Puis, il effectue une analyse syntaxique de surface qui fournit une annotation des relations syntaxiques et entités nommées. Enfin, il réalise une analyse syntaxique profonde (Hagège & Tannier, 2007).

L'analyse syntaxique de XIP a été affinée par un post-traitement, afin de la rendre plus adaptée aux traitements ultérieurs dans le cadre de la thèse d'Hatier ; certaines relations ont été regroupées (Baroni & Lenci, 2010) et des dépendances (indiquées par le préfixe *U3_*) ont été ajoutées: coordination, sujet profond, coréférence, etc.

Les fichiers XML résultants indiquent, pour chaque phrase, la liste de ses tokens avec, pour chacun, son id, sa catégorie syntaxique, son lemme, ses traits morpho-syntaxiques et ses relations de dépendances avec d'autres tokens. Un exemple de phrase du corpus est consultable en Annexe 1.

[38] Voir le point 3.2. du Chapitre 11

2. Description et évaluation de la liste initiale

La liste initiale dont nous disposions comporte 8308 entrées sous forme d'expressions lemmatisées extraites à partir de Wikitionnaire[39] et/ou du Dictionnaire Electronique des Mots[40] (DEM) (Dubois & Dubois-Charlier, 2011). Pour chacune d'entre elles, plusieurs informations étaient indiquées (sens, type (collocation ou expression figée),etc.). Afin d'estimer dans quelle mesure il était possible d'utiliser cette liste pour projeter les expressions qu'elle contenait sur le corpus, il nous fallait évaluer son taux de couverture des expressions polylexicales fréquentes.

Dans cette optique, nous avons tout d'abord sélectionné 20 verbes, dont certains sont très productifs et d'autres plus rares[41]. Pour chacun, nous avons extrait les expressions polylexicales de la liste initiale qui le contenait. Puis, grâce à l'outils Lexicoscope (Kraif, 2016), nous avons extrait sur un corpus littéraire et un corpus journalistique les cooccurrences du verbe et de ses objets directs (cette relation est la plus fréquente des constructions verbales)[42]. Ensuite, nous avons comparé, pour chaque verbe, la liste obtenue par Lexicoscope à celle obtenue par extraction sur la liste initiale.

Les résultats de cette comparaison montraient que seulement 38,64% des expressions extraites sur Lexicoscope étaient présentes dans la liste initiale. De plus, nous avons constaté que les collocations étaient très peu présentes dans cette liste, tandis que les expressions figées l'étaient davantage (respectivement, 3,36% et 55,49% de présence). Il n'était donc pas envisageable de l'utiliser pour projeter les expressions qu'elle comportait sur le corpus.

Les informations contenues dans cette liste ne nous ont pas pour autant été inutiles. Elle nous a en effet permis de lister les structures possibles des expressions polylexicales françaises, utilisées dans notre processus d'extraction présenté au chapitre suivant.

[39] https://fr.wiktionary.org/wiki/Wiktionnaire:Page_d%E2%80%99accueil

[40] http://rali.iro.umontreal.ca/rali/dem/

[41] *prêter, tenir, tirer, battre, risquer, rendre, casser, dire, passer, jeter, mettre, avoir, porter, avaler, faire, prendre, donner, fermer, craindre, garder.*

[42] Un filtrage manuel de la liste initiale a montré que plus de la moitié des expressions qui y figuraient sont formées autour de cette structure Verbe-Objet direct, ce qui tend à confirmer cette affirmation

Chapitre 9. Processus d'extraction

Dans ce chapitre, nous présentons tout d'abord les principes généraux de notre méthode d'extraction. Puis, nous détaillons le processus mis en place pour déterminer les patrons catégoriels que nous avons utilisés. Enfin, nous expliquons les grandes lignes du fonctionnement de notre script d'extraction.

1. Principes généraux de la méthode employée

Nous avons décidé de concentrer notre travail sur les expressions polylexicales construites autour d'une relation VERBE-OBJET DIRECT NOMINAL, qui est la structure verbale la plus productive. Cette focalisation de notre travail sur un sous-type précis d'expressions a pour conséquence de produire une extraction, certes, bien moins diversifiée et complète que si nous avions décidé de prendre en compte toutes les structures d'expressions contenant un verbe, mais certainement aussi plus précise et détaillée. Nous avons ainsi préféré axer notre travail davantage sur une réflexion méthodologique et théorique que sur la volonté de produire une ressource quantitativement importante, ce que permettait le choix de nous limiter à une seule structure principale.

Afin de réaliser l'extraction de ces expressions à partir de notre corpus, nous avons choisi d'utiliser une méthode principalement basée sur des patrons de suites catégorielles. Ce choix de méthode n'exclut pas l'exploitation occasionnelle d'informations syntaxiques contenues dans les relations de dépendance. Nous avons en effet pris le parti de considérer que toutes les informations que comporte le corpus sont utiles à certaines tâches. Aussi, nous avons estimé que la bonne réussite de notre travail résiderait en partie dans la sélection des informations les plus pertinentes pour chacune de ces tâches.

Les éléments suivants ont motivé notre choix d'utiliser des patrons catégoriels :

- L'étiquetage morphosyntaxique est moins susceptible de contenir des erreurs que l'analyse syntaxique.
- Nous disposions de modèles de patrons (dans la liste initiale).
- Le problème majeur que peut poser une méthode par patrons réside dans la possible discontinuité des unités lexicales et dans la modification de leur ordre linéaire (dû aux alternances syntaxiques). Si l'objectif de la tâche est d'extraire l'intégralité des expressions et de leurs occurrences, ces éléments peuvent constituer des freins importants. Cependant, tel n'était pas le cas dans le cadre de notre travail ; il s'agissait d'extraire des expressions hautement fréquentes et nous estimions que ces cas de figure pouvaient être omis de nos décomptes sans que cela n'altère les résultats de manière significative. De plus, comme nous l'expliquerons, certaines structures avec des insertions typiques ont été intégrées à nos patrons, ce qui a permis d'atténuer le nombre d'omissions dues à ce problème.

D'autre part, nous avons choisi de ne pas utiliser de mesure d'association. La première raison à cela est que nous ne souhaitions pas nous limiter aux expressions binaires, ce qu'imposent généralement ces mesures. Les conclusions de Benigno (2012, pp.147-150) indiquent, en outre, que sur les quatre mesures utilisées dans le cadre de son travail, une seule a été réellement utile dans le filtrage des expressions candidates. De plus, cette dernière a eu en partie recours à ces mesures afin de repérer les collocations peu fréquentes mais « *pertinentes* » (potentiellement disponibles) (p.121). Une enquête auprès de locuteurs natifs a été nécessaire pour confirmer ou infirmer cette supposition sur chaque collocation extraite, ce qui nous semble être une démarche tout à fait intéressante. Cependant, ce travail s'est limité à dix mots pivots et le nombre des expressions était donc restreint. Il n'était pas envisageable pour nous, dans la limite du temps de réalisation de ce travail, de mener une enquête qui aurait dû être assez large pour permettre de tester de nombreuses expressions auprès de nombreux locuteurs. Aussi, nous avons décidé que les expressions peu fréquentes mais disponibles ne feraient pas partie des objets que nous allions traiter. Par conséquent, l'utilisation de mesures d'association pour repérer des expressions peu fréquentes mais éventuellement disponibles était donc inutile, dû au choix de ne pas entreprendre cette démarche.

Ces principes et choix généraux ayant été explicités, il convient désormais de détailler la méthode que nous avons mise en place, en commençant par la définition des patrons catégoriels utilisés.

2. Définition des patrons catégoriels

La liste des patrons catégoriels que nous avons utilisés a été établie à partir de notre liste initiale et résulte de deux étapes de traitement.

2.1. Annotation

Nous avons commencé par extraire de la liste initiale les différentes structures catégorielles des expressions construites autour d'une relation entre un verbe et un complément d'objet direct nominal. Pour cela, nous avons annoté manuellement les suites de catégories qui composaient chacune de ces expressions. Par exemple, pour *serrer la vis*, nous annotions [V DET N].

Le but de cette annotation étant de projeter ces patrons sur le corpus, nous avons cherché à anticiper les erreurs qui pourraient être provoquées par un mauvais étiquetage de certaines unités lexicales ambigües. C'est notamment le cas de *du/des* et *de la/de l'* qui peuvent être des déterminants mais aussi des structures PREP+DET. Nous avons donc choisi de ne pas indiquer leur catégorie dans notre annotation, mais plutôt leur forme. Ainsi, nous

avons annoté par [V **de DET(la|l')** N PREP **des** N][43] l'expression *donner **de la** confiture à **des** cochons*. Nous avons ainsi patrons différentes composés d'étiquettes de catégories syntaxiques et de quelques formes spécifiques.

2.2. Structuration

Nous avons ensuite séparé, dans chacune des suites extraites, ce que nous appellerons les « *noyaux* » et les « *éléments périphériques* ». Les premiers correspondent au verbe et à son COD réduit au nom et, éventuellement, à son déterminant. Les seconds sont les suites catégorielles qui peuvent être 1) antéposées au noyau, 2) postposées au noyau, 3) insérées entre le verbe et le GN du noyau. Par exemple, le patron [V DET N PREP N] a été décomposé ainsi : $[V\ DET\ N]_{NOYAU}\ [PREP\ N]_{ELEMENT\_PERIPHERIQUE.}$ Il nous faut préciser qu'il là s'agit d'une représentation des structures purement surfacique ; les relations syntaxiques entre le nom ou le verbe et les éléments périphériques n'influent pas sur la description de la structure. Par exemple, le patron que nous venons de citer décrit si bien l'expression *enterrer sa vie de garçon* que *changer son fusil d'épaule*, même si l'élément périphérique [PREP N] n'a pas la même fonction syntaxique dans les deux cas.

Nous avons ainsi obtenu 6 noyaux différents :

- [V N]
- [V DET N]
- [V du N]
- [V des N]
- [V de DET(la|l') N]
- [V DET DET N][44]

Les éléments périphériques étaient, quant à eux, au nombre de 89.

Les combinaisons possibles entre noyaux et éléments périphériques sont présentés dans l'Annexe 2 qui donne, pour chaque combinaison, un exemple d'expression de la liste initiale. L'Annexe 3, quant à lui, présente les nombres d'expressions contenues dans cette liste pour chaque combinaison possible.

Nous disposions, à l'issue de cette étape, d'assez d'éléments pour réaliser notre script d'extraction dont nous allons décrire le fonctionnement dans le chapitre suivant.

---

[43] Le choix de transformer *de la* en *de DET(la|l')* va permettre de repérer les cas où cette suite aura été étiquetée à tort *PREP DET*. Les occurrences correctement étiquetées (i.e. où *de la* sera identifié comme un déterminant partitif composé de deux éléments) seront repérés grâce au patron V DET N.

[44] DET DET décrit principalement les déterminants du type *les NUM*, comme dans *dire **ses quatres** vérités à* quelqu'un, ou celles du type *tous les* (qui est normalement étiqueté comme un déterminant unique dans le corpus, mais le décrire ainsi permet d'extraire les occurrences où cela ne serait pas le cas)

3. Script d'extraction

Le script d'extraction, dont le pseudo-algorithme est présenté en Annexe 4, a été réalisé en Python avec l'aide du module lxml pour parser les fichiers du corpus. Il prend en entrée le corpus ainsi que des seuils fréquentiels et donne en sortie une liste d'expressions candidates accompagnées de plusieurs informations. Son fonctionnement peut être résumé, dans les grandes lignes, à un procédé en trois étapes.

3.1. Etape de repérage

La première étape est celle du repérage des expressions candidates et du décompte statistique de leur fréquence et de leur dispersion. Pour chaque phrase du corpus, on repère tous les couplets d'éléments liés par une relation OBJ et :

- dont le dépendant est un nom
- dont la tête est un verbe qui n'est pas un verbe d'état (*être, paraître, apparaître, sembler, devenir, demeurer, sembler, rester*) car, dans le corpus, l'attribut du sujet est relié au verbe par la même relation OBJ qui relie un verbe d'action à son objet direct. Un filtre de ce type est donc nécessaire pour ne garder que les objets directs.
- qui ne sont pas précédés de *y* si le verbe est *avoir,* car *il y a* n'est pas traité comme un objet spécifique dans le corpus.

Chaque couplet V NObj trouvé est stocké en mémoire. Puis, l'expression est lemmatisée sous forme *verbe_VERB nom_NOUN* ou *verbe_NOUN DET nom_NOUN* s'il contient un déterminant. Par exemple, pour l'occurrence *j'ai eu peur*, on stockera *avoir_VERB peur_NOUN* comme couplet et comme expression lemmatisée ; pour *il a pris sa décision*, le couplet sera *prendre_VERB décision_NOUN* et l'expression lemmatisée sera *prendre_VERB DET décision_NOUN*. On remarque ici que le lemme du déterminant est supprimé. Cette opération de généralisation a pour but de ne pas créer de sorties différentes pour une même expression qui permettrait une variation de cet élément[45].

Ensuite une recherche des éléments périphériques est effectuée. Pour les suites postposées, le script va chercher si les mots qui le noyau forment des suites catégorielles qui se trouvent dans notre liste de patrons. On recherche d'abord des suites de 7 éléments, si aucune n'est trouvée, on cherche des suites de 6 éléments, etc. Lorsqu'une suite valide est ainsi repérée, les lemmes qui la composent sont concaténés au noyau de l'expression lemmatisée. Pour l'antéposition, on concatène le mot précédent si celui-ci est : un pronom réflexif, *ne, pas, plus, y* et *en* (sauf si le verbe termine par *–ant*). Il ne s'agit donc pas d'une recherche par catégories mais par lemmes. Ces derniers sont ceux qui sont les plus fréquents en antéposition dans la liste initiale. Un tel choix permet de limiter le bruit dans les

[45] Le lemme du déterminant est tout de même stocké en mémoire afin d'apparaître en sortie dans la liste des déterminants possibles de l'expression

expressions candidates en restreignant 1) les adverbes antéposés à la négation, 2) les pronoms à deux clitiques également fréquents ou à un pronom réflexif. Enfin, les éléments périphériques insérés entre le verbe et l'objet direct sont concaténés dans l'expression lemmatisée s'il s'agit d'un adjectif ou de l'adverbe *pas*.

Les expressions lemmatisées ainsi repérées sont stockées en mémoire avec les informations suivantes :

- La liste des différents déterminants du noyau, regroupés par classes (définis, indéfinis, possessifs, démonstratifs et autres).
- La liste des phrases contenant l'expression ainsi que leur identifiants.
- La fréquence et la dispersion de l'expression.
- La fréquence et la dispersion du couplet V NObj sur lequel est basée l'expression. Ces mesures incluent celles de toutes les expressions formées autour du couplet. Imaginons par exemple que les expressions *rendre visite*, *rendre DET visite* et *rendre DET visite de courtoisie* aient été repérées avec des fréquences respectives de 100, 110 et 20 ; la fréquence du couplet *rendre_VERB visite_NOUN* sera donc égale à 230[46]. Imaginons ensuite que ces trois expressions aient été repérées, respectivement, dans les sous-corpus [1,5,7,8],[2,5,7] et [3] ; elles sont donc éparpillées dans 6 sous-corpus différents, ce qui sera la valeur de dispersion du couplet.

3.2. Etape de redimensionnement des expressions longues à basse fréquence

La seconde étape consiste en un redimensionnement des expressions longues à basse fréquence. La liste des expressions repérées lors de l'étape précédente contient en effet beaucoup de suites qui ne sont pas des expressions polylexicales mais qui peuvent en contenir une. Par exemple, étant donné qu'un de nos patrons est [V N]$_{\text{NOYAU}}$ [PREP N]$_{\text{ELEMENT\_PERIPHERIQUE}}$, si le corpus contient le syntagme *a besoin de Paul*, ce dernier a été stocké comme expression candidate. Or, ce n'est pas une expression polylexicale, mais *avoir besoin* en est bel et bien une. Le problème est donc que cette occurrence de l'expression *avoir besoin* ne sera pas prise en compte dans le décompte de sa fréquence, mais dans celui de l'expression candidate *avoir besoin de Paul*. Pour remédier à ce problème, on cherche dans la liste toutes les expressions dont la fréquence est inférieure ou égale à 15 et on supprime par itérations le dernier mot de la suite jusqu'à ce que l'expression obtenue soit déjà présente dans la liste. Lorsque l'on arrive à une expression déjà présente dans le dictionnaire, on fusionne les informations de l'expression longue[47] à celle de l'expression plus courte, puis on supprime l'entrée de l'expression longue.

[46] On imagine que seules ces trois expressions aient été repérées pour ce couplet

[47] Mesures statistiques, liste des déterminants et liste des phrases contenant l'expression

	Entrée *avoir besoin*	**Entrée *avoir besoin de Paul***
Couplet	*avoir_VERB besoin_NOUN*	*avoir_VERB besoin_NOUN*
Fréquence du couplet	3622	3622
Sous-corpus contenant le couplet	1,2,3,5,6,7 → dispersion du couplet : 6	1,2,3,5,6,7 → dispersion du couplet : 6
Expression lemmatisée	*avoir_VERB besoin_NOUN*	*avoir_VERB besoin_NOUN de_PREP Paul_NOUN*
Déterminants	NULL	NULL
Fréquence de l'expression	3000	3
Sous-corpus contenant l'expression	1,2,5,6 → dispersion de l'expression : 4	2,3 → dispersion de l'expression : 2

Tableau 4 : Exemple de deux entrées dans la liste des extractions avant le redimensionnement des expressions à basse fréquence

	Entrée *avoir besoin*	**Entrée *avoir besoin de Paul***
Couplet	*avoir_VERB besoin_NOUN*	*avoir_VERB besoin_NOUN*
Fréquence du couplet	3622	3622
Sous-corpus contenant le couplet	1,2,3,5,6,7 → dispersion du couplet : 6	1,2,3,5,6,7 → dispersion du couplet : 6
Expression lemmatisée	*avoir_VERB besoin_NOUN*	*avoir_VERB besoin_NOUN de_PREP Paul_NOUN*
Déterminants	NULL	NULL
Fréquence de l'expression	**3003**	3
Sous-corpus contenant l'expression	1,2,5,6,**3** → dispersion de l'expression : **5**	2,3 → dispersion de l'expression : 2

Tableau 5 : Exemple de deux entrées dans la liste des extractions après le redimensionnement des expressions à basse fréquence

Pour reprendre notre exemple, soient, dans la liste des expressions candidates, les deux entrées schématisées dans le Tableau 4[48] :

[48] La liste des phrases contenant les expressions a été omise pour plus de lisibilité

- L'expression *avoir besoin de Paul* a une fréquence inférieure à 15.
- La première itération de suppression du dernier mot produit l'expression *avoir besoin de*. Cette dernière ne correspond à aucune entrée de notre liste, car [PREP]$_{\text{ELEMENT\_PERIPHERIQUE}}$ n'est pas un patron postposée potentiel.
- La deuxième itération produit *avoir besoin*. Cette expression est présente dans la liste.
- Le processus de redimensionnement crée alors la restructuration des données schématisée par le Tableau 5.

3.3. Etape de création de la sortie en fonction des seuils fréquentiels

Enfin, la dernière étape consiste en la création de la sortie en fonction des seuils fréquentiels donnés en entrée du script (fréquence minimum du couplet, fréquence minimum de l'expression, dispersion minimum du couplet et dispersion minimum de l'expression). La sortie prend la forme d'un fichier .csv contenant les expressions candidates dont les mesures statistiques sont supérieures aux seuils souhaités, accompagnées de leurs informations.

La méthode ayant été choisie et le script réalisé, il nous a ensuite fallu évaluer leur efficacité.

Chapitre 10. Tests et évaluations de la méthode appliquée

Nous avons réalisé trois évaluations du script et, plus généralement, de la méthode choisie. Premièrement, nous avons évalué les résultats d'une première extraction en nous attardant sur l'impact de la fréquence et de la dispersion sur les résultats, puis sur la qualité des sorties consistant en des expressions longues, et enfin, sur une observation des phénomènes linguistiques qui compliquaient la tâche d'évaluation. Suite à cela, nous avons comparé les résultats permis par notre script à ceux fournis par une méthode syntaxique.

1. Evaluation d'une première version du script d'extraction et recalibrage

Nous avons réalisé une extraction grâce à une première version opérationnelle du script. L'objectif de cette étape était d'évaluer les résultats de l'extraction en repérant le taux d'expressions candidates valides parmi celles fournies en sortie. Un premier intérêt de cette évaluation était d'ajuster les paramètres du script pour l'extraction finale en repérant ceux qui amélioraient ou, au contraire, affectaient la qualité des résultats. Un deuxième intérêt était de pouvoir repérer les phénomènes linguistiques inhérents aux expressions extraites, afin d'entamer une réflexion sur les critères à utiliser par la suite pour déterminer si une expression candidate était valide ou non.

1.1. Brève description de l'extraction réalisée

L'extraction a été réalisée avec un seuil minimum de fréquence de 30 et une dispersion minimum de 10 (cette première version du script calculait la dispersion sur la base du nombre de fichiers (540) et non du nombre de sous-corpus (9)). Ces seuils bas permettaient d'obtenir un large échantillonnage d'expressions candidates afin de mieux repérer les phénomènes que nous souhaitions observer. Ce premier script effectuait en outre un redimensionnement des expressions longues ayant une fréquence inférieure à 5 (et non à 15 comme indiqué dans le chapitre précédent). Enfin, les patrons d'éléments périphériques étaient plus nombreux que ceux décrits précédemment[49]. Nous avons ainsi obtenu en sortie 8429 expressions candidates et en avons évalué manuellement 3110.

1.2. Méthode d'évaluation

Nous avons évalué la validité des expressions candidates fournies par notre script d'extraction. Nous entendons par « *validité* » le fait qu'une expression candidate soit une expression polylexicale et non une association libre.

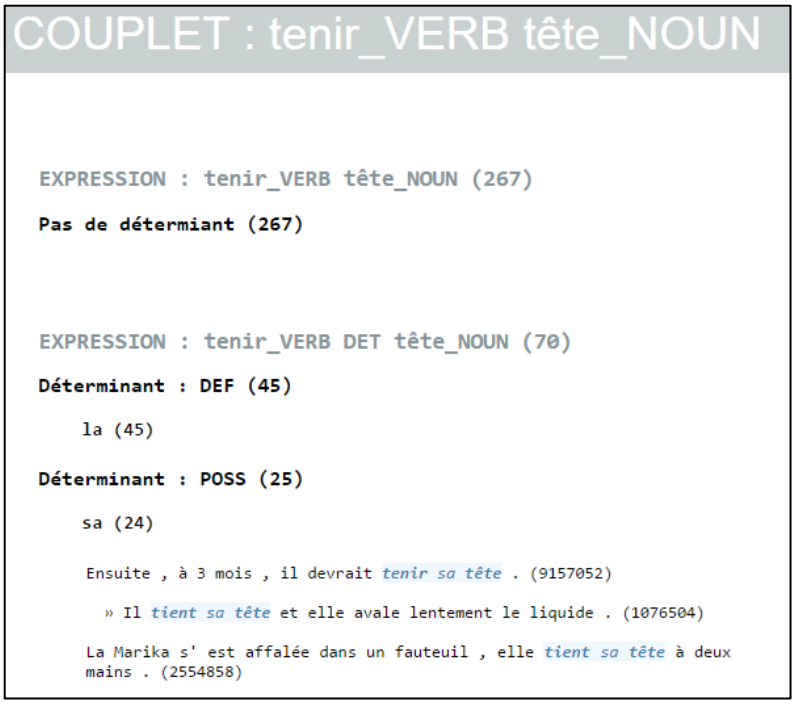


Figure 4 : Capture d'écran d'un fichier html contenant les occurrences en contexte d'un couplet V NObj (*tenir_VERB tête_NOUN*)

Pour nous aider à réaliser cette évaluation, nous avons mis en place un script qui permettait de créer, pour chaque couplet V NObj repéré, un fichier html contenant les

[49] Ils incluaient, par exemple, la possibilité d'antéposer *bien* au noyau des expressions. Ce patron a par la suite été supprimé (voir le point 1.5.2. de ce chapitre)

phrases qui comportaient les expressions correspondant au couplet[50]. Dans chaque fichier, les phrases apparaissaient classées par expressions, puis par classes de déterminants, puis par lemmes de déterminant. Un exemple est donné en Figure 4[51]. Ces fichiers nous permettaient, si besoin, de dénombrer les différents sens de l'expression en cas de polysémie et d'observer ses différentes variations.

Nous avons ainsi évalué la validité des expressions candidates en annotant pour chacune :

- **Valide** = La suite de mots forme toujours une expression polylexicale (qui peut être polysémique)
    - Par exemple, *avoir_VERB DET pain_NOUN sur_PREP DET planche_NOUN*, apparaissait tout le temps avec le déterminant *du* et correspondait toujours à l'expression dont le sens est *avoir une quantité importante de travail à faire*.
- **Partiellement valide** = la suite de mots donne lieu à des expressions polylexicales, mais toutes les occurrences de cette suite de mots n'en sont pas.
    - Par exemple, la suite *avoir_VERB DET nez_NOUN* produit des occurrences qui sont des expressions polylexicales comme dans le cas de la phrase : « *Pour œuvrer dans cette confrérie, mon gars, il faut* ***avoir du nez****, beaucoup de nez, et encore plus de souffle !* ». Cependant, il s'agit d'une simple association libre dans le cas de la phrase : *« Il* ***avait un nez*** *en forme de patate […]* » .
- **Invalide** = La suite de mots ne donne jamais lieu à une expression polylexicale
    - Par exemple, *continuer_VERB DET enquête_NOUN* ne constitue jamais une expression polylexicale.
- **?** = Le cas est difficile à évaluer. Rappelons que cette étape de notre travail nous a servi entre autres à repérer les cas difficiles à évaluer afin de les répertorier et de mettre ensuite en place des critères précis dont nous ne disposions donc pas encore.

1.3. Résultats globaux

Comme le démontre le Tableau 6, sur les 3110 expressions candidates évaluées, seules 25,43% étaient des expressions valides, 6.21% de ces expressions étaient partiellement valides, et 63.37% ne l'étaient pas du tout.

---

[50] Les phrases associées à chaque expression avaient été conservées par une compilation du dictionnaire Python contenant les expressions et leurs informations. Cette compilation est une tâche que réalise le script à chaque extraction et nous a été utile pour plusieurs tâches dont celle de la création des fichiers html.

[51] Les phrases contenant les expressions s'affichent par un clic sur le titre du type de déterminant. Dans la capture d'écran présentée en Figure 4, on a cliqué sur *sa (24)*. C'est la raison pour laquelle les phrases contenant les occurrences sans déterminant et avec le déterminant *la* ne sont pas affichées.

Valide	791	**25,43%**
Partiellement valide	193	6,21%
Invalide	1969	63,31%
?	157	5,05%
TOTAL	3110	

Tableau 6 : Résultats globaux de l'évaluation d'une première version du script d'extraction

1.4. Evaluation de l'impact des mesures de fréquences et de dispersion validité des expressions candidates

Nous avons évalué les expressions contenues dans les tranches de fréquences et de dispersion les plus hautes et les plus basses. Le but de cette étape était de tenter de répondre aux questions suivantes : Comme indiqué généralement dans la littérature, les expressions candidates ayant des taux de fréquences et de dispersion plus élevées sont-elles plus susceptibles d'être des expressions valides ? Si tel est le cas, laquelle de ces deux mesures est la plus utile pour déterminer la probabilité de validité des expressions candidates ?

1.4.1. Résultats par tranches de fréquence

Les résultats obtenus semblent confirmer l'hypothèse selon laquelle une fréquence haute implique un meilleur taux de validité des expressions candidates. En effet, les expressions candidates faisant partie des 10% des fréquences les plus hautes présentaient une majorité d'expressions valides ou partiellement valides. Les expressions candidates les moins fréquentes comportaient quant à elles une majorité d'expressions invalides. Les chiffres précis sont donnés dans les Tableaux 7(a) et (b).

Sur les 10% ayant les fréquences les plus hautes
(de 186 à 18 687)

Valide	406	**47,82%**
Partiellement valide	105	**12,37%**
Invalide	291	34,28%
?	47	5,54%
Total	849	

Sur les 10% ayant les fréquences les plus basses
(de 30 à 32)

Valide	86	9,84%
Partiellement valide	21	2,40%
Invalide	726	**83,07%**
?	41	4,69%
Total	874	

Tableau 7 : Résultats de l'évaluation d'une première version du script d'extraction sur les expressions candidates contenues (a) dans les 10% ayant les fréquences les plus hautes, (b) les 10% ayant les fréquences les plus faibles

1.4.2. Résultats par tranches de dispersion

Comme le montrent les Tableaux 8 (a) et (b), les résultats obtenus en fonction de la mesure de dispersion sont similaire à ceux obtenus en fonction de la fréquence ; plus cette mesure est faible, moins l'expression candidate a de chances d'être valide.

Sur les 10% ayant les valeurs de dispersion les plus hautes
(de 176 à 539)

Valide	402	**46,85%**
Partiellement valide	119	**13,87%**
Invalide	288	33,57%
?	49	5,71%
Total	858	

Sur les 10% ayant les valeurs de dispersion les plus hautes
(de 10 à 25)

Valide	68	7,84%
Partiellement valide	17	1,96%
Invalide	756	**87,20%**
?	26	3,00%
Total	867	

Tableau 8 : Résultats de l'évaluation d'une première version du script d'extraction sur les expressions candidates contenues dans (a) les 10% ayant taux de dispersion les plus hauts, (b) 10% ayant les taux de dispersion les plus faibles

1.4.3. Différence entre fréquence et dispersion

Un point à préciser est celui de la corrélation entre les expressions contenues dans les tranches (hautes ou basses) de fréquence et de dispersion. Nous avons constaté que la majorité des expressions ayant une fréquence parmi les plus hautes avaient également une dispersion parmi les plus hautes. Mais, cela n'était pas le cas pour les valeurs les plus basses. En effet, nous trouvions 74,36% d'expressions communes entre les tranches de fréquence haute et les tranches de dispersion haute. En revanche, cette proportion descendait à 28,58% lorsque l'on étudiait les tranches les plus basses.

En ce qui concerne l'utilisation de l'une ou l'autre des deux mesures à utiliser pour augmenter les probabilités d'obtenir des expressions valides, il semble, au vu de nos résultats, que la fréquences soit plus fiable. La différence entre les deux mesures est cependant minime (environ 1% de moins d'expressions valides dans les tranches de dispersion haute par rapport aux fréquences hautes).

1.5. Evaluation de l'inclusion des éléments périphériques

Nous avons ensuite réalisé une évaluation qui visait à repérer quels patrons permettaient de fournir des expressions valides et quels autres n'en produisaient pas. Il s'agissait aussi de quantifier la proportion de bruit et de silence que provoquait la prise en compte des éléments périphériques. Par exemple, *avoir_VERB hâte_NOUN* constitue une expression valide et avait une fréquence de 794. En revanche, *avoir_VERB hâte_NOUN de_PREP PRON rencontrer_VERB* était repérée comme expression ayant une fréquence de 36. Elle est invalide et sa fréquence devrait être ajoutée à celle de *avoir_VERB hâte_NOUN*. Elle est donc considérée comme facteur de bruit de par le fait qu'elle soit incorrecte, mais également comme source de silence car son repérage diminue la fréquence de l'expression valide qu'elle contient.

1.5.1. Méthode

Afin d'étudier cette question, nous avons extrait les expressions qui contenaient des éléments périphériques. Nous avons également extrait les expressions correspondant à leurs

noyaux seuls si ces dernières avaient été repérées. Par exemple, nous avons extrait *avoir_VERB DET impression_NOUN de_PREP être_NOUN* et son noyau, *avoir_VERB DET impression_NOUN*, qui est une des sorties du script. Parfois, le noyau n'était pas présent, comme dans le cas de *avoir_VERB DET longueur_NOUN de_PREP avance_NOUN*, pour lequel l'expression *avoir_VERB DET longueur_NOUN* ne faisait pas partie de la liste de sortie.

Pour chacune des expressions, nous avons annoté leurs caractéristiques avec les mentions suivantes :

- **Noyau** : lorsqu'il s'agissait du noyau de l'expression seul
- **[suite d'étiquettes] [antéposé/postposé/inséré] fait partie de l'expression** : lorsque le noyau et les éléments périphériques formaient une expression valide
    - Par exemple, pour *avoir_VERB DET main_NOUN libre_ADJ* nous annotions *ADJ postposé fait partie de l'expression*
- **[suite d'étiquettes] [antéposé/postposé/inséré] ne fait pas partie de l'expression** : lorsque le noyau et les éléments périphériques ne formaient pas une expression valide, tandis que le noyau seul en formait une.
    - Par exemple, pour *faire_VERB confiance_NOUN à_PREP DET homme_NOUN* nous annotions *PREP DET NOUN postposé ne fait pas partie de l'expression*
- **Tout invalide** : lorsque ni le noyau seul ni le noyau et les éléments périphériques ne formaient d'expression valide
    - Par exemple, pour *plaquer_VERB DET main_NOUN sur_PREP DET bouche_NOUN* nous annotions *Tout invalide* car ni cette expression ni son noyau *plaquer_VERB DET main_NOUN* ne forment d'expressions valides.
- **Erreur étiquetage** : lorsque l'étiquette de l'élément périphérique était fausse
    - Par exemple, pour *avoir_VERB bien_ADJ lieu_NOUN* nous annotons *Erreur étiquetage* car, dans cette structure, *bien* est adverbe et non adjectif.

1.5.2. Résultats quantitatifs et modification de la liste des patrons

Cette évaluation a démontré que, globalement, l'ajout d'éléments périphériques créait plus de bruit et de silence qu'il ne permettait d'extraire d'expressions valides ; la quantification de chacun des cas de figure est donnée dans le Tableau 9.

Il s'est alors agi de déterminer quels patrons produisaient le plus de bruits et de silences, et quels autres étaient au contraire utiles au repérage d'expressions longues valides. La liste des patrons et de leurs incidences respectives sur les sorties est présentée en Annexe 5.

Cas où noyau + élément périphérique = expression valide	249	26,55%
Cas où noyau seul = expression valide mais noyau + élément périphérique ≠ expression valide	372	**39,66%**
Cas où ni le noyau seul ni le noyau + élément périphérique ne forment d'expression valide	206	21,96%
Cas invalides pour cause de mauvais étiquetage	42	4,48%
Doute	69	7,36%
TOTAL	938	

Tableau 9 : Résultats de l'évaluation de la prise en compte des éléments périphériques – quantification des occurrences pour chaque cas de figure

On remarque tout d'abord que tous les patrons inclus dans le script ne donnaient pas forcément de résultat en sortie. Par exemple, le patron *ADV ADV postposé* (repéré sur 6 expressions dans la liste initiale comme par exemple *placer la barre trop haut*), n'apparaissait jamais. Nous avons donc décidé de supprimer ces patrons de notre script pour l'extraction finale. Ensuite, on peut voir que certains patrons étaient le plus souvent utiles au repérage d'expressions valides tandis que d'autres ne l'étaient pas du tout. Nous avons donc décidé de supprimer les patrons qui provoquaient exclusivement des sorties invalides. En revanche, nous avons décidé de conserver ceux qui provoquaient majoritairement des sorties non valides mais également quelques résultats acceptables ; ce choix, accompagné de celui d'une étape de redimensionnement manuelle décrite au point 2. du Chapitre 11, nous a permis de conserver les sorties valides permises par ces patrons.

1.6. Etude des phénomènes de polysémie repérés lors de l'évaluation

Cette évaluation nous a permis de faire un premier point sur les phénomènes linguistiques, inhérents aux expressions candidates extraites, qui rendent difficile la décision de leur validation. Le principal problème auquel nous avons été confrontée est celui de la polysémie ; une même expression candidate peut en effet donner lieu à :

- plusieurs expressions, par exemple, *jouer_VERB DET comédie_NOUN* donne lieu à la collocation *jouer la comédie* dont le sens est *faire une performance au théâtre/cinéma* mais aussi à l'expression figée *jouer la comédie* dont le sens est *faire semblant de*.
- une expression et des colligations comme *accorder_VERB DET violons_NOUN* qui donne lieu à l'expression *accorder ses violons* mais aussi à des colligations comme *il a accordé son violon avant de jouer*.
- plusieurs expressions et des colligations. Tel est le cas de *avoir_VERB DET cœur_NOUN* qui peut relever de colligations (*il a un cœur solide*) et de plusieurs expressions valides différentes comme *avoir le cœur à/de* (=*avoir envie de*) ou *avoir du cœur* (=*être de nature bienveillante et généreuse*).

