Approches multidisciplinaires pour l'étude du lexique et la construction de ressources lexicales nouvelles

Núria Gala

Approches multidisciplinaires pour l'étude du lexique et la construction de ressources lexicales nouvelles

---

Mémoire d'Habilitation à Diriger des Recherches

Spécialité Linguistique (7e section CNU)

Núria Gala Pavia

Volume I : Synthèse

05 Juin 2015

Jury :

Philippe Blache (LPL-CNRS, Aix en Pce)	Directeur de recherche
Nabil Hathout (ERSS, Toulouse)	Rapporteur
Alain Polguère (ATILF, Nancy)	Rapporteur
Horacio Saggion (UPF, Barcelone)	Rapporteur
Agnès Tutin (LIDILEM, Grenoble)	Examinatrice
Patrice Bellot (LSIS, Marseille)	Examinateur

Remerciements

Le travail que je présente dans ce mémoire n'aurait pas pu aboutir sans l'aide, le support et les échanges réguliers avec plusieurs personnes à qui je voudrais ici exprimer ma gratitude.

Premièrement, je tiens à remercier Philippe Blache d'avoir accepté d'être mon directeur de recherche (et de l'avoir fait avec enthousiasme !). Je le remercie pour ses conseils judicieux et pour la confiance qu'il m'a accordée.

Je remercie Nabil Hathout, Alain Polguère et Horacio Saggion de m'avoir fait l'honneur d'être rapporteurs de mon travail. Un grand merci pour des échanges constructifs et enrichissants, prometteurs -je l'espère- d'échanges futurs. Merci également à Agnès Tutin et à Patrice Bellot d'avoir accepté cordialement mon invitation à faire partie du jury.

Mes remerciements s'adressent très spécialement à Michael Zock pour son soutien et son œil bienveillant, et à Laurent Tichit, pour son aide précieuse : merci à deux collègues et amis d'exception. De même, et surtout, merci à Jaume Gala, mon père, pour d'inestimables conseils philologiques.

Plus globalement, je remercie les membres du CENTAL à Louvain-la-Neuve, spécialement Thomas François et Cédrick Fairon, ainsi que les membres de l'équipe Langues en Contact et Typologie (LCT) à Aix en Provence, particulièrement Médéric Gasquet-Cyrus, Sylvie Wharton, Adam Wilson et Nicolas Tournadre, pour des échanges toujours fructueux. Un grand merci également à Mathieu Mangeot du LIG à Grenoble, pour une relecture attentive et des commentaires avisés. Enfin, merci à tous les membres de mon équipe TALEP pour des collaborations précieuses, spécialement Alexis Nasr, Elisabeth Godbert et Carlos Ramisch, et merci au directeur du LIF Jean Marc Talbot et aux *LIF angels* (Nadine, Sylvie et Martine) pour leur aide, leur support et leur bonne humeur !

Un dernier merci, mais pas le moindre, à mes enfants et à mes proches, pour leurs encouragements constants et pour avoir suivi avec patience et confiance le processus de rédaction de ce mémoire.

Résumé

L'intérêt pour les ressources lexicales n'a cessé d'évoluer en fonction des besoins et des technologies. Le lexique se trouve, ainsi, au cœur de nombreuses recherches dans des domaines variés : construction de dictionnaires, apprentissage du vocabulaire, aide à la lecture, etc. Il est également le socle des outils de traitement automatique des langues et des technologies du langage au sens large.

La construction et l'enrichissement de ressources reste une tâche coûteuse et requiert des compétences dans différentes disciplines. À ce jour, les traitements automatiques permettent d'améliorer la couverture des lexiques et le caractère explicite, détaillé et approfondi des informations qu'ils contiennent. Les méthodes de construction sont ainsi diversifiées (semi-automatiques, collaboratives) et les lexiques qui en résultent sont de plus en plus dynamiques, dans une perspective de partage de données à grande échelle.

Enfin, les ressources lexicales représentent un enjeu sociétal important, parce qu'elles sont nécessaires pour développer des applications d'assistance à l'apprentissage des langues comme dans l'aide à la remédiation de pathologies de la lecture et de la parole, etc.