Nous avons également noté qu'il était possible d'utiliser des critères de distinction pour repérer dans le corpus les occurrences des différents cas pour une expression donnée.

Cependant, cela n'est pas possible pour toutes les expressions polysémiques. Dans les exemples donnés, la présence combinée du déterminant *le* et des préposition *à*/*de* permet d'extraire les occurrences de *avoir_VERB DET cœur_NOUN* qui correspondent à *avoir le cœur à/de*. En revanche aucun critère structurel ne permet d'effectuer une telle distinction pour *jouer_VERB DET comédie_NOUN*. Par exemple, dans les phrases du corpus (1) et (2) ci-dessous, on trouve le sens *faire une performance au théâtre/cinéma* en (1) et *faire semblant de* en (2) ; aucun critère syntaxique ou morphologique ne distingue ces deux occurrences.

(1) *« Tous ont eu la capacité de chanter ou jouer la comédie. »*
(2) *« Ce n' est pas la peine de jouer la comédie. »*

Une typologie plus fine des cas de figure et des critères de distinction possibles estimés à partir de cette évaluation est présentée en Annexe 6.

Cette prise en compte de la polysémie des expressions polylexicales, au-delà de son intérêt lexicographique, nous a permis dans la suite de notre travail d'ajuster les mesures de fréquence et de dispersion (voir point 2.2. du Chapitre 11).

Cette évaluation nous a donc permis de réaffirmer l'hypothèse selon laquelle des seuils de fréquence et de dispersion élevés favorisaient l'extraction d'expressions valides. Elle nous a également permis de constater que les patrons préétablis n'étaient pas tous utiles, et que certains ne produisaient que des mauvais résultats, ce qui nous a amené à en supprimer un nombre important. Enfin, nous avons pu avoir un aperçu des problèmes que la polysémie des expressions créait lorsqu'il s'agissait d'évaluer leur validité.

Les performances et problèmes de notre méthode ayant été ainsi, en partie au moins, cernés, nous avons jugé bon de la comparer à une autre, basé sur la syntaxe.

2. Comparaison de la méthode avec celle d'une extraction par ALR

Afin d'estimer si notre choix de méthode principalement basée sur des patrons catégoriels était adapté à notre tâche, nous avons décidé de comparer la sortie de notre système à celle issue d'une extraction basée sur la syntaxe. Cette dernière consiste en une extraction par ALR (arbres lexico-syntaxiques récurrents) qui utilise les arbres syntaxiques des corpus arborés ; comme dans le cadre du travail de Corman (2012) que nous avons précédemment évoqué, il s'agit d'identifier les sous-arbres fréquents présents dans le corpus.

2.1. Brève présentation de la méthode d'extraction par ALR

Une des caractéristiques majeures de l'approche par ALR est l'importante possibilité combinatoire des arbres (Corman, 2012; Tutin & Kraif, 2017). En outre, l'intérêt tout à fait

notable qu'elle présente par rapport à notre méthode est de permettre le repérage des expressions dont les éléments sont séparés par une incise ou ayant subi une inversion syntaxique (Tutin & Kraif, 2016). Le fonctionnement itératif de la méthode est lui aussi particulier. Ce dernier consiste tout d'abord à partir d'un mot (nœud) et à extraire la liste de ses collocatifs privilégiés (identifiés par une mesure d'association). Chaque collocation ainsi identifiée et dépassant un certain de seuil de mesure d'association constitue alors un sous-arbre pour lequel on va chercher les collocatifs privilégiés. Le processus est ainsi répété jusqu'à ce que les collocatifs trouvés dépassent un certain seuil de mesure d'association et que la taille maximum du sous-arbre (paramétrable) soit atteinte (Tutin & Kraif, 2017). Lors de ce processus, les sous-arbres inclus dans des sous-arbres plus grands sont éliminés des sorties sauf si leur fréquence est supérieure ou égale à 25% de la fréquence de leurs enfants et qu'ils représentent au moins 3 occurrences de plus.

2.2. Méthode de comparaison des sorties

Une extraction par ALR a été réalisée par Olivier Kraif sur notre corpus. Afin de limiter la combinatoire, nous nous en tenons, pour les verbes, aux 1000 les plus fréquemment présents avec un objet direct nominal dans le corpus. Les seuils utilisés sont de 50 pour la fréquence et 8/9 pour la dispersion. La taille des ALR peut varier de 2 à 5 éléments (i.e. les expressions extraites peuvent être composées de 2 à 5 mots). En outre, seuls 10 cooccurrents possibles par verbe sont retenus.

Nous avons pour notre part réalisé une extraction avec notre méthode et utilisant les mêmes seuils fréquentiels et la même liste de 1000 verbes en entrée.

Une différence notable entre les deux méthodes est que nous n'utilisons que la relation OBJ pour repérer les couplets V NObj dans la méthode par patrons catégoriels tandis que celle par ALR utilise également la relation d'objet profond DEEPOBJ. Cette dernière relie le sujet d'un verbe au passif à ce verbe et permet donc d'indiquer que ce sujet correspond à l'objet direct du verbe dans une structure à la voix active. Ainsi, dans le syntagme *Marie a été invitée par Paul*, la relation DEEPOBJ serait utilisée pour relier *Marie* à *invitée* ; si on transforme cette phrase à la voix active (*Paul a invité Marie*), *Marie* est donc l'objet direct de *invité*.

Les résultats des deux extractions ont alors été comparés.

2.3. Résultats de la comparaison

Nous présentons ci-dessous les résultats de la comparaison réalisée sur plusieurs types de données et critères.

2.3.1. Tests de la présence des expressions de chaque sortie dans l'autre sortie

Afin de comparer la couverture des sorties, nous avons extrait 30 expressions dans chacune d'entre elles, réparties régulièrement sur les échelles de fréquence de chaque liste. Nous avons ensuite cherché, pour chaque expression extraite, si elle était présente en sortie de l'autre extraction.

Nous trouvons ainsi que 83,33% des expressions extraites par ALR sont présentes dans la sortie de l'extraction par patrons catégoriels. En revanche, seulement 13,33% des expressions extraites par patrons catégoriels sont également présentes dans la sortie de l'extraction par ALR. Les deux listes consignant le détail de cette comparaison sont présentées en Annexe 7.

Ce faible rappel des ALR par rapport aux patrons catégoriels est étonnant et peut provenir de plusieurs facteurs :

- Les filtrages par mesure d'association (Log Likelihood) ont pu éliminer des sorties qui sont présentes dans notre extraction. Cela est notamment le cas avec des verbes très fréquents.
- La limite de 10 cooccurrents possibles par verbe a réduit le nombre de sorties.
- La méthode par patrons catégoriels rassemble tous les déterminants sous une même classe ; celle par ALR les regroupe uniquement par lemmes, et ceci diminue les mesures statistiques de chaque sortie générée.

Ainsi, si nous prenons l'expression *prendre DET précaution*, absente dans les sorties ALR, et que nous utilisons l'outil Lexicoscope (Kraif, 2016) sur notre corpus pour observer ses statistiques, nous nous apercevons que :

- *prendre* et *précaution* ont une mesure de Log Likelihood élevée (3265,8894), une fréquence de 443 et une dispersion de 9.
- Mais, *précaution* est seulement le 18ème collocatif le plus fortement associé à *prendre*. Il n'a donc certainement pas été retenu à cause de la limite de 10 collocatifs par verbe.
- Même si on suppose que ce couplet ait été retenu, la fréquence et la dispersion de l'expression *prendre DET précaution* ont été fractionnées selon les diverses entrées correspondant aux différents lemmes de déterminants possibles (*prendre une/des précaution(s), prendre cette/ces précaution(s), (ne pas) prendre de précaution, prendre ses précautions, etc.*) ; cela est susceptible de faire passer ces diverses entrées sous les seuils de fréquence et de dispersion. On observe effectivement sur Lexicoscope que, par exemple, *prendre quelque(s) précaution(s)* a une dispersion de 6 et une fréquence de 50, ce qui la fait passer sous les seuils acceptables.
- En outre, le rattachement du déterminant à l'expression est lui aussi soumis à un seuil de mesure d'association. Or, le score de Log Likelihood entre le sous-arbre *prendre*

→ *OBJ* → *précaution* et le nœud *un/une/des* est très faible (0,2043) ; ainsi, l'expression *prendre une/des précaution(s)* n'aurait pas été retenue.

Les silences dans l'extraction par patrons syntaxiques par rapport à celle par ALR sont quant à eux dus à l'incapacité qu'a cette méthode à repérer les éléments d'une expression linéairement discontinue ou ayant subi des alternances syntaxiques. On observe en effet que, par exemple, l'expression *arborer sourire* pour laquelle l'extraction par ALR donne une fréquence de 165 et une dispersion de 9 est présente dans le corpus sous la forme VERBE DET NOM sans insertion seulement 86 fois et avec une dispersion de 7. Elle apparait 2 fois, et dans un seul sous-corpus sous la forme VERBE NEG DET NOM, et jamais sous la forme VERBE ADJ NOM, qui sont les deux seuls cas d'insertion pris en compte par notre script.

Nous pouvons donc conclure que la méthode par ALR produit largement plus de silences en termes de repérage d'items que la méthode par patrons catégoriels. Ces silences pourraient cependant être réduits en utilisant des seuils plus permissifs (par exemple, 40 cooccurrents possibles par verbe). De plus, notre méthode, quant à elle, a bel et bien tendance à sous-mesurer les scores statistiques des éléments repérés, de par son absence de prise en compte des séquences discontinues ou ayant subi des inversions.

2.3.2. Estimation de la précision des deux méthodes

Afin d'estimer la probabilité de chacune des deux techniques de fournir des expressions polylexicales valides en sortie, nous avons extraits les 50 expressions les plus fréquentes de chaque sortie et avons annoté leur validité. Les résultats, représentés dans le Tableau 10, montrent qu'il ne semble pas y avoir de différence significative en matière de précision entre les deux méthodes comparées.

	ALR	PC
Valides	66%	66%
Partiellement valides	16%	22%
Invalides	18%	12%

Tableau 10 : Taux d'entrées valides, partiellement valides et invalides sur les cinquante expressions les plus courantes extraites par la méthode par ALR et par patrons catégoriels

2.3.3. Comparaison des capacités respectives des deux méthodes à extraire des expressions longues

Nous avons enfin effectué une dernière comparaison, afin d'estimer si une des deux méthodes permettait d'extraire davantage d'expressions dont la longueur excède celle du noyau V (DET) NObj. Les deux méthodes ont des approches diamétralement opposées quant au repérage de ces structures. Les ALR tendent en effet à limiter l'agglutination d'éléments par des seuils de mesure d'association, mais leur ordre linéaire n'a aucun effet sur le

processus. De son côté, la méthode par patrons catégoriels tend à limiter cette agglutination par un recours aux catégories syntaxiques, sans utiliser de mesure d'association, et l'apparition des éléments dans un ordre linéaire précis est nécessaire à leur repérage. En outre, la méthode par ALR suit une logique d'augmentation itérative d'une séquence courte et élimine cette séquence lorsqu'une séquence plus longue satisfaisante est rencontrée. A l'inverse, notre méthode de redimensionnement consiste à diminuer itérativement une séquence longue, et à l'éliminer lorsqu'une séquence plus courte satisfaisante est rencontrée.

Nous avons extrait de chaque sortie les expressions longues qu'elle contenait. Nous avons ainsi observé que la méthode par patrons catégoriels a permis d'extraire davantage d'expressions de ce type avec 65 expressions extraites contre 10 avec la méthode par ALR. De plus, 4 expressions était communes aux deux listes[52]. En outre, les taux de précision des deux méthodes (i.e., le nombre d'expressions valides ou partiellement valide par rapport au nombre total d'items extraits) sur ce type d'expressions sont relativement similaires avec 70% pour les ALR[53] et 80% pour les patrons catégoriels. Il semble donc que les contraintes imposées par les ALR ne favorisent pas particulièrement la pertinence des extractions. Cette constatation n'est cependant vraie que dans les limites de notre définition de ce qu'est une expression valide. En effet, dans la liste ALR nous trouvons *donner un petit coup* et *donner un grand coup*, que nous considérons comme invalides car elles sont deux variantes de la même expression (*donner un coup*) et ne constituent pas à elles seules de nouvelles expressions (i.e., créant un sens différent de l'expression de base)[54]. Un autre système d'évaluation qui les considérerait comme valides donnerait une précision plus élevée pour les ALR, la faisant passer à 90%[55].

2.3.4. Conclusions

La méthode par ALR semble donc donner des résultats quantitativement inférieurs à la méthode par patrons catégoriels. Cela est également vrai dans le cas des expressions comportant un noyau et des éléments périphériques. En revanche, il ne semble pas y avoir d'écarts significatifs entre les deux méthodes au niveau de la précision des extractions. Ces constatations faites, il faut ajouter qu'elles seraient sans doute différentes si les seuils utilisés par la méthode par ALR avaient été plus permissifs. Une nouvelle extraction serait donc utile pour vérifier cette hypothèse et réaliser une comparaison qui soit plus précise et qui rende compte des pleines capacités de la méthode ALR, paramétrée avec des seuils différents.

---

[52] *donner DET coup de pied*, *lever les yeux au ciel*, *donner DET coup de poing* et *donner DET coup de pouce*

[53] Pour significative que soit une telle mesure calculée sur seulement 10 items

[54] Voir point 1. du Chapitre 11

[55] Idem à la note 53

Il nous faut préciser aussi qu'une différence importante entre les deux méthodes comparées est que la nôtre a été conçue spécifiquement pour la tâche que nous effectuons, à savoir, l'extraction spécifique d'expressions polylexicales formées autour d'un couplet V NObj et qui peuvent être des collocations ou des expressions figées, mais pas des locutions ou des proverbes (nous ne prenons pas en compte le sujet par exemple). La méthode par ALR est initialement conçue quant à elle pour repérer un nombre bien plus important et divers de séquences polylexicales. Ces deux optiques initiales ont certainement favorisé les meilleurs scores de notre méthode et montrent qu'elle est susceptible de fournir des résultats acceptables.

En outre, la supériorité d'une méthode syntaxique en termes de précision de décompte des occurrences pour une expression donnée a été démontrée par cette comparaison.

A l'issue de ces deux évaluations, nous disposions d'un script recalibré ainsi qu'une estimation de ses capacités. Nous pouvions donc entamer la constitution de notre liste d'expressions fondamentales, qui fera l'objet de la prochaine partie.

Partie V

-

CONSTITUTION DE LA LISTE

Chapitre 11. Sélection des items

Le script recalibré dont nous disposions nous a permis d'extraire du corpus des expressions polylexicales candidates accompagnées de leurs mesures statistiques et, si elles en comportaient, de la liste des déterminants avec lesquelles elles apparaissent dans le corpus. Il allait donc s'agir de sélectionner les éléments à inclure dans notre liste, par une suite de redimensionnement et de filtrages successifs que nous allons détailler dans cette partie.

1. Filtrage de la liste sur le critère de la validité des expressions candidates

L'extraction finale que nous avons réalisée nous a fourni une liste de 17 229 entrées, avec des fréquences par expressions candidates allant de 15 à 27 337 et des mesures de dispersion allant de 4 à 9[56]. En outre, 1505 d'entre elles comportaient des éléments périphériques postposés, 774 des pronoms antéposés, 100 des éléments insérés et 21 présentaient à la fois un pronom antéposé et un élément inséré.

La première étape était de sélectionner, parmi ces sorties, lesquelles étaient des expressions polylexicales valides. Nous avons donc annoté chacune d'entre elles avec les mentions *valide, partiellement valide* et *invalide* utilisées lors du test du premier script d'extraction et présentées au point 1.2. du Chapitre 10. Nous avons eu recours des critères précis, que nous allons à présent détailler, pour juger la validité d'une expression candidate. Cela nous a permis de mettre en place une certaine systématisation de la sélection. Il n'en reste pas moins qu'il aurait été difficile, et peut-être peu judicieux, de se restreindre à ce système sans que ce processus de validation ne soit réalisé, au moins en partie, à partir d'une certaine « intuition linguistique ».

1.1. Critères utilisés

1.1.1. Critères d'élimination des erreurs d'étiquetage

Nous avons tout d'abord filtré les erreurs résultant de fautes d'étiquetage et d'analyse syntaxiques dans le corpus. Nous avons en effet trouvé des cas où l'étiquetage était incorrect, comme dans le cas de l'expression candidate *!bouger pas*[57] où l'adverbe de négation est confondu avec le nom homonyme et étiqueté comme tel. De plus, nous avions des séquences dans lesquelles les noms analysés comme objets directs n'en étaient pas ; c'était par exemple

[56] Ces seuils volontairement bas vont permettre l'inclusion dans les sorties d'expressions longue à redimensionner (voir point 2. de ce chapitre), et ils permettent aussi de produire une liste de fréquence plus complète que celle élaborée dans le cadre de notre travail en vue de possibles réutilisations futures.

[57] L'indication *!* précédant une expression candidate signifiera dans la suite que cette dernière est invalide

le cas de certains compléments circonstanciels de temps comme dans les expressions candidates *!se taire un moment* et *!arrêter un instant*.

1.1.2. Critère de l'absence du déterminant

Une caractéristique syntaxique utile a été l'absence de déterminant avant le nom objet direct. La majeure partie des expressions candidates ayant cette structure sont des expressions valides, comme *avoir besoin*, *demander pardon* ou *faire erreur*. Il faut cependant ajouter deux exceptions à cette règle :

1) Elle n'est pas valable si le nom est un nom propre (*!regarder Tom*, *!quitter Paris*)

2) Elle n'est pas valable non plus si l'expression candidate comporte des éléments périphériques ; dans la majorité des cas, les expressions candidates ayant ces deux caractéristiques étaient en fait composées d'une expression valide (le noyau) et d'un complément qui apparait fréquemment après cette expression sans qu'il n'en fasse partie ou que son ajout ne constitue de nouvelle expression (*!avoir besoin de repos*, *!avoir lieu vendredi*)

Il faut également préciser que si cette règle et ses exceptions ont grandement facilité le filtrage, une vérification manuelle était nécessaire car certains cas y dérogeaient. On trouvait par exemple dans liste la séquence *!voir papa*, qui aurait été retenue comme valable selon ces critères bien qu'elle ne constitue pas une expression valide. De la même façon, de nombreuses expressions sans déterminant et qui comportaient de éléments périphériques étaient valides, bien qu'a priori exclues par l'exception 2) (*faire machine arrière* ou *avoir pignon sur rue*).

1.1.3. Critère de distinction entre les superpositions, les enchâssements et les collocations ternaires

Les expressions comportant des éléments périphériques sont susceptibles d'être le fruit de trois phénomènes différents : la « *superposition* », l' « *enchâssement* » (ou « *enchainemen*t ») ou les vraies expressions ternaires (Tutin, 2010b, pp.41-44). Ces trois cas de figure sont définis dans le Tableau 11 qui traite du cas des collocations ternaires (mais cette description s'applique plus généralement aux expressions polylexicales n-aires).

Nous avons rejeté les expressions consistant en des superpositions comme *!perdre le contrôle du véhicule* (*perdre le contrôle* + *contrôle du véhicule*), mais avons retenu les expressions n-aires et enchaînements comme *franchir la ligne d'arrivée* ou *dérouler le tapis rouge*.

Phénomène	Définition	Exemple (les collocatifs sont soulignés)
Superposition de collocations (quand il y a compatibilité syntaxique)	Deux collocations qui ont la même base et qui peuvent se combiner syntaxiquement.	*peur bleue* + *avoir une peur* (+ Modif) → *avoir une peur bleue*
Enchaînement de collocations (collocations récursive)	Cas 1 : la base est une collocation.	*Se mettre* (*en*(*colère*))
Enchaînement de collocations (collocations récursive)	Cas 2 : le collocatif est une collocation	(*freshly baked*) *bread*
Vraies collocations ternaires (ou plus)	Deux collocatifs (ou plus) peuvent être associés à la base (et la séquence ne peut pas être décomposée)	*En dernière analyse*

Tableau 11 : Description des différents types de collocations ternaires (emprunté à Tutin, 2010b, p.44)

1.1.4. Critère de non-compositionnalité

Nous avons retenu comme valides les expressions non-compositionnelles, qu'elles soient relativement transparentes (*regarder la vérité en face*) ou complètement opaques (*tirer les vers du nez*).

1.1.5. Constructions à verbe support

Il nous a semblé raisonnable de retenir comme valides toutes les constructions à verbe support. En effet, dans ces structures, la prédication est portée par le nom et ce, de par sa cooccurrence avec le verbe ; il s'agit donc d'un phénomène linguistique qui entre dans le domaine de la phraséologie. Comme pour le critère de l'absence du déterminant, celui-ci ne s'applique pas au cas où l'expression candidate comporte des éléments périphériques (*!rendre visite à un ami*).

1.1.6. Critères d'inclusion ou de rejet des pronoms antéposés

Nous avons choisi de considérer comme valides les expressions contenant des pronoms réfléchis ou les proclitiques *y* et *en*, à condition que ces expressions :

- N'existent pas sans le pronom ; nous avons ainsi validé *se frayer un chemin* et *en avoir sa claque* (**frayer un chemin*, **avoir sa claque*) et rejeté *!en faire la promesse*.
- Existent sans le pronom mais n'ont pas le même sens (*rendre compte* vs. *se rendre compte*), ce qui revient à dire que ce pronom en fait une expression nouvelle. Ce n'est pas le cas, par exemple, de *se faire peur*, dont le sens est similaire à celui de *faire peur*.

1.1.7. Critère d'affinité lexicale

Certaines expressions ont été validées de par l'affinité lexicale particulière entre leurs composants. Ainsi, l'expression *se frayer un chemin* n'est pas compositionnelle, elle comporte un déterminant et n'est pas une expression à verbe support, mais elle est tout de

même valide. En effet, le verbe *se frayer* n'a de cooccurrent objet direct que *chemin* et certains de ses quasi-synonymes dans le corpus (*passage, voie* et très rarement *place*)[58].

1.2. Résultats quantitatifs du filtrage

A l'aide de ces critères, nous avons donc filtré la liste. Un aperçu de l'annotation est donné en Annexe 8. Cette opération a résulté en l'élimination de 13 372 entrées non valides et la conservation de 2 937 expressions valides et 940 expressions partiellement valides. Si l'on considère ces deux dernières catégories comme une seule classe de résultats satisfaisants, la précision de notre script est donc de 22,5%. Ce score est relativement faible, mais il faut prendre en compte le fait que nous avons utilisé des seuils très bas ; si, par exemple, on ne conserve dans notre liste de sortie que les expressions candidates ayant une fréquence minimum de 50 et une dispersion minimum de 7 (afin de simuler l'utilisation de ces paramètres lors de l'utilisation du script), la précision s'élève à 40,17%. De même, si l'on choisit des seuils de 200 de fréquence et 8 de dispersion, ceux que nous utiliserons par la suite[59], nous obtenons un score de 63,73%.

Ce processus de filtrage ayant été réalisé, il nous a paru judicieux de tenter d'estimer si, contrairement au choix que nous avions fait de ne pas utiliser de mesures d'association[60], ces dernières auraient pu fournir une aide intéressante ou non dans le cadre de notre travail. Nous avons donc réalisé la petite expérience que nous allons à présent exposer.

1.3. Estimation de l'apport qualitatif de deux mesures d'association

Nous avons extrait de la liste de sortie 60 expressions : 30 valides ou partiellement valides et 30 non valides. Dans chacun de ces deux groupes, le premier tiers présente des valeurs de fréquence élevées, le deuxième des valeurs moyennes, et le troisième des valeurs faibles. Nous nous limitons aux expressions binaires. Nous avons ensuite utilisé l'outil Lexicoscope (Kraif, 2016) sur notre corpus pour calculer les mesures de Log Likelihood et de Z-score de ces expressions. La première mesure est très répandue et est utilisée par défaut sur Lexicoscope. La deuxième est celle qui avait été la plus utile à Benigno (2012, p.149) dans le cadre de son travail. Précisons également que les seuils de fréquence et de dispersion que nous avons utilisés sur Lexicoscope sont les mêmes que ceux utilisés dans le cadre de notre extraction (15 pour la fréquence, 4 pour la dispersion).

---

[58] Information extraite à partir d'une recherche de cooccurrence sur notre corpus avec l'outil Lexicoscope (Kraif, 2016) sans aucun filtre statistique ; certains autres nom objets directs apparaissent en sortie mais cela était dû à des analyses syntaxiques incorrectes.

[59] Voir 3.1. de ce chapitre

[60] Voir point 1. du Chapitre 9

Les listes ainsi produites sont consultables en Annexes 9 et 10. On peut y voir que, si on avait utilisé un filtre de Log Likelihood de 10,83[61], 3/30 expressions invalides et 1/30 expressions valides ou partiellement valides ne seraient pas apparues en sortie. On note aussi que ces 4 expressions ne passant pas le seuil de Log Likelihood ont des fréquences basses ; une première estimation est donc que ce seuil n'aurait certainement eu aucun ou peu d'impact sur les expressions aux fréquences hautes, qui constituent l'objet de notre recherche. Il aurait en revanche pu limiter le nombre d'entrées invalides de cette phase d'extraction avec des seuils de fréquence et dispersion bas, et aurait, dans des proportions moindres, généré quelques silences.

Plus généralement, en fusionnant les deux listes (valides/invalides) et en classant les entrées par ordre décroissant de mesure de Log Likelihood puis de Z-score, nous avons pu obtenir des estimations des taux de rappel et de précision que nous aurions obtenus si nous avions utilisé certains seuils de ces deux mesures (Tableau 12). Il faut souligner ici que nos données de test consistent en une liste finie et parfaitement connue, ce qui permet de donner une mesure de rappel ; cela n'est pas possible lorsqu'il s'agit du corpus, pour lequel nous ne pouvons dresser avec certitude une liste exhaustive des couplets V NObj qu'elle contient autre que par le type de procédé-même que nous étudions. Ainsi, la mesure de 100% qui est donnée pour les extractions sans utilisation de seuil de mesure d'association signifie simplement que toutes les données sur lesquelles nous nous basons ont été retenues. En aucun cas cela ne n'affirme que le script identifie correctement et exhaustivement les cooccurrences que le corpus donne à identifier. Nous utilisons donc uniquement la mesure de rappel dans l'optique de comparer son évolution avec celle de la précision selon les seuils considérés, i.e., d'estimer la proportion dans laquelle nous perdrions des expressions candidates en sortie au fur et à mesure que la précision augmenterait.

Seuil Log Likelihood	Rappel	Précision	F-mesure	Seuil Z-Score	Rappel	Précision	F-mesure
AUCUN	100%	50,00%	66,67%	AUCUN	100,00%	50,00%	66,67%
10,83	96,67%	50,88%	66,67%	5	90,00%	49,09%	63,53%
30	90,00%	51,92%	65,85%	10	83,33%	51,02%	63,29%
50	83,33%	50,00%	62,50%	15	80,00%	53,33%	64,00%
100	76,67%	53,49%	63,01%	20	76,67%	41,00%	53,43%
300	60,00%	56,25%	58,06%	30	66,67%	60,61%	63,49%
500	46,67%	51,85%	49,12%	50	56,67%	60,71%	58,62%

Tableau 12 : Estimation de la précision et du rappel obtenus selon différents seuils de Log Likelihood et de Z-score sur une répartition égale de colligations et d'expressions, pour des seuils de fréquences de 15 et de dispersion de 4

[61] Seuil paramétré par défaut sur Lexicoscope, utilisé notamment par Tutin & Kraif (2017), et qui détermine le fait que la cooccurrence ait moins d'une chance sur 1000 d'être due au hasard

De plus, nos données étaient également réparties entre expressions valides et invalides, ce qui n'est pas le cas dans le corpus ; se baser sur des données également réparties reviendrait à partir du principe que, dans le corpus, exactement la moitié des couplets V NObj (de fréquence > 15 et dispersion > 4) forment des expressions polylexicales, ce qui est un postulat très certainement inexact. Nous avons donc pris comme proportions celles données par les résultats de notre extraction en conservant 10 des expressions valides (toujours réparties sur l'échelle de fréquence) et les 30 colligations, pour un rapport de 25/75%[62] (Tableau 13)[63].

Seuil Log Likelihood	Rappel	Précision	F-mesure	Seuil Z-Score	Rappel	Précision	F-mesure
0	100,00%	25,00%	40,00%	AUCUN	100,00%	25,00%	40,00%
10,83	90,00%	24,32%	38,30%	5	90,00%	24,32%	38,30%
30	90,00%	26,47%	40,91%	10	90,00%	27,27%	41,86%
50	90,00%	26,47%	40,91%	15	80,00%	27,59%	41,03%
100	60,00%	21,43%	31,58%	20	80,00%	30,77%	44,44%
300	60,00%	28,57%	38,71%	30	80,00%	38,10%	51,61%
500	60,00%	33,33%	42,86%	50	60,00%	35,29%	44,44%

Tableau 13 : Estimation de la précision et du rappel obtenus selon différents seuils de Log Likelihood et de Z-score sur une répartition de 75% de colligations et 25% d'expressions valides dans le corpus, pour des seuils de fréquences de 15 et de dispersion de 4

Ces estimations tendent à montrer, sans surprise, que la précision et la F-mesure sont augmentées par l'utilisation de seuils de mesures d'association. Leur utilisation aurait donc effectivement facilité notre tâche de filtrage manuel. Cependant, la perte de données valides aurait été considérable.

Une hypothèse serait alors de penser que les expressions perdues par l'utilisation de ces mesures d'association sont en majeure partie des expressions peu fréquentes et que, notre travail consistant *in fine* à produire une liste d'expressions fréquentes, le recours à ces seuils aurait été utile. Or, comme le montrent les tableaux en Annexe 11 qui présentent les expressions les plus fréquentes de notre liste classées par ordre croissant de mesures d'association, cela ne semble pas être le cas à moins d'utiliser des seuils très bas (et donc peu filtrants). En effet, au-dessus de 447 pour le Log Likelihood et de 33 pour le z-score, on retrouve des expressions valides et invalides globalement réparties équitablement sur les ranges de mesures d'association.

Ces estimations ayant été présentées et expliquées, il convient de rappeler qu'il s'agit uniquement d'un test basé sur un nombre restreint de données et que les résultats et

[62] Rapport proche de celui obtenu lorsqu'on ne prend en compte que les expressions binaires de la liste

[63] Rappelons qu'il est tout à fait normal le rappel ne varie pas entre les deux tableaux (12 et 13) car cette mesure ne tient pas compte des faux positifs (i.e., des expressions candidates non valides) et que le rapport de proportion entre ces derniers et les vrais positifs (i.e., expressions candidates valides) n'a donc pas d'impact.

conclusions auraient éventuellement pu être différents si ce nombre et/ou le jeu de données avaient été différents. En outre, nous précisons que ce test avait pour seul et unique but d'estimer l'apport potentiel qu'aurait pu représenter l'utilisation de deux mesures d'association dans le cadre de notre travail et en tant que paramètre de notre script ; en aucun cas il ne s'agissait de donner de quelconque conclusion sur l'utilité des mesures d'association en général.

Quoi qu'il en soit, le travail de filtrage selon le critère de la validité des expressions ayant été réalisé, il nous a ensuite fallu redimensionner les mesures statistiques de certaines expressions.

2. Redimensionnement des mesures statistiques

Afin d'obtenir des mesures de fréquence et de dispersion les plus proches possibles de la réalité du corpus, nous avons effectué un travail de redimensionnement des fréquences organisé en deux parties. La première consistait en l'addition de ces mesures dans les cas d'entrées différentes pour une seule expression. La deuxième était basée sur une attribution de valeurs correctes aux entrées différentes produite par une même expression polysémique.

2.1. Cumul des mesures statistiques de deux entrées différentes pour une seule expression

Parmi les entrées jugées invalides lors de l'étape présentée au point 1. du présent chapitre se trouvaient des expressions candidates composées d'un noyau valide et d'éléments périphériques dont la combinaison ne produisant pas une expression polylexicale[64]. Se contenter de les éliminer serait revenu à ignorer les occurrences des expressions valides que ces expressions candidates contenaient. Pour remédier à ce problème, nous avons repéré dans la liste d'extraction non filtrée les expressions de ce type et avons fusionné leur mesures statistiques et listes de déterminants ; il s'agit du même processus de redimensionnement que celui que réalise le script pour les expressions longues ayant une fréquence inférieure à 15[65]. Continuer ce processus manuellement au lieu de paramétrer un seuil fréquentiel automatique plus élevé nous a permis de limiter les cas de non repérage des expressions longues, valides, et à fréquence basse. Cela nous a également permis d'effectuer un redimensionnement qui prenait en compte les éléments périphériques antéposés et insérés, ce que le script ne fait pas. Un exemple de l'application de ce procédé est consultable en Annexe 12.

[64] Voir point 1.5.1. du Chapitre 10

[65] Voir 3.2. du Chapitre 9

2.2. Attributions de mesures correctes aux entrées différentes d'une même expression polysémique

Nous avons considéré que les expressions polysémiques devaient donner lieu à une entrée pour chaque sens. Il fallait pour cela être capable de déterminer les mesures statistiques et la liste des déterminants correspondant à chacun de ces sens.

Nous désignerons à présent par le terme « *désambiguïsation des expressions polysémiques* », ou simplement, « *désambiguïsation* », ce processus de notre travail. Pour le mettre en œuvre, nous avons créé, pour chaque couplet V NObj repéré, un formulaire PHP comportant les phrases qui contenaient le couplet. Comme pour les fichiers html créés lors de l'étape d'évaluation du script[66], les phrases étaient classées par expressions formées, puis par classes de déterminants, puis par déterminants. Il nous suffisait alors de sélectionner les phrases qui contenaient le sens précis d'une expression, et nous obtenions la fréquence et la dispersion correspondantes[67]. Un aperçu de ces formulaires et des résultats affichés et donné en Annexe 13.

Ce processus étant assez lourd, nous nous sommes limitée à la désambiguïsation des expressions qui avaient des fréquences initiales supérieures à 200. A chaque désambiguïsation, nous créions donc des entrées différentes pour chaque sens repéré accompagnées de :

- leurs mesures statistiques
- la liste des déterminants ; lorsqu'un des critères que nous nous donnions pour la désambiguïsation était l'exclusivité d'un déterminant, nous adaptions la liste. Par exemple, *donner_VERB DET parole_NOUN* dans le sens de *promettre* nécessitait la présence d'un possessif singulier. Nous ne laissions donc pour cette entrée que la liste *POSS(sa,ma,leur,votre,ta,notre)*.
- Le sens
- Les critères de désambiguïsation utilisés.

Des exemples du résultat de cette opération sur trois entrées sont donnés dans le Tableau 14[68]. Ajoutons que la même technique a également servi à désambiguïser le cas d'expressions candidates pouvant être des expressions valides ou des colligations.

---

[66] Voir 1.2. du Chapitre 10

[67] La fréquence correspondait au nombre de phrases sélectionnées. Pour le calcul de la dispersion, le script du formulaire que nous avons mis en place utilisait les identifiants des phrases ; nous avions préalablement listé les id de début et de fin de chaque sous-corpus.

[68] Le sigle *VM* signifie *vérification manuelle* et indique que la désambiguïsation a été réalisée phrase par phrase sans critère préétabli.

Expression lemmatisée	Sens	Condition de validité	Début de la liste des déterminants	Fréquence expression	Dispersion expression	Fréquence couplet	Dispersion couplet
avoir_VERB DET rôle_NOUN	au cinéma/théâtre	VM	des(des) /DEF(le,	52	5	434	9
avoir_VERB DET rôle_NOUN	avoir la fonction de	VM	des(des) /DEF(le,	294	9	434	9
avoir_VERB soif_NOUN	avoir besoin de boire	VM		470	9	581	9
avoir_VERB soif_NOUN	être avide de	Suivi de "de" + VM		58	8	581	9
faire_VERB DET course_NOUN	achats	Nom au pluriel + VM	des(des) /DEF(les	643	9	1136	9
faire_VERB DET course_NOUN	concours de rapidité	VM	des(des) /DEF(la,	180	9	1136	9
faire_VERB DET course_NOUN	excursion avec un but	Det=une + VM	IND(une)	51	8	1136	9

Tableau 14 : Exemple de trois expressions désambiguïsées

Il nous faut préciser que si ce procédé nous a été très utile, son fonctionnement et son résultat sont certainement loin d'être parfaits. En effet, la sémantique des expressions était parfois complexe, et le degré de finesse dans la séparation en plusieurs sens était souvent difficile à établir. De plus, la démarche reposait sur une anticipation des différents sens que nous nous attendions à trouver dans le corpus ; à de nombreuses reprises, nous trouvions dans les phrases extraites des sens supplémentaires à ceux que nous avions prévus. Cependant, une description visant à la précision la plus parfaite possible nécessiterait la vérification de l'intégralité des occurrences de toutes les expressions extraites, même dans le cas de celles qui ne semblent pas ambigües. Cela représenterait un travail intéressant mais qui requerrait une quantité de temps et de travail considérable. Enfin, nos contextes se limitant à la phrase, certaines occurrences étaient difficiles à désambiguïser ; des contextes plus larges auraient sans doute été plus adaptés.

Ces redimensionnements des mesures statistiques effectués, nous pouvions alors sélectionner les expressions à inclure dans notre liste finale.

3. Filtrage de la liste sur le critère de l'aspect fondamental des expressions – Méthode adoptée et évaluations

A l'issu de ce redimensionnement, nous disposions des éléments nécessaires pour décider des items à inclure dans notre liste d'expressions fondamentales.

3.1. Seuils choisis et liste finale obtenue

Nous avons choisi de ne retenir comme fondamentales que les expressions valides ayant une fréquence minimale de 200 et une dispersion minimale de 8. Ces seuils ont été choisis de manière un peu arbitraire, à partir des observations linguistiques et quantitatives réalisées sur les données extraites lors des phases que nous avons détaillées précédemment. En effet, les expressions respectant ces seuils nous ont paru, de manière globale, stylistiquement neutres et, bien entendu, fréquentes. De plus, cela nous permettait d'obtenir des données quantitativement intéressantes pour les traitements futurs visant à décrire les expressions retenues, et ce, pour les raisons suivantes :

- Le nombre d'entrées s'élevait à 423, ce qui rendait possible le recours à des vérifications manuelles de leurs caractéristiques si cela s'avérait nécessaire.
- La liste comprenait, à première vue, si bien des collocations que des expressions figées, permettant d'éventuelles comparaisons entre ces deux types sur certaines caractéristiques.
- Elle comportait également :
    - De nombreuses constructions à verbe support
    - Des expressions avec et sans déterminant
    - Quelques verbes pronominaux
    - Plusieurs expressions polysémiques dont deux des sens avaient été conservés par le filtrage statistique

En revanche, seules deux expressions comportant des compléments postposés avaient les mesures statistiques requises pour figurer dans cette liste. Aucune ne comportait, en outre, d'élément périphérique inséré entre le verbe et l'objet. Cette constatation porte à croire que les expressions ayant ses caractéristiques sont plus rares et/ou plus stylistiquement marquées dans la langue que celles qui sont uniquement formées par une suite V (det) NObj. Bien sûr, cette hypothèse reste à confirmer ; notre script s'est montré capable d'extraire de nombreuses expressions longues, mais d'autres techniques donneraient peut-être des sorties de ce types différentes.

Quoi qu'il en soit, la liste que nous avons obtenue nous semblait acceptable et il s'est alors agi de vérifier si ces seuils, et plus généralement notre méthode, donnaient des résultats conformes à notre objectif de création d'un lexique fondamental.

Chapitre 12. Estimation de la capacité de la méthode à produire une liste d'expressions fondamentales

Afin d'estimer si notre méthode permettait effectivement de produire une liste d'expressions que l'on peut considérer comme fondamentales, nous avons tout d'abord évaluer l'impact de la sous-représentation de l'oral spontané dans le corpus. Puis, nous avons cherché à savoir si les seuils que nous avons utilisés étaient adaptés. Enfin, nous nous sommes interrogée quant à la significativité de la mesure de dispersion que nous avons utilisée.

1. Evaluation de l'impact de la sous-représentation de l'oral spontané

Comme nous l'avons évoqué au point 1.1.2 du Chapitre 8, la quantité d'oral spontané est nettement moins importante dans notre corpus comparé aux autres genres[69]. Il nous est paru utile d'évaluer dans quelle mesure cette caractéristique nuisait à notre but qui était de produire un lexique que se veut représentatif de la langue générale. Nous disposions alors, à cette étape de notre travail, d'assez de données pour réaliser une estimation de l'ampleur de l'impact qu'a eu cette caractéristique du corpus sur les résultats.

1.1. Méthode

Nous avons tout d'abord réalisé deux extractions différentes : une sur le sous-corpus constitué des trois parties de transcription oral (TCOF, ESTER et CFPP) et une exclusivement sur le TCOF qui est la partie la plus susceptible de contenir de la parole conversationnelle plus ou moins spontanée. L'objectif était d'avoir à notre disposition des résultats produits à partir de corpus exclusivement oraux afin de les comparer à ceux obtenus sur l'intégralité du corpus ; nous voulions ainsi voir si certaines expressions plus fréquentes à l'oral qu'à l'écrit avaient été quantitativement sous-représentées. Les seuils de fréquence utilisés pour les extractions sur l'oral sont proportionnels au rapport de 200 occurrences pour 117M de mots du corpus général ; pour le TCOF seulement, ce seuil proportionnel était inférieur à 1, et nous avons donc décidé d'utiliser le seuil de 2 pour obtenir un minimum de significativité. En outre, nous ne gardons pour la comparaison que les expressions valides.

1.2. Résultats

Les résultats de cette comparaison, présentés dans le Tableau 15, montrent que l'impact de la sous-représentation de l'oral dans le corpus est relativement négligeable. En effet, 86% des expressions extraites sur le corpus oral sont présentes dans la liste des expressions retenues comme fondamentales avec des seuils statistiques proportionnels. Le corpus oral susceptible de contenir le plus de parole conversationnelle spontanée produit des résultats sensiblement équivalents (81% de présence). En outre la quasi-totalité des expressions en sortie de cette extraction sur l'oral était présente dans les sorties de l'extraction à large spectre sur le corpus général. Une seule expression (*construire_VERB DET famille_NOUN*) n'est jamais apparue dans nos sorties antérieures.

L'impact de cette caractéristique de notre corpus sur les résultats obtenus est donc considéré comme négligeable. Une comparaison avec des données issues de corpus contenant davantage de conversations spontanées portant sur des sujets divers serait sans doute utile à une meilleure estimation de ce constat. Elle permettrait, d'autre part, de fournir

[69] L'oral spontané représente environ 1,5% du corpus

une liste plus complète et représentative des spécificités de l'oral concernant les expressions polylexicales et leurs fréquences.

	Nombre de mots	Seuil de fréquence	Nombre d'expressions valides extraites	Liste finale +	Liste finale - Liste seuils 15/4 +	Liste finale - Liste seuils 15/4 -
TCOF, CFPP, ESTER	1 736 156,00	3	111	96	15	0
TCOF	341 244,00	2	55	45	9	1

Tableau 15 : Résultats de l'extraction d'expressions réalisée sur les corpus oraux – Nombre d'expressions valides et indications de leur présence ou absence dans les extractions réalisées sur le corpus général[70]

Cette conséquence d'une des caractéristiques de notre corpus sur la représentativité des données extraites ayant été mesurée, nous pouvions évaluer si les seuils utilisés correspondaient au but que nous nous étions fixé.

2. *Evaluation des seuils utilisés*

Nous avons demandé à des enseignants spécialistes de Français Langue Etrangère (FLE) de répondre à un petit questionnaire comportant un échantillon des expressions extraites. Les objectifs de cette enquête étaient 1) de juger si les seuils que nous avons choisis permettent de constituer un lexique fondamental, 2) d'obtenir une estimation de l'utilité de la liste obtenue dans une optique d'enseignement des langues. Quatre professeurs ont répondu à ce questionnaire, ce qui nous a permis d'établir quelques estimations quant aux réponses à ces questions. Une enquête à plus large échelle serait toutefois nécessaire pour obtenir des conclusions plus précises.

2.1. Items et questions

Nous avons sélectionné 16 expressions extraites sur notre corpus ; ces dernières sont de fréquences variées et ont des dispersions différentes. Nous avons inclus si bien des

[70] Légende :

- *Liste finale* + : expressions apparaissant dans la liste des expressions retenues comme fondamentales (fréquence > 200, dispersion > 8) à partir des extractions réalisées sur le corpus général
- *Liste finale - / Liste seuils 15/4* + : expressions n'apparaissant pas dans la liste des expressions retenues fondamentales à partir du corpus générale mais qui étaient présente dans les sorties de l'extraction initiale avec des seuils de fréquence > 15, dispersion > 4
- *Liste finale - / Liste seuils 15/4* + : expressions n'étant pas apparues en sortie des extractions réalisées sur le corpus général

collocations que des expressions figées et des expressions métaphoriques[71]. La moitié des expressions dépassaient les seuils que nous avons utilisés pour constituer notre liste finale. De plus, nous avons volontairement inclus des expressions qui nous paraissaient assez familières comme *foutre le camp* et d'autres qui nous semblaient relativement formelles comme *porter une accusation*. De la même manière, nous avons inclus des expressions transparentes comme *avoir besoin* et d'autres, bien plus opaques comme *pousser le bouchon*.

Pour chacune d'entre elle, nous demandions à chaque personne :

- De trouver un synonyme à l'expression ; cette question permettait de vérifier l'hypothèse selon laquelle plus une expression est fondamentale, plus il est aisé de lui trouver un synonyme
- D'indiquer le registre de langue auquel il estimait que l'expression appartenait
- De préciser si elle jugeait que l'expression avait sa place dans un lexique fondamental du français au niveau B1, qui est un niveau seuil dans l'apprentissage des langues secondes.

Le questionnaire ne comportait pas les informations de types[72] et ni les mesures statistiques relatives à chaque expression. Les résultats obtenus sont présentés en Annexe 14.

2.2. Résultats

Dans la majorité des cas, des synonymes corrects ont été trouvés pour chaque expression. Il est cependant intéressant de noter que les deux seules expressions pour lesquelles aucun synonyme n'a été donné sont très fréquentes ; il s'agit de *avoir raison* (11 953 occurrences) et *hausser les épaules* (2 563 occurrences). La deuxième décrit un geste particulier qu'il n'est pas aisé d'expliquer sans le mimer, ce qui explique la difficulté qu'ont rencontrée les spécialistes à donner un synonyme. La raison de cette difficulté est plus difficile à expliquer dans le cas d'*avoir raison*. Il semble donc difficile d'établir un rapport direct entre synonymie et aspect fondamental et cette petite enquête porterait même à croire que, si la fréquence est un bon indicateur de cet aspect, il est alors plus difficile de trouver un synonyme à une expression fondamentale qu'à une expression non fondamentale. Ceci va donc à l'encontre de notre hypothèse de départ.

D'autre part, des registres similaires sont globalement attribués à chaque expression par les différents spécialistes. Ceci démontre que les différentes réponses données à la question suivante (la place des items dans un lexique fondamental) partent d'un accord plutôt général quant au registre des expressions proposées.

Il est en revanche intéressant de noter que de nombreux désaccords apparaissent entre les quatre spécialistes quant à la question de légitimité de la place accordée aux expressions

[71] Voir 1.1.1. du Chapitre 13

[72] *Ibid*.

dans un lexique fondamental ; plus des trois quarts des réponses données ne sont pas unanimes.

En outre, il semble que le seuil que nous avons choisi est relativement adapté à la tâche. En effet, seules 2 expressions de fréquence supérieure à 200 et de dispersion supérieure à 9 sont majoritairement jugées inappropriées à un lexique fondamental de niveau B1 ; c'est en revanche le cas de six expressions étant en dessous de ces seuils. Il est également intéressant de noter que les trois expressions unanimement acceptées sont des collocations et ont des valeurs de fréquence élevées. De plus, une dispersion faible semble provoquer le rejet d'une expression car toutes les expressions pour lesquelles cette valeur est inférieure à 9 sont majoritairement jugées comme n'ayant pas leur place dans un lexique fondamental.

Enfin, il nous a semblé intéressant de relever le fait que, pour deux expressions, à cette question de la place dans un lexique fondamental, les spécialistes répondent oui, mais dans un lexique spécialisé ou universitaire (*explorer la possibilité* et *porter une accusation*). Or, ces expressions ont toutes deux une dispersion maximale, ce qui signifie qu'elles ont été rencontrées dans tous les genres des différents sous-corpus et ce, malgré leur fréquence basse (20 et 29). Il semblerait donc que la fréquence d'une expression soit plus encline à la faire ressentir comme spécialisée que sa dispersion, ce qui semble contradictoire avec nos attentes et l'utilisation-même de la dispersion. Bien entendu, comme pour chaque élément de réponse apporté par cette petite enquête, cette hypothèse mériterait d'être confirmée à l'aide d'une expérience plus vaste, contenant davantage d'expressions et réalisée auprès de locuteurs/spécialistes plus nombreux. D'autre part, comme nous allons immédiatement l'expliquer, il est possible que la mesure de dispersion que nous avons utilisée ne soit pas assez finement élaborée pour garantir absolument la neutralité du genre des expressions.

3. Relativisation de la valeur informative de la mesure de dispersion utilisée

Partant ainsi de la constatation du fait que des experts du FLE jugeaient comme spécialisées des expressions ayant obtenu le score de dispersion maximum, il nous a semblé opportun de tenter d'estimer dans quelle mesure ce score donnait une indication de neutralité du genre de l'expression. Nous avons fait l'hypothèse que, si une expression donnée figurait effectivement dans tous nos sous-corpus, cela n'excluait pas le fait que la majorité de ses occurrences se trouve cantonnée à un seul genre, ce qui rendrait la valeur informative de cette mesure nettement moins significative. Nous avons donc sélectionné 10 expressions non polysémiques ; nous estimions que 4 d'entre elles[73] étaient assez neutres et que les 6 autres[74]

[73] *rendre compte, résoudre DET problème, avoir besoin* et *poser DET question*

[74] *faire gaffe, porter DET accusation, porter plainte, traverser DET esprit, explorer DET possibilité* et *hausser DET épaules*

étaient davantage susceptibles d'être largement plus présentes dans un des genres que dans les deux autres. Ces 6 expressions comportaient parmi elles les 2 qui avaient été identifiées comme appartenant à des registres spécialisés par les experts de FLE ; elles ont été incluses à ce test pour cette raison, bien que n'ayant pas été retenues comme fondamentales contrairement aux autres items de test.

Nous avons ensuite extrait la distribution de ces 10 objets dans le corpus et avons obtenu les résultats représentés en Figure 5 ; ce graphique donne, en pourcentage, pour chacune des expressions sélectionnées, sa distribution sur les trois genres du corpus.

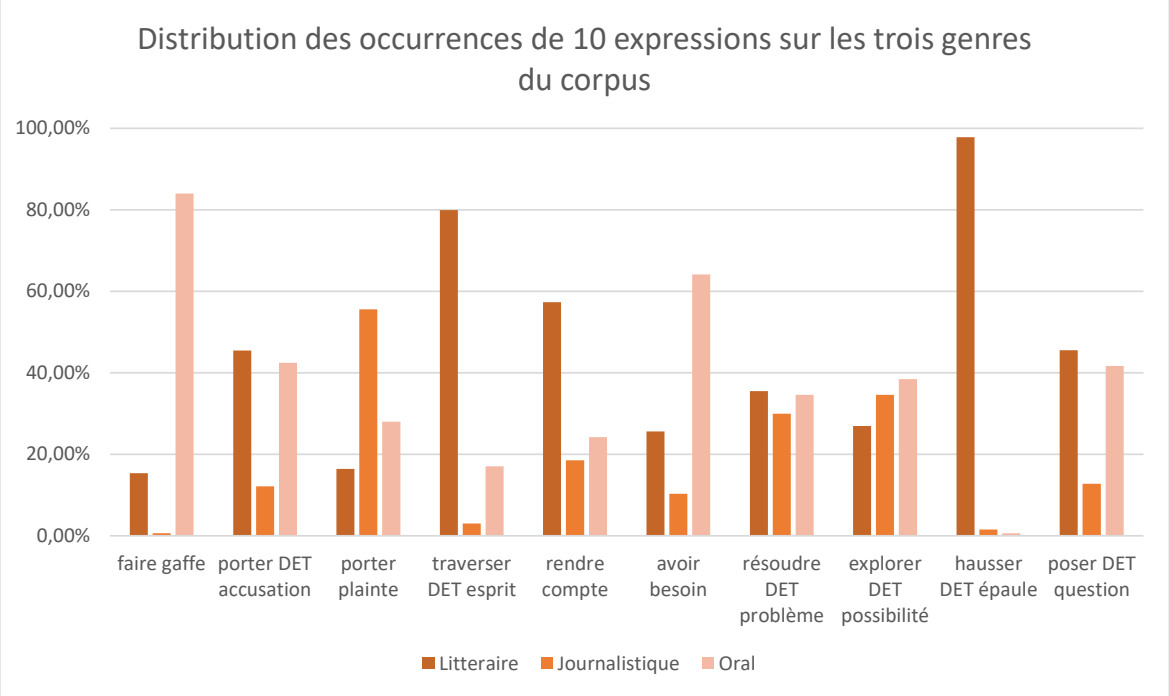


Figure 5 : Graphique représentant la distribution sur les trois genres du corpus des occurrences de 10 expressions tests

Nous constatons que la distribution de certaines expressions est effectivement très inégale, et que ce n'est pas nécessairement seulement le cas des expressions que nous avions envisagées ; notamment, plus de 60% des occurrences d'*avoir besoin* sont présentes dans le sous-corpus d'oral et de sous-titres contre environ 10% dans le sous-corpus journalistique. Les trois cas les plus caractéristiques de ce phénomène sont *hausser les épaules*, *faire gaffe* et *traverser l'esprit*.

Ces constatations signifient que notre mesure de dispersion ne remplit pas pleinement son rôle d'indicateur de neutralité du genre. Une solution aurait été d'établir un nombre ou pourcentage d'occurrences minimum par genre ou par sous-corpus. Cependant, appliquer ce correctif à cette étape de notre travail aurait signifié devoir recommencer intégralement le processus de désambiguïsation des expressions polysémiques afin d'obtenir des données cohérentes ; en effet, il aurait fallu réitérer l'attribution de chaque phrase comportant une

expression polysémique au sens qu'elle portait afin de recalculer la distribution en tenant compte de ce nouveau paramètre. Nous avons donc choisi de laisser cette mesure en l'état, mais précisons que sa valeur informative est à prendre avec précaution. Elle a tout de même été utile, ainsi calculée, à notre démarche de filtrage en permettant d'éliminer de nombreuses expressions. D'autre part, il faut rappeler que les résultats obtenus par ce test se basent sur des données que nous soupçonnions majoritairement d'être problématiques ; rien ne permet d'indiquer la proportion de ces expressions dans notre liste.

Une autre remarque, beaucoup plus anecdotique, est que nous remarquons que les deux expressions décrites comme spécialisées par certains enseignants de FLE ont des distributions bien moins inégales que d'autres.

A cette étape de notre étude, la liste finale des expressions considérées comme fondamentales était donc établie. Le nombre des items retenus permettait de fournir pour chacune une série d'informations descriptives que nous devions donc créer et modéliser. Cette démarche constituera l'objet de la dernière partie de ce mémoire.

Partie VI

-

DESCRIPTION DES EXPRESSIONS ET MODELISATION

Chapitre 13. Annotations et extractions de certaines caractéristiques des expressions polylexicales fondamentales

La description des caractéristiques propres à chaque expression a consisté en deux processus ; certaines informations ont été annotées, d'autres, extraites à partir du corpus.

1. Caractéristiques annotées

Les catégories d'informations que nous avons annotées relèvent de trois ordres : la classification des expressions, l'indication de certaines de leurs propriétés syntaxico-sémantiques et leur(s) définition(s). Cette annotation nous a permis de réaliser quelques observations générales sur la caractérisation des expressions polylexicales que nous avons retenues comme étant fondamentales. Nous exposons dans ce point les catégories annotées et les résultats obtenus.

1.1. Catégories d'informations

1.1.1. Classification

Nous avons tout d'abord jugé qu'il fallait annoter le type des expressions. Nous avons choisi une classification relativement simple, basée sur les trois types suivants :

- **Les collocations.** Cette classe regroupe les expressions qui ont les caractéristiques, largement décrites dans la littérature (voir notamment Grossmann & Tutin (2002)) :
    - D'être transparentes et non figées sémantiquement
    - D'être constituées de deux éléments dont un (la base) conserve son sens et impose une sélection du deuxième élément (le collocatif) et en définit le sens.

  Quelques exemples d'expressions annotées comme collocations sont *prendre_VERB DET risque_NOUN*, *mettre_VERB DET terme_NOUN* ou encore *avoir_VERB honte_NOUN*.
- **Les expressions figées.** Elles ont comme caractéristiques principales d'être relativement figées sur le plan syntaxique et d'avoir un sens non-compositionnel (même si parfois déductible par association métaphorique). Par exemple, ont été annotées comme expressions figées : *jeter_VERB DET œil_NOUN*, *perdre_VERB connaissance_NOUN* ou *tenir_VERB tête_NOUN*.
- **Les expressions métaphoriques.** Cette classe est constituée d'expressions qui, dans les contextes que nous avons conservés lors de la désambiguïsation, dénotent un sens métaphorique d'une action qui peut être interprétée au sens propre dans d'autres contextes. Par exemple, l'expression *tourner_VERB DET page_NOUN* signifie *passer à autre chose* dans les contextes retenus pour calculer ses valeurs statistiques. Mais, dans les contextes non retenus, cette séquence dénotait l'action de tourner

littéralement une page d'un livre. Certaines structures dans cette catégorie peuvent être apparentées à des expressions figées (*enfoncer_VERB DET clou_NOUN*), d'autres, davantage à des collocations (*marquer_VERB DET point_NOUN*). Un sous-découpage en deux classes (collocations métaphoriques/expressions figées métaphoriques) pourrait donc être envisagé pour affiner la classification.

1.1.2. Propriétés syntaxico-sémantiques

Deux propriétés syntaxico-sémantiques ont été annotées. La première est l'indication, lorsqu'il y avait lieu, de la présence d'un verbe support dans les collocations[75]. Des exemples de collocations à verbe support présents dans notre liste sont *avoir_VERB DET effet_NOUN*, *faire_VERB DET pause_NOUN* ou *prêter_VERB DET attention_NOUN*.

Les constructions à verbe support ayant ainsi été repérées, une deuxième annotation, celle des fonctions lexicales mel'čukiennes ,[76] a été réalisée sur notre liste par Agnès Tutin. Les FL proposent une approche intéressante à la notion de verbes supports en permettant de les formaliser par le triplet suivant (Alonso Ramos, 1999; Mel'čuk, 2003) :

- *Oper$_i$* qui prend pour mot clé le COD et donne pour valeur le verbe support dont le mot clé est COD
- *Func$_i$* qui prend pour mot clé le sujet et donne pour valeur le verbe support dont le mot clé est le sujet
- *Labor$_{ii}$*, qui prend pour mot clé le COI et donne pour valeur le verbe support dont le mot clé est COI

On obtient alors des formalisations tels que *Oper$_1$(attention)=prêter [à N]*, *Func$_2$(liste)=contient, comprend [N]* ou *Labor$_{12}$(défi)=mettre[N à ART]*[77] (*Ibid.*).

Etant donné que notre lexique ne prend en compte ni les sujets ni les COI, c'est la fonction *Oper* seulement qui a été utilisée pour donner un début de description à nos expressions. Ainsi, pour *avoir_VERB DET accident_NOUN*, l'annotation *Oper$_1$* est indiquée

---

[75] Voir 1.2. du Chapitre 5

[76] Les « *fonctions lexicales* » (FL), chez Mel'čuk, permettent de définir les préférences lexicales d'association, i.e. d'expliquer pourquoi un locuteur anglophone dira volontiers *heavy smocker* mais n'emploiera certainement pas l'expression *big smocker* (exemple emprunté à Granger et al., 2008). Une FL décrit la relation sémantico-lexicale entre la base et le collocatif sous la forme suivante : *f(base)=collocatif* (Mel'čuk, 2013, p.7). Ces FL sont nombreuses ; parmi les plus citées, la fonction « *Magn* » exprime l'intensité. Ainsi, *Magn(shave)=close* dénote la collocation dans laquelle on *rase* avec intensité. De la même manière, et pour un exemple en français, *Magn(blessé)=gravement,grièvement* représente les collocations dans lesquelles la blessure est intense (Granger et al., 2008; Mel'čuk, 2013).

[77] Les valeurs des indices attribués aux fonctions est ce que Mel'čuk appelle « *indice actanciel* ». Elles renvoient, dans ce cas, aux trois actants syntaxiques profonds du mot clés. Par exemple, dans le cas de *Oper$_1$(attention)=prêter [à N ~]*, l'indice *1* est utilisé pour indiquer que le sujet de la valeur est le premier actant syntaxique profond du mot clé *attention*.

et signifie qu'*accident* est le COD du verbe support *avoir* et que le sujet profond de l'expression est le premier actant syntaxique[78]. Certaines fonctions lexicales plus complexes ont été annotées, comme *IncepOper*$_1$ pour *prendre_VERB DET contact_NOUN*, et dans laquelle *Incep* signifie que l'expression désigne le début d'un état.

1.1.3. Définitions

Les dernières informations ajoutées sont des définitions des expressions. Nous avons utilisé pour cela les deux sources libres qui avaient été employées pour la création de la liste initiale[79], c'est-à-dire Wikitionnaire[80] et le DEM[81]. Si le DEM a été créé par des spécialistes, Wikitionnaire est en revanche une plateforme collaborative ; les définitions proposées ont donc été produites par des personnes nombreuses, anonymes, et dont on ne peut juger le niveau d'expertise.

De ces diverses annotations, nous pouvons tirer les quelques conclusions que nous présentons au sous-point suivant.

1.2. Conclusions sur les caractéristiques annotées : quantifications des types et analyse des possibilités de définition des expressions par les ressources utilisées

Comme le montre le Tableau 16, la majorité des expressions retenues comme fondamentales sont des collocations, ce qui tend à confirmer leur fréquence plus élevée dans la langue par rapport aux expressions figées. De plus, le 2/3 d'entre elles sont construites avec un verbe support.

Collocations	**339**
Dont verbes supports	*213*
Expressions figées	**65**
Expressions métaphoriques	**19**

Tableau 16 : Répartition des expressions fondamentales par type

En revanche, nous constatons que pour la majorité des collocations, aucune définition n'est donnée dans les deux ressources libres que nous avons utilisées (Tableau 17). A l'inverse, les expressions figées et métaphoriques, qui sont données par notre étude comme ne représentant qu'un tiers des expressions les plus fréquemment utilisées dans la langue, font majoritairement l'objet de définitions.

[78] Voir 2.4. du présent chapitre

[79] Voir point 2. du Chapitre 8

[80] https://fr.wiktionary.org/wiki/Wiktionnaire:Page_d%E2%80%99accueil

[81] http://rali.iro.umontreal.ca/rali/dem/

	Collocations	Expressions figées	Expressions métaphoriques	Total
DEM + Wkitionnaire	57	27	8	92
DEM seulement	38	9	2	49
Wikitionnaire seulement	65	23	3	91
Ni DEM ni Wikitionnaire	179	6	6	191

Tableau 17 : Nombres de définitions trouvées dans le DEM et/ou Wikitionnaire par types d'expressions

Ce manque de couverture dans le DEM est certainement dû au fait que cette ressource a été construite par une collecte de données sur des ressources majoritairement antérieures aux années 80 de siècle dernier (Dubois & Dubois-Charlier, 2011) ; l'intérêt lexicographique pour les collocations et leurs intégration à des dictionnaires étaient alors peu courants (Tutin, 2010b, p.109). Concernant Wikitionnaire, une explication possible à ce manque de couverture des collocations par rapport aux expressions figées est que les premières, ayant un sens plus transparent que les deuxièmes, seraient considérées comme n'ayant pas besoin d'être définies par les créateurs de la ressource. Pourtant, on y trouve les définitions de collocations dont on peut difficilement remettre en question la transparence comme *avoir envie, avoir honte, faire une apparition* ou *perdre la vie*. Il est alors difficile d'expliquer les raisons qui poussent ou non les collaborateurs à inclure une collocation dans cette gigantesque base de données qu'est Wikitionnaire.

Il faut en outre rappeler que ces deux ressources ne sont pas spécialisées dans la définition des expressions polylexicales ; Tutin (2005) et Netzlaff (2005, pp. 83-151) ont montré que les collocations les plus fréquentes dans les corpus étaient davantage couvertes par les dictionnaires dédiés à ces expressions. Cependant, les dictionnaires de collocations (comme par exemple, le *Lexique Actif du Français* (LAF) (Mel'čuk & Polguère, 2007)), s'ils proposent généralement un inventaire très fourni et structuré des collocatifs possibles pour une base donnée, ne donnent pas en revanche de définition facile d'accès pour les collocations formées. Dans le cas du LAF, le sens des collocations peut en effet être déduit à condition de s'approprier le formalisme relativement complexe qui est utilisé, mais une présentation simple du type *collocation : définition* n'est pas donnée. Autre ressource spécialisée dans les collocations, le *Dictionnaire des combinaisons de mots* (*Le Robert*)[82], n'est quant à elle pas basée sur un formalise complexe ; le sens des collocations est accessible par un regroupement synonymique des collocatifs de chaque base ainsi que (parfois) des

[82] *Dictionnaire des combinaisons de mots : les synonymes en contexte*, sous la direction de Dominique Le Fur, 2007, Paris : Dictionnaire le Robert

phrases d'exemple[83]. Cependant, ce dictionnaire (grand public à visée didactique) ne donne pas non plus des définitions pour chaque collocation. Ainsi, prenons l'exemple hypothétique d'un apprenant du français anglophone qui serait confronté à l'expression *passer un examen (scolaire)* et qui devrait comprendre que le verbe *passer* dénote ici le fait de se soumettre à l'examen et non de le réussir comme en anglais *to pass an exam* (qui est d'ailleurs une acception possible de ce verbe en français, ce qui rend la collocation moins transparente de par cette ambiguïté). S'il utilisait le LAF, il devrait commencer par comprendre le codage complexe, basé sur les fonctions lexicales, que propose cet ouvrage. S'il utilisait le *Dictionnaire des combinaisons de mots*, il devrait déduire le sens de la collocation à partir du regroupement de *passer* avec des verbes comme *se présenter à* ou *venir à* (p.148 de l'ouvrage). En revanche, une recherche sur Wikitionnaire lui fournirait directement le sens de la collocation.

Aussi, nous trouvons regrettable qu'il n'existe pas pour les collocations, à notre connaissance tout du moins, de ressources ayant la couverture des dictionnaires de collocations et la simplicité d'accès au sens que proposent les dictionnaires de langue générale. Nous estimons que cela serait utile dans une visée didactique et aurait permis en outre à notre liste de contenir des informations plus complètes, car c'est cette facilité d'accès au sens des expressions que nous voulions y inclure.

Après cette présentation des caractéristiques que nous avons annotées, et dont un exemple de résultat est consultable en Annexe 15, nous allons immédiatement introduire les caractéristiques que nous avons dû, quant à elles, extraire à partir du corpus.

2. Caractéristiques extraites

Nous souhaitions ajouter à notre description quelques caractéristiques dont nous avons préféré extraire les éléments de réponse à partir du corpus plutôt que de se fier à d'éventuelles sources lexicographiques et/ou suppositions. Chacune de ces extractions de caractéristiques a été réalisée par une projection de nos expressions sur le corpus à partir duquel elles avaient été extraites, à la recherche des informations souhaitées dans les contextes où elles apparaissaient. Nous avons donc procédé ainsi afin de déterminer, pour chaque expression, quelles étaient sa forme la plus courante, la variabilité numérique de son NObj et sa flexibilité syntaxique.

[83] Voir notamment Tutin (2010a) pour une description de ce dictionnaire et, plus généralement, du traitement dans collocations dans les dictionnaires qui y sont dédiés

2.1. Extraction des formes les plus courantes de chaque expression

Nous avons tout d'abord cherché à savoir quelle était la forme la plus fréquente sous laquelle une expression apparaissait. Les variations prises en compte entre les formes étaient le nombre du nom et les nombre et lemme du déterminant ; nous ne nous intéressions donc pas à la conjugaison du verbe. Notre extraction donnait également, lorsqu'il y avait lieu, la préposition ou conjonction qui apparaissait le plus fréquemment après l'expression ; elles devaient être présentes au moins 3 fois pour pouvoir apparaître en sortie. Nous avons ainsi trouvé, par exemple, la forme *prendre ses distances avec* pour *prendre_VERB DET distance_NOUN*. Les sorties qui nous paraissaient étranges ou incorrectes ont été vérifiées et corrigées à partir d'une observation des occurrences en contexte[84].

2.2. Extraction de la variabilité du nombre du complément d'objet nominal

Nous avons ensuite observé si les noms des expressions pouvaient être numériquement variables ; nous avons donc repéré pour chacune d'entre elles si au moins une occurrence au pluriel et une au singulier apparaissaient. Le script d'extraction fournissait en sortie les indications *SG*, *PL* ou *SG-PL* selon les résultats obtenus. Il créait, en outre, pour chaque couplet V NObj, un fichier qui contenait les contextes des occurrences au singulier, et un autre pour les occurrences au pluriel ; nous pouvions ainsi corriger d'éventuelles erreurs dues à l'étiquetage morphosyntaxique dans le corpus et donner des résultats corrects pour les expressions polysémiques (en repérant pour le sens voulu si des occurrences avec l'un ou l'autre des nombres étaient présentes).

Nous avons ainsi pu déterminer que 12 expressions[85] ne permettaient pas au nom d'être au singulier. Les cas où le nom ne pouvait pas être au pluriel étaient bien plus nombreux avec 181 expressions. La grande majorité de ces dernières ne comportaient pas de déterminant, mais ce n'était pas le cas de toutes ; on trouvait par exemple des expressions comme *avoir (une/de l') importance/*avoir des importances*.

2.3. Extraction et étude de la flexibilité syntaxique

Nous avons ensuite choisi d'analyser la possibilité de chaque expression de subir les trois alternances syntaxiques suivantes : passivation, relativisation et construction moyenne. A titre d'exemple, le verbe *arroser* et son complément d'objet direct *fleurs* se trouvent dans une construction passive dans la phrase (3), relative dans la phrase (4) et moyenne dans la phrase (5).

---

[84] Grâce aux fichiers html que nous avions créés (1.2. du Chapitre 10)

[85] *baisser les bras , baisser les yeux, dépasser les bornes , écarquiller les yeux , ne pas en croire ses yeux, faire les courses, faire les frais de, faire ses preuves, lever les yeux aux ciel, porter ses fruits , reprendre ses esprits, tourner les talons*

(3) *Les fleurs ont été arrosées hier*

(4) *Les fleurs qu'il arrosait tous les jours ont fané.*

(5) *Ces fleurs s'arrosent une fois par semaine*

Afin de chercher des occurrences de ces constructions, nous avons utilisé certaines relations de dépendances du corpus, à savoir :

- La relation *DEEPOBJ* pour la passivation. Cette dernière relie un verbe passif (tête) à son objet profond (dépendant), c'est-à-dire le sujet de sa forme passive. On aurait donc, pour la phrase (3), la relation *arrosées → DEEPOBJ → fleurs.*
- La relation *U3_OBJ* pour la relativisation. Elle relie le verbe (tête) à l'antécédent (dépendant) du pronom relatif *que*. Nous aurions donc pour la phrase (4) *arrosait → U3_OBJ → fleurs*.

Nous cherchions donc si chacun de nos couplets V N apparaissaient en étant reliés par ces relations.

En ce qui concerne la construction moyenne, nous cherchions des occurrences où :

- Le NObj d'un couplet était relié au verbe de ce couplet par une relation *SUBJ*.
- Le verbe était précédé du pronom réflexif *se*.

Dans cette configuration, nous aurions, pour la phrase (5), *arrosent → SUBJ → fleurs*, et le mot *se* serait présent à l'indice précédent celui du verbe *arrosent*.

Chaque alternance repérée au moins une fois pour chaque expression donnait lieu à l'indication « *oui* » et à une phrase d'exemple en sortie (la première extraite). De plus, pour cette étape également, nous avons créé pour chaque couplet des fichiers de vérification contenant toutes les phrases qui présentaient des occurrences de chaque alternance.

Nous avons vérifié les sorties en examinant si chaque phrase d'exemple donnée constituait une preuve de possibilité d'alternance (i.e., si elle n'était pas due à une erreur d'analyse syntaxique, et si le sens de l'expression en contexte était bien celui correspondant à l'entrée traitée). Si la phrase donnée ne permettait pas de certifier la possibilité de l'alternance, nous vérifiions dans les fichiers extraits si d'autres phrases extraites le permettaient. Ainsi, nous corrigions la sortie soit en supprimant l'indication « *oui* », soit en remplaçant la phrase d'exemple incorrecte par une phrase acceptable.