Dans ce mémoire pour l'obtention de l'Habilitation à Diriger des Recherches, c'est par le biais de la multidisciplinarité que nous abordons le lexique et la construction de ressources. Outre la description de quelques ressources que nous avons eu l'occasion de créer et/ou enrichir, nous apportons une mise en perspective historique et méthodologique de quelques approches et applications où le lexique reste au centre des problématiques.

Abstract

The interest in lexical resources is evolving continuously as a result of different needs and technologies. The Lexicon is central to research in various domains : lexicography, vocabulary learning, reading tools, etc. By and large, it is also the basis for natural language processing tools and language technologies in general.

Building and enriching lexical resources remains a costly and time-consuming task that requires competencies in different disciplines. At present, language processing tools enable better coverage of lexicons as well as the specific, explicit and detailed linguistic information contained within them. In addition, the methods used to build the resources have become diversified (automatic, collaborative) and the resulting lexicons tend to be increasingly dynamic, designed with a view towards large-scale linked data.

Lastly, lexical resources are a major issue for society, because they are essential in developping tools for learning languages, assistive technologies for reading and writing, etc.

With a view to obtaining the *Habilitation à Diriger des Recherches*, this thesis focuses on the Lexicon and on lexical resources in general. The issue is addressed through an interdisciplinary approach : as well as describing various resources that we have created and/or enriched, we also offer historical and methodological insight into a number of approaches and applications where the lexicon plays a central role.

Table des matières

Table des figures

Introduction

Ce document décrit nos travaux de recherche en linguistique et en traitement automatique des langues (TAL) depuis une dizaine d'années (2004-2014). Cette période correspond à notre parcours scientifique en tant que chercheur *autonome*, c'est-à-dire, au delà des travaux pour l'obtention du doctorat.

Dans le cadre de la thèse [Gal03], notre intérêt était centré sur la grammaire et la syntaxe par le biais de la formalisation de règles que nous avions intégrées à un analyseur syntaxique robuste. Ce travail nous avait également amenées à modéliser certains aspects liés à l'hétérogénéité des corpus. Enfin, il nous avait aussi sensibilisées à l'importance du lexique (ajout d'informations lexicales dans les grammaires). La suite de nos recherches a alors été guidée par une volonté d'approfondir nos connaissances sur le lexique et ses propriétés, toujours dans un cadre de traitement automatique et de construction de ressources pour des applications variées.

Bien évidemment, la notion d'*autonomie* dans la recherche scientifique est très relative car nous n'aurions pas pu avancer sans la collaboration et l'implication, à différents degrés, de plusieurs collègues qui nous ont constamment stimulée, remise en question et ouvert de nouvelles perspectives dans des disciplines connexes. Pour ne citer que ceux avec qui nous avons eu l'occasion de publier et/ou mener des projets (par ordre plus ou moins chronologique) :

- Véronique Rey (linguistique), sur l'apprentissage du vocabulaire et la morpho-phonologie
- André Valli (linguistique), sur les lexiques syntaxiques et les valences des verbes français
- Nabil Hathout (linguistique et TAL), sur la morphologie constructionnelle et l'analyse morphologique
- Michael Zock (psycholinguistique et TAL), sur les ressources lexicales et le lexique mental
- Mathieu Lafourcade (TAL), sur les vecteurs sémantiques et l'acquisition de données lexicales

- Philippe Gambette (TAL), sur les arbres de mots
- Alexis Nasr (TAL), sur les clusters sémantiques et l'analyse morpho-syntaxique
- Caroline Brun (TAL), sur l'extraction d'opinions
- Thomas François (statistique et TAL), sur les statistiques lexicales, la lisibilité et la simplification
- Delphine Bernhard (TAL), sur le traitement automatique de la morphologie et la simplification
- Anne Laure Ligozat (TAL), sur la lisibilité et la simplification
- Amalia Todirascu (TAL), sur l'analyse du discours pour la simplification
- Serge Pinto (neurolinguistique), sur la dysarthrie et la maladie de Parkinson
- Johannes Ziegler (psychologie cognitive), sur l'apprentissage de la lecture et la dyslexie

Toujours avec *le lexique comme fil conducteur*, ces collaborations nous ont menée dans des domaines et des disciplines interreliées. Comme une évidence, c'est dans cette approche multidisciplinaire que nous avons choisi de décrire nos travaux et nos perspectives de recherche dans le cadre de cette Habilitation.