Cette étape de caractérisation nous a montré que seules 27 expressions supportaient les trois alternances, tandis que 143 n'en supportaient aucune. En outre, comme nous pouvons le voir dans le Tableau 18[86], toutes les combinaisons d'alternances permises et non permises sont possibles parmi les expressions que nous avons décrites.

[86] Pour plus de lisibilité, le tableau ne présente que des syntagmes d'exemple et non les phrases entières que nous avons extraites et conservées dans nos données

Expression lemmatisée	Passivation	Passivation - exemple	Relativisation	Relativisation - exemple	Construction moyenne	Construction moyenne - exemple
prendre_VERB DET habitude_NOUN	Oui	*[…] l' **habitude** était **prise** et rien ne l' aurait modifiée*	Oui	*[…] l' **habitude** que **prirent** rapidement nos […]*	Oui	*Des **habitudes se prennent** de tenir des listes .*
donner_VERB DET coup_NOUN	Oui	*Des **coups** auraient été **donnés** .*	Oui	*Le **coup** qu' il **donna** à la porte […]*	Non	
tourner_VERB DET page_NOUN	Oui	*La **page** du coup de froid , en tout cas , serait **tournée** .*	Non		Oui	*Avec lui une **page se tourne** […]*
dépasser_VERB DET borne_NOUN	Oui	*[…] quand les **bornes** sont **dépassées** , il n' y a plus de limites*	Non		Non	
éprouver_VERB DET besoin_NOUN	Non		Oui	*[…] le **besoin** irrépressible qu' il **éprouvait** de […]*	Non	
faire_VERB DET cuisine_NOUN	Non		Oui	*La cuisine qu' on **fait** dans ta ville pourrie […]*	Oui	*[…] la **cuisine se fait** dehors .*
faire_VERB appel_NOUN (2)	Non		Non		Oui	*L' **appel se fait** devant la Cour Suprême américaine .*
fermer_VERB DET oeil_NOUN	Non		Non		Non	

Tableau 18 : Différentes combinaisons de flexibilité syntaxique possibles avec exemples de syntagmes extraits du corpus

D'autre part, le tableau d'exemples qui est ici présenté contient des expressions ambigües. C'est notamment le cas de *fermer_VERB DET œil_NOUN*, dont l'entrée est celle qui a le sens de *ignorer volontairement/faire comme si on n'avait pas vu*. On voit que sa flexibilité syntaxique est nulle et dépendante de son sens. En effet, on trouve par exemple dans le corpus des occurrences du couplet dans des constructions relatives (« *[…]à l'angle d'un **oeil** que Cecilia **fermait** de temps en temps […]* »). Mais, ces occurrences ne correspondent pas au sens voulu. Ce type de phénomènes démontre, à notre sens, l'importance d'une désambiguïsation la plus précise possible dans un processus de caractérisation des expressions polylexicales.

Il nous faut également noter que, si certaines restrictions d'alternances syntaxiques sont déterminées par l'expression comprise dans son ensemble, d'autres sont dues aux propriétés syntaxiques du verbe seulement. Ainsi, de nombreuses constructions passives sont interdites non pas par les expressions elles-mêmes, mais par la présence du verbe *avoir* qui ne supporte pas la passivation.

Enfin, la dernière constatation qui peut être établie à partir des résultats de cette caractérisation est que de nombreuses expressions que nous avions établies comme figées sur le plan sémantique semblent faire preuve d'une certaine souplesse syntaxique. En effet 13 (sur 65) des expressions que nous avons identifiées comme figées (non-compositionnelles) semblent supporter au moins une alternance. Par exemple, on voit dans

le Tableau 18 que *dépasser les bornes* peut être passivé. A l'inverse, des collocations tout à fait compositionnelles et transparentes apparaissent figées dans leur flexibilité syntaxique. C'est le cas de *risquer sa vie*, qui n'apparait jamais dans d'autres constructions que la voix active standard. Ces éléments tendent donc à confirmer, à partir d'une observation sur corpus, que figements syntaxique et sémantique ne vont pas nécessairement de pair.

Il faut tout de même préciser les limites de l'approche *corpus-driven* que nous avons adoptée. En effet, si cette méthode nous a permis d'extraire des occurrences d'alternances que nous n'aurions peut-être pas considérées possibles de prime abord, nous pensons qu'il est peu probable que toutes les possibilités d'alternances aient été repérées.

Nous avons donc recueilli plusieurs informations à propos de nos expressions. Une caractérisation plus complexe que nous avons réalisée, celle de la sous-catégorisation, fera l'objet du chapitre suivant.

Chapitre 14. Extractions comparées de la sous-catégorisation

L'exploitation de corpus pour extraire la sous-catégorisation a été utilisée dès le début des années 90 pour l'anglais, mais avec des possibilités quantitatives limitées (Messiant, Gábor, & Poibeau, 2010). Aujourd'hui de nombreux travaux ont été réalisés dans cette optique. Les corpus étiquetés et analysés syntaxiquement permettent plusieurs modalités d'extraction. Kupsc (2007) répertorie trois grandes approches :

- L'exploitation des « *fonctions* » comme <SUJ, OBJ>, ce qui revient à dire, dans le cas de corpus analysés en dépendances, à l'exploitation des relations de dépendance.
- L'exploitation des suites de catégories.
- Les approches hybrides.

Nous avons choisi d'utiliser les deux premières parallèlement afin de les comparer, puis d'exploiter les deux sorties obtenues afin d'obtenir *in fine* des informations plus nombreuses et complémentaires. Après une brève présentation des repères théoriques autour du concept de sous-catégorisation et des ressources existantes traitant de ce sujet, nous présenterons les méthodes utilisées et leurs résultats.

1. Repères théoriques et ressources existantes

1.1. Définitions et ressources

La « *sous-catégorisation* » ou « *valence* »[87] d'un lexème correspond au « *cadre catégoriel* » dans lequel il apparaît (Jacob, 1984) ; il s'agit de la description de la

[87] Les deux termes sont hérités de traditions linguistiques différentes, à savoir, respectivement, les courants générativiste (Chomsky) ou structural (Tesnière)

construction standard dans laquelle il s'inscrit. Cette description est donnée à travers l'inventaire des arguments syntaxiques du lexème considéré. Par exemple, les « *traits de sous-catégorisation* » du verbe *hésiter* sont *SN1_à SVinf*, ce qui signifie que le verbe a deux « *arguments* » (ou « *actants* ») : un sujet, réalisé comme groupe nominal, et un complément d'objet indirect réalisé comme groupe prépositionnel verbal infinitif introduit par *à*. Ces traits permettent ainsi de déterminer que la phrase *J'ai hésité à partir* est correcte, tandis que **J'ai hésité de partir*, **J'ai hésité le départ* ou **J'ai hésité qu'il parte* ne le sont pas.

En TAL, les informations de sous-catégorisation sont très utiles. Elles permettent notamment d'améliorer considérablement la qualité de l'analyse syntaxique (voir, par exemple, Carroll & Fang, (2004)) et, par conséquent, celle des autres traitements qui en découlent comme la désambiguïsation. Plusieurs ressources existent pour le français, comme le *Lexique-grammaire* de Gross (1975), *Les Verbes français* (LVF) de Dubois & Dubois-Charlier (1997), ou encore le *Lexique des Formes Fléchies du Français* (Lefff) de Sagot (2010) qui décrit la sous-catégorisation à travers l'architecture *Alexina*.

En outre, on peut ajouter à la sous-catégorisation des « *traits de restriction de sélection* ». Ces derniers permettent de déterminer des caractéristiques sémantiques (par exemple, *animé/inanimé*) d'un argument sous-catégorisé. Ainsi, le verbe *dormir*, de par le trait *animé* qui accompagne son sujet (*SN1[animé]*) ne peut donner lieu à la phrase **La chaise dort*, étant donné que *chaise* n'est pas un être animé. De nombreuses ressources pour le français incluent des traits de restriction de sélection dans les informations syntaxiques qu'elles proposent ; c'est le cas, par exemple, du LVF ou, de façon plus rudimentaire, du DICOVALENCE (Mertens, 2010).

1.2. La sous-catégorisation des expressions polylexicales dans les ressources pour le français

Les ressources et formalismes que nous avons cités dans les paragraphes précédents ne comportent pas tous des expressions polylexicales. Elles sont par exemple absentes dans DICOVALENCE.

Des tables du Lexique-grammaire traitent quant à elles d'expressions dites « *figées* » ; les entrées sont nombreuses et incluent, en réalité, aussi bien des expressions effectivement figées que des collocations, traitées de manière indifférenciée. La sous-catégorisation des expressions que nous souhaitons réaliser sera associée au typage préalablement réalisé, ce qui permettra d'étudier les différences syntaxiques entre collocations et expressions figées à ce niveau. De plus, cette ressource a été réalisée manuellement et par expertise, sans recours à un corpus, ce qui constitue une différence notable avec notre travail.

On peut également citer le *Dictionnaire explicatif et combinatoire* (Mel'čuk & Arbatchewsky-Jumarie, 1999) qui, dans sa description des collocations traite séparément de la combinatoire collocative (grâce aux fonctions lexicales) et de la combinatoire syntaxique ;

ces deux descriptions sont réalisées au niveau des lexies simples et nous avons donc, pour chacune d'entre elles, d'un côté sa sous-catégorisation et, de l'autre, la liste de ses collocatifs. On ne trouve pas, en revanche, la sous-catégorisation des collocations résultantes. Par exemple, les différentes fonctions lexicales appliquées au nom *besoin* permettent de donner les verbes collocatifs *avoir* ou *éprouver*, et la sous-catégorisation de *besoin* est également donnée. Mais, les sous-catégorisations des expressions *avoir besoin* et *éprouver le besoin* ne sont pas présentées. Dans le cas de certaines collocations à verbe support, ces dernières peuvent être déduites à partir de deux niveaux de description dans la mesure où la sous-catégorisation de l'expression correspond à celle du nom prédicatif (dans l'exemple donné, *de SN/de Vinf/que VS*). Mais, cela n'est pas le cas des collocations qui n'ont pas cette caractéristique (par exemple, le complément *à SN* de la collocation *adresser la parole* n'est pas régi par la lexie simple *parole*). Nous souhaitons, pour notre part, décrire la sous-catégorisation de nos expressions en les envisageant comme des éléments syntaxiques formant des unités capables de régir des compléments.

Ayant ainsi présenté quelques caractéristiques de la sous-catégorisation et défini l'objet que nous souhaitions extraire, nous présenterons à présent les deux méthodes comparées qui ont été employées à cet escient.

2. Extraction par méthodes comparées

Les deux approches qui ont été utilisées sont celles que nous avons déjà opposées dans cette étude, à savoir, la syntaxe et les patrons catégoriels. Nous présentons dans ce point les fonctionnements respectifs de ces deux méthodes, puis la comparaison des résultats.

2.1. Méthodes et fonctionnement des algorithmes

Les deux méthodes utilisées reposent sur des principes généraux identiques et se distinguent de par l'information qu'ils exploitent.

2.1.1. Principes généraux

Les deux algorithmes utilisés fonctionnent tous deux, dans les grandes lignes, de la même manière ; ils prennent en entrée la liste des expressions et parcourent le corpus à leur recherche. Chaque fois qu'une d'entre elles est rencontrée, le sujet est recherché en distinguant les noms et les pronoms. Lorsqu'il s'agit d'un pronom, son lemme est stocké en mémoire. Puis, les deux algorithmes recherchent les compléments prépositionnels et les complétives liés à l'expression.

2.1.2. Différence entre les deux méthodes

La **méthode par syntaxe** utilise les relations de dépendance. Pour identifier le sujet, nous avons cherché dans les phrases le nom ou le pronom relié au verbe de l'expression traitée par une relation *SUBJ*. Pour les compléments prépositionnels, nous avons utilisé les relations *U3_(lemme de la préposition)_NMOD* présentes entre le nom de l'expression et le nom ou le verbe du groupe complétif. Nous avons également entrait le lemme de la préposition en utilisant la relation *PREPOBJ* qui relie le nom ou le verbe à l'infinitif du groupe prépositionnel à sa préposition. Par exemple, le syntagme *il avait besoin d'aide* est analysé dans le corpus de la manière représentée en Figure 6. Nous utilisons donc la relation *SUBJ* pour identifier le pronom *il*. Puis, nous utilisons *U3_DE_NMOD* pour identifier *aide*. Enfin, nous utilisons *PREPOBJ* pour identifier la préposition *d'* à partir du nom *aide*.

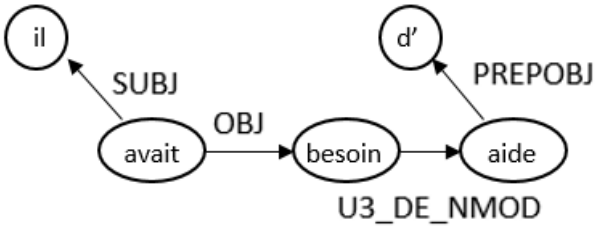


Figure 6 : Analyse syntaxique du segment « *il avait besoin d'aide* » dans le corpus

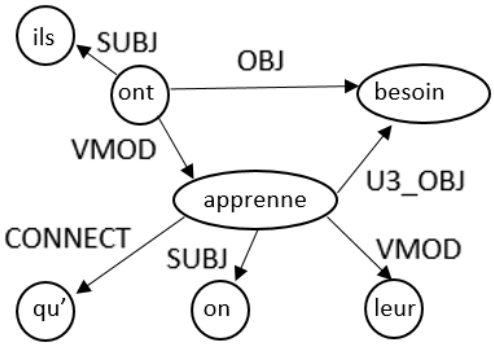


Figure 7 : Analyse syntaxique du segment « *ils ont besoin qu'on leur apprenne* » dans le corpus

En ce qui concerne les complétives, nous avons utilisé soit la relation *VMOD* qui lie le verbe de la proposition principale à celui de la complétive, soit la relation *U3_OBJ* qui relie le verbe de la complétive au nom de la principale ; utiliser ces deux possibilités permet d'obtenir davantage de résultats malgré les erreurs d'analyse syntaxique possibles. Puis, nous avons utilisé la relation *CONNECT* qui relie le verbe de la complétive à la conjonction *que*, en vérifiant que le lemme est bien *que*. Par exemple, le syntagme *ils ont besoin qu'on leur apprenne* est analysé comme représenté en Figure 7. Le verbe *apprenne* est donc repéré soit par la relation *VMOD* à partir du verbe *ont*, soit par la relation *U3_OBJ* à partir du nom

besoin (dans ce cas, le verbe *apprenne* est tête de la relation). Ensuite, on utilise la relation *CONNECT* pour trouver *qu'* et on vérifie que son lemme est bien *que*.

La **méthode par patrons catégoriels** consiste à chercher linéairement des suites de mots ayant des catégories syntaxiques définies et qui gravitent autour de l'expression concernée.

Pour les sujets, nous avons cherché, placés avant le verbe, les patrons suivants :

- [PRONSuj (ne)] : par exemple *[il (ne)] mange (pas)*
- [PRONSuj (ne) PRON] : par exemple *[il (ne) le] mange (pas)*
- [NOM (ne)] : par exemple *le jeune [homme (ne)] mange (pas)* ou *[Jean (ne] mange (pas)*
- [NOM (ne) PRON] : par exemple *le jeune [homme (ne) le] mange (pas)* ou *[Jean (ne) le] mange (pas)*

où *PRONSuj* ne peut être que *je, tu, il, elle, on, nous, vous, ils, elles, ça* et *cela* (tandis que *PRON* peut être n'importe quel pronom), et où l'adverbe de négation *ne* peut être présent ou non dans ces suites.

PATRONS	EXEMPLE DE REALISATION
1. PREP Vinf	J'ai besoin **de manger**
2. PREP NOM	J'ai besoin **de Jean**
3. PREP PRON	J'ai besoin **de toi**
4. PREP DET NOM	J'ai besoin **d'un toit**
5. PREP DET (ADJ NOM/NOM ADJ)	J'ai besoin **d'une (petite table/table verte)**
6. que PRONSuj VERBE	J'ai besoin **que tu viennes**
7. que NOM VERBE	J'ai besoin **que Jean vienne**
8. que DET NOM VERBE	J'ai besoin **que le garçon vienne**
9. que DET (ADJ NOM/NOM ADJ) VERB	J'ai besoin **que la (petite fille/fille joyeuse) vienne**
10. que PRONSuj ne VERBE	J'ai besoin **que tu ne sois** (pas ...)
11. que NOM ne VERBE	J'ai besoin **que Jean ne soit** (pas...)
12. que DET NOM ne VERBE	J'ai besoin **que la fille ne soit** (pas...)
13. que DET (ADj NOM/NOM ADJ) ne VERBE	J'ai besoin **que la (petite fille/fille joyeuse) ne soit** (pas...)
14. que PRONSuj PRON VERBE	J'ai besoin **que tu lui dises**
15. que NOM PRON VERBE	J'ai besoin **que Jean y aille**
16. que DET NOM PRON VERBE	J'ai besoin **que la fille nous voie**
17. que DET (ADJ NOM/NOM ADJ) PRON VERBE	J'ai besoin **que la (petite fille/fille joyeuse) nous voie**
18. que PRONSuj ne PRON VERBE	J'ai besoin **que tu ne lui parles** (pas)
19. que NOM ne PRON VERBE	J'ai besoin **que Jean n'y aille** (pas)
20. que DET NOM ne PRON VERBE	J'ai besoin **que la fille ne nous voie** (pas)
21. que DET (ADJ NOM/NOM ADJ) ne PRON VERBE	J'ai besoin **que la (petite fille/fille joyeuse) ne nous voie** (pas)

Tableau 19 : Liste des patrons de compléments prépositionnels et complétives utilisés pour extraire la sous-catégorisation des expressions – Patrons et exemples de réalisations correspondants

Pour les compléments prépositionnels et les complétives, nous utiliserons les patrons définis dans le Tableau 19. Ainsi, pour reprendre les phrases données en exemple pour la méthode syntaxique, la sous-catégorisation dans *il avait besoin d'aide* sera repérée grâce au

patron 2) *[PREP NOM]*, tandis que *ils ont besoin qu'on leur apprenne* le sera grâce au patron 14) *[que PRONSuj PRON VERBE]*.

2.1.3. Sortie

Quelle que soit la méthode utilisée, les scripts fournissaient en sortie, pour chaque expression :

- pour le sujet :
    - la mention *SNnominal* lorsqu'au moins un sujet nominal a été rencontré
    - la mention *SNpronominal*(liste des pronoms rencontrés) ; par exemple *SNPronominal(je, tu, cela, nous)*.

  Nous obtenions, en outre un fichier qui comportait les phrases de chaque couplet V NObj classées par pronom sujet.
- Pour les compléments, nous indiquons la catégorie de la tête du complément et la préposition ou conjonction, par exemple *NOUN(sur), PVinf(de)* ou *VERB(que)*. Il est à noter que les verbes à l'infinitif sont repérés comme tels par les scripts grâce aux informations morphologiques présentes dans le corpus.

De plus, il était fortement souhaitable que chaque complément rencontré n'apparaisse pas en sortie. En effet, la sous-catégorisation étant basée sur le caractère quasi obligatoire de la complémentation, la fréquence d'apparition devait être élevée. Afin de prendre en compte les possibles erreurs d'étiquetage, d'analyse syntaxique et de non linéarité de la complémentation pour le cas de l'analyse par patrons, nous procédions de la façon suivante : chaque complémentation devait apparaitre un certain nombre de fois, celui-ci étant relatif à la fréquence de l'expression traitée dans le corpus. Afin de tester ces seuils, nous procédions, pour chacune des deux méthodes, à une extraction ou chaque complémentation devait apparaître dans au moins 10% des cas où apparaissait l'expression et une autre où elle devait apparaître dans 1/3 des cas.

Enfin, nous créions, pour chaque complémentation de chaque expression, un fichier contenant les phrases dans lesquelles l'extraction avait eu lieu, afin de procéder à des vérifications.

2.2. Comparaison des résultats

Afin de comparer les résultats des deux méthodes, nous avons extrait les résultats de 24 expressions de fréquences plus ou moins élevées.

Cette comparaison (dont le détail est donné en Annexe 16) a permis de constater que, concernant les sujets, les résultats étaient sensiblement similaires quelle que soit la méthode utilisée. La méthode par syntaxe permettait en revanche de fournir un nombre plus important

de pronoms, ce qui est tout à fait normal étant donné que nous avions restreint les possibilités dans la méthode par patrons syntaxiques.

Les résultats concernant les compléments sont quant à eux détaillés en Annexe 17, et résumés dans le Tableau 20. Comme le montre ce tableau, il semble que la méthode donnant les meilleurs résultats soit celle par patrons catégoriels avec un seuil fréquentiel de 10% de la fréquence de l'expression. De plus, le Tableau 21 démontre, comme il était prédictible, que les deux méthodes appliquées avec des seuils de fréquence élevés provoquaient davantage de silences tandis qu'appliquées avec des seuils plus bas, elles engendraient plus de bruits. Cependant, pour un même seuil de 10%, la méthode par patrons provoquait moins de silences et moins de bruits que la méthode par syntaxe.

	Correct	Partiellement correct	Incorrect
Patrons - Tiers	12	5	7
Patrons - Dixième	**18**	**2**	**4**
Syntaxe - Tiers	14	3	7
Syntaxe - Dixième	14	3	7

Tableau 20 : Résultats des comparaisons d'extraction des compléments des expressions par différentes méthodes et seuils de fréquence

	Silence	Bruit
Patrons - Tiers	37,50%	8,33%
Patrons - Dixième	4,17%	25,00%
Syntaxe - Tiers	37,50%	4,17%
Syntaxe - Dixième	12,50%	33,33%

Tableau 21 : Taux de bruits et de silences produits dans les sorties d'extraction des compléments des expression par différentes méthodes et seuils de fréquence

En outre, si l'on s'intéresse à la nature des groupes linguistiques repérés ou non par les différentes méthodes, il ne semble pas y avoir de différences significatives ; on aurait pu, par exemple, s'attendre à ce qu'une de ces méthodes soit plus adaptée à la recherche des complétives, or, dans l'échantillon étudié, le taux de repérage est équivalent.

Ces deux méthodes ayant été évaluées, il s'est alors agi d'exploiter leurs résultats pour sélectionner les informations pertinentes à utiliser pour caractériser la sous-catégorisation de nos expressions.

3. Sélection des informations de sous-catégorisation pertinentes

Afin d'exploiter au mieux les résultats fournis par les différentes méthodes, nous avons sélectionné, pour chaque expression, les compléments régis pertinents en vérifiant, lorsque nécessaire, les réalisations de ces compléments dans les phrases extraites par les scripts.

Lorsque nous rencontrions une expression ayant plusieurs compléments régis qui correspondaient à des actants différents, nous le spécifions en numérotant différemment les deux actants. Ainsi, l'expression *donner accès* a deux compléments régis (*donner accès à quelque chose à quelqu'un* comme dans la phrase *Il a donné accès au coffre à Marie*) ; nous annotions donc *SN2(à)/SN3(à)*. A l'inverse, nous gardions le même numéro lorsque les compléments régis correspondent à des réalisations différentes du même actant. Par exemple, nous annotions *VS2(que)/SN2(de)* pour l'expression *prendre conscience*.

Ainsi, nous avons par exemple sélectionné :

- *SN2(de)/SN3(avec)* pour *aborder_VERB DET sujet_NOUN*
- *SN2(sur)* pour *faire_VERB pression_NOUN*
- *VInf2(de)* pour *valoir_VERN DET peine_NOUN*

Pour un aperçu des sous-catégorisations retenues à partir de celles proposées, nous renvoyons le lecteur à l'Annexe 18.

En ce qui concerne les sujets, l'intégralité de ceux qui ont été extraits consistait en des groupes nominaux ou pronominaux. Jugeant qu'il est intéressant d'ajouter des traits de restriction de sélection, avons utilisé pour cela les pronoms extraits. La méthode par syntaxe donnait des résultats plus pertinents pour cette tâche dans la mesure où elle permettait de fournir une liste plus complète des sujets pronominaux. En utilisant cette sortie donc, et en vérifiant les occurrences extraites, nous avons procédé de la manière suivante :

- Les expressions ayant pour sujets *je, tu, nous* ou *vous* et jamais *cela* ou *ceci* ont été considérées comme ayant seulement des sujets humains ; nous annotions *SN1[hum]*
- A l'inverse, les expressions ayant parmi leurs sujets *ceci* ou *cela*[88] et jamais *je, tu, nous* ou *vous* ont été considérées comme ayant seulement des sujets inanimés ; nous annotions *SN1[ina]*
- Enfin, pour les expressions ayant des occurrences de chacun des deux groupes de sujets, nous annotions *SN1[hum|ina]*

Quelques exemples de cette annotation à partir des pronoms extraits sont présentés dans le Tableau 22.

[88] Le pronom *ça* n'était pas exploitable à cause des nombreuses occurrences, de registres de langue familiers, dans lesquelles il est utilisé pour parler d'une personne, comme dans la phrase « *Les hommes [...] faut toujours* ***que ça fasse les malins***. »

Expression lemmatisée	Actant 1	SN1
attirer_VERB DET attention_NOUN	SN1[hum|ina]	nous,il,je,cela,on,vous,celui,l'un,quelque chose,celui-ci,qui,tout,ça,chacun,me,tu,rien,le,certains,un peu,qui est-ce qui,
avoir_VERB DET sens_NOUN	SN1[ina]	tout,cela,ça,il,ce,celui-ci,tout ce,peu,lui,qui,rien,l'un,celui-là,celui,ceci,
faire_VERB DET bien_NOUN	SN1[hum|ina]	cela,je,ça,il,on,tout,celui-là,nous,tu,ce,vous,quelque chose,leur,un peu,tout ce,
avoir_VERB DET raison_NOUN	SN1[hum|ina]	il,tu,vous,le,je,tout,l'un,nous,quiconque,même,chacun,celui,leur,on,moi,celui-ci,rien,ce,personne,qui,lui,quelqu'un,cela,quelque chose,ça,quel,tout ce,toi,beaucoup,
avoir_VERB DET importance_NOUN	SN1[hum|ina]	cela,il,tout,ça,rien,on,je,ce,tu,vous,ceci,tout ce,aucun,qui est-ce qui,nous,quelqu'un,celui-ci,le nôtre,
rendre_VERB visite_NOUN	SN1[hum]	personne,il,celui,nous,l'un,moi,je,on,tout,vous,d'autres,tu,ce,lui,leur,le,celui-ci,toi,ça,qui,

Tableau 22 : Extrait des résultats de l'ajout des traits de restriction de sélection aux actants sujets à partir des pronoms sujets extraits dans le corpus

A cette étape du travail, une dernière extraction nous a paru nécessaire ; afin d'obtenir des données modélisables selon le formalisme que nous avons choisi d'adopter, et qui sera détaillé dans le point 1. du prochain chapitre, nous avons décidé d'extraire les possibilités de cliticisation des compléments régis.

4. Extraction des possibilités de cliticisation des pronoms régis

4.1. Motivations

La ressource DICOVALENCE, qui propose des cadres de sous-catégorisation de nombreux verbes, a été réalisée dans le cadre de l'Approche Pronominale (par exemple, Blanche-Benveniste (1984)). Cette dernière utilise la pronominalisation des compléments afin de décrire la valence des prédicats. Pour en comprendre les bases, il faut partir du constat que tous les compléments régis ne sont pas cliticisables par les mêmes pronoms (i.e. remplaçables pas les mêmes pronoms clitiques) et, dans certains cas, ils ne sont pas cliticisables du tout. Prenons le verbe simple *boire* ; son complément objet régi est un groupe nominal qui peut être réalisé, par exemple, par *un verre d'eau*. Ce complément peut être remplacé par un pronom clitique accusatif (*le* dans *le boire*), mais pas par le pronom clitique locatif *y* (**y boire*) ni par un datif (*lui* dans **lui boire*). Dans le modèle de DICOVALENCE, les verbes sont regroupés en paradigmes qui définissent les pronoms avec lesquels ils peuvent apparaître en cooccurrence. Par exemple, les verbes qui s'inscrivent dans le paradigme P2 régissent un complément d'objet indirect introduit par *à* et sa cliticisation est possible soit par un clitique datif, soit par le clitique locatif *y*, soit impossible (Sagot & Danlos, 2008). Le verbe *boire* ne fait donc pas partie du paradigme P2, les cliticisations permises par ce paradigme étant refusées à son complément régi, qui n'est donc pas un COI introduit par *à*. Ces paradigmes ont en partie inspiré la création de la liste des fonctions

syntaxiques du Lefff (*Ibid.*) que nous avons utilisées par la suite pour modéliser la sous-catégorisation de nos expressions (voir point 1. du Chapitre 15).

Nous avons donc jugé utile d'extraire les possibilités de cliticisation des compléments régis de nos expressions.

4.2. Extraction

Nous avons réalisé un script d'extraction basé sur des patrons catégoriels ; cette méthode ayant déjà fait ses preuves et les pronoms clitiques ne pouvant pas, par définition, être éloignées du verbe dans la phrase, nous n'avons pas jugé utile d'avoir recours à la syntaxe. Nous recherchions donc dans le corpus, autour des verbes de nos expressions, des clitiques datifs, le génitif *en* et le locatif *y*. Lorsqu'ils apparaissaient en sortie pour une expression, nous vérifiions la pertinence de l'extraction et annotions la possibilité de cliticisation.

4.3. Exemple de possibilité offerte par l'extraction

L'extraction de cette information pourrait servir, dans une certaine mesure, à l'ajout de traits de restriction de sélection. Par exemple, lorsqu'un objet indirect introduit par *à* n'est pas cliticisable par un pronom datif, il est permis d'en déduire que le trait humain lui est refusé. C'est le cas par exemple de *prendre part* dont le complément régi est *SN2(à)*. Ce dernier n'est pas cliticisable par le datif *lui* (**lui prendre part*), et on peut donc en déduire que ce complément ne peut avoir un référent humain.

Dans ce chapitre et dans le précédent, nous avons présenté les annotations et extractions de diverses caractéristiques de nos expressions polylexicales fondamentales. Une modélisation de ces informations s'est avérée nécessaire.

Chapitre 15. Modélisation

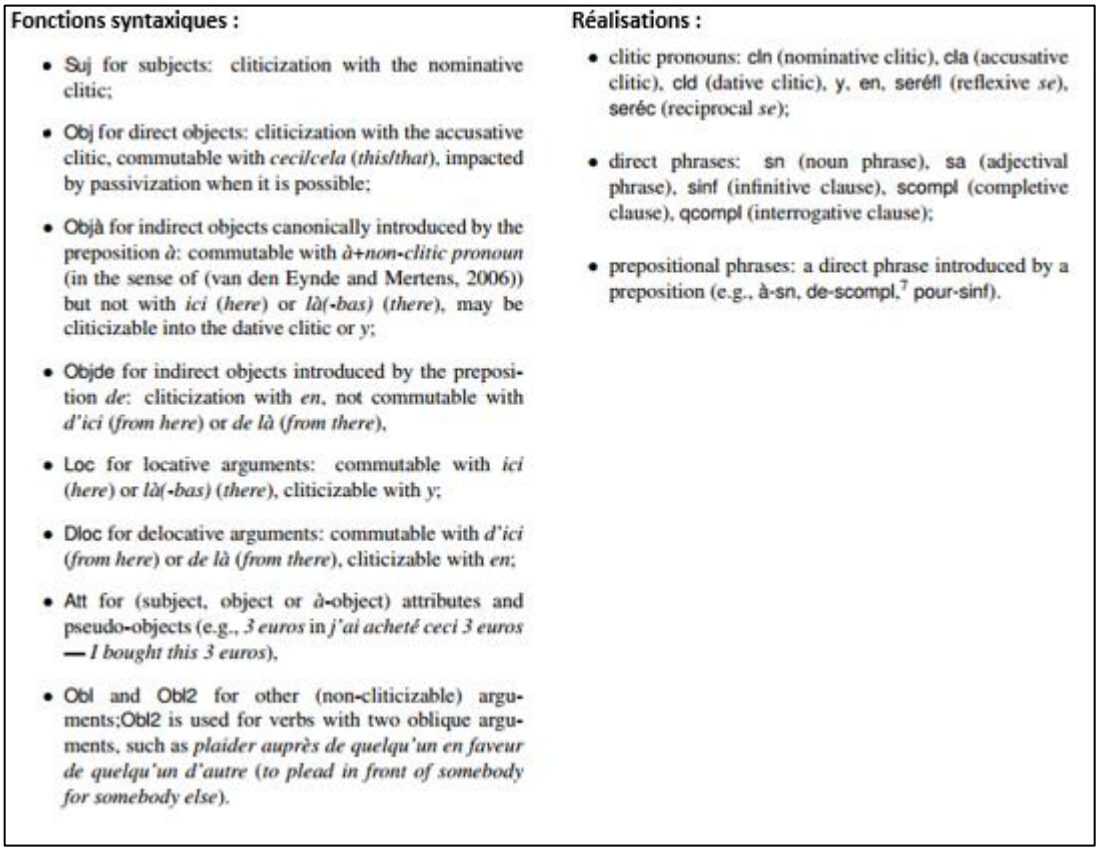
Dans ce chapitre nous présenterons premièrement la modélisation que nous avons adoptée pour décrire la sous-catégorisation des expressions. Puis, nous proposerons un modèle de représentation de toutes les informations que nous avons récoltées dans le cadre de notre étude.

1. Modélisation de la sous-catégorisation à partir du modèle du Lefff et de Dicovalence et observations permises

1.1. Présentation du formalisme Alexina et conversion des annotations

Comme nous l'avons précédemment évoqué, nous avons choisi d'utiliser le modèle Alexina du Lefff (Sagot, 2010) afin de retranscrire l'information de sous-catégorisation. Ce

modèle requiert d'indiquer les fonctions syntaxiques des compléments régis (COD, COI, etc.) accompagnées de leurs possibles réalisations ; le jeu d'étiquettes prévu à cet escient est présenté dans la capture d'écran réalisée sur la publication de Sagot (2010) et reproduit en Figure 8.



Fonctions syntaxiques :

- Suj for subjects: cliticization with the nominative clitic;
- Obj for direct objects: cliticization with the accusative clitic, commutable with *cecilcela* (*thislthat*), impacted by passivization when it is possible;
- Objà for indirect objects canonically introduced by the preposition *à*: commutable with *à+non-clitic pronoun* (in the sense of (van den Eynde and Mertens, 2006)) but not with *ici* (*here*) or *là(-bas)* (*there*), may be cliticizable into the dative clitic or *y*;
- Objde for indirect objects introduced by the preposition *de*: cliticization with *en*, not commutable with *d'ici* (*from here*) or *de là* (*from there*),
- Loc for locative arguments: commutable with *ici* (*here*) or *là(-bas)* (*there*), cliticizable with *y*;
- Dloc for delocative arguments: commutable with *d'ici* (*from here*) or *de là* (*from there*), cliticizable with *en*;
- Att for (subject, object or *à*-object) attributes and pseudo-objects (e.g., *3 euros* in *j'ai acheté ceci 3 euros* — *I bought this 3 euros*),
- Obl and Obl2 for other (non-cliticizable) arguments;Obl2 is used for verbs with two oblique arguments, such as *plaider auprès de quelqu'un en faveur de quelqu'un d'autre* (*to plead in front of somebody for somebody else*).

Réalisations :

- clitic pronouns: cln (nominative clitic), cla (accusative clitic), cld (dative clitic), y, en, seréfl (reflexive *se*), seréc (reciprocal *se*);
- direct phrases: sn (noun phrase), sa (adjectival phrase), sinf (infinitive clause), scompl (completive clause), qcompl (interrogative clause);
- prepositional phrases: a direct phrase introduced by a preposition (e.g., à-sn, de-scompl,[7] pour-sinf).

Figure 8 : Fonctions syntaxiques et réalisations dans le Lefff – capture d'écran réalisée sur la publication de Sagot (2010)

Avant de montrer en quoi ce modèle résulte lorsqu'il est appliqué à nos expressions, nous jugeons utile de donner un exemple avec un verbe simple pour bien expliciter en quoi il consiste. Nous utiliserons le verbe *boire* précédemment donné comme exemple. Dans les données du Leff issues de la conversion du DICOVALENCE, sa sous-catégorisation est donnée ainsi : <*Suj:sn|cln,Obj:sn|cla,Obl:dans-sn*>[89]. Elle est en outre accompagnée de la phrase d'exemple « *il boit de la soupe dans un verre* ». Cette annotation signifie donc que :

- la fonction de sujet (*Suj*), peut être réalisée par un groupe nominal ou un clitique nominatif (*sn|cln*).
- la fonction d'objet direct (*Obj*) peut être réalisée par un groupe nominal ou un clitique accusatif (*sn|cla*)

[89] L'annotation a été simplifiée par le retrait de certaines parenthèses qui indiquaient le caractère optionnel des compléments et que nous n'avons pas utilisées dans notre modélisation des données.

- la fonction de complément oblique (*Obl*) est réalisée par un groupe prépositionnel nominal introduit par *dans* (*dans-sn*), comme pour « *dans un verre* »

Pour donner à présent un exemple de modélisation d'une expression selon ce modèle, prenons la sous-catégorisation de l'expression *rendre visite*. Elle était, selon notre annotation, décrite de la manière suivante *SN1[hum|ina]_SN2(à)*. Nous avions en outre extrait et annoté le fait que le complément régi était cliticisable par le pronom datif *lui*. Nous obtenons donc le cadre de sous-catégorisation modélisé : *<Suj:sn|cln,Objà:a-sn|cld>*, ce qui signifie que son sujet peut être un groupe nominal ou un clitique nominatif (*Suj:sn|cln*, anciennement *SN1*) et qu'elle régit un complément d'objet indirect introduit par la préposition *à* cliticisable par un pronom datif (*Objà :a-sn|cld*, anciennement *SN2(à)* + l'information de possibilité de cliticisation en pronom datif).

Une remarque à émettre est que le modèle Alexina n'intègre pas les traits de restriction de sélection. Ainsi, lorsque les tables du Lexique-grammaire et les sous-catégorisations du DICOVALENCE ont été converties et intégrées au Lefff (Danlos, Sagot, *et al.*, 2007), cette information a été supprimée des données converties. Nous avons fait de même en passant à ce formalisme, et avons donc supprimé les traits de restriction des sujets. Ce choix a été motivé par une volonté de rester fidèle au règles d'Alexina, mais il ne serait pas exclus d'y déroger en annotant, par exemple, pour *rendre visite*, *<Suj[hum|ina]:sn|cln,Objà:a-sn|cld>*.

Pour l'instant, dans l'optique de conserver cette information, nous garderons dans nos données les deux représentations produites. Cela permettra également d'indiquer le numéro des différents actants, ce que Leff ne prévoit pas. Par exemple, les deux actants de *donner accès* précédemment évoqués résultent en une seule annotation dans la transcription au format Leff ; on passe de *SN1[hum|ina]_SN2(à)_SN3(à)* à *<Suj:sn|cln,Objà:à-sn|cld>*, perdant ainsi l'information de la présence de deux actants compléments distincts. Ainsi, les informations complètes conservées donnent lieu aux quelques descriptions d'exemple données dans le Tableau 23.

Expression lemmatisée	Actant 1	Actant 2	Actant 3	Pronominalisation des compléments	Alexina
jeter_VERB DET coup d'oeil_NOUN	SN1[hum]	SN2(à)		PRON_DATIF/y	<Suj:sn|cln,Objà:à-sn|cld|y>
avoir_VERB DET mal_NOUN	SN1[hum|ina]	Vinf2(à)			<Suj:sn|cln,Objà:à-sn>
avoir_VERB DET problème_NOUN	SN1[hum|ina]	SN2(avec)			<Suj:sn|cln,Obl:avec-sn>
faire_VERB attention_NOUN	SN1[hum]	SN2(à)		y	<Suj:sn|cln,Objà:a-sn|y>
avoir_VERB tort_NOUN	SN1[hum]	Vinf2(de)/SN2(sur)			<Suj:sn|cln,Objde:de-sinf,Obl:sur-sn>
hausser_VERB DET épaule_NOUN	SN1[hum]				<Suj:sn|cln>
avoir_VERB DET choix_NOUN	SN1[hum]	SN2(de)/SVinf2(de)		en	<Suj:sn|cln,Objde:de-sn|de-sinf|en>
avoir_VERB DET intention_NOUN	SN1[hum]	SVinf2(de)		en	<Suj:sn|cln,Objde:de-sinf|en>
faire_VERB DET mal_NOUN	SN1[hum|ina]	SN2(à)		PRON_DATIF	<Suj:sn|cln,Objà:a-sn|cld>
prendre_VERB DET décision_NOUN	SN1[hum]	SVinf2(de)			<Suj:sn|cln,Objde:de-sinf>
faire_VERB DET tour_NOUN (1)	SN1[hum|ina]	SN2(de)		en	<Suj:sn|cln,Objde:de-sn|en>
sauver_VERB DET vie_NOUN	SN1[hum|ina]	SN2(de)/SN2(à)		PRON_DATIF	<Suj:sn|cln,Objede:de-sn,Objà:à-sn|cld>
faire_VERB part_NOUN	SN1[hum]	SN2(de)	SN3(à)	PRON_DATIF/en	<Suj:sn|cln,Objde:de-sn|en,Objà:à-sn|cld>
prendre_VERB place_NOUN	SN1[hum|ina]	SN2(dans)/		y	<Suj:sn|cln,Loc:dans-sn|y>
détourner_VERB DET oeil_NOUN	SN1[hum]	SN2(de)			<Suj:sn|cln,Dloc:de-sn>

Tableau 23 : Exemple d'informations complètes de sous-catégorisation de quelques expressions

Même si les cadres de sous-catégorisations modélisés à partir du modèle de Leff manquent d'inclure certaines informations, ils nous semblent tout à fait intéressants de par le fait qu'ils consistent en des codes précis et établis autour de modèles théoriques solides, dont celui de l'Approche Pronominale que nous avons précédemment évoqué ; c'est donc à travers le prisme de cette représentation que nous avons choisi d'observer les phénomènes inhérents à la sous-catégorisation des expressions polylexicales. Nous allons immédiatement exposer les résultats de ces observations.

1.2. Observation des phénomènes révélés par les cadres de sous-catégorisation modélisés

1.2.1. Nombre des cadres de sous-catégorisation

Les annotations que nous avions réalisées, de par la présence des traits de restriction de sélection et la dualité des compléments qu'elles incluaient, produisaient une combinatoire importante de suites d'actants, et donc des cadres de sous-catégorisation nombreux. Simplifiés et standardisés par le format Leff, les cadres de sous-catégorisation sont moins nombreux mais on en distingue tout de même 76 différents pour nos 423 expressions. Leur liste est présentée en Annexe 19. Nous notons que 140 d'entre elles n'acceptent pas de complément régi.

1.2.2. Remarque sur la dépendance du complément régi aux compléments de l'expression

Nous remarquons en outre que le cadre de sous-catégorisation le plus fréquent est *<Suj :sn|cln,Objà :à-sn|cld>*. En observant les expressions qu'il décrit, on se rend compte que les verbes qu'elles contiennent ont majoritairement un cadre de sous-catégorisation possible qui ressemble à *<Suj :sn|cld,Obj:sn,Obj :à-sn/cld>*. Ce cadre correspond à la structure dans laquelle le verbe apparaît lorsqu'il fait partie d'une de nos expressions (qui contient forcément un *Obj :sn*). C'est le cas notamment des expressions formées autour du verbe *donner* comme *donner_VERB DET leçon*_NOUN. Cela porte à croire que la sous-catégorisation de l'expression serait déterminée par celle du verbe. Cependant, on trouve aussi avec ce cadre de sous-catégorisation des expressions comme *tenir_VERB compagnie_NOUN* pour laquelle ce n'est pas le cas ; la structure sujet-verbe-COD-COI introduit par *à* ne constitue aucun des cadres de sous-catégorisation possibles pour le verbe *tenir*[90]. Il semblerait donc que, dans ce type de cas, la sous-catégorisation soit déterminée par l'expression entière ou par le nom.

[90] Vérifier dans la partie DICOVALENCE du Lefff

Cette question de la dépendance des compléments régis au nom ou au verbe a été maintes fois débattue, et nous ne la développerons pas davantage. Il nous semblait tout de même que son existence se devait d'être soulignée.

1.2.3. Rapport entre type et sous-catégorisations

En observant quels cadres de sous-catégorisation correspondent à quels types d'expressions, on se rend compte que seuls 8 cadres sur les 76 permettent de décrire à la fois des collocations et des expressions figées[91]. Il s'agit, pour la plupart, des cadres les plus fréquents et les plus simples (i.e. ne comportant qu'un seul complément régi). Il nous faut préciser que la différence quantitative entre les deux types d'expressions dans notre liste est importante (les collocations y sont très largement plus nombreuses), et que nous avons trouvé en tout 20 cadres pour les expressions figées ; cela signifie que près de la moitié des cadres de sous-catégorisation des expressions figées sont identiques à ceux des collocations. Nous ajoutons également que les cadres les plus complexes et spécifiques (par exemple, *<Suj:sn|cln,Objde:de-sn|de-sinf|en|de-scompl,Obl:pour-sinf>* pour l'expression *avoir DET idée*) sont presque exclusivement associés à des collocations. Cela renforce pour nous l'idée selon laquelle la description de la sous-catégorisation des collocations est au moins aussi importante et riche que celle des expressions figées, or, elle est largement moins présente dans les ressources lexicales.

1.2.4. Influence du déterminant sur la sous-catégorisation

Le fait de traiter les expressions polylexicales comme unités à sous-catégoriser fait apparaître la problématique de la variation du déterminant et de son impact sur la modélisation de la sous-catégorisation.

Le modèle que nous avons adopté dans le cadre de notre étude présente les déterminants comme des variables pouvant être réalisées par un nombre limité d'unités lexicales. Or, la présence d'un déterminant parmi ceux possibles pour une expression peut changer sa sous-catégorisation. Nous avons par exemple indiqué que le complément *de SN* était possible pour la collocation *aborder DET sujet ;* cela n'est pas vrai si le déterminant est un indéfini (*il a abordé le sujet de la religion* /**il a abordé un sujet de la religion*). Sur ce point donc, il semble que le formalisme choisi ne soit pas adapté à la structure générale de nos données. Ce problème prend d'ailleurs toute son ampleur quand on observe que, si l'on s'en tient à décrire nos expressions avec la liste de leurs déterminants et leur sous-catégorisation comme deux éléments indépendants et cumulables, on ne tient pas compte du fait que le complément *de SN* ne puisse pas apparaître en cooccurrence avec un déterminant

[91] *<Suj:sn|cln,Objà:à-sn|cld,Obl:sur-sn>*, *<Suj:sn|cln,Objà:à-sn|cld>*, *<Suj:sn|cln,Objde:de-sinf|en>*, *<Suj:sn|cln,Objde:de-sinf>*, *<Suj:sn|cln,Objde:de-sn|en,Objà:à-sn|cld>*, *<Suj:sn|cln,Objde:de-sn|en>*, *<Suj:sn|cln,Objde:de-sn>*, et *<Suj:sn|cln,Obl:pour-sinf>*.

possessif si ces deux ont le même référent. On obtient ainsi une modélisation qui tend à indiquer que *briser son cœur* est une variation de *briser_VERB DET cœur_NOUN*, au même titre que *briser le cœur de Jean*, sans pour autant expliciter que **briser son cœur de Jean* n'est pas possible.

Ce phénomène mériterait donc d'être intégré à la modélisation des cadres de sous-catégorisation. Par exemple, pour *aborder_VERB DET sujet_NOUN,* on pourrait indiquer : *<Suj :sn|cln,[DET=le,les ;Objde :de-sn],[DET=un,des ;NULL]>*. Ce codage permettrait de donner des sous-catégorisations différentes en fonction du déterminant.

Après avoir présenté la modélisation de la sous-catégorisation, nous terminerons notre exposé par une proposition de modélisation pour l'ensemble des informations présentes dans nos données.

2. Proposition de modélisation de l'intégralité des informations extraites

Afin de modéliser les informations que nous avons recueillies lors de notre étude, nous présentons dans ce point une proposition de conversion des données dans un format xml, puis des pistes de regroupement d'informations par configurations communes de caractéristiques.

2.1. Proposition de conversion des données dans une architecture balisée

Les informations collectées ont été stockées dans différents fichiers csv/tableaux Excel puis rassemblées dans un seul tableau. Pour chaque entrée, les informations sont au nombre de 28.

Le Tableau 24 donne l'exemple de l'entrée correspondant à l'expression *avoir_VERB DET mal_NOUN* dont le sens est *rencontrer des difficultés pour*. Pour une question de lisibilité, nous avons tronqué la liste des pronoms sujets et la phrase d'exemple des alternances, et n'avons pas inclus d'exemples en contexte.

Verbe	avoir_VERB
Nom	mal_NOUN
Expression lemmatisée	avoir_VERB DET mal_NOUN
Forme la plus courante	avoir du mal à
Type	Collocation
Verbe support?	oui
FL	Oper1
Définition Wikitionnaire	Connaître des difficultés, des embarras à faire quelque chose.
Définition DEM	
Condition de validité	Det=du/aucun + suivi de à
Déterminants	PART(du) /AUTRES(aucun)
Variabilité nombre nom	SG
Passivation	Non
Passivation - exemple	
Relativisation	Oui
Relativisation - exemple	*[…]Tout le mal que j' ai eu à m' en souvenir , il fallait que ça serve !*
Construction moyenne	Non
Construction moyenne - exemple	
Cadre de sous-catégorisation	<Suj:sn|cln,Objà:à-sn>
Actant 1	SN1[hum|ina]
SN1	je,il,on,beaucoup,personne,cela,ça,tout,d'autres,celui,[…]
Actant 2	Vinf2(à)
Actant 3	
Pronominalisation des compléments	
Fréquence expression	3077
Dispersion expression	9
Fréquence couplet	4341
Dispersion couplet	9

Tableau 24 : Exemple de l'ensemble des informations consignées pour une expression

Nous avons jugé qu'il serait utile que ces informations puissent être structurées. Pour cela, nous proposons leur transposition dans un format xml. Ce dernier pourrait avoir la forme de l'exemple présenté dans l'encadré ci-dessous, qui décrit l'expression du Tableau 24.

```
<exp idExp="50" frequenceExp="3077" dispersionExp="9" frequenceCoup="4341"
dispersionCoup="9">
   <lemme>avoir_VERB DET mal_NOUN</lemme>
   <formeCourante>avoir du mal à</formeCourante>
   <type>Collocation</type>
   <fl>Oper1</fl>
   <def source="Wiktionnaire">Connaître des difficultés, des embarras à faire
   quelque chose.</def>
   <contextes>
      <contexte idPhrase="5232158">En tout cas j' ai du mal à descendre au bon
      poids .</contexte>
      <contexte idhrase="5810802">Sa diplomatie a pourtant eu du mal , ces
      jours - ci , à cacher la satisfaction d' avoir doublé les États-Unis dans
      leur rôle d' arbitre . </exemple>
   </contextes>
   <desambiguisation>Det=du/aucun + suivi de à</desambiguisation>
   <elements>
      <element id="1" c="VERBE" vs="verbeSupport" l="avoir"/>
```

```
      <element id="2" c="DET" />
      <element id="3" c="NOM" l="mal" nb="SG "/>
   </elements>
   <variations>
      <variation id="2" classeDet="PART" l="du" nb="SG" val="du"/>
      <variation id="2" classeDet="AUTRES" l="aucun" nb="SG" val="aucun"/>
   </variations>
   <dependances>
      <d t="OBJ" h="1" d="3">
      <d t="DETERM" h="3" d="2">
   </dependances>
   <alternances>
      <alt type="relativisation">À une époque où il avait tant de peine à
      apprendre le fameux poème ... Adulte , il en a fait une chanson qui l' a
      rendu célèbre et lui a valu une Victoire de la musique : « Tout le mal
      que j' ai eu à m' en souvenir , il fallait que ça serve ! </alt>
   </alternances>
   <souscat>
      <actants>
         <actant actNum="1" trait="hum|ina">SN</actant>
         <actant actNum="2" clitique="NULL">Vinf2(à)</actant>
      <actants>
      <cadre>Suj:sn|cln,Objà:à-sn</cadre>
   <souscat>
</exp>
```

Cette représentation des données pourrait bien entendu être améliorée, mais elle a l'avantage de rendre compte de l'intégralité des informations dont nous disposons pour chaque expression en les regroupant en différentes catégories (liste des éléments, variations du déterminant, alternances, etc.).

2.2. Propositions de regroupement des informations par configurations similaires

Afin de limiter la taille des fichiers xml produits et de proposer un degré supérieur de structuration des informations, nous pourrions créer des classes de configurations possibles numérotées, et se contenter d'indiquer le numéro de la configuration correspondant à chaque expression.

Ainsi, on pourrait remplacer l'énumération des alternances par des codes ; par exemple, on indiquerait *PCm* pour les expressions qui permettent la passivation et les constructions moyennes, *PR* pour celles qui permettent la passivation et la relativisation, etc.

Il serait également possible de pousser plus loin la démarche en créant une seule nomenclature pour décrire toutes les propriétés syntaxiques présentées. Par exemple, on pourrait numéroter les cadres de sous-catégorisation et agglutiner ce numéro au trait de restriction de sélection du sujet puis au code des alternances pour former un code à attribuer aux expressions. Par exemple, si *<Suj:sn|cln,Objde:de-sn|en>* avait le numéro 1, il serait possible du supprimer toutes les balises contenant ces informations et les remplacer par une balise unique qui consisterait par, exemple, pour l'expression donnée au point 2.2., en une seule ligne du type *<syntaxe config="1SujHumPR">*.

Un tel travail de factorisation des informations, qui reste à ce jour à réaliser, permettrait en outre de pouvoir établir des liens de similitudes entre les expressions aux

codes similaires et d'obtenir une estimation du nombre de configurations de propriétés syntaxiques possibles qu'offre la variété des expressions polylexicales du français.

CONCLUSION ET PERSPECTIVES

Le travail que nous avons présenté a démontré qu'il était possible de constituer un lexique polylexical verbal fondamental pour le français basé sur un repérage à partir d'un corpus de langue générale.

Il a tout d'abord été nécessaire de définir l'objet linguistique sur lequel notre étude allait porter. Dans cette optique, nous avons premièrement dû imposer un cadre précis à la sélection d'associations considérées comme constituant des expressions polylexicales. Nous avons démontré que cette tâche n'est pas triviale et qu'il est préférable qu'elle soit guidée par l'utilisation de critères spécifiques afin de maintenir une certaine cohérence linguistique dans la liste des unités lexicales retenues à l'issue de la sélection. La définition du deuxième aspect caractéristique de l'objet linguistique sur lequel a porté notre étude, celui de son appartenance à un vocabulaire fondamental, a été établie à partir de critères purement statistiques. Nous avons en effet pris le parti de considérer que l'aspect fondamental d'une expression pouvait être déduit à partir de deux indications données par deux mesures spécifiques, à savoir 1) le fait que l'expression soit fréquemment utilisée dans la langue, indiqué par une fréquence élevée dans un corpus, 2) le fait qu'elle n'appartienne pas à un genre textuel particulier, indiqué par une mesure de dispersion qui est calculée à partir d'une segmentation du corpus par genres. Les résultats de la petite enquête menée auprès de spécialistes du FLE ont montré que ces deux critères permettaient, dans une certaine mesure, de délimiter les contours d'un vocabulaire fondamental. Cependant, elle a aussi révélé que la perception de la nature fondamentale d'une unité lexicale est soumise à de nombreuses variations individuelles.

A l'instar de Benigno (2012), nous pensons que la classe d'expressions retenues pourrait être plus large en incluant des objets sélectionnés sur le critère de la disponibilité. Cet ajout d'expressions disponibles constituerait alors, selon nous, une extension possible à notre étude.

Une autre extension consisterait en l'intégration d'expressions verbales construites autour de relations autres que V NObj ; nous nous y sommes restreinte, non pas pour des raisons théoriques mais pour des besoins méthodologiques et techniques. En effet, la méthode que nous avons utilisée n'était pas motivée par une volonté de créer un lexique large. Il s'agissait plutôt de démontrer qu'il était possible de créer une liste d'expressions et d'y consigner de nombreuses informations descriptives établies empiriquement selon une approche *corpus-driven*. Le choix de nous limiter à des expressions fréquentes formées autour d'une structure V NObj fréquente rendait alors cela plus abordable.

Notre méthode a donc davantage été basée sur des processus visant à une certaine précision, permise par plusieurs vérifications manuelles, qu'à une automatisation maximum des traitements. Elle a en outre été construite à partir d'un modèle structurel que nous avons

établi, celui la séparation noyaux/éléments périphériques des expressions, et d'autres que nous avons adaptés et interprétés en fonction de nos besoins, comme par exemple lorsque nous avons choisi quels éléments nous devions intégrer à nos cadres de sous-catégorisation.

De plus, nous avons essayé de faire en sorte que notre méthode prenne en compte les phénomènes de polysémie, afin que la liste résultante ne rende pas seulement compte de suites de mots fréquentes, mais bien d'expressions à part entière, ayant chacune leurs mesures et caractéristiques propres même si certaines étaient formées par des mots identiques. Les variations que nous avons trouvées dans les descriptions des propriétés syntaxiques d'expressions polysémiques ont d'ailleurs démontré la nécessité de la prise en compte de ce phénomène.

Nous pensons, d'autre part, que si une liste assez fournie en termes de description des propriétés syntaxiques a été produite, bien d'autres informations pourraient venir la compléter. Notamment, une étude des possibilités d'insertions de modifieurs entre le verbe et le nom n'a pas été réalisée. De plus, quelques biais dans la modélisation de nos données pourraient être corrigés. Nous pensons en particulier à une meilleure mise en perspective de la variation du déterminant et de la sous-catégorisation. Un autre biais qui pourrait être corrigé serait, par exemple, le fait que nous ayons considéré les expressions formées par les mêmes couplets V NObj comme différentes si l'une comportait un déterminant et l'autre non. Or, cela peut être le cas (*avoir conscience* ≠ *avoir une conscience*), mais pas nécessairement (*avoir accès* = *avoir un accès*).

Enfin, nous pensons qu'il serait intéressant d'appliquer notre méthode sur d'autres corpus afin d'en vérifier l'efficacité et de comparer les résultats obtenus. Il pourrait notamment s'agir d'un corpus contenant davantage d'oral et/ou une partie consacrée aux communications électroniques.

Bibliographie

Aït-Mokhtar, S., Chanod, J.-P., & Roux, C. (2002). Robustness beyond shallowness: incremental deep parsing. *Natural Language Engineering*, *8*(2–3), 121–144.

Alonso Ramos, M. (1999). *Étude sémantico-syntaxique des constructions à verbe support*. (Thèse de doctorat), Université de Montréal.

Arnaud, P., Ferragne, E., Lewis, D., & Maniez, F. (2008). Adjective+ noun sequences in attributive or NP-final positions. *Phraseology: An Interdisciplinary Perspective*, 111–125.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. *Text and Technology: In Honour of John Sinclair*, *233*, 250.

Baroni, M., Bernardini, S., Ferraresi, A., & Picci, G. (2010). Web corpora for bilingual lexicography: a pilot study of English/French collocation extraction and translation. *Using Corpora in Contrastive and Translation Studies. Newcastle: Cambridge Scholars Publishing*, 337–362.

Benigno, V. (2012). *La notion de collocation fondamentale: étude de corpus en vue d'une exploitation didactique*. (Thèse de doctorat en cotutelle), Università degli Studi di Palermo et Université Stendhal de Grenoble.

Benzitoun, C., Fort, K., & Sagot, B. (2012). TCOF-POS: un corpus libre de français parlé annoté en morphosyntaxe. *JEP-TALN 2012-Journées d'Études sur la Parole et conférence annuelle du Traitement Automatique des Langues Naturelles*, 99–112.

Biber, D. (2012). Corpus-based and corpus-driven analyses of language variation and use. *International Journal of Corpus Linguistics*, *3*, 275-311.

Blanche-Benveniste, C. (1984). *Pronom et syntaxe: l'approche pronominale et son application au français*. Paris : SELAF.

Bonin, P., Chalard, M., Méot, A., & Fayol, M. (2001). Age-of acquisition and word frequency in the lexical decision task: Further evidence from the French language. *Current Psychology of Cognition*, *20*(6), 401–444.

Branca-Rosoff, S., Fleury, S., Lefeuvre, F., & Pires, M. (2009). Corpus de français parlé parisien des années 2000 (CFPP). *Discours Sur La Ville*.

Carroll, J., & Fang, A. C. (2004). The automatic acquisition of verb subcategorisations and their impact on the performance of an HPSG parser. A l'*International Conference on Natural Language Processing* (pp. 646–654). Sanya City, Chine.

Carter, R. (1998). *Vocabulary: Applied linguistic perspectives*, (2e èd.). London et New York: Routledge.

Corman. (2012). Extraction d'expressions polylexicales sur corpus arboré. (Mémoire de Master), Université Stendhal de Grenoble.

Cortier, C. (2006). De quelques enjeux et usages historiques du Français fondamental. Présentation. *Documents Pour L'histoire Du Français Langue Étrangère Ou Seconde*, (36), 9–12.

Cowie, A. (1998). *Phraseology, Theory, Analysis, and Applications*, Oxford : Clarendon Press.

Cowie, A. P. (1988). Stable and creative aspects of vocabulary. *Vocabulary and Language Teaching*, *139*, 126-137.

Danlos, L., Sagot, B., & Signes-INRIA, P. (2007). Comparaison du Lexique-Grammaire des verbes pleins et de DICOVALENCE: vers une intégration dans le Lefff. Dans *Actes du TALN*, Toulouse.

Dubois, J., & Dubois-Charlier, F. (1997). *Les verbes français*. Paris:Larousse-Bordas.

Dubois, J., & Dubois-Charlier, F. (2011). La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration,, *Langages*, (179–180), 31–56.

Duchet, J.-L., Kraif, O., & Castillo, M. T. (2008). Corpus massifs et corpus bilingues alignés : leur impact sur la recherche linguistique. *Bulletin de La Société de Linguistique de Paris*, *103*(1), 129–150.

Dupont, M., (2014). L'environnement collocationnel de mais: approche diaphasique. Dans *A l'articulation du lexique, de la grammaire et du discours: marqueurs grammaticaux et marqueurs discursifs*, Paris.

Evert, S. (2005). *The statistics of word cooccurrences: word pairs and collocations*, (Thèse de doctorat), Université de Stuttgart.

Evert, S., & Krenn, B. (2001). Methods for the qualitative evaluation of lexical association measures. *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, 188–195.

Firth, J. . (1957). A synopsis of linguistic theory, 1930-1955. Dans *Firth J. R. et al., Studies in Linguistic Analysis*, *Special volume of the Philological Society*, Oxford, Blackwell, 1–32.

Gaatone, D. (1997). La locution: analyse interne et analyse globale. *La Locution Entre Langue et Usages*, 165–177.

Galisson, R. (1976). *Inventaire thématique et syntagmatique du français fondamental*. Paris:Hachette-Larousse, Coll. Le Français dans le Monde/BELC.

Galliano, S., Geoffrois, E., Mostefa, D., Choukri, K., Bonastre, J.-F., & Gravier, G. (2005). The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. *Interspeech*, 1149–1152.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. *Translation Studies in Scandinavia*, *1*, 88–95.

Gougenheim, G. (1964). *L'élaboration du français fondamental (ler degré): Etude sur l'établissement d'un vocabulaire et d'une grammaire de base (Vol. 1).* Paris:Didier.

Gougenheim, G. (1971). *Dictionnaire fondamental de la langue française*. Paris:Didier Edition Internationale.

Granger, S., Paquot, M., & others. (2008). Disentangling the phraseological web. *Phraseology. An Interdisciplinary Perspective*, 27–50.

Gross, G. (1996). *Les expressions figées en français: noms composés et autres locutions*. Paris:Ophrys.

Gross, M. (1975). *Méthodes en syntaxe*. Paris:Hermann.

Gross, M. (1981). Les bases empiriques de la notion de prédicat sémantique. *Langages*, (63), 7–52.

Grossmann, F. (2011). Didactique du lexique: état des lieux et nouvelles orientations. *Pratiques. Linguistique, Littérature, Didactique*, (149–150), 163–183.

Grossmann, F., & Tutin, A. (2002). Collocations régulières et irrégulières: esquisse de typologie du phénomène collocatif. *Revue Française de Linguistique Appliquée*, *7*(1), 7–25.

Hagège, C., & Tannier, X. (2007). XRCE-T: XIP temporal module for TempEval campaign. Dans *Proceedings of the 4th International Workshop on Semantic Evaluations*, 492–495, Prague.

Hatier, S. (2016). *Identification et analyse linguistique du lexique scientifique transdisciplinaire. Approche outillée sur un corpus d'articles de recherche en SHS*. (Thèse de doctorat), Université Grenoble Alpes.

Hatim, B., Mason, I. (2005). *The translator as communicator*. Routledge.

Hausmann, F. J., & Blumenthal, P. (2006). Présentation: collocations, corpus, dictionnaires. *Langue Française*, (2), 3–13.

Heid, U. (2008). Computational phraseology. An overview. *Phraseology: An Interdisciplinary Perspective; Granger, S., Meunier, F., Eds*, 337–360.

Jackendoff, R. S. (1995). The boundaries of the lexicon. *Idioms: structural and psychological perspectives*, 133-165, Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Jacob, P. (1984). La syntaxe peut-elle être logique? *Communications*, *40*(1), 25–96.

Joseph, A. (2013). Améliorer l'extraction et la description d'expressions polylexicales grâce aux règles transformationnelles. Dans *TALN-RÉCITAL 2013*, 42-55, Les Sables d'Olonnes.

Kraif, O. (2011). Les concordances pour l'observation des corpus: utilité, outillage, utilisabilité. *Le Langage et Ses Niveaux D'analyse*, 67–80.

Kraif, O. (2016). Le lexicoscope : un outil d'extraction des séquences phraséologiques basé sur des corpus arborés. *Cahiers de Lexicologie: Revue Internationale de Lexicologie et Lexicographie*, (108), 91–106.

Kraif, O. (à paraître). Traduire le polar : une étude textométrique comparée de la phraséologie duroman policier en français source et cible. *Synergie Pologne*.

Kupsc, A. (2007). Extraction automatique de cadres de sous-catégorisation verbale pour le français à partir d'un corpus arboré. Dans *TALN 2007*, Toulouse.

Laviosa, S. (2002). *Corpus-based translation studies: theory, findings, applications*. Amsterdam:Rodopi.

Lété, B., Sprenger-Charolles, L., & Colé, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, *36*(1), 156–166.

Levelt, W. J. (1999). Producing spoken language: A blueprint of the speaker. Dans *The neurocognition of language*, 83–122. Oxford University Press.

Lo Cascio, V. (2000). La théorie des profils textuels et la compétence lexicale : les collocations. In *Didactique des langues romanes : le développement des compétences chez l'apprenant, Actes du colloque de Louvain-la-Neuve* (De Boeck & Duculot, pp. 349–359). Bruxelles: Collès L., Dufays J.-L., Fabry G., Maeder C.

Mel′čuk, I. A., & Arbatchewsky-Jumarie, N. (1999). *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques*. Montréal:PUM.

Mel′čuk, I. A., & Polguère, A. (2007). *Lexique actif du français: l'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. Bruxelles:De Boeck.

MEL'ĆUK, I. (1998). Collocations and lexical functions. *Phraseology, Theory analysis, and applicatons*, 23–54.

Mel'čuk, I. (2003). Collocations dans le dictionnaire. *Les Écarts Culturels Dans Les Dictionnaires Bilingues*, 19–64.

Mel'čuk, I. (2013). Tout ce que nous voulions savoir sur les phrasèmes, mais…. *Cahiers de Lexicologie*, *102*, 129–149.

Mel'čuk, I. (1995). Phrasemes in Languages and Phraseology, *Linguistics Idioms: Structural and Psychological Perspectives*, 167-232, New York:Psychology Press.

Mertens, P. (2010). Restrictions de sélection et réalisations syntagmatiques dans Dicovalence: conversion vers un format utilisable en TAL. Dans *TALN*, Montréal.

Messiant, C., Gábor, K., & Poibeau, T. (2010). Acquisition de connaissances lexicales à partir de corpus: la sous-catégorisation verbale en français. *Traitement Automatique Des Langues*, *51*(1), 65–96.

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford University Press.

Netzlaff, M. (2005). *La collocation adjectif-adverbe et son traitement lexicographique*. (Thèse de doctorat) Université de Erlangen.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE^{TM}//A lexical database for contemporary french: LEXIQUE^{TM}. *L'année Psychologique*, *101*(3), 447–462.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 1–15). Springer.

Sagot, B. (2010). The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *7th international conference on Language Resources and Evaluation (LREC 2010)*, La Vallette.

Sagot, B., & Danlos, L. (2008). Améliorer un lexique syntaxique à l'aide des tables du lexique-grammaire—Constructions impersonnelles et expressions verbales figées. *Cahiers Du CENTAL*, *5*, 107–126.

Seretan, V. (2011). *Syntax-based collocation extraction*. Dordrechet:Springer.

Storjohann, P. (2005). Corpus-driven vs. corpus-based approach to the study of relational patterns. Dans *Proceedings of the Corpus Linguistics conference 2005,* Birmingham

Tiedemann, J., & Nygaard, L. (2004). The OPUS corpus-parallel and free. Dans *The International Conference on Language Resources and Evaluation,* 93-96.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: Publishing.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press.

Tutin, A. (2005). Le dictionnaire de collocations est-il indispensable? *Revue Française de Linguistique Appliquée*, *10*(2), 31–48.

Tutin, A. (2010a). Le traitement des collocations dans les dictionnaires monolingues de collocations du français et de l'anglais. Dans *2ème Congrès Mondial de Linguistique Française,* 1075-1090, La Nouvelle-Orléans.

Tutin, A. (2010b). *Sens et combinatoire lexicale: de la langue au discours*. (Habilitation à diriger des recherches), Université de Toulouse 2.

Tutin, A. & Kraif, O. (2016). Routines sémantico-rhétoriques dans l'écrit scientifique de sciences humaines: l'apport des arbres lexico-syntaxiques récurrents. *Lidil. Revue de Linguistique et de Didactique Des Langues*, (53), 119–141.

Tutin, A., & Kraif, O. (2017). Comparing Recurring Lexico-Syntactic Trees (RLTs) and Ngram Techniques for Extended Phraseology Extraction: a Corpus-based Study on

French Scientific Articles. Dans *13th Workshop on Multiword Expressions-EACL*, Valence.

Zykova, I. V. (2013). Phraseological meaning as a mechanism of cultural memory. *Research on Phraseology Across Continents.–Poland: University of Bialystok Publishing House*, *2*, 422–441.