Ainsi, dans ce document, on y trouvera aussi bien une synthèse qu'un approfondissement de nos travaux sur le lexique. En effet, nous avons considéré que cet exercice universitaire nous laisse suffisamment de liberté pour explorer des domaines qui nous tiennent à cœur et que nous n'avons pas pu approfondir auparavant. Sur la base de la multidisciplinarité, et tout en étant consciente que tous les domaines liés à une "science du mot" ne pourront pas être abordés, nous avons choisi d'articuler ce mémoire sous le prisme de quatre disciplines principales [1] :

- la *lexicologie* (**chapitre 1**), où nous avons voulu mener une réflexion sur le 'mot' [2], sur les unités lexicales et sur les différentes approches existantes pour appréhender ces concepts. Notre objectif a été d'approfondir ces notions au niveau théorique et via différentes analyses aussi

1. Bien que tout à fait nécessaires pour considérer le lexique dans sa globalité, des travaux sur des approches anthropologiques, ethnographiques ou socio-culturelles se trouvent en dehors du cadre de nos recherches.

2. L'utilisation de 'mot' est ici un pis-aller. Nous précisons cette notion dans le chapitre 1 de ce mémoire.

bien anciennes que contemporaines. Ainsi, nous proposons un recueil de quelques notions fondamentales pour mieux comprendre le lexique, en termes morphosyntaxiques, sémantiques et statistiques. En somme, ce chapitre constitue un état de l'art abordé sous différents prismes, notre objectif étant de montrer à quel point la description lexicale a suscité et suscite encore de l'intérêt dans des approches variées.

- la *lexicographie* (**chapitre 2**), où nous nous penchons inévitablement sur cette "activité pratique (qui) existe depuis l'antiquité, dont l'objet est précisément les unités lexicales, et qui est destinée à répertorier commodément les signes, selon un ordre convenu, et a apporter des informations à leur sujet" [Rey70a]. Notre objectif est ici de présenter les dictionnaires d'aujourd'hui et de mettre en lumière ce que l'informatique et les traitements automatiques du lexique ont entraîné au niveau des contenus de ces ressources et de la façon d'y accéder.

- le *traitement automatique des langues* (**chapitre 3**), où nous décrivons ce domaine qui, en France, "commence par la mécanisation du lexique" [Léo04]. Nous postulons ici que les traitements du lexique constituent le fondement de ce domaine car ils sont indispensables à la plupart des applications. Nous montrons quels sont ces traitements automatiques et quels sont les principaux écueils auxquels ils se heurtent. Notre objectif est également de décrire et de caractériser les ressources lexico-sémantiques existantes, créées à des fins de traitement automatique des langues, et pouvant aussi être utilisées dans d'autres applications.

- la *psycholinguistique* (**chapitre 4**), où nous abordons des aspects liés à la compréhension et à la perception des mots en tant qu'unités qui véhiculent le sens. Notre intérêt se focalise ici, principalement, sur des aspects liés à la notion de 'complexité linguistique' et 'complexité lexicale'. Enfin, nous faisons état des apports issus de l'étude de la parole pédagogique (français langue maternelle et seconde) et pathologique (Parkinson, dyslexie) dans la construction de certaines ressources.

Dans une deuxième partie de ce mémoire, le **chapitre 5** décrit le dernier aspect de nos recherches en cours. Nous y développons les grandes lignes des perspectives scientifiques que nous envisageons dans trois domaines d'application où le lexique et la description lexicale sont au cœur des préoccupations : la complexité linguistique, la lisibilité et la simplification lexicale et la construction de ressources pour des langues peu dotées.

Nous terminons ce document par une synthèse finale et un bilan global de nos contributions (recouvrant la période 2004-2014).

Première partie

Synthèse

Chapitre 1

De l'étude du lexique à sa formalisation

« On peut encore distinguer les mots d'après leur nature ou leur emploi. D'après leur nature, les uns sont plus sonores, plus nobles, plus harmonieux et en quelque sorte plus brillants, ou inversement (...). »

Cicéron. *Divisions de l'art oratoire*. 16-21.
Paris, "Les Belles Lettres", 1924 (p. 8-10).

« Without grammar very little can be conveyed, without vocabulary, nothing can be conveyed. »

Wilkins, D. (1972) *Linguistics in Language Teaching*. Edward Arnold.

Dans ce chapitre, notre objectif est d'évoquer les principales notions théoriques liées au lexique et à ses unités. Pour cela, nous avons voulu, d'une part, nous attarder sur quelques points terminologiques fondamentaux (section 1.1, glossaire terminologique à l'annexe A). D'autre part, nous avons tenu à regarder les différentes approches qui, depuis l'Antiquité, se sont penchées sur son étude (sections suivantes). Les descriptions proposées au travers des différents courants témoignent de l'évolution des perceptions que l'on a eues sur le lexique et constituent des prémisses pour sa formalisation et la construction de ressources diverses.

1.1 Précisions terminologiques

Nous avons considéré fondamental de mettre en lumière différentes notions liées aux "mots" en tant qu'unités, au Lexique en tant qu'ensemble

théorique (par rapport à Vocabulaire) et, enfin, en tant que composant de base d'une langue (par rapport à Grammaire).

1.1.1 L'unité lexicale (le mot ?)

> « Les mots sont au cœur de la connaissance linguistique puisque parler une langue consiste avant tout à combiner des mots au sein des phrases en vue de communiquer. »[Pol02]

Lorsqu'on s'intéresse au mot 'mot' on s'accorde très vite sur le fait que cette appellation est trop ambiguë et possède, intuitivement, un sens trop général dans la langue. Son usage prête ainsi à confusion (comparons, par exemple, `piano` / `piano droit` / `piano à queue` : trois mots ? quatre ? six ?). La lexicologie, et aussi la morphologie, dans une démarche scientifique, se sont dotées de termes plus précis pour caractériser et mieux appréhender ce concept.

Dans sa définition la plus générale, le 'mot' correspond au **signe linguistique** [1] : une association entre une forme et un contenu (une idée, un sens). Concrètement, tel que proposé par F. de Saussure dans son *Cours de Linguistique Générale* (1916), le signe linguistique aurait cinq propriétés :

1. L'association d'un signifiant (image acoustique ou écrite) et d'un signifié (le sens)
2. L'arbitraire du signe (l'association signifiant-signifié n'est pas motivée)
3. Son caractère figé (immutabilité du signe)
4. Son caractère évolutif (mutabilité du signe, en apparente contradiction avec le point précédent)
5. Son caractère linéaire (au niveau de la réalisation, un signe est une suite linéaire de sons ou de caractères)

Hormis les aspects sémantiques, sur lesquels nous reviendrons plus loin, la définition intuitive de 'mot' a longtemps été liée à la tradition écrite des langues occidentales (séparation par des espaces). De ce fait, l'unité lexicale est souvent plus précisément définie par rapport à sa présence dans les textes ou les discours. Le **mot-forme** (*wordform* ou *token*) désigne alors une unité linguistique "pourvue de toutes les marques que requiert la syntaxe et prosodiquement autonome" [ABD+14]. Il s'agit d'unités libres dotées (a) d'une certaine autonomie de fonctionnement et (b) de cohésion interne (intégrité lexicale, c'est-à-dire : on ne peut insérer d'autres mots à l'intérieur d'un mot-forme, on ne peut pas appliquer, à des sous-parties d'un mot-forme,

1. Dans la suite du chapitre, nous mettons en gras les notions qui nous semblent clefs.

des procédés qui s'appliquent à d'autres unités syntaxiques -coordination, extraction, etc.). D'un point de vue sémantique, un mot-forme peut contenir un ou plusieurs morphèmes (unités minimales de signification). Selon le nombre et le type de ses morphèmes (lexicaux ou grammaticaux), on peut distinguer des mots-forme avec :

- un seul morphème lexical (par exemple, 'grand', 'marche', 'pied')
- un morphème lexical et un morphème grammatical ('grande', 'marchent', 'pieds')
- un morphème lexical et plusieurs morphèmes grammaticaux ('grandes', 'marcheriez')
- deux morphèmes lexicaux ('marchepied')
- etc.