Table des illustrations

Figure 1: Extrait de la liste de fréquence du projet Lexique telle que téléchargeable à partir de la plateforme http://www.lexique.org.....15
Figure 2 : Extrait de la base de données lexicale Manulex telle que téléchargeable à partir de la plateforme http://www.manulex.org/.....16
Figure 3: Exemples de syntagmes autour du thème de la maison puis de la famille construits par Galisson, tels que présentés dans l'ouvrage *Inventaire thématique et syntagmatique du français fondamental*.....29
Figure 4 : Capture d'écran d'un fichier html contenant les occurrences en contexte d'un couplet V NObj (*tenir_VERB tête_NOUN*).....50
Figure 5 : Graphique représentant la distribution sur les trois genres du corpus des occurrences de 10 expressions tests.....77
Figure 6 : Analyse syntaxique du segment « *il avait besoin d'aide* » dans le corpus.....91
Figure 7 : Analyse syntaxique du segment « *ils ont besoin qu'on leur apprenne* » dans le corpus..91
Figure 8 : Fonctions syntaxiques et réalisations dans le Lefff – capture d'écran réalisée sur la publication de Sagot (2010).....98

Tableau 1 : Composition du sous-corpus littéraire.....38
Tableau 2 : Composition du sous-corpus de parole transcrite.....39
Tableau 3 : Segmentation du corpus.....40
Tableau 4 : Exemple de deux entrées dans la liste des extractions avant le redimensionnement des expressions à basse fréquence.....48
Tableau 5 : Exemple de deux entrées dans la liste des extractions après le redimensionnement des expressions à basse fréquence.....48
Tableau 6 : Résultats globaux de l'évaluation d'une première version du script d'extraction.....52
Tableau 7 : Résultats de l'évaluation d'une première version du script d'extraction sur les expressions candidates contenues (a) dans les 10% ayant les fréquences les plus hautes, (b) les 10% ayant les fréquences les plus faibles.....52
Tableau 8 : Résultats de l'évaluation d'une première version du script d'extraction sur les expressions candidates contenues dans (a) les 10% ayant taux de dispersion les plus hauts, (b) 10% ayant les taux de dispersion les plus faibles.....53
Tableau 9 : Résultats de l'évaluation de la prise en compte des éléments périphériques – quantification des occurrences pour chaque cas de figure.....55
Tableau 10 : Taux d'entrées valides, partiellement valides et invalides sur les cinquante expressions les plus courantes extraites par la méthode par ALR et par patrons catégoriels.....59
Tableau 11 : Description des différents types de collocations ternaires (emprunté à Tutin, 2010b, p.44).....65

Tableau 12 : Estimation de la précision et du rappel obtenus selon différents seuils de Log Likelihood et de Z-score sur une répartition égale de colligations et d'expressions, pour des seuils de fréquences de 15 et de dispersion de 4 67
Tableau 13 : Estimation de la précision et du rappel obtenus selon différents seuils de Log Likelihood et de Z-score sur une répartition de 75% de colligations et 25% d'expressions valides dans le corpus, pour des seuils de fréquences de 15 et de dispersion de 4 68
Tableau 14 : Exemple de trois expressions désambiguïsées 71
Tableau 15 : Résultats de l'extraction d'expressions réalisée sur les corpus oraux – Nombre d'expressions valides et indications de leur présence ou absence dans les extractions réalisées sur le corpus général 74
Tableau 16 : Répartition des expressions fondamentales par type 82
Tableau 17 : Nombres de définitions trouvées dans le DEM et/ou Wikitionnaire par types d'expressions 83
Tableau 18 : Différentes combinaisons de flexibilité syntaxique possibles avec exemples de syntagmes extraits du corpus 87
Tableau 19 : Liste des patrons de compléments prépositionnels et complétives utilisés pour extraire la sous-catégorisation des expressions – Patrons et exemples de réalisations correspondants. 92
Tableau 20 : Résultats des comparaisons d'extraction des compléments des expressions par différentes méthodes et seuils de fréquence 94
Tableau 21 : Taux de bruits et de silences produits dans les sorties d'extraction des compléments des expression par différentes méthodes et seuils de fréquence 94
Tableau 22 : Extrait des résultats de l'ajout des traits de restriction de sélection aux actants sujets à partir des pronoms sujets extraits dans le corpus 96
Tableau 23 : Exemple d'informations complètes de sous-catégorisation de quelques expressions .. 99
Tableau 24 : Exemple de l'ensemble des informations consignées pour une expression 103

Table des annexes

ANNEXE 1 Exemple de phrase du corpus : script html et représentation graphique des relations de dépendance 117
ANNEXE 2 Combinaisons Noyaux-Eléments périphériques repérées dans la liste initiale : exemples d'expressions produites par les structures décrites 119
ANNEXE 3 Combinaisons Noyaux-Eléments périphériques repérées dans la liste initiale : nombre d'entrées dans la liste initiales décrites par chaque patron 127
ANNEXE 4 Pseudo-algorithme du script d'extraction 131
ANNEXE 5 Liste des patrons d'éléments périphériques extraits par la version test du script d'extraction et quantification de leurs apports en termes d'utilité à extraire des expressions valides 141
ANNEXE 6 Typologie des cas de figure de polysémies des expressions candidates extraites par le script de test – Exemples et critères envisagés pour distinguer dans le corpus les occurrences des différents sens de chaque expression 142
142
ANNEXE 7 Comparaison de la sortie des deux extractions (par patrons catégoriels et par ALR) .. 145
ANNEXE 8 Extrait de l'annotation de la validité des expressions candidates en sortie du script d'extraction (premières entrées par ordre de fréquence décroissant) 147
ANNEXE 9 Mesures de Log Likelihood et de Z-score de 30 expressions invalides extraites 149
ANNEXE 10 Mesures de Log Likelihood et de Z-score de 30 expressions valides ou partiellement valides extraites 150
ANNEXE 11 Expressions utilisées pour les tests de mesures d'association les plus fréquentes classées par ordre croissant de Log Likelihood et de Z-score 151
ANNEXE 12 Exemple de redimensionnement manuel des expressions longues incluant des expressions valides – Cas de l'expression faire_VERB DET tour_NOUN 152
ANNEXE 13 Capture d'écran d'une partie d'un formulaire de désambiguïsation 153
ANNEXE 14 Réponses recueillies dans le cadre de l'enquête menée auprès de spécialistes de l'enseignement de FLE 155
ANNEXE 15 Exemple sur quatre expressions des résultats des caractéristiques annotées 157
ANNEXE 16 Comparaison des extractions des sujets des expressions par méthodes syntaxiques et de patrons catégorielles 158
ANNEXE 17 Comparaison des extractions des compléments des expressions par méthodes syntaxique et par patrons catégoriels 159
ANNEXE 18 Premières lignes (par ordre alphabétique) du tableau de sélection des compléments sous-catégorisés à partir des résultats des différentes extractions réalisées 161
ANNEXE 19 Différents cadres de sous-catégorisation trouvés pour les expressions fondamentales selon le modèle Lefff 162

ANNEXE 1

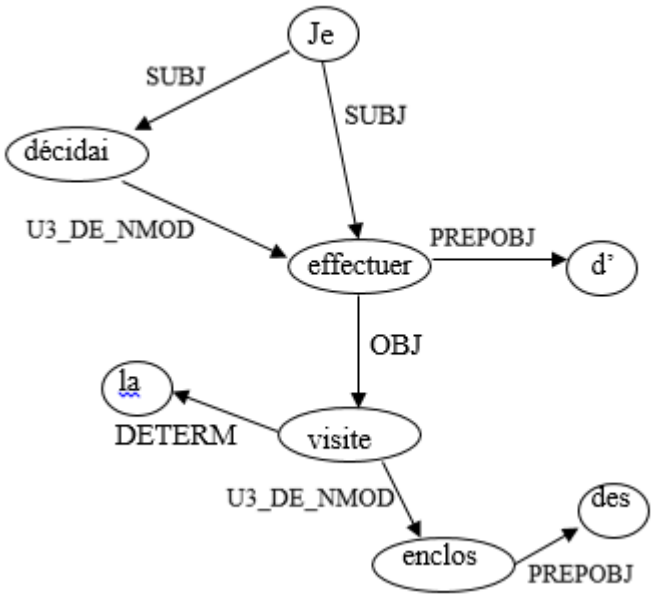
Exemple de phrase du corpus : script html et représentation graphique des relations de dépendance

La phrase présentée est « *Je décidai d'effectuer la visite des enclos.* » (dans J.C. Grangé, *Le Vol des cigognes*, 1994)

Le code XML correspondant est présenté ci-dessous. La partie du code comprise entre les balises *<tc></tc>* donne les informations sur chaque lemme (id, catégorie morphosyntaxique, lemme, traits morphologiques). La partie suivante, comprise entre les balises *<dc></dc>*, donne la liste des relations de dépendance de la phrase en indiquant pour chacune les id de la tête et du dépendant.

```
<s id="463129">
     <tc>
        <t id="0" c="PRON" l="je" f="STARTBIS MAJ CLOSED MASC FEM SG P1 CLIT
NOM PRON START LAST FIRST">Je</t>
        <t id="2" c="VERB" l="décider" f="SNAINF_SO SE QUEP DESVINF DESN AVOIR
ACEQUEP ASVINF ASN SVINFDIR SN REFLEXTYPE REFLEXIVE SFDE SFA SG P1 PAS IND
VERB LAST FIRST">décidai</t>
        <t id="4" c="PREP" l="de" f="PREPINF SFDE FORM PREP DIR FIRST">d'</t>
        <t id="6" c="VERB" l="effectuer" f="SE AVOIR SN REFLEXTYPE REFLEXIVE
INF VERB LAST">effectuer</t>
        <t id="8" c="DET" l="le" f="CLOSED FEM SG DEF DET FIRST">la</t>
        <t id="10" c="NOUN" l="visite" f="PARSN DESN SFPAR SFDE CLOSED FEM SG
P3 NOUN LAST">visite</t>
        <t id="12" c="PREP" l="de" f="SFDE FORM MASC FEM PL DEF PREP DET
FIRST">des</t>
        <t id="14" c="NOUN" l="enclos" f="CLOSED MASC PL SG P3 NOUN LAST
FIRST">enclos</t>
        <t id="16" c="SENT" l="." f="TOUTMAJ1 SENT END LAST">.</t>
     </tc>
     <dc>
        <d t="SUBJ" h="2" d="0" />
        <d t="SUBJ" h="6" d="0" />
        <d t="OBJ" h="6" d="10" />
        <d t="PREPOBJ" h="14" d="12" />
        <d t="PREPOBJ" h="6" d="4" />
        <d t="DETERM" h="10" d="8" />
        <d t="U3_DE_NMOD" h="10" d="14" />
        <d t="U3_DE_VMOD" h="2" d="6" />
     </dc>
        </s>
```

La représentation graphique des relations de dépendance de cette phrase est la suivante :



ANNEXE 2

Combinaisons Noyaux-Eléments périphériques repérées dans la liste initiale : exemples d'expressions produites par les structures décrites

		NOYAUX :	V N	V DET N	V du N	V des N	V de DET(la|l') N	V DET DET N
	Noyaux sans éléments périphériques		*avoir faim*	*lever l'ancre*	*mettre du temps*	*faire des manières*	*prendre de la vitesse*	*faire les cent coups*
Suites d'éléments périphériques antéposées au noyau		**Nombre de lexies**						
	ADV	1	*ne pas voir cure de*	*ne plus avoir vingt ans*		*ne pas casser des briques*		*ne pas avoir toute sa tête*
	ADV PREP	2		*ne pas en perdre une bouchée*				
	PREP	1	*en avoir marre*	*en valoir la peine*				*en faire tout un plat*
	PRON	1		*s'y casser les dents*		*y laisser des plumes*	*y mettre de conscience*	
Suites d'éléments périphériques postposées au noyau								
	ADJ	1	*faire peau neuve*	*avoir les dents longues*		*se faire des cheveux blancs*	*manger de la vache enragée*	
	ADJ DET N	3		*manger son pain blanc le premier*				
	ADJ N	2		*mettre une grosse tête*				
	ADJ PREP DET N	4		*avoir le ventre plat comme une punaise*				

ADJ PREP DET N du N	6		*avoir un pois chiche à la place du cerveau*				
ADJ PREP N	3		*avoir les yeux gros de larmes*				
ADV	1	*faire appel a mixima*	*mettre le nez dehors*				
ADV ADJ	2		*avoir l'esprit mal tourné*				
ADV ADJ que DET N	5		*avoir les yeux plus gros que le ventre*				
ADV ADV	2		*porter sa vue bien loin*				
ADV DET N	3		*avoir cent pieds par-dessus la tête*				
COORD des N	3				*mélanger des choux et des carottes*		
COORD DET ADJ N	4		*faire la pluie et le beau temps*				
COORD DET N	3		*prendre ses cliques et ses claques*				
COORD DET N du N	5		*vouloir le beurre et l'argent du beurre*				
COORD N	2	*remuer ciel et terre*					
des ADJ N	3		*faire la tournée des grands ducs*				
des DET N	3	*faire feu des deux fuseaux*					
des N	2	*faire partie des meubles*	*prendre la clef des champs*				
DET DET N PREP DET N	6			*avoir du front tout le tour de la tête*			
DET N	2		*appeler un chat un chat*				

DET N PREP DET N	5		*tourner sept fois sa langue dans sa bouche*				
du ADJ N	3		*prendre la vie du bon côté*				
du DET N	3		*tirer une épine du pied*				
du N	2		*faire le tour du propriétaire*				*avoir toutes les peines du monde*
du N CONJ PRON V ADJ	6		*baiser le cul du Diable quand il est frette*				
du N PREP N	4		*avoir le trou du cul en chou-fleur*				
du N PREP PRON du N	6		*coudre la peau du renard à celle du lion*				
du N qui V	4		*avoir les dents du fond qui baignent*				
N	1	*faire moit moit*	*faire la rue Michel*				
N N	2	*faire pan pan cucul*	*avoir les jambes pâté foie*				
ni N	2		*ne craindre Dieu ni diable*				
NUM	1		*prendre la ligne 11*				
PREP ADJ	2		*voir la vie en rose*				
PREP ADJ N	3	*faire appel au bon cœur*	*voir la Lune en plein jour*				
PREP ADV	2		*voir la mort de près*				
PREP ADV ADJ	3		*prendre le carême de trop haut*				
PREP des N	3		*ne pas attacher son chien avec des saucisses*		*prendre des vessies pour des lanternes*		
PREP DET ADJ N	4						*avoir les deux pieds dans le même sabot*
PREP DET DET N	4		*brûler la chandelle par les deux bouts*				

PREP DET N	3	*mettre la clef sous la porte*	*prendre ses désirs pour des réalités*	*mettre du beurre dans les épinards*	*mettre des bâtons dans les roues*	*jeter de l'huile sur le feu*	*pleurer toutes les larmes de son corps*
PREP DET N ADJ	4	*parler français comme une vache espagnole*					
PREP DET N COORD DET N ADJ	6		*avoir le cœur sur la main et le poignet coupé*				
PREP DET N COORD DET N PREP N	8		*mettre les points sur le i et les barres aux t*				
PREP DET N du N	5		*jeter le bébé avec l'eau du bain*				
PREP DET N PREP DET N	6		*gagner son pain à la sueur de son front*				
PREP DET N PREP N	5		*crier famine sur un tas de blé*				
PREP DET N PREP PREP N	6		*se fourrer le doigt dans l'œil jusqu'au coude*				
PREP N	2	*donner signe de vie*	*avoir la boule au ventre*	*avoir du cœur au ventre*	*tirer des plans sur la comète*	*jeter de la poudre aux yeux*	*avoir les deux pieds sur terre*
PREP N ADJ	3		*avoir un œil au beurre noir*		*faire des yeux de merlan frit*		
PREP N du N	4		*garder l'église au milieu du village*				
PREP N PREP DET N	5		*donner un coup de canif dans le contrait*				

PREP PRON	2	*faire arme de tout*	*tirer la couverture à soi*	*avoir du temps devant soi*		*prendre de l'empire sur soi-même*	
PREP PRON PRON V V	5		*scier la branche sur laquelle on est assis*				
PREP PRON V	3					*faire de la prose sans le savoir*	
PREP PRON V V	4		*plumer la poule sans la faire crier*				
PREP V	2		*avoir son mot à dire*	*donner du grain à moudre*			
PREP V ADV	3		*avoir un nom à coucher dehors*				
PREP V DET N	4		*tuer la poule pour avoir l'œuf*				
PREP V N	3	*déshabiller Paul pour habiller Jean*					
PREP V PREP DET N PREP N	7		*avoir trouvé son permis de conduire dans une boîte de Cracker Jack*				
PREP V V	3		*donner le bâton pour se faire battre*				
PRON	1	*faire justice soi-même*					
que N	2	*faire signe que non*					
qui V	2		*avoir les fils qui se touchent*				
qui V DET N	4		*avoir les dents qui rayent le parquet*				
qui V N PREP DET N	6		*avoir un œil qui dit zut à l'autre*				
qui V PREP DET N	5		*avoir les yeux qui sortent de la tête*	*avoir le sans qui bout dans les veines*			

Structures avec des éléments périphériques avant et après le noyau *(X=noyau)*							
ADV X ADJ	1;1		*ne pas avoir inventé l'eau chaude*				
ADV X ADV	1;1		*ne pas mettre le nez dehors*				
ADV X des N	1;2	*ne pas avoir peur des mots*					
ADV X PREP DET DET N	1;4		*ne pas avoir la lumière à tous les étages*				
ADV X PREP DET N	1;3		*ne pas casser trois pattes à un canard*				
ADV X PREP N	1;2	*ne pas avoir voix au chapitre*	*ne pas valoir un pet de lapin*				
ADV X PREP N des N	1;4		*ne pas avoir les yeux en face des trous*				
ADV X PREP N PREP DET N	1;5		*ne pas avoir une goutte de sang dans les veines*				
ADV X PREP PRON V	1;3		*ne pas avoir l'air d'y toucher*				
ADV X PREP V DET N	1;4		*ne pas avoir inventé le fil à couper le beurre*				
ADV X PREP V N	1;4		*ne pas avoir le temps de dire ouf*				
ADV X que PRON V	1;3		*ne pas valoir le pain qu'on mange*				
ADV X-là	1;1					*ne pas manger de ce pain-là*	
PREP X COORD DET N	1;3		*en oublier le boire et le manger*				
PREP X PREP V	1;2		*en donner sa tête à couper*				

Insertion d'éléments périphériques entre V et GN-OBJET							
ADJ	1	*faire long feu*					
ADV	1		*mener bien sa barque*	*se donner bien du mouvement*			
ADV ADV	2		*placer trop haut la barre*				
PRON de	2		*faire une de ces têtes*				
que	1	*comprendre que dalle*					
Autres structures							
ADV V DET ADJ N			*ne pas avoir le premier sou*				
ADV V PREP N DET N PREP DET N			*n'avoir jamais perdu de vue le clocher de son village*				
ADV V que des N					*ne pas sucer que des glaçons*		
PREP V DET ADJ N			*en avoir une sacrée couche*				
PREP V PREP ADJ N				*n'y voir que du feu*			
PRON V que du N		*n' y comprendre que couic*					
PRON V que N		*avoir bon pied, bon œil*					
V ADJ N ADJ N		*découvrir saint Pierre pour couvrir saint Paul*					
V ADJ N PREP V ADJ N		*avoir plus grands yeux que grand ventre*					

V ADV ADJ N que ADJ N		*chanter toujours la même chanson*				
V ADV PREP DET N PREP DET N		*avoir plus d'un tour dans son sac*				
V ADV que DET N PREP V		*n'avoir plus que ses yeux pour pleurer*				
V COORD N COORD N	*n'avoir ni queue ni tête*					
V des ADJ N				*prendre des grands airs*		
V DET ADJ N		*attendre un heureux événement*				
V DET ADJ N du N		*se croire le premier moutardier du pape*				
V DET ADJ N PREP DET N		*mettre les petits plats dans les grands*				
V DET ADJ N PREP N		*aller son petit bonhomme de chemin*				
V DET ADJ N PREP PRON		*se faire une haute idée de soi*				
V du ADJ N			*se donner du bon temps*			
V PREP ADJ N ADJ N	*faire contre mauvaise fortune bon cœur*					
V PREP N DET N ADJ		*ramener au bercail une brebis égarée*				
V PREP N N PREP N	*faire à Dieu barbe de feurre*					

ANNEXE 3

Combinaisons Noyaux-Eléments périphériques repérées dans la liste initiale : nombre d'entrées dans la liste initiales décrites par chaque patron

		NOYAUX :	V N	V DET N	V du N	V des N	V de DET(la|l') N	V DET DET N	TOTAL
	Noyaux sans éléments périphériques		469	1953	86	61	69	21	2659
		Nombre de lexies							
Suites d'éléments périphériques antéposées au noyau	ADV	1	5	46		1		3	55
	ADV PREP	2		7					7
	PREP	1	2	36				2	40
	PRON	1		1		1	1		3
									0
									0
Suites d'éléments périphériques postposées au noyau	ADJ	1	31	189		3	1		224
	ADJ DET N	3		1					1
	ADJ N	2		1					1
	ADJ PREP DET N	4		2					2
	ADJ PREP DET N du N	6		1					1
	ADJ PREP N	3		4					4
	ADV	1	2	7					9
	ADV ADJ	2		9					9
	ADV ADJ que DET N	5		2					2
	ADV ADV	2		6					6
	ADV DET N	3		2					2
	COORD des N	3				2			2
	COORD DET ADJ N	4		1					1
	COORD DET N	3		18					18
	COORD DET N du N	5		1					1

COORD N	2	12						**12**
des ADJ N	3		1					**1**
des DET N	3	4						**4**
des N	2	1	7					**8**
DET DET N PREP DET N	6			1				**1**
DET N	2		2					**2**
DET N PREP DET N	5		1					**1**
du ADJ N	3		2					**2**
du DET N	3		1					**1**
du N	2		21				1	**22**
du N CONJ PRON V ADJ	6		1					**1**
du N PREP N	4		1					**1**
du N PREP PRON du N	6		1					**1**
du N qui V	4		1					**1**
N	1	2	1					**3**
N N	2	1	1					**2**
ni N	2		1					**1**
NUM	1		1					**1**
PREP ADJ	2		2					**2**
PREP ADJ N	3	1	3					**4**
PREP ADV	2		2					**2**
PREP ADV ADJ	3		1					**1**
PREP des N	3		1		1			**2**
PREP DET ADJ N	4						3	**3**
PREP DET DET N	4		2					**2**
PREP DET N	3	16	193	12	6	14	2	**243**
PREP DET N ADJ	4	1						**1**
PREP DET N COORD DET N AD	6		1					**1**
PREP DET N COORD DET N PREF	8		1					**1**
PREP DET N du N	5		2					**2**
PREP DET N PREP DET N	6		1					**1**
PREP DET N PREP N	5		1					**1**
PREP DET N PREP PREP N	6		1					**1**
PREP N	2	59	228	6	7	2	1	**303**
PREP N ADJ	3		3		1			**4**
PREP N du N	4		1					**1**
PREP N PREP DET N	5		3					**3**

	PREP N PREP N	4		6					**6**
	PREP PREP DET N	4		2					**2**
	PREP PRON	2	3	5	1		1		**10**
	PREP PRON PRON V V	5		1					**1**
	PREP PRON V	3					1		**1**
	PREP PRON V V	4		1					**1**
	PREP V	2		9	4				**13**
	PREP V ADV	3		1					**1**
	PREP V DET N	4		3					**3**
	PREP V N	3	3						**3**
	PREP V PREP DET N PREP N	7		1					**1**
	PREP V V	3		3					**3**
	PRON	1	1						**1**
	que N	2	2						**2**
	qui V	2		5					**5**
	qui V DET N	4		1					**1**
	qui V N PREP DET N	6		1					**1**
	qui V PREP DET N	5		4	1				**5**
Structures avec									**0**
des éléments	ADV X ADJ	1;1		5					**5**
périphériques	ADV X ADV	1;1		3					**3**
avant et après le	ADV X des N	1;2	2						**2**
noyau	ADV X PREP DET DET N	1;4		1					**1**
(X=noyau)	ADV X PREP DET N	1;3		6					**6**
	ADV X PREP N	1;2	4	8					**12**
	ADV X PREP N des N	1;4		1					**1**
	ADV X PREP N PREP DET N	1;5		1					**1**
	ADV X PREP PRON V	1;3		2					**2**
	ADV X PREP V DET N	1;4		1					**1**
	ADV X PREP V N	1;4		1					**1**
	ADV X que PRON V	1;3		1					**1**
	ADV X-là	1;1					1		**1**
	PREP X COORD DET N	1;3		1					**1**
	PREP X PREP V	1;2		3					**3**

Insertion									**0**
d'éléments	ADJ	1	54						**54**
périphériques	ADV	1		6	1				**7**
entre V et GN-	ADV ADV	2		1					**1**
OBJET	PRON de	2		1					**1**
	que	1	4						**4**
Autres structures									**0**
	ADV V DET ADJ N			1					**1**
	ↄV V PREP N DET N PREP DET N			1					**1**
	ADV V que des N					1			**1**
	PREP V DET ADJ N			2					**2**
	PREP V PREP ADJ N		2						**2**
	PRON V que du N				2				**2**
	PRON V que N		1						**1**
	V ADJ N ADJ N		1						**1**
	V ADJ N PREP V ADJ N		1						**1**
	V ADV ADJ N que ADJ N		1						**1**
	V ADV PREP DET N PREP DET N			1					**1**
	V ADV que DET N PREP V			1					**1**
	V COORD N COORD N		10						**10**
	V des ADJ N					3			**3**
	V DET ADJ N			102					**102**
	V DET ADJ N du N			1					**1**
	V DET ADJ N PREP DET N			2					**2**
	V DET ADJ N PREP N			2					**2**
	V DET ADJ N PREP PRON			1					**1**
	V du ADJ N				5				**5**
	V PREP ADJ N ADJ N		1						**1**
	V PREP N DET N ADJ			1					**1**
	V PREP N N PREP N		1						**1**
		TOTAL	**697**	**2976**	**119**	**87**	**90**	**33**	**4002**

ANNEXE 4
Pseudo-algorithme du script d'extraction

```
(Rappel utile à la compréhension de ce pseudo-algorithme : les
identifiants des mots dans le corpus sont incrémenté de deux en deux)


#INITIALISATIONS

#Définition des arguments à indiquer lors du lancement du script
freq_couplet = argument[1]     //seuil de fréquence des couplets VN
freq_expression = argument[2]  //seuil de fréquence des expressions
dispersion = argument[3]       //seuil de dispersion des couplets VN
dispersion_exp = argument[4]   //seuil de dispersion des expressions
nb_exemples = argument[5]      //nombre de phrases d'exemple à donner en
                               sortie pour chaque expression


#Initialisations des variables
expressions[] //Tableau associatif utilisé pour stocker les expressions
              extraites
motsDeLaPhrase[]   //Tableau associatif utilisé pour stocker les
                     informations lexicales et les identifiants de chaque
                     mot pour chaque phrase
dependances[] //Tableau associatif utilisé pour stocker les informations
              des relations de dépendance pour chaque phrase

dossier =""; //variable contenant le nom du sous-corpus en cours de
           traitement



#EXTRACTION
Pour chaque fichier xml dans le dossier du corpus :
  dossier = nom du sous-corpus;
  Pour chaque phrase :
    numPh = id de la phrase;
    nbmots = 0;

    #Stockage des informations lexicales dans motsDeLaPhrase[], avec les
    ids de chaque mot comme clef
    Pour chaque élément xml dans une balise <t> :
      nbmots = nbmots+2;
      idmot = id contenu dans la balise <t>;
      motsDeLaPhrase[idmot]=[lemme du mot,mot forme,catégorie];
                          //exemple :
                          motsDeLaPhrase[2]=["manger","mangeras","VERB"]

    #Stockage des informations de dépendance dans dependances[] avec les
    ids de la tête et du dépendant comme clefs et la relation de dépendance
    comme valeur
    Pour chaque élément xml dans une balise <d> :
      dependances[h,d]=nom de la dépendance;  //exemple :
      dépendances[14,12]="SUBJ"
```

```
#Parcours du tableau des dépendances
Pour chaque clef de forme [h,d] de dependances[] :
  determinant = "";
  idDebDet="";
  idFinDet=""
  phraseentiere = "";

  #Repérage et stockage des occurrences de V-N OBJET DIRECT
  Si[
  (dependances[h,d]=="OBJ") et
  (h est une des clefs de motsDeLaPhrase[]) et
  (d est une des clefs de motsDeLaPhrase[]) et
  (motsDeLaPhrase[h][2]=="VERB") et
  (motsDeLaPhrase[d][2]=="NOUN") et
  (motsDeLaPhrase[h][0] n'est pas un verbe d'état) et    //testé par
                                          expression régulière
  ((motsDeLaPhrase[h-2][0]≠"y")et(motsDeLaPhrase[h][0]≠"avoir")) et
                          //élimination des occurrences de "il y a"
  ((d-h==2) ou          //la tête et le dépendant peuvent
                          apparaître consécutivement

     ((d-h==)4) et               // la tête et le dépendant peuvent
                              être séparés par un mot qui peut :
       (motsDeLaPhrase[h+2][2]=="DET") ou  - être un déterminant
       (motsDeLaPhrase[h+2][2]=="ADV") ou  - être un adverbe
       (motsDeLaPhrase[h+2][2]=="ADJ") ou  - être un adjectif
       (motsDeLaPhrase[h+2][2]=="NUM") ou  - être un numéro
       (motsDeLaPhrase[h+2][1]=="pas") ou  - avoir pour lemme la
                                             négation "pas"
       (motsDeLaPhrase[h+2][1]=="du") ou   - avoir pour lemme "du"
       (motsDeLaPhrase[h+2][2]=="de")      - avoir pour lemme "de"
     ) ou
     (d-h=)6) et           //la tête et le dépendant peuvent être
                           séparés par deux mots qui peuvent être :
     (
     (motsDeLaPhrase[h+2][0]=="de") et - de + la/l'
     (motsDeLaPhrase[h+4][0]=="le"))
     ou
     ((motsDeLaPhrase[h+2][0]=="pas") et - pas DET
     (motsDeLaPhrase[h+4][2]=="DET"))
     ou
     ((motsDeLaPhrase[h+2][0]=="que") et - que DET
     (motsDeLaPhrase[h+4][2]=="DET")) ou

     ((motsDeLaPhrase[h+2][2]=="DET") et - DET DET
     (motsDeLaPhrase[h+4][2]=="DET"))
     )
  )
  ]:

    expression="";//initialisation de la variable du texte de
                l'expression
    expressionLemmatisee=""; //initialisation de la variable du
                  texte de la forme lemmatisée de l'expression

    #Concaténation des mots de l'expression
    Pour i de h à d :
      expression = expression += motsDeLaPhrase[i][1]+" "
```

```
        #généralisation du déterminant par DET dans le nom de
        l'expression lemmatisée
        Si (motsDeLaPhrase[i][2]=="DET") ou
        (motsDeLaPhrase[i][2]=="NUM") :

            expressionlemmatisee = expressionlemmatisee + "DET ";
            determinant = motsDeLaPhrase[i][1]+" " ; //stockage du mot
                                                    forme du déterminant
        #Suppression de "pas", "que" et des adverbes dans le nom de
        l'expression lemmatisée
        Sinon si (motsDeLaPhrase[i][1]=="pas") ou
        (motsDeLaPhrase[i][1]=="que") ou
        (motsDeLaPhrase[i][0]=="ADV") :

            expressionLemmatisee = expressionLemmatisee+"";

        #Concaténation des éléments retenus dans la chaine
        expressionLemmatisee sous forme lemme_CAT (ex: avoir_VERB)
        Sinon :
            expressionLemmatisee = expressionLemmatisee +
            motsDeLaPhrase[i][0]+"_"+motsDeLaPhrase[i][2]+" " ;

            determinant = determinant + motsDeLaPhrase[i][1]+" ";

    #Généralisation de DET DET par DET dans la chaine de
    l'expression lemmatisée (par expression régulière)
    Si expressionLemmatisee contient "DET DET":
        Remplacer "DET DET" par "DET";

    #Stockage des id de début et de fin d'expression
    idDebut=h;
    idFin=d;

    Si d-h==4 :
        idDebDet=h+2;
        idFinDet=h+2;
    Sinon d-h==6 :
        idDebDet=h+2;
        idFinDet=h+4;

    #Repérage des éléments périphériques postposés - Exemple avec un
    élément périphérique composé de 7 éléments ; si les conditions
    ne sont pas remplies, on passe au test avec 6 éléments, puis 5,
    etc.
    Si [
       (d+2 est une clef de motsDeLaPhrase[]) et
       (d+4 est une clef de motsDeLaPhrase[]) et
       [ETC JUSQU'A d+14] et
       #Test de V DET N PREP V DET N
       (
          (motsDeLaPhrase[d+2][2]=="VERB") et
          (motsDeLaPhrase[d+4][2]=="DET")et
          (motsDeLaPhrase[d+6][2]=="NOUN")et
          (motsDeLaPhrase[d+8][2]=="PREP")et
          (motsDeLaPhrase[d+10][2]=="VERB")et
          (motsDeLaPhrase[d+12][2]=="DET")et
          (motsDeLaPhrase[d+14][2]=="NOUN")
       )ou (
```

```
      #Test de PREP V PREP DET N PREP N
      (motsDeLaPhrase[d+2][2]=="PREP") et
      (motsDeLaPhrase[d+4][2]=="VERB")et
      (motsDeLaPhrase[d+6][2]=="PREP")et
      (motsDeLaPhrase[d+8][2]=="DET")et
      (motsDeLaPhrase[d+10][2]=="N")et
      (motsDeLaPhrase[d+12][2]=="PREP")et
      (motsDeLaPhrase[d+14][2]=="N")
     )
   )
]:

   #Concaténation des éléments périphériques dans les chaines du
   texte de l'expression et de l'expression lemmatisée
   expression = expression + " " +motsDeLaPhrase[d+2][1]+"
   "+motsDeLaPhrase[d+4][1]+" "+motsDeLaPhrase[d+6][1]+"
   "+motsDeLaPhrase[d+8][1]+" "+motsDeLaPhrase[d+10][1] +"
   +motsDeLaPhrase[d+12][1]+" "+motsDeLaPhrase[d+14][1];

   expressionLemmatisee = expressionLemmatisee +
   motsDeLaPhrase[d+2][0] + "_"+motsDeLaPhrase[d+2][2]+" "+
   motsDeLaPhrase[d+4][0]+"_"+motsDeLaPhrase[d+4][2]+" "+
   motsDeLaPhrase[d+6][0]+"_"+motsDeLaPhrase[d+6][2]+" "+
   motsDeLaPhrase[d+8][0]+"_"+motsDeLaPhrase[d+8][2]+" "+
   motsDeLaPhrase[d+10][0]+"_"+motsDeLaPhrase[d+10][2]+" "+
   motsDeLaPhrase[d+12][0]+"_"+motsDeLaPhrase[d+12][2]+" "+
   motsDeLaPhrase[d+14][0]+"_"+motsDeLaPhrase[d+14][2];

   #Modification de l'id de fin stocké
   idFin=d+14;

      [TESTS DES AUTRES PATRONS]

#Inclusion de l'élément précédent si c'est pas, plus, bien, y
Si [(h-2 est une clef de motsDeLaPhrase[]) et
   (
      (motsDeLaPhrase[h-2][0]=="en") et
      (motsDeLaPhrase[h][1] ne finit pas en "ant")
   ) ou
   (motsDeLaPhrase[h-2][0]=="y")ou
   (motsDeLaPhrase[h-2][0]=="plus")
]:

   expression = motsDeLaPhrase[h-2][0] +" "+ expression;
   expressionLemmatisee = motsDeLaPhrase[h-2][0]+"_" +
   motsDeLaPhrase[h-2][2] + " " + expressionLemmatisee;


   #Instanciation de la variable "couplet" qui va contenir le
   verbe et le nom de l'expression avec l'élément périphérique
   antéposé
   couplet = "("+motsDeLaPhrase[h-2][0]+
   "_"+motsDeLaPhrase[h-2][2]+") ";

   #Modification de l'id de fin stocké
   idDebut=h-2;
```

```
[Même test pour les pronoms réflexifs placés avant le verbe :
recherche de "me", "te", "se" et "nous"/"vous" dans les séquences
"nous nous" et "vous vous"]


Généraliser les déterminants en "DET" et les pronoms en "PRON"
des éléments périphériques dans la chaine d'expression
lemmatisée;   //substitution par expressions régulières

#Création de la chaine de caractères "phraseentière" par
concaténation des mots de la phrase et en ajoutant :
-  Les balises <exp></exp>  atour de l'expression
-  Les balises <head></head> autour du verbe
-  Les balises <dep></dep> autour du nom
-  Les balises <det></det> autour du déterminant
Pour i de 0 à nbmots :

  Si (i==idDebut) et (i==h) :
    phraseentiere = phraseentiere + <exp><head>"+
    motsDeLaPhrase[i][1]+"</head> ";

  Sinon si (i==idDebut) et (i≠h) :
    phraseentiere = phraseentiere +
    "<exp>"+motsDeLaPhrase[i][1]+" ";

  Sinon si (i≠idDebut) et (i==h) :
    phraseentiere = phraseentiere +
    "<head>"+motsDeLaPhrase[i][1]+"</head> ";

  Sinon si (i==idFin) et (i==d) :
    phraseentiere = phraseentiere +
    "<dep>"+motsDeLaPhrase[i][1] +"</dep></exp> ";

  Sinon si (i==idFin) et (i≠d) :
    phraseentiere = phraseentiere +
    motsDeLaPhrase[i][1]+"</exp> "

  Sinon si (i≠idFin) et (i==d) :
    phraseentiere = phraseentiere +
    "<dep>"+motsDeLaPhrase[i][1]+"</dep> ";

  Sinon si (idDebDet ≠"") et (idFinDet≠"")
  et(idFinDet==idDebDet) et (i==idFinDet) et
  (motsDeLaPhrase[i][2]≠"ADV") et
  (motsDeLaPhrase[i][2]≠"ADJ") :
    phraseentiere = phraseentiere +
    "<det>"+motsDeLaPhrase[i][1]+"</det> ";

  Sinon si (idDebDet ≠"") et (idFinDet≠"") et
  (idFinDet≠idDebDet) et
  (i==idDebDet):
    phraseentiere = phraseentiere +
    "<det>"+motsDeLaPhrase[i][1]+" ";

  Sinon si (idDebDet ≠"") et (idFinDet≠ "") et
  (idFinDet≠idDebDet) et (i==idFinDet):
    phraseentiere = phraseentiere +
    motsDeLaPhrase[i][1]+"</det> ";
```

```
    Sinon :
        phraseentiere = phraseentiere + motsDeLaPhrase[i][1]+" ";

#Concaténation de l'id de la phrase à la fin de la chaine
phraseentiere = phraseentiere + " ("+numPh+")"


#Instanciation de la variable couplet pour le stockage de
l'expression
couplet = couplet +
motsDeLaPhrase[h][0]+"_"+motsDeLaPhrase[h][2]+ " " +
motsDeLaPhrase[d][0]+"_"+motsDeLaPhrase[d][2];



#STOCKAGE DANS LE TABLEAU "expressions[]" DONT LA STRUCTURE
EST:
-  expressions[couplet][frequence]=fréquence du couplet

-  expressions[couplet][dispersion]= liste des sous-corpus
   contenant le couplet
-  expressions[couplet][expression lemmatisée 1]=[fréquence de
l'expression,[liste  des  sous-corpus],[liste  des  déterminants
possibles],listes des phrases contenant l'expression]
-  expressions[couplet][expression lemmatisée 2]=[fréquence de
l'expression,[liste       des       sous-corpus       contenant
l'expression],[liste  des  déterminants  possibles],  liste  des
phrases contenant l'expression]
etc. pour chaque expression lemmatisée formée autour du même
couplet V N

Si couplet est une des clefs de expressions[] :
    Si expressionLemmatisee n'est pas une des clefs  de
    expressions[couplet][] :

        expressions[couplet][expressionLemmatisee][0] =
        expressions[couplet][expressionLemmatisee][0]+1;

        Ajouter phraseentiere à la suite de la liste
        expressions[couplet][expressionLemmatisee];

        Si dossier n'est dans la liste
        expressions[couplet][expressionLemmatisee][1]:
            Ajouter dossier à la liste
            expressions[couplet][expressionLemmatisee][1];

        expressions[couplet]["frequence"] = expressions[couplet]
        ["frequence"]+1;

        Si dossier n'est pas dans
        expressions[couplet]["dispersion"]:
            Ajouter dossier à expressions[couplet]["dispersion"];

        Si (determinant n'est pas dans la liste
        expressions[couplet][expressionLemmatisee][2]) et
        (determinant ≠ ""):
            Ajouter determinant à
            expressions[couplet][expressionLemmatisee][2];
```

```
                Sinon :

                    expressions[couplet][expressionLemmatisee]=
                    [1,[dossier],[determinant],phraseentiere];

                    expressions[couplet]["frequence"] =
                    expressions[couplet]["frequence"]+1;

                    Si dossier n'est pas dans
                    expressions[couplet]["dispersion"] :
                        Ajouter dossier à expressions[couplet]["dispersion"];

            Sinon :

                Créer expressions[couplet];
                expressions[couplet]["dispersion"]=[dossier];
                expressions[couplet]["frequence"]=1
                expressions[couplet][expressionLemmatisee]=
                [1,[dossier],[determinant],phraseentiere];

        #Réinitilisation des tableaux contenant les informations lexicales
        et de dépendance avant le passage à la phrase suivante
        Réinitialiser motsDeLaPhrase[];
        Réinitialiser dependances[];


#REDIMENSIONNEMENT DES EXPRESSIONS LONGUES A FREQUENCE INFERIEURE A 15
Pour chaque clefCouplet de expressions[] :
    Pour chaque clefExpLem de expressions[clefCouplet]:
        Si (clefExpLem ≠ "frequence") et (clefExpLem ≠ "dispersion") :
            Si expressions[clefCouplet][clefExpLem][0] <= 15 :
                minim=Faux; //booléen qui prend la valeur Vrai lorsqu'une
                            expression plus courte a été trouvée ou que la
                            taille minimum d'expression a été atteinte sans
                            avoir trouvé d'expression plus courte
                cpt=7;
                nvelleEx=clefExpLem;
                Tant que (minim==Faux)et(cpt>0):
                    //Cas où l'expression fait ou est réduite à la taille d'un
                    noyau :
                    Si nvelleEx fait la taille d'un noyau :    //Reconnaissance
                                                de la taille d'un noyau par expression
                                                régulière qui décrit V (det) N
                        //Cas où l'expression réduite à la taille minimum
                        correspond à une expression stockée :
                        Si nvelleEx est une des clefs de
                        expressions[clefCouplet][]:
                            expressions[clefCouplet][nvelleEx][0]=
                            expressions[clefCouplet][nvelleEx][0]+1;//Incrémentation
                                                de la fréquence de l'expression

                            Pour chaque nom_de_dossier dans la liste //Ajout des
                            expressions[clefCouplet][clefExpLem][1]: noms sous-
                                                corpus si besoin
                                Si nom_de_dossier n'est pas dans la liste
                                expressions[clefCouplet][nvelleEx][1] :
                                    Ajouter nom_de_dossier à
                                    expressions[clefCouplet][nvelleEx][1];
```

```
                Pour chaque determinant dans la liste   //Ajout des
                expressions[clefCouplet][nvelleEx][2]:  déterminants si
                                                          besoin
                    Si determinant n'est pas dans la liste
                    expressions[clefCouplet][nvelleEx][2] :
                        Ajouter determinant à
                        expressions[clefCouplet][nvelleEx][2];

                Ajouter expressions[clefCouplet][clefExpLem][3])à
                expressions[clefCouplet][nvelleEx];   //Ajout des
                                                  phrases d'exemple

                expressions[clefCouplet][clefExpLem][0]=-1; /Attribution
                de la valeur de fréquence -1 à l'entrée de l'expression
                redimensionnée
            minim=Vrai;

        //Cas où la taille de l'expression excède celle d'un noyau :
        Sinon :
            Suppression du dernier mot de nvelleEx ;
            //Cas où la réduction de la taille de l'expression produit
            une expression déjà existante :
            Si nvelleEx est une clef de la liste
            expressions[clefCouplet][]:

                expressions[clefCouplet][nvelleEx][0] =
                expressions[clefCouplet][nvelleEx][0]+1;

                Pour chaque nom_de_dossier dans la liste
                expressions[clefCouplet][clefExpLem][1]:

                    Si nom_de_dossier n'est pas dans la liste
                    expressions[clefCouplet][nvelleEx][1] :

                        Ajouter nom_de_dossier à
                        expressions[clefCouplet][nvelleEx][1];

                Pour chaque determinant dans la liste
                expressions[clefCouplet][clefExpLem][2]:

                    Si determinant n'est pas dans la liste
                    expressions[clefCouplet][nvelleEx][2] :

                        Ajouter determinant à
                        expressions[clefCouplet][nvelleEx][2];

                Ajouter expressions[clefCouplet][clefExpLem][3] à
                expressions[clefCouplet][nvelleEx];

                expressions[clefCouplet][clefExpLem][0]=-1;
                minim=Vrai;

            //Cas ou le redimensionnement de l'expression ne l'amène
            pas à correpondre à une autre expression :
            Sinon :
                minim=Faux;
```

```
#SUPPRESSION DES EXPRESSIONS AYANT ETE REDIMENSIONNEES
Pour chaque clefCouplet de expressions[] :
  Pour chaque clefExpLem de expressions[clefCouplet][] :
    Si expressions[clefCouplet][clefExpLem][0]==-1 :
      Supprimer expressions[clefCouplet][clefExpLem];


#REGROUPEMENT DES DETERMINANTS
dets[] //Initialisations

Pour chaque clefCouplet de expressions[] :
  Pour chaque clefExpLem de expressions[clefCouplet][] :
    Si (clefExpLem ≠ "frequence") et (clefExpLem ≠ "dispersion") :
      deters=""
      Si expressions[clefCouplet][clefExpLem][2] ≠ liste vide:
        Pour chaque det dans la liste
        expressions[clefCouplet][clefExpLem][2] :
          Si (det=="le") ou (det=="la") ou(det=="les") ou (det=="l'"):
            Si 'DEF' est une clef de dets[]:
              Ajouter det à la liste dets['DEF'];
            Sinon :
              dets['DEF']=[det]
          Sinon si (det=="un") ou (det=="une") ou (det=="de") ou
          (det=="d\'") :
            Si 'IND' est une clef de dets[] :
              Ajouter det à dets['IND']
            Sinon :
              dets['IND']=[det]

            [Etc. pour tous les types de déterminants]

        //Suppression de la liste des déterminants et remplacement par
        une liste de forme DEF(le,la,les),IND(un),DEM(ce,ces)
        expressions[clefCouplet][clefExpLem][2]=[];
        Pour chaque clefDet dans dets[]:
          deters = deters + clefDet +"("
          Pour chaque det dans la liste dets[clefDet][det] :
            deters = det+",";
          deters = deters+") ";
          expressions[clefCouplet][clefExpLem][2]=deters;
          deters="";
        Réinitialiser dets[];

[Compilation de expressions[] pour sauvegarde des données]

#ECRITURE DE LA SORTIE AVEC FILTRAGE DES EXPRESSIONS A ECRIRE EN FONCTION DES SEUILS
DONNES EN ARGUMENTS AU LANCEMENT DU SCRIPT
fichSortie=fichier de sortie;
Ouvrir fichSortie;
sortie="Couplet\tExpressions lemmatisées\tDéterminants\tFréquence
couplet\tFréquence expression\tDispersion couplet\tDispersion
expression\tPhrases correspondantes\n";

Pour chaque clefCouplet de expressions[]:
  Pour chaque clefExpLem de expressions[clefCouplet]:
    Si (clefExpLem ≠ "frequence") et (clefExpLem ≠ "dispersion") :
      Si [(expressions[clefCouplet]["frequence"] >= freq_couplet) et
      (expressions[clefCouplet][clefExpLem][0] >= freq_expression)) et
      (longueur de expressions[clefCouplet]["dispersion"])>=dispersion)
      et
```

```
        (longueur de expressions[clefCouplet][clefExpLem][1]
        >=dispersion_exp)
        ]:
            sortie = sortie + clefCouplet+"\t"+clefExpLem+"\t";

            Si expressions[clefCouplet][clefExpLem][0]≠"":

                Pour chaque det dans la liste
                expressions[clefCouplet][clefExpLem][2] :
                    sortie = sortie +det+"/";

            sortie = sortie + "\t" +
            expressions[clefCouplet]["freq_couplet_minimale"];

            sortie = sortie +"\t"+ expressions[clefCouplet][clefExpLem][0];

            sortie = sortie +"\t"+ longueur de la liste
            expressions[clefCouplet]["dispersion"];

            sortie = sortie +"\t"+ longueur de la liste
            expressions[clefCouplet][clefExpLem][1];

            i=3
            Tant que (i <= nb_exemples+2) et (i < longueur de la liste
            expressions[clefCouplet][clefExpLem]) :

                sortie = sortie +
                expressions[clefCouplet][clefExpLem][i])+' //;'

                i=i+1;

            sortie = sortie +"\n";

            Ecrire sortie dans fichSortie
            sortie="";

fermer fichSortie;
```