Comme son appellation l'indique, la notion de mot-forme caractérise l'unité lexicale selon des propriétés principalement formelles (traits morpho-syntaxiques, structure interne). Si la notion est importante pour les traitements automatiques au sens large (étiquetages morphologiques, désambiguïsations, etc.), en linguistique de corpus on utilise plus fréquemment le terme **mot-occurrence** pour faire référence à chaque apparition d'un mot-forme dans un corpus (dénombrement de formes, statistique lexicale).

Or l'identification des morphèmes au sein d'une unité linguistique ne suffit pas pour caractériser la notion de 'mot'. En effet, il existe des variations de forme qui ne modifient pas, pour autant, le contenu sémantique de base de l'unité (au sens du référent). Ainsi, entre `piano` et `pianos`, la notion de 'pluriel' portée par la flexion, n'entraîne pas de différence sémantique. La notion de **lexème** tient compte de cette caractéristique : il s'agit d'une "unité abstraite de ses variations flexionnelles" [ABD+14] [2]. Egalement, d'après le TLFi, le lexème est une "unité minimale de signification appartenant au lexique". L'idée est donc celle d'une unité autonome dépourvue de marques morphosyntaxiques : les mots-formes seraient, ainsi, les instanciations des lexèmes dans les textes et les discours. Lorsque la signification n'est pas lexicale (pas de référent) mais relève plutôt de la grammaire (informations morphosémantiques concernant le fonctionnement des signes linguistiques) certains courants en morphologie utilisent le terme **grammème** [ABD+14]. Il s'agit d'une différence entre lexèmes (au sens de 'mot plein' : noms, verbes, adverbes et adjectifs) et mots grammaticaux (au sens de 'mots vides' -sans référent- : conjonctions, déterminants, prépositions, etc.) [3].

2. Pour [Pol02] et selon l'approche Sens-Texte [MCP95], le lexème regroupe des mots-formes ne se distinguant que par la flexion. Il s'agit donc d'un ensemble de mots-formes dans l'axe paradigmatique (axe des substitutions).

3. La façon d'exprimer ce contenu morphosémantique varie selon les langues.

Par ailleurs, en traitement automatique des langues, on aura une préférence pour utiliser la notion de **lemme** ou même de **forme de base**. Il s'agit d'une notion fondamentale pour l'étiquetage morphologique, par exemple. Un lemme est alors l'unité de référence d'un point de vue morphologique (le choix de sa forme est totalement arbitraire : infinitif pour les verbes en français, 3e personne du présent en latin et en arabe ; forme masculine et singulier pour les noms, etc.). Un lemme ne contient que des morphèmes lexicaux [4].

Les lexèmes formellement complexes, dotés de plusieurs mots-forme regroupés dans l'axe syntagmatique (axe linéaire), sont par ailleurs appelés **locutions** ('à la veille de'), **collocations** ou **expressions polylexicales** ('en bonne voie'). Enfin, la **lexie** est un terme qui désigne un hyperonyme regroupant les notions de lexème et de locution [Pol02].

À cette vision lexicologique s'ajoute parfois une vision lexicographique : les unités sont alors appelées **vocables**, **mots-vedette** ou encore **entrées** [5]. Il s'agit des unités faisant partie d'une ressource.

1.1.2 Lexique et Vocabulaire

De façon générale, le **lexique** est l'ensemble de tous les mots d'une langue [Pol02]. À ce titre, il s'agit d'une liste riche [6], ouverte et en constante évolution (néologismes, emprunts, créations, mots vieillis, etc.). Dans cette perspective, le lexique reste une entité théorique, en quelque sorte insaisissable dans sa totalité.

Par ailleurs, le **vocabulaire** serait l'ensemble d'unités lexicales d'un individu, d'un groupe d'individus, d'un texte, etc. Il s'agit clairement d'un inventaire plus ou moins défini et limité par rapport à un usage déterminé (domaine de spécialité, zone géographique, groupe social, etc.).