ANNEXE 5

Liste des patrons d'éléments périphériques extraits par la version test du script d'extraction et quantification de leurs apports en termes d'utilité à extraire des expressions valides

	Catégorie du/des mot(s)	Nombre de cas où l'ajout au noyau est légitime	Nombre de cas où l'ajout au noyau n'est pas légitime	Exemples
Antéposé(s) au noyau	ADV	0	1	bien_ADV avoir_VERB lieu_NOUN
	PRONOM *en*	19	51	en_PRON croire_VERB DET oreille_NOUN en_PRON faire_VERB usage_NOUN
	PRONOM *se*	52	12	se_PRON refaire_VERB DET santé_NOUN se_PRON donner_VERB rendez-vous_NOUN
	PRONOM *y*	4	23	y_PRON passer_VERB DET journée_NOUN y_PRON faire_VERB attention_NOUN
Postposés au noyau	ADJ	52	56	faire_VERB profil_NOUN bas_ADJ jeter_VERB DET regard_NOUN furieux_ADJ
	ADV	14	27	mettre_VERB DET grappin_NOUN dessus_ADV avoir_VERB lieu_NOUN ici_ADV
	que PRON VERB	0	7	avoir_VERB DET impression_NOUN que_CONJQUE PRON aller_VERB
	COORD DET NOM	1	0	souffler_VERB DET chaud_NOUN et_COORD DET froid_NOUN
	COORD NOM	1	0	remuer_VERB ciel_NOUN et_COORD terre_NOUN
	DET NOM	0	14	avoir_VERB lieu_NOUN DET semaine_NOUN
	NOM ADJ	1	0	passer_VERB DET vitesse_NOUN supérieur_ADJ
	NOM	0	7	avoir_VERB lieu_NOUN jeudi_NOUN
	PREP ADJ	0	2	faire_VERB rien_NOUN de_PREP autre_ADJ
	PREP ADJ NOM	0	2	avoir_VERB DET air_NOUN en_PREP plein_ADJ forme_NOUN
	PREP DET NOM ADJ		1	assurer_VERB DET garde_NOUN dans_PREP DET cabinet_NOUN médical_ADJ
	PREP DET NOM	28	65	apporter_VERB DET pierre_NOUN à_PREP DET édifice_NOUN rendre_VERB visite_NOUN à_PREP DET ami_NOUN
	PREP NOM	59	45	mettre_VERB DET point_NOUN de_PREP honneur_NOUN hocher_VERB DET tête_NOUN en_PREP silence_NOUN
	PREP PRON VERBE	1	52	avoir_VERB DET souci_NOUN à_PREP PRON faire_VERB faire_VERB mine_NOUN de_PREP PRON lever_VERB
	PREP VERBE DET NOM	1	0	avoir_VERB DET droit_NOUN de_PREP garder_VERB DET silence_NOUN
	PRON VERB	1	0	avoir_VERB DET tête_NOUN PRON tourner_VERB
Insérés entre le verbe et l'objet	ADJ	15	7	faire_VERB faux_ADJ bond_NOUN avoir_VERB grand_ADJ besoin_NOUN
	TOTAL	**249**	**372**	
	Pourcentage sur le total des extractions	26,55%	39,66%	

ANNEXE 6

Typologie des cas de figure de polysémies des expressions candidates extraites par le script de test – Exemples et critères envisagés pour distinguer dans le corpus les occurrences des différents sens de chaque expression

Phénomène observé	Exemples d'expressions relevées	Critère discriminant (expression/pas expression)?	Critère de distinction entre plusieurs expressions?	Eléments pouvant servir de critère
Le déterminant définit si l'expression candidate est une expression polylexicale ou non	en_PRON avoir_VERB DET autre_NOUN	✓		Oui quand det = *d'*
	en_PRON faire_VERB DET affaire_NOUN	✓		Oui quand det = POSS
	avoir_VERB DET allure_NOUN	✓		Oui quand det=PART
	trouver_VERB DET mort_NOUN	✓		Oui quand det=*la*
C'est une expression lorsque le sens est figuré et un critère permet de trancher entre les deux possibilités	se_PRON renvoyer_VERB DET balle_NOUN	✓		Oui quand det=*la* (toutes les occurrences qui présentent ce déterminant sont au sens figuré)
	attirer_VERB DET foudre_NOUN	✓		Oui quand le nom est au pluriel
	avoir_VERB DET cafard_NOUN	✓		Oui quand det=*le*
	avoir_VERB DET dent_NOUN	✓		Oui quand det=*une* et suivi de *contre*
C'est une expression lorsque le sens est figuré et il y a des critères qui peuvent aider à filtrer les occurrences au sens figuré mais ils ne sont pas suffisants	en_PRON faire_VERB DET plat_NOUN	✓		Peut-être quand det = *tout un* ou *un*
	se_PRON arracher_VERB DET cheveu_NOUN	✓		Peut-être quand det=les
	se_PRON frotter_VERB DET main_NOUN	✓		Peut-être quand det=les
	y_PRON passer_VERB DET nuit_NOUN	✓		Peut-être quand det=*la* et négatif
	abandonner_VERB DET navire_NOUN	✓		Peut-être quand det=le
	allumer_VERB DET feu_NOUN			Peut-être quand det=le
	avoir_VERB DET absence_NOUN	✓		Peut-être quand det = IND

C'est une expression si le sens est figuré et un critère permet de ne fournir que des expressions au sens figuré. Mais, d'autres critères permettent de fournir la même expression mais également des colligation	atteindre_VERB DET sommet_NOUN	✓		Oui quand det=*des* Peut-être quand det=autre déterminant
	attendre_VERB DET feu_NOUN vert_ADJ	✓		Oui quand suivi de *de* Peut-être lorsque suivi d'autre chose ou de rien du tout
C'est une expression seulement si le sens est figuré et il n'y a pas de critère de distinction entre occurrences au sens propre et celles au sens figuré	se_PRON salir_VERB DET main_NOUN	✓		
	avoir_VERB besoin_NOUN de_PREP air_NOUN	✓		
	apprécier_VERB DET spectacle_NOUN	✓		
	faire_VERB DET bond_NOUN	✓		
	sortir_VERB DET tête_NOUN de_PREP DET eau_NOUN	✓		
Cas de polysémie : la même expression a plusieurs significations (propres et/ou figuré) et des critères permettent de distinguer les occurrences de chaque sens	avoir_VERB DET avantage_NOUN		✓	Distinction des cas où : - det=*l'* + pas suivi de *de* → sens = *être en train de gagner* - Autres cas → *avoir un atout*
	avoir_VERB DET compte_NOUN		✓	Distinction des cas où : - det = POSS → *en avoir eu pour son grade* - det = *des* → *avoir des problèmes à régler* - det = *le* → *avoir la somme/le nombre requis*
	avoir_VERB DET parole_NOUN		✓	Distinction des cas où : - det=POSS → *avoir la promesse de quelqu'un* - det=*la* → *être dans la position de locuteur* - det=*de* et voie négative → *ne pas tenir ses promesses*

L'expression est polysémique et il existe des critères pour distinguer en les divers sens de l'expression, mais ces critères ne sont pas suffisants	régler_VERB DET compte_NOUN		✓	Distinction des cas où : - det=POSS → *tuer quelqu'un* mais pas tous les cas où det=POSS a ce sens (peut aussi être *payer* ou *remédier à un différend*) - autres cas → ne peut pas être *tuer quelqu'un mais peut-être mes deux autres sens*
L'expression est polysémique (qui a une sens propre et un sens figuré), et il n'existe aucun critère de distinction	jouer_VERB DET comédie_NOUN		✓	
	faire_VERB DET bruit_NOUN		✓	
C'est une expression lorsque le sens est figuré, et l'expression figurée est polysémique, et il existe des critères pour distinguer les différentes expressions	avoir_VERB DET coeur_NOUN	✓	✓	Distinction des cas où : det=*du* det=*un* det=*le* + suivi de *à*
C'est une expression lorsque le sens est figuré. L'expression figurée est polysémique. Il existe des critères pour distinguer les différentes expressions polysémiques, mais ils ne sont pas siffusants pour trancher.	prendre_VERB DET main_NOUN	✓	✓	Oui quand det=*la* mais pas toujours bon (*tenir la main de quelqu'un → non / dans un jeu* → oui). Oui quand det=POSS mais pas toujours bon (*tenir la main de quelqu'un →* non / *épouser* → oui)
L'expression est valide lorsqu'il s'agit d'une expression plus longue dont un mot est élidé. Ce n'en est pas une lorsqu'il s'agit de la même suite de mots mais qu'aucune élision n'est présente	avoir_VERB DET casier_NOUN	✓		

ANNEXE 7

Comparaison de la sortie des deux extractions (par patrons catégoriels et par ALR)

Extraction par ALR			
Expressions	**Fréquence**	**Dispersion**	**Présent dans la sortie des patrons catégoriels ?**
poser questions	7712	9	oui
sauver la vie	1801	9	oui
prêta attention	1007	9	oui
joué le rôle	678	9	oui
résolu le problème	503	9	oui
échangèrent regards	402	8	non
élever enfant	369	9	oui
donnait des coups de pied	315	9	oui
perdre votre temps	246	8	oui
faciliter la tâche	222	9	oui
rompit le silence	201	9	oui
éveilla soupçons	182	9	oui
arborant sourire	165	9	non
contester décision	144	8	oui
racontait cette histoire	133	9	oui
décrocha son téléphone	123	8	oui
brossa les dents	114	8	oui
redonner vie	109	9	oui
bloquait l' accès	100	9	oui
fournir des informations	95	9	oui
prédisant l' avenir	90	9	oui
purgera sa peine	83	8	oui
encaisser coups	77	9	oui
aborda ce sujet	71	9	oui
tâter le terrain	66	9	oui
me remonter le moral	64	8	oui
savourait victoire	60	8	non
collectes fonds	57	9	non
surmonter obstacles	54	9	non
attirer leur attention	51	9	oui

1) Liste de 30 expressions extraites par ALR avec l'indication de leur présence ou absence dans les sorties de l'extraction par patrons catégoriels

Extraction par patrons catégoriels			
Expressions	**Fréquence**	**Dispersion**	**Présent dans la sortie ALR ?**
poser_VERB DET question_NOUN	3883	9	oui
prendre_VERB DET verre_NOUN	658	9	non
prendre_VERB congé_NOUN	407	9	non
prendre_VERB DET précaution_NOUN	314	9	non
donner_VERB accès_NOUN	260	9	non
manger_VERB DET morceau_NOUN	210	8	non
remplir_VERB DET mission_NOUN	186	9	non
baisser_VERB DET ton_NOUN	167	9	non
porter_VERB chance_NOUN	150	9	non
effacer_VERB DET trace_NOUN	138	9	non
crever_VERB DET oeil_NOUN	128	9	oui
rendre_VERB DET service_NOUN	119	9	non
élire_VERB président_NOUN	109	9	oui
faire_VERB surface_NOUN	102	9	non
mener_VERB DET action_NOUN	96	8	non
perdre_VERB DET face_NOUN	92	9	non
donner_VERB DET chair de poule_NOUN	87	9	non
rétablir_VERB DET ordre_NOUN	83	9	oui
reprendre_VERB DET service_NOUN	79	9	non
atteindre_VERB DET limite_NOUN	75	9	non
poser_VERB DET lapin_NOUN	71	8	non
faciliter_VERB DET vie_NOUN	68	9	non
défendre_VERB DET cause_NOUN	65	9	non
mordre_VERB DET poussière_NOUN	63	9	non
donner_VERB confiance_NOUN	60	8	non
décrocher_VERB DET titre_NOUN	58	9	non
faire_VERB faux_ADJ bond_NOUN	56	9	non
y_PRON laisser_VERB DET peau_NOUN	54	9	non
renverser_VERB DET gouvernement_NOUN	52	9	non
reprendre_VERB DET rêne_NOUN	50	9	non

2) Liste de 30 expressions extraites par patrons catégoriels avec l'indication de leur présence ou absence dans les sorties de l'extraction par ALR

ANNEXE 8

Extrait de l'annotation de la validité des expressions candidates en sortie du script d'extraction (premières entrées par ordre de fréquence décroissant)

Couplet	Expressions lemmatisées	Validité	Déterminants	Fréquence expression	Dispersion expression	Fréquence couplet	Dispersion couplet
avoir_VERB besoin_NOUN	avoir_VERB besoin_NOUN	Valide		27373	9	37733	9
avoir_VERB raison_NOUN	avoir_VERB raison_NOUN	Valide		11755	9	13885	9
avoir_VERB peur_NOUN	avoir_VERB peur_NOUN	Valide		10554	9	13662	9
avoir_VERB air_NOUN	avoir_VERB DET air_NOUN	Valide	des(des) /DEF(l') /POSS(son,mon,ton) /PART(de l') /IND(un,d') /AUTRES(tout l',tous l',toutes l',aucun,L',tous un,toutes un) /DEM(cet) /	9632	9	13629	9
avoir_VERB lieu_NOUN	avoir_VERB lieu_NOUN	Valide		8162	9	10488	9
avoir_VERB envie_NOUN	avoir_VERB envie_NOUN	Valide		5520	9	9264	9
avoir_VERB an_NOUN	avoir_VERB DET an_NOUN	Invalide	AUTRES(quatre,94,100,etc.)	5235	9	5728	9
avoir_VERB temps_NOUN	avoir_VERB DET temps_NOUN	Valide	des(des) /DEF(le,l',les) /POSS(leur,son,mon) /PART(du) /IND(un,de) /AUTRES(tout le,tout leur,tout ton,tout son,tout mon,tout notre,tout votre,tout ce,tous le,toute le,plusieurs,tout le) /DEM(ce) /	4898	9	6366	9
avoir_VERB impression_NOUN	avoir_VERB DET impression_NOUN	Valide	des(des) /DEF(l',les) /POSS(votre) /IND(une) /AUTRES(tous l',toutes l',L',aucune,un l') /DEM(cette) /	4752	9	5990	9
faire_VERB plaisir_NOUN	faire_VERB plaisir_NOUN	Valide		4309	9	2727	9
poser_VERB question_NOUN	poser_VERB DET question_NOUN	Valide	des(des) /DEF(la,les) /POSS(votre,leurs,sa,ses,vos,ma,mes,nos,tes,ta) /IND(une,de,d') /AUTRES(LA,toutes les,d'autres,plusieurs,aucune,quelques,toutes leurs,toutes nos,toutes ces,quelque,toutes mes) /DEM(cette,ces) /	3883	9	6275	9
hocher_VERB tête_NOUN	hocher_VERB DET tête_NOUN	Valide	AUTRES(tous la) /DEF(la) /POSS(sa,leurs) /PART(de la) /IND(une) /	3823	9	4238	9

avoir_VERB chance_NOUN	avoir_VERB DET chance_NOUN	Valide	des(des) /DEF(la) /POSS(ses,sa,mes,ma,leur,leurs,nos,notre,tes,ta,votre,vos) /PART(de la) /IND(de,une,d') /AUTRES(aucune,toutes les,toute,quelques,quelque,toutes mes,toutes ses,toutes nos,toutes tes,toutes vos,d' la,toutes leurs,toutes,toutes leur) /DEM(cette) /	3543	9	4474	9
secouer_VERB tête_NOUN	secouer_VERB DET tête_NOUN	Invalide	DEM(cette) /AUTRES(tous la) /DEF(la) /POSS(leur,ma,sa,leurs,votre) /IND(une) /	3521	9	3869	9
ouvrir_VERB porte_NOUN	ouvrir_VERB DET porte_NOUN	Partiellement valide	des(des) /DEF(les,la,le) /POSS(ses,leur,leurs,sa,notre,ta,votre,ma,tes,nos,vos,mon) /IND(une) /AUTRES(aucune,toutes les,d'autres,chaque) /DEM(cette,ces) /	3420	9	4279	9
avoir_VERB droit_NOUN	avoir_VERB DET droit_NOUN	Valide	des(des) /DEF(le,les,l',la) /POSS(mes,mon) /IND(un,de) /AUTRES(tous les,aucun,tout le,tout les,tous,tous le) /DEM(ce) /	3399	9	4331	9
avoir_VERB idée_NOUN	avoir_VERB DET idée_NOUN	Valide	des(des) /DEF(l',les) /POSS(son,ses,mes,leurs,mon,ton,leur,vos,votre,notre) /PART(de l') /IND(une,d') /AUTRES(aucune,d'autres,de ces,quelques,de telles,toutes ses,toutes ces,toutes les,plusieurs,quelque) /DEM(cette,ces) /	2997	9	4546	9
avoir_VERB problème_NOUN	avoir_VERB DET problème_NOUN	Valide	des(des) /DEF(les,le) /POSS(leurs,vos,ses,son,mes,tes,nos,mon) /PART(du) /IND(de,un,d',une) /AUTRES(aucun,d'autres,quelques,tous leurs,tous nos,tous des,plusieurs,tous ses,tous un,aucuns) /DEM(ce,ces) /	2933	9	3637	9
avoir_VERB mal_NOUN	avoir_VERB DET mal_NOUN	Partiellement valide	des(des) /DEF(le) /PART(du) /IND(de,d',un) /AUTRES(aucun,tous,toutes du,AUCUN) /DEM(ce) /	2872	9	4341	9
avoir_VERB tort_NOUN	avoir_VERB tort_NOUN	Valide		2723	9	3051	9
faire_VERB confiance_NOUN	faire_VERB confiance_NOUN	Valide		2717	9	4611	9
fermer_VERB oeil_NOUN	fermer_VERB DET oeil_NOUN	Partiellement valide	DEM(ces) /AUTRES(toutes vos) /DEF(les,l') /POSS(leurs,son,vos,mes,ses,tes,votre) /IND(un) /	2678	9	2798	9
faire_VERB attention_NOUN	faire_VERB attention_NOUN	Valide		2582	9	2970	9
faire_VERB peur_NOUN	faire_VERB peur_NOUN	Valide		2031	9	3315	9

ANNEXE 9

Mesures de Log Likelihood et de Z-score de 30 expressions invalides extraites

	Expressions	Fréquence	Présente en sortie Lexicoscope avec filtre LogLikelihood >10,83 ?	Log Likelihood	Z-score
Expressions fréquentes	avoir_VERB DET an_NOUN	5235	oui	14322,4773	147,0703
	secouer_VERB DET tête_NOUN	3521	oui	42188,3039	968,769
	avoir_VERB DET enfant_NOUN	1571	oui	1040,1675	33,8357
	appeler_VERB DET police_NOUN	1041	oui	7972,9607	289,0135
	voir_VERB DET visage_NOUN	787	oui	2322,8661	71,4093
	connaître_VERB DET nom_NOUN	639	oui	3061,8945	96,4785
	entendre_VERB DET bruit_NOUN	525	oui	6277,5623	221,6799
	lire_VERB DET journal_NOUN	435	oui	4029,3219	244,1402
	voir_VERB DET mère_NOUN	349	oui	447,7842	26,5254
	faire_VERB DET café_NOUN	216	oui	222,8815	17,1064
Expressions moyennement fréquentes	prendre_VERB DET livre_NOUN	128	oui	89,0875	10,9826
	sentir_VERB DET parfum_NOUN	100	oui	791,7997	74,864
	connaître_VERB DET mot_NOUN	90	oui	61,3398	25,6284
	rendre_VERB DET argent_NOUN	90	oui	317,5288	8,9952
	porter_VERB DET pantalon_NOUN	81	oui	1445,9283	116,3411
	sentir_VERB DET colère_NOUN	80	oui	257,6171	26,1415
	entendre_VERB DET parole_NOUN	78	oui	154,682	16,6515
	rencontrer_VERB DET président_NOUN	72	oui	599,8532	48,3228
	prêter_VERB DET argent_NOUN	69	oui	737,5046	68,5699
	voir_VERB DET avion_NOUN	40	non	0,2916	0,5419
Expressions peu fréquentes	retrouver_VERB DET maison_NOUN	25	non	9,8586	3,467
	guetter_VERB DET retour_NOUN	21	oui	133,746	28,9388
	récupérer_VERB DET objet_NOUN	20	oui	53,593	11,0186
	obliger_VERB DET entreprise_NOUN	19	oui	105,7098	23,4362
	enlever_VERB DET partie_NOUN	18	oui	23,2577	6,3515
	réprimer_VERB DET bâillement_NOUN	17	oui	281,6556	232,0465
	prendre_VERB DET canne_NOUN	16	oui	29,8371	7,2512
	faire_VERB DET hamburger_NOUN	15	non	24,6304	5,9401
	tenir_VERB DET bâton_NOUN	15	oui	51,7512	11,96
	voler_VERB DET cheval_NOUN	15	oui	77,5951	16,3671

ANNEXE 10

Mesures de Log Likelihood et de Z-score de 30 expressions valides ou partiellement valides extraites

	Expressions	Fréquence	Présente en sortie Lexicoscope avec filtre LogLikelihood >10,83 ?	Log Likelihood	Z-score
Expressions fréquentes	avoir_VERB besoin_NOUN	30748	oui	168992,6731	558,6072
	hausser_VERB DET épaule_NOUN	2563	oui	39378,858	2030,9182
	faire_VERB signe_NOUN	1241	oui	6774,3962	123,3295
	valoir_VERB DET peine_NOUN	943	oui	9715,8722	434,3758
	avoir_VERB conscience_NOUN	867	oui	1838,9268	50,903
	faire_VERB gaffe_NOUN	785	oui	5294,4209	133,136
	donner_VERB lieu_NOUN	552	oui	1589,9868	59,7803
	avoir_VERB recours_NOUN	457	oui	1749,6931	52,8792
	emboîter_VERB DET pas_NOUN	345	oui	7624,5614	858,61
	retenir_VERB DET attention_NOUN	229	oui	1698,5384	113,9691
Expressions moyennement fréquentes	ficher_VERB DET paix_NOUN	130	oui	2618,6941	317,2192
	remonter_VERB DET pente_NOUN	101	oui	1244,9197	227,7694
	faire_VERB mention_NOUN	99	oui	260,3658	22,6752
	prendre_VERB DET élan_NOUN	97	oui	411,4081	38,7763
	servir_VERB DET intérêt_NOUN	81	oui	818,1079	70,7225
	étouffer_VERB DET cri_NOUN	79	oui	920,5798	148,9296
	reprendre_VERB haleine_NOUN	72	oui	541,4307	86,185
	avoir_VERB DET répercussion_NOUN	70	oui	407,5187	27,0058
	suivre_VERB DET rythme_NOUN	67	oui	368,0177	42,7177
	prendre_VERB effet_NOUN	66	oui	35,7707	6,6675
Expressions peu fréquentes	occuper_VERB DET scène_NOUN	22	oui	59,569	11,5923
	faire_VERB DET esclandre_NOUN	20	oui	94,6589	16,8288
	avancer_VERB DET argument_NOUN	19	oui	257,362	55,4926
	éplucher_VERB compte_NOUN	17	oui	114,6208	31,2131
	aiguiser_VERB DET curiosité_NOUN	17	oui	208,354	123,362
	avoir_VERB DET répondant_NOUN	16	oui	36,7635	7,5115
	porter_VERB conseil_NOUN	16	oui	12,4793	4,1728
	faire_VERB DET poubelle_NOUN	16	non	0,7997	0,9012
	recevoir_VERB confirmation_NOUN	15	oui	128,6087	25,6249
	tenir_VERB DET conversation_NOUN	15	oui	14,9085	4,6285

ANNEXE 11

Expressions utilisées pour les tests de mesures d'association les plus fréquentes classées par ordre croissant de Log Likelihood et de Z-score

Expressions	Fréquences	Log Likelihood
avoir_VERB besoin_NOUN	30748	168992,6731
secouer_VERB DET tête_NOUN	3521	42188,3039
hausser_VERB DET épaule_NOUN	2563	39378,858
avoir_VERB DET an_NOUN	5235	14322,4773
valoir_VERB DET peine_NOUN	943	9715,8722
appeler_VERB DET police_NOUN	1041	7972,9607
emboîter_VERB DET pas_NOUN	345	7624,5614
faire_VERB signe_NOUN	1241	6774,3962
entendre_VERB DET bruit_NOUN	525	6277,5623
faire_VERB gaffe_NOUN	785	5294,4209
lire_VERB DET journal_NOUN	435	4029,3219
connaître_VERB DET nom_NOUN	639	3061,8945
voir_VERB DET visage_NOUN	787	2322,8661
avoir_VERB conscience_NOUN	867	1838,9268
avoir_VERB recours_NOUN	457	1749,6931
retenir_VERB DET attention_NOUN	229	1698,5384
donner_VERB lieu_NOUN	552	1589,9868
avoir_VERB DET enfant_NOUN	1571	1040,1675
voir_VERB DET mère_NOUN	349	447,7842
faire_VERB DET café_NOUN	216	222,8815

Expressions	Fréquences	Z-score
hausser_VERB DET épaule_NOUN	2563	2030,9182
secouer_VERB DET tête_NOUN	3521	968,769
emboîter_VERB DET pas_NOUN	345	858,61
avoir_VERB besoin_NOUN	30748	558,6072
valoir_VERB DET peine_NOUN	943	434,3758
appeler_VERB DET police_NOUN	1041	289,0135
lire_VERB DET journal_NOUN	435	244,1402
entendre_VERB DET bruit_NOUN	525	221,6799
avoir_VERB DET an_NOUN	5235	147,0703
faire_VERB gaffe_NOUN	785	133,136
faire_VERB signe_NOUN	1241	123,3295
retenir_VERB DET attention_NOUN	229	113,9691
connaître_VERB DET nom_NOUN	639	96,4785
voir_VERB DET visage_NOUN	787	71,4093
donner_VERB lieu_NOUN	552	59,7803
avoir_VERB recours_NOUN	457	52,8792
avoir_VERB conscience_NOUN	867	50,903
avoir_VERB DET enfant_NOUN	1571	33,8357
voir_VERB DET mère_NOUN	349	26,5254
faire_VERB DET café_NOUN	216	17,1064