Par abus de langage pour le premier, les deux termes (lexique et vocabulaire) font également référence à des ressources lexicales : *Lexique des Verbes Français*, *Vocabulaire de l'informatique*, etc.

1.1.3 Lexique et Grammaire

Le lexique et la grammaire sont les deux composants majeurs des connaissances que tout locuteur possède sur sa langue. Ainsi, d'une part, le lexique contient les unités ("briques de base"). D'autre part, la **grammaire** définit

4. En général, un seul morphème lexical, sauf dans les cas des noms composés (tire-bouchon) ou des formes lexicalisées (comme "rendez-vous", "sot-l'y-laisse", etc.), et dans les infinitifs si on considère que leur terminaison en français (-er, -re, -ir, -oir) est un morphème grammatical.

5. *C.f.* Annexe A pour des définitions précises.

6. Une langue contiendrait facilement un million d'unités lexicales, un dictionnaire en recenserait entre cinquante et cent mille, alors qu'entre trois et cinq mille suffiraient pour qu'un individu puisse communiquer normalement.

les règles qui permettent de combiner ces unités pour former des structures plus grandes (des syntagmes, des phrases). Il s'agit des deux composantes fondamentales de toute langue naturelle, mais aussi des langues artificielles ou langages de programmation [7].

1.2 Approches anciennes

Les réflexions théoriques sur le lexique s'appuient sur une longue tradition. Nous avons voulu aller jusqu'aux origines de façon à retracer un parcours historique des différents courants et savoirs qui se sont intéressés aux unités lexicales en tant que briques de base des langues.

1.2.1 Premières traces écrites

L'histoire de la linguistique est forcément liée à l'histoire de l'écriture. En effet, pour appréhender la langue et les unités qui la composent, il semble nécessaire, dans un premier temps, de nous pencher du côté des premières écritures, d'autant plus que, d'après A. Rey [Rey70a], le premier stade de l'écriture est fondamentalement lexicologique. L'adjectif 'lexicologique' doit être compris ici dans un sens plus strict : les signes graphiques correspondent à l'expression d'un objet ou, plus tard historiquement parlant, d'un concept.

Les premiers systèmes d'écriture, 3600 ans av. J. C., sont de nature pictographique, puis deviennent idéographiques lorsque les pictogrammes se constituent en système (liés les uns aux autres, dérivés les uns des autres). La picturalité se justifie par les besoins : moyen de communication et de pérennisation du message (trace). Pour les Sumériens, chaque pictogramme correspondait à une notion concrète (objets, animaux, etc.), comme en font preuve les deux mille pictogrammes des tablettes du site d'Uruk en Mésopotamie (les premiers répertoires attestés de signes linguistiques sont le Tablettes d'Ebla, listes bilingues sumérien-éblaïte datées entre 2500 et 2250 avant J. C.).

L. J. Calvet [Cal96] soutient que les changements techniques font évoluer cette "écriture des choses" vers une écriture moins picturale : perte de la ressemblance, de la "motivation" des signes graphiques qui deviennent conventionnels. Egalement, ils commencent à avoir une valeur symbolique, parfois via la combinaison de deux pictogrammes. Les idéogrammes se transforment avec le temps, ils encodent alors des sons et non pas des objets ou des concepts. L'écriture dévient alors phonétique, elle est définie ainsi par rapport à la langue (regard phonologique de l'écriture).

7. Beaucoup de principes d'analyse des langages informatiques se font par analogie aux langues naturelles. Nous verrons plus loin qu'il existe d'autres analogies entre langues naturelles et artificielles, par exemple, pour ce qui concerne la notion d'ordre alphabétique (2.1.1).

1.2.2 Descriptions morphologiques

Les premières descriptions linguistiques systématiques sont dues à Panini, grammairien de l'Inde antique (IVe siècle av. J. C.). Ses analyses du sanskrit ont donné lieu à tout un ensemble de *sutras* (grammaires ou règles formelles) avec une approche analytique des mots en tant que formes. Il distinguait les 'mots vrais' des 'mots fictifs' et la forme du contenu. Dans sa théorie morphologique, il identifiait les plus petites unités de sens qu'il distinguait des unités lexicales.