ANNEXE 12

Exemple de redimensionnement manuel des expressions longues incluant des expressions valides – Cas de l'expression faire_VERB DET tour_NOUN

Avant le redimensionnement

Couplet	Expressions lemmatisées	Validité	Début de la liste des déterminants	Fréquence expression	Dispersion expression	Fréquence couplet	Dispersion couplet
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN	Valide	des(des) /DEF(le,les,l') /	1885	9	3112	9
(en_PRON) faire_VERB tou	en_PRON faire_VERB DET tour_NOUN	Invalide	DEF(le) /IND(un) /	59	9	59	9
(y_PRON) faire_VERB tour	y_PRON faire_VERB DET tour_NOUN	Invalide	des(des) /DEF(le) /IND(	33	8	33	8
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN complet_ADJ	Invalide	AUTRES(plusieurs) /DEF	43	9	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP DET maison_NOUN	Invalide	AUTRES(tout le) /DEF(le	54	5	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP DET pièce_NOUN	Invalide	DEF(le) /	33	5	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP DET question_NOUN	Invalide	DEF(le) /	17	6	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP DET salle_NOUN	Invalide	DEF(le) /	16	6	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP DET table_NOUN	Invalide	DEF(le) /	41	4	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP DET ville_NOUN	Invalide	DEF(le) /	18	8	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP DET voiture_NOUN	Invalide	DEF(le) /	28	5	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP France_NOUN	Invalide	DEF(le) /POSS(son) /IND	17	6	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP globe_NOUN	Invalide	DEF(le) /	16	4	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP magie_NOUN	Invalide	des(des) /DEM(ce) /POS	21	6	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP monde_NOUN	Invalide	DEF(le) /IND(un) /	149	9	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP pâté_NOUN de_PREP r	Invalide	DEF(le) /	20	4	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP propriétaire_NOUN	Invalide	DEF(le,l') /	20	6	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP quartier_NOUN	Invalide	DEF(le) /	18	7	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN de_PREP table_NOUN	Invalide	DEF(le) /IND(un) /	21	6	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN dehors_ADV	Invalide	IND(un) /	15	5	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN en_PREP ville_NOUN	Invalide	IND(un) /	20	7	3112	9
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN en_PREP voiture_NOUN	Invalide	DEM(ce) /IND(un) /	23	5	3112	9

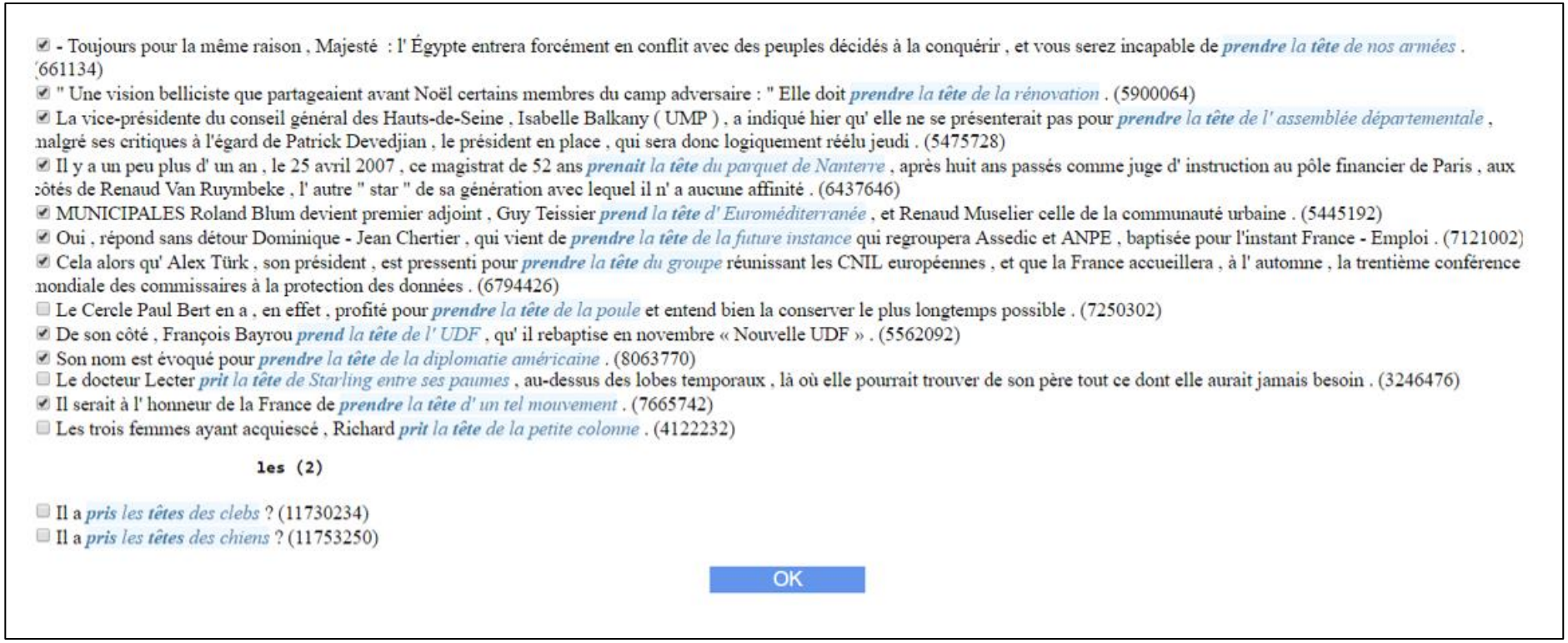
Après le redimensionnement

Couplet	Expressions lemmatisées	Validité	Début de la liste des déterminants	Fréquence expression	Dispersion expression	Fréquence couplet	Dispersion couplet
faire_VERB tour_NOUN	faire_VERB DET tour_NOUN	Valide	des(des) /DEF(le,les,l') /	**2567**	9	**3204**	9

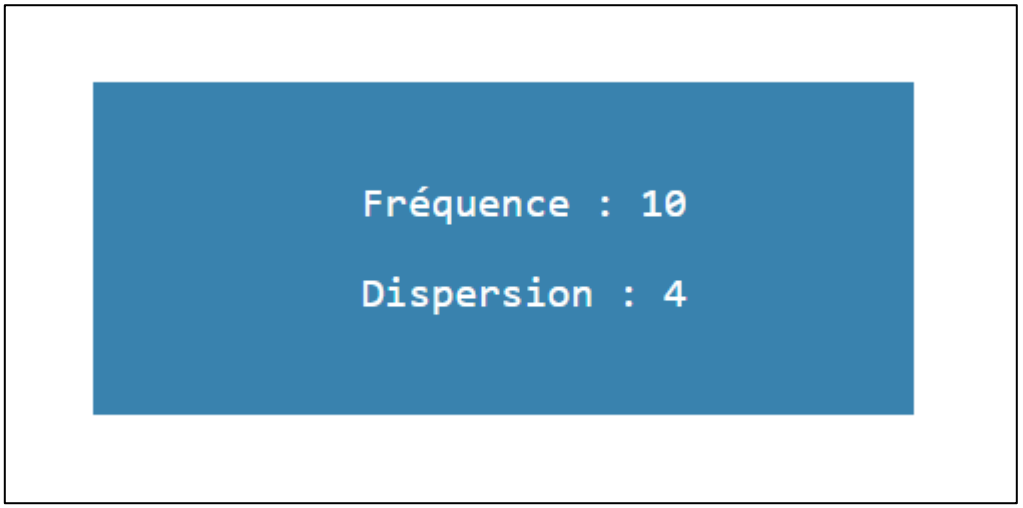
ANNEXE 13

Capture d'écran d'une partie d'un formulaire de désambiguïsation

1) Sélection de 10 phrases contenant l'expression *prendre_VERB DET tête_NOUN* avec le sens de *prendre les commandes* :



2) Affichage du résultat à partir des 10 phrases cochées



ANNEXE 14

Réponses recueillies dans le cadre de l'enquête menée auprès de spécialistes de l'enseignement de FLE

Expressions proposées				Réponses obtenues		
Expression	Fréquence	Dispersion	Type	Synonymie	Registre	Place dans lexique fondamental
avoir besoin	30748	9	Collocation	oui (4)	oral soutenu courant familier soutenu courant, écrit et oral standard, écrit et oral	oui (3)/oui mais apparaît déjà au niveaux A1/A2 (1)
avoir raison	11953	9	Collocation	non (3)/oui(1)	courant soutenu oral et écrit familier courant standard oral et écrit	oui (4)
hausser les épaules	2563	9	Collocation	non (4)	courant oral? Familier soutenu courant formel oral et écrit	non (1) / oui(3)
rendre visite	1244	9	Collocation	oui (4)	courant (oral et écrit) familier soutenu courant oral et écrit standard	oui (4)
foutre le camp	544	9	Expression figée	oui (4)	familier oral familier familier oral oral familier	oui (1) / non (3)
jeter un œil	382	9	Expression figée	oui (4)	courant (oral et écrit) familier familier oral oral familier	oui (2)/nsp(1)/non(1)
briser le cœur	353	9	Expression métaphorique	oui (4)	courant soutenu oral et écrit familier courant formel oral et écrit	peut-être (1) /nsp(1)/ oui (2)

avoir raison de (triompher	217	9	Collocation	non (1) / oui(3)	soutenu écrit et oral soutenu formel écrit formel oral et écrit	non (3)/oui(1)
mettre les bouchée doubl	84	9	Expression figée	oui (4)	familier/courant oral familier courant formel oral et écrit	non (3) / oui(1)
porter une accusation	23	9	Collocation	oui (4)	courant/soutenu (oral et écrit) soutenu standard écrit formel	oui(1)/oui dans une spécialité (2)/non (1)
avoir élan	22	5	Collocation	non (1) / oui(3)	courant/soutenu oral et écrit soutenu écrit formel oral et écrit	non (3)/oui(1)
déjouer un piège	22	5	Collocation	oui (4)	soutenu courant oral et écrit familier courant formel oral et écrit	non (3)/oui(1)
explorer la possibilité	20	9	Collocation	oui (4)	courant/soutenu à l'écrit soutenu écrit, écrit formel formel oral ou écrit	oui(1)/oui pour des universitaires (2) / non (1)
pousser le bouchon	19	9	Expression métaphorique	oui (4)	familier oral familier oral familier familier, oral ou écrit informel	non (3)/oui(1)
avoir un grain	10	4	Expression métaphorique	oui (4)	familier oral familier familier oral oral familier	non (3) / oui(1)
avoir les épaules	4	1	Expression figée	oui (4)	courant écrit oral familier familier familier oral	non (3)/oui(1)

ANNEXE 15

Exemple sur quatre expressions des résultats des caractéristiques annotées

Expression lemmatisée	avoir_VERB besoin_NOUN	dépasser_VERB DET borne_NOUN	fermer_VERB DET oeil_NOUN	reconnaître_VERB DET fait_NOUN
Type	Collocation	Expression figée	Expression métaphorique	Collocation
Verbe support?	oui			
FL	Oper1			
Définition Wikitionnaire	*Être dans le besoin de quelque chose ; être dans un lien de nécessité, de dépendance.*	*Aller trop loin dans ses réactions, aller au-delà du raisonnable, en faire trop.*	*Faire comme si l'on avait pas vu ; feindre l'ignorance.*	
Définition DEM	*manquer de / ê ds nécessité de*	*exagérer*	*refuser d voir*	
Déterminants		DEF(les) /	DEM(ces) /AUTRES(toutes vos) /DEF(les,l') /POSS(leurs,son,vos,mes,ses,tes,votre) /IND(un) /	DEM(ces,ce) /des(des) /AUTRES(tous les,d'autres) /DEF(les,le) /POSS(ses,leurs) /
Fréquence expression	30748	212	300	387
Dispersion expression	9	9	9	9
Fréquence couplet	37733	217	2798	396
Dispersion couplet	9	9	9	9

ANNEXE 16

Comparaison des extractions des sujets des expressions par méthodes syntaxiques et de patrons catégorielles

Expression lemmatisée	Sujet_Patrons	Sujet_Syntaxe	Fréquence
avoir_VERB besoin_NOUN	SNnominal /SNpronominal(je,nous,il,on,tu,vous,ça,cela,)	SNnominal /SNpronominal(je,nous,il,celui-ci,on,ce,tu,vous,quelqu'un,celui,d'aut	30748
avoir_VERB peur_NOUN	SNnominal /SNpronominal(il,je,nous,on,vous,tu,ça,cela,)	SNnominal /SNpronominal(il,je,on,nous,vous,me,beaucoup,ce,tu,celui-ci,le nôtr	12233
donner_VERB DET impression_NOUN	SNnominal /SNpronominal(il,on,cela,tu,ça,nous,vous,je,)	SNnominal /SNpronominal(il,cela,on,je,tu,ça,vous,tout,celui-ci,nous,l'un,certain	1129
avoir_VERB tort_NOUN	SNnominal /SNpronominal(il,on,je,nous,vous,tu,cela,)	SNnominal /SNpronominal(nous,je,il,on,celui,l'un,vous,le nôtre,celui-ci,tu,cela,c	2723
avoir_VERB DET chance_NOUN	SNnominal /SNpronominal(je,on,il,vous,tu,nous,cela,ça,)	SNnominal /SNpronominal(je,il,on,vous,nous,qui,chacun,celui,tout,peu,tu,quelc	4064
faire_VERB DET amour_NOUN	SNnominal /SNpronominal(il,je,on,nous,vous,tu,cela,)	SNnominal /SNpronominal(il,tu,je,vous,on,celui-ci,nous,le,ça,lui,beaucoup,tout,	1415
faire_VERB DET tour_NOUN (2)	SNnominal /SNpronominal(on,il,je,tu,nous,vous,ça,)	SNnominal /SNpronominal(nous,il,on,je,leur,lui,vous,tu,certains,me,un peu,l'un	1084
prendre_VERB DET photo_NOUN	SNnominal /SNpronominal(on,je,il,tu,vous,nous,)	SNnominal /SNpronominal(on,il,vous,je,nous,quelqu'un,l'un,qui,tu,le,te,chacun,	847
avoir_VERB conscience_NOUN	SNnominal /SNpronominal(je,nous,on,il,vous,tu,cela,ça,)	SNnominal /SNpronominal(je,tout,nous,on,il,chacun,beaucoup,vous,celui,celui-	867
faire_VERB état_NOUN	SNnominal /SNpronominal(il,vous,on,nous,je,)	SNnominal /SNpronominal(il,lui-même,celui-ci,vous,d'autres,on,tout,l'un,certai	697
porter_VERB DET nom_NOUN	SNnominal /SNpronominal(il,vous,cela,je,nous,tu,ça,)	SNnominal /SNpronominal(il,cela,vous,je,l'un,chacun,celui,tu,qui,certains,celui-	597
risquer_VERB DET vie_NOUN	SNnominal /SNpronominal(il,on,je,nous,tu,vous,)	SNnominal /SNpronominal(il,on,je,vous,certains,tu,nous,lui,le,chacun,ça,quelqu	533
tendre_VERB DET oreille_NOUN	SNnominal /SNpronominal(il,on,je,vous,nous,tu,)	SNnominal /SNpronominal(il,on,je,tout,personne,vous,lui,moi,tu,quelque chose	460
prendre_VERB position_NOUN	SNnominal /SNpronominal(il,je,on,vous,nous,tu,)	SNnominal /SNpronominal(il,je,on,vous,leur,l'un,celui-ci,nous,qui,moi,tu,lui,cha	423
reconnaître_VERB DET fait_NOUN	SNnominal /SNpronominal(je,il,)	SNnominal /SNpronominal(il,je,l'un,tout,celui-ci,vous,)	387
avoir_VERB DET avantage_NOUN	SNnominal /SNpronominal(il,vous,ça,nous,je,on,cela,tu,)	SNnominal /SNpronominal(il,vous,ça,je,on,celui-ci,cela,tu,celui,quelqu'un,ce,cha	351
faire_VERB office_NOUN	SNnominal /SNpronominal(il,nous,on,vous,ça,je,tu,)	SNnominal /SNpronominal(il,ça,celui-ci,on,cela,nous,celui-là,je,vous,tu,)	332
donner_VERB DET avis_NOUN	SNnominal /SNpronominal(vous,il,nous,je,on,tu,)	SNnominal /SNpronominal(il,chacun,vous,nous,on,d'autres,je,tu,moi,)	311
avoir_VERB DET rôle_NOUN	SNnominal /SNpronominal(il,on,je,vous,nous,tu,)	SNnominal /SNpronominal(il,on,je,quel,celui-ci,tout,nous,cela,chacun,vous,tu,)	294
faire_VERB pipi_NOUN	SNnominal /SNpronominal(je,il,on,nous,tu,)	SNnominal /SNpronominal(je,le,nous,il,tu,on,vous,quelqu'un,me,ça,)	276
reprendre_VERB DET esprit_NOUN	SNnominal /SNpronominal(je,il,tu,vous,on,)	SNnominal /SNpronominal(on,lui,je,il,le,tu,vous,celui,)	265
reprendre_VERB DET route_NOUN	SNnominal /SNpronominal(il,je,on,tu,vous,)	SNnominal /SNpronominal(il,celui-ci,je,on,vous,tu,nous,)	248
faire_VERB DET unanimité_NOUN	SNnominal /SNpronominal(il,cela,ça,je,vous,)	SNnominal /SNpronominal(il,un,celui,ça,je,vous,)	232
faire_VERB mouche_NOUN	SNnominal /SNpronominal(il,on,ça,je,)	SNnominal /SNpronominal(il,qui,cela,vous,on,nous,l'un,je,chacun,tu,ça,celui,)	220
faciliter_VERB DET tâche_NOUN	SNnominal /SNpronominal(nous,cela,il,vous,ça,tu,je,)	SNnominal /SNpronominal(cela,ça,il,on,je,celui-ci,qui,vous,tu,tout ce,personne,	202

ANNEXE 17

Comparaison des extractions des compléments des expressions par méthodes syntaxique et par patrons catégoriels

Expression lemmatisée	Comp_Patrons_Tiers	Comp_Patrons_dixième	Comp_syntaxe_Tiers	Comp_Syntaxe_dixième	Fréquence expression
avoir_VERB peur_NOUN		SV(que)/SN(de)/Vinf(de)/		SV(que)/SN(de)/SVinf(de)/	12233
avoir_VERB DET chance_NOUN	Vinf(de)/	Vinf(de)/	SVinf(de)/SV(que)/	SV(que)/SVinf(de)/	4064
avoir_VERB tort_NOUN		Vinf(de)/			2723
faire_VERB DET amour_NOUN		SN(avec)		SN(avec)/	1415
donner_VERB DET impression_NOUN	Vinf(de)/	Vinf(de)/SN(de)/SV(que)/	SVinf(de)/	SN(de)/SV(de)/SVinf(de)/	1129
faire_VERB DET tour_NOUN (2)	SN(de)/	SN(à)/SN(de)/SN(dans)/	SN2(de)/	SN(dans)/SN(en)/SVinf(pour)/SN(de)/	1084
avoir_VERB conscience_NOUN	SN(de)/	SV(que)/Vinf(de)/SN(de)/	SN2(de)/	SV(que)/SN(de)/	867
prendre_VERB DET photo_NOUN		SN(de)/		SN(de)/	847
faire_VERB état_NOUN	SN(de)/	SN(de)/	SN(de)	SN(dans)/SN(de)/	697
porter_VERB DET nom_NOUN	SN(de)/	SN(de)/	SN(de)/	SN(de)/	597
risquer_VERB DET vie_NOUN		Vinf(pour)/SN(de)/SN(pour)/		SN(pour)/SN(de)/SVinf(pour)/	533
tendre_VERB DET oreille_NOUN				SVinf(pour)/	460
prendre_VERB position_NOUN		SN(sur)/		SN(dans)/	423
reconnaître_VERB DET fait_NOUN					387
avoir_VERB DET avantage_NOUN	Vinf(de)/	Vinf(de)/SN(de)/	SVinf(de)/	SN(de)/SN(sur)/SVinf(de)/	351
faire_VERB office_NOUN	SN(de)/	SN(de)/	SN(de)/	SN(de)/	332
donner_VERB DET avis_NOUN	SN(sur)/	SN(sur)/			311
avoir_VERB DET rôle_NOUN	Vinf(à)/	SN(dans)/SN(de)/Vinf(à)/		SN(de)/	294
faire_VERB pipi_NOUN		SN(à)/SN(dans)/		SN(dans)/	276
reprendre_VERB DET esprit_NOUN					265
reprendre_VERB DET route_NOUN		SN(de)/		SN(de)/	248
faire_VERB DET unanimité_NOUN					232
faire_VERB mouche_NOUN					220
faciliter_VERB DET tâche_NOUN		SN(de)/		SN(de)/	202

Les cases en vert signifient que la complémentation de l'expression nous semble correcte et complète. Celles en jaune sont correctes mais partiellement complètes ou comportent du bruit. Celles en rouge sont fausses.

Quelques explications sont les suivantes :

- Les compléments de l'expression *faire un tour* sont considérés comme faux car il s'agit de l'entrée de notre liste pour laquelle l'expression signifie *aller se promener*. Or, *faire un tour de SN*, par exemple, est possible pour l'expression lorsqu'elle exprime l'idée de *contourner, parcourir une distance circulaire*. De plus, les compléments ayant pour préposition *dans* et *en* sont possibles pour cette expression, mais ils sont modifieurs et ne font pas partie de la sous-catégorisation de l'expression.
- Pour l'expression *risquer DET vie*, la liste *Vinf(pour)/SN(de)/SN(pour)* est considérée comme juste. Cependant, cette liste comporte deux actants ; la sous-catégorisation de l'expression est en effet *SN1,SN2(de),(SN3(pour)|Vinf(pour) ;* par exemple *il risque la vie de chacun de ses hommes pour gagner cette bataille*.
- Le complément *Vinf(pour)* de l'expression *tendre l'oreille* est considéré comme faux car il s'agit d'un modifieur.
- Le complément *SN(sur)* pour l'expression *avoir un avantage* est considéré comme bruit car cette expression a été désambiguïsée ; elle a le sens de *posséder un avantage, contraire d'un inconvénient*, tandis qu'*avoir l'avantage sur* dans le sens *être en train de gagner, dominer* fait l'objet d'une entrée différente.
- Nous rejetons *Vinf(à)* pour l'expression *avoir DET rôle* car l'observation des phrases comportant cette complémentation révèle qu'il s'agit majoritairement de cas où le verbe à l'infinitif est *jouer* ; il s'agit donc d'une alternance de l'expression *jouer un rôle* et non d'une sous-catégorisation de *avoir un rôle*.

ANNEXE 18

Premières lignes (par ordre alphabétique) du tableau de sélection des compléments sous-catégorisés à partir des résultats des différentes extractions réalisées

Expression lemmatisée	Syntaxe_Tiers	Syntaxe_dixième	Patrons_Tiers	Patrons_dixième	Compléments retenus
aborder_VERB DET sujet_NOUN		SN2(avec)/SN2(de)/		SN2(de)/	**SN2(de)/SN3(avec)/**
adresser_VERB DET parole_NOUN				SN2(à)/	**SN2(à)**
apporter_VERB DET soutien_NOUN		SN2(à)/SN2(dans)/	SN2(à)/	SN2(à)/	**SN2(à)/SN2(dans)/**
apprendre_VERB DET nouvelle_NOUN					
arranger_VERB DET chose_NOUN					
assumer_VERB DET responsabilité_NOUN	SN2(de)/	SN2(de)/		SN2(de)/	**SN2(de)/**
assurer_VERB DET sécurité_NOUN	SN2(de)/	SN2(de)/	SN2(de)/	SN2(de)/	**SN2(de)/**
atteindre_VERB DET objectif_NOUN		SN2(de)/		SN2(de)/	**SN2(de)/**
attirer_VERB DET attention_NOUN	SN2(de)/	SN2(de)/		SN2(de)/SN2(sur)/	**SN2(de)/**
avoir_VERB accès_NOUN	SN2(à)/	SN2(à)/	SN2(à)/	SN2(à)/ADJ(à)/	**SN2(à)/**
avoir_VERB affaire_NOUN		SN2(à)/	SN2(à)/	SN2(à)/PRON(à)/ADJ(à)/	**SN2(à)/**
avoir_VERB besoin_NOUN	SN2(de)/	SN2(de)/SVinf2(de)/	SN2(de)/	Vinf2(de)/SN2(de)/PRON(de)/	**SN2(de)/SVinf2(de)/SV2(que)**
avoir_VERB confiance_NOUN	SN2(en)/	SN2(dans)/SN2(en)/	PRON(en)/	PRON(en)/SN2(en)/	**SN2(dans)/SN2(en)/**
avoir_VERB connaissance_NOUN	SN2(de)/	SN2(de)/	SN2(de)/	SN2(de)/	**SN2(de)/**
avoir_VERB conscience_NOUN	SN2(de)/	SV2(que)/SN2(de)/	SN2(de)/	VS2(que)/Vinf2(de)/SN2(de)/	**SV(que)/SN2(de)/**
avoir_VERB DET accident_NOUN		SN2(de)/		SN2(de)/	
avoir_VERB DET air_NOUN		SN2(de)/		SN2(de)/Vinf2(de)/	**SN2(de)/**
avoir_VERB DET autorisation_NOUN	SVinf(de)/	SVinf2(de)/SN2(de)/	Vinf(de)/	Vinf2(de)/SN2(de)/	**SVinf2(de)/SN2(de)/**
avoir_VERB DET avantage_NOUN	SVinf(de)/	SN2(de)/SVinf2(de)/	Vinf(de)/	Vinf2(de)/SN2(de)/	**SN2(de)/SVinf2(de)/**
avoir_VERB DET capacité_NOUN		SN2(de)/	Vinf(de)/	Vinf2(de)/SN2(de)/Vinf2(à)/	**Vinf2(de)/NOUN(de)/Vinf2(à)/**
avoir_VERB DET certitude_NOUN			VS2(que)/	VS2(que)/	**VS2(que)**
avoir_VERB DET cesse_NOUN		SV2(que)/	Vinf(de)/	Vinf2(de)/VS2(que)/	**Vinf2(de)/**
avoir_VERB DET chance_NOUN	SVinf(de)/	SV2(que)/SVinf2(de)/	Vinf(de)/	Vinf2(de)/	**SV2(que)/SV2inf2(de)/**

ANNEXE 19

Différents cadres de sous-catégorisation trouvés pour les expressions fondamentales selon le modèle Lefff

Cadres de sous-catégorisation		
<Suj:sn|cln,Dloc:de-sn|en>	<Suj:sn|cln,Objd:de-sinf|en>	<Suj:sn|cln,Objde:de-sn|en|scompl>
<Suj:sn|cln,Dloc:de-sn>	<Suj:sn|cln,Objde: de-sn|de-sinf>	<Suj:sn|cln,Objde:de-sn|en>
<Suj:sn|cln,Loc:à-sn|dans-sn|y>	<Suj:sn|cln,Objde:de-sinf,Objà:à-sn|cld>	<Suj:sn|cln,Objde:de-sn>
<Suj:sn|cln,Loc:dans-sn|y>	<Suj:sn|cln,Objde:de-sinf,Obl:pour-sinf>	<Suj:sn|cln,Objde:en|scompl,Objà:à-sn|cld>
<Suj:sn|cln,Loc:dans-sn>	<Suj:sn|cln,Objde:de-sinf,Obl:sur-sn|à propos de-sn>	<Suj:sn|cln,Objede:de-sn,Objà:à-sn|cld>
<Suj:sn|cln,Loc:loc-sn|y>	<Suj:sn|cln,Objde:de-sinf,Obl:sur-sn>	<Suj:sn|cln,Objede:de-sn|en,Objà:à-sn|cld>
<Suj:sn|cln,Loc:sur-sn>	<Suj:sn|cln,Objde:de-sinf|en,Objà:à-sinf>	<Suj:sn|cln,Objde:de-sn,Objà:à-sn|cld>
<Suj:sn|cln,Objà:à-sinf,Obl:avec-sn>	<Suj:sn|cln,Objde:de-sinf|en,Objà:à-sn|cld>	<Suj:sn|cln,Obl:à-sn|dans-sn>
<Suj:sn|cln,Objà:à-sinf,Obl:pour-sinf>	<Suj:sn|cln,Objde:de-sinf|en|scompl>	<Suj:sn|cln,Obl:à-sn>
<Suj:sn|cln,Objà:à-sinf|à-scompl>	<Suj:sn|cln,Objde:de-sinf|en>	<Suj:sn|cln,Obl:avec-sn,Obl2:entre-sn>
<Suj:sn|cln,Objà:à-sn,Obl:avec-sn>	<Suj:sn|cln,Objde:de-sinf|scomp,Objà:à-sn|cld>	<Suj:sn|cln,Obl:avec-sn|entre-sn>
<Suj:sn|cln,Objà:à-sn|à-sinf|à-scompl,Obl:dans-sn>	<Suj:sn|cln,Objde:de-sinf|scompl>	<Suj:sn|cln,Obl:avec-sn>
<Suj:sn|cln,Objà:à-sn|à-sinf|y>	<Suj:sn|cln,Objde:de-sinf>	<Suj:sn|cln,Obl:contre-sn|pour-sn>
<Suj:sn|cln,Objà:à-sn|cld,Obl:avec-sn>	<Suj:sn|cln,Objde:de-sn,Obl:avec-sn>	<Suj:sn|cln,Obl:dans-sn>
<Suj:sn|cln,Objà:à-sn|cld,Obl:pour-sinf>	<Suj:sn|cln,Objde:de-sn,Obl:dans-sn|y>	<Suj:sn|cln,Obl:en-sn,Obl2:dans-sn>
<Suj:sn|cln,Objà:à-sn|cld,Obl:pour-sn>	<Suj:sn|cln,Objde:de-sn,Obl:dans-sn>	<Suj:sn|cln,Obl:entre-sn>
<Suj:sn|cln,Objà:à-sn|cld,Obl:sur-sn,Obl2:dans-sn>	<Suj:sn|cln,Objde:de-sn,Obl:pour-sinf|pour-sn>	<Suj:sn|cln,Obl:pour-sinf>
<Suj:sn|cln,Objà:à-sn|cld,Obl:sur-sn>	<Suj:sn|cln,Objde:de-sn|cld>	<Suj:sn|cln,Obl:pour-sn,Obl2:contre-sn>
<Suj:sn|cln,Objet:à-sn|cld|y,Obl:sur-sn>	<Suj:sn|cln,Objde:de-sn|de-sinf|en,Objà:à-sinf>	<Suj:sn|cln,Obl:pour-sn|pour-sinf>
<Suj:sn|cln,Objà:à-sn|cld,Obl:vers-sn>	<Suj:sn|cln,Objde:de-sn|de-sinf|en,Objà:à-sn|cld>	<Suj:sn|cln,Obl:pour-sn>
<Suj:sn|cln,Objà:à-sn|cld|y>	<Suj:sn|cln,Objde:de-sn|de-sinf|en,Obl:dans-sn>	<Suj:sn|cln,Obl:scompl|sur-sn>
<Suj:sn|cln,Objà:à-sn|cld>	<Suj:sn|cln,Objde:de-sn|de-sinf|en|de-scompl,Obl:pour-sinf>	<Suj:sn|cln,Obl:scompl>
<Suj:sn|cln,Objà:à-sn|y|à-scompl>	<Suj:sn|cln,Objde:de-sn|de-sinf|en|scompl>	<Suj:sn|cln,Obl:sur-sn,Obl2:pour-sn|contre-sn>
<Suj:sn|cln,Objà:à-sn|y>	<Suj:sn|cln,Objde:de-sn|de-sinf|en>	<Suj:sn|cln,Obl:sur-sn>
<Suj:sn|cln,Objà:à-sn>	<Suj:sn|cln,Objde:de-sn|en,Objà:à-sn|cld>	<Suj:sn|cln,Obl:vers-sn,Obl2:sur-sn>
		<Suj:sn|cln>

Table des matières

Remerciements 4

Sommaire 6

INTRODUCTION 8

PARTIE I - LA NOTION DE LEXIQUE FONDAMENTAL 10

CHAPITRE 1. DEFINITION 11

CHAPITRE 2. HISTORIQUE 12

CHAPITRE 3. CRITERES DEFINISSANT L'ASPECT « FONDAMENTAL » DU LEXIQUE 17

PARTIE II - LE DOMAINE DE LA PHRASEOLOGIE 19

CHAPITRE 4. CARACTERISATION DES EXPRESSIONS POLYLEXICALES 20

CHAPITRE 5. VERS UN LEXIQUE POLYLEXICAL VERBAL FONDAMENTAL 26

PARTIE III - POSSIBILITES D'EXTRACTION DES UNITES PHRASEOLOGIQUES 32

CHAPITRE 6. PRINCIPES GENERAUX ET MESURES D'ASSOCIATIONS 33

CHAPITRE 7. APPROCHES ET METHODES 34

PARTIE IV - REPERAGE DES EXPRESSIONS POLYLEXICALES DANS LE CORPUS 37

CHAPITRE 8. RESSOURCES ET DONNEES 38

CHAPITRE 9. PROCESSUS D'EXTRACTION 43

CHAPITRE 10. TESTS ET EVALUATIONS DE LA METHODE APPLIQUEE 49

PARTIE V - CONSTITUTION DE LA LISTE 62

CHAPITRE 11. SELECTION DES ITEMS 63

CHAPITRE 12. ESTIMATION DE LA CAPACITE DE LA METHODE A PRODUIRE UNE LISTE D'EXPRESSIONS FONDAMENTALES 72

PARTIE VI - DESCRIPTION DES EXPRESSIONS ET MODELISATION 79

CHAPITRE 13. ANNOTATIONS ET EXTRACTIONS DE CERTAINES CARACTERISTIQUES DES EXPRESSIONS POLYLEXICALES FONDAMENTALES 80

CHAPITRE 14. EXTRACTIONS COMPAREES DE LA SOUS-CATEGORISATION 88

CHAPITRE 15. MODELISATION 97

CONCLUSION ET PERSPECTIVES 106

Bibliographie 108

Table des illustrations 114

Table des annexes 116

Table des matières 163

MOTS-CLÉS : expressions polylexicales, phraséologie, lexique fondamental, extraction sur corpus

RÉSUMÉ

Ce mémoire de Master porte sur la création d'un lexique français fondamental d'expressions polylexicales verbales à partir d'une extraction sur corpus. La notion de vocabulaire fondamental a été théorisée à de nombreuses reprises et a donné lieu à plusieurs lexiques. Le domaine de la phraséologie a quant à lui amplement décrit ce phénomène linguistique particulier et omniprésent dans la langue que sont les expressions polylexicales. Cependant, rares ont été les tentatives de production d'un lexique qui soit à la fois polylexical et fondamental. Aussi, un premier axe de notre travail est de proposer une définition pour ces objets linguistiques.

Le deuxième axe porte sur le processus d'extraction des expressions fondamentales. Nous nous concentrons dans cette étude sur les expressions formées autour d'une relation Verbe-Objet direct nominal, que nous extrayons à partir d'un large corpus de genres diversifié. Nous proposons pour cela une méthode principalement basée sur des patrons catégoriels.

Enfin, le troisième axe porte sur la description et la modélisation des propriétés syntaxico-sémantiques de chacune des expressions du lexique. Elles consistent notamment en un repérage sur corpus des possibilités d'alternances syntaxiques et des cadres de sous-catégorisation des expressions.

KEYWORDS : multiword expressions, core lexicon, extraction from corpora

ABSTRACT

This Master thesis focuses on the creation of a French core lexicon of verbal multiword expressions. The concept of core lexicon has been theorized numerous times and several lexicons have been produced. On the other hand, the linguistic field of phraseology has yet given a quite detailled description of multiword expressions, which are omnipresent in language. Though, attemps to create lexicons of core multiword expressions have been few. For that reason, the first main line of our study is to give a definition of these linguistic objects.

The second main line is the creation of an extraction process for the multiword expressions. We restrain the scope of our study to expressions based on a Verb-Nominal Object construction, that we extract from a large and multi-genre corpus. The extraction method we use is mainly based on POS patterns.

The third and last main line is the description and modelisation of the core expressions' syntactical and semantical properties. Among others, it consists in extracting from the corpus the subcategorization frames and syntactical alternations allowed by each expression.