L'intérêt des grammairiens latins était lexicologique et non pas philosophiques comme c'était le cas pour les Grecs (1.2.3) : ils s'intéressaient aux régularités du système (ex. déclinaisons, flexions), distinguaient 'mots primitifs' *vs* 'mots dérivés', flexion et dérivation. Ils cherchaient à expliquer les différences entre les possibilités du système et ses réalisations (respectivement 'compétence' et 'performance' selon la terminologie de Chomsky). On ne peut qu'être surpris par cette vision moderne : (a) étude formelle et linguistique (morphologique) et (b) considération des signes lexicaux dans l'expérience humaine (usage).

> « La nature a voulu que les mots primitifs fussent en très petit nombre, afin qu'on pût les apprendre très vite ; et que les mots déclinés fussent en très grand nombre, afin qu'on pût exprimer très facilement toutes les nuances de la pensée. Pour connaître l'origine des mots primitifs, nous avons besoin de l'histoire (...) ; mais à l'égard des mots déclinés, c'est l'art qui doit nous servir de guide, et cet art repose sur un petit nombre de préceptes, qui sont très simples. »

Varron. *De la langue latine*. Livre VIII, 3-6. Traduction française. Paris, Dubochet, 1850. (Dans [Rey70a].)

> « En matière de vocabulaire tout se ramène donc à deux principes fondamentaux, l'application et la transformation (...). Pour les noms appliqués originellement aux choses leurs auteurs n'ent ont voulu qu'un minimum, mais ils ont souhaité un maximum de dérivés. »

Varron. *De la langue latine*. Livre VIII, 1-44. Traduction française. Paris, Baratin-Desbordes, 1981. (Dans [Aur89].)

À cette époque, les préoccupations étaient fondamentalement pédagogiques (enseignement du latin).

1.2.3 Considérations philosophico-sémantiques

Dans la Grèce classique, et surtout avec Aristote, l'intérêt pour les mots est principalement sémantique : ils sont considérés en tant qu'outils pour penser les choses. Aristote cherche ainsi à concilier le classement en parties du discours et les classes conceptuelles. Les mots sont alors analysés en termes de contenu sémantique, dotés de signification, représentants des idées. Les mots existent pour exprimer les idées et les concepts.

> « Spoken words are the symbols of mental experience and written words are the symbols of spoken words. Just as all men have not the same writing, so all men have not the same speech sounds, but the mental experiences, which these directly symbolize, are the same for all, as also are those things of which our experiences are the images. »

Aristote. *De l'interprétation*. 1 et 2 (pp.77-80). Traduction J. Tricot. Paris, Bibliothèque des textes philosophiques, 1966. (Dans [Rey70a].)

Plus tard, Platon se pose la question essentielle du rapport entre les mots et les choses, et conclut que l'unité lexicale ne peut pas conduire à la connaissance de la 'chose' en soi. On peut trouver un point commun entre Platon et Aristote : une certaine notion de l'arbitraire des signes naturels, "relations inexplicables et apparemment anarchiques que l'on rapporte à une puissance mystérieuse : celle de l'usage" [Rey70a].

Outre ces analyses sémantiques, d'autres formes de description lexicale ont coexisté dans l'Antiquité. Par exemple, des études étymologiques (recherche de sens primitifs chez le philosophe grec Plotin) et rhétoriques (arts oratoires, aspects 'artistiques' des mots chez Cicéron). Également, au cours de la pensée médiévale, l'étymologie a une place importante : notion antique du mot, perfectionnée par Isidore de Séville (VIe et VIIe siècle, *Etymologia est origo vocabulorum*, étymologie comme origine du vocabulaire, livre 10, *Les étymologies des mots*). Le mot est ainsi perçu comme l'unique chemin pour parvenir à la connaissance du monde [8].

> « La connaissance du mot est la clé de la connaissance de la chose. La maîtrise du mot passe par l'étymologie, et celle-ci devient la base de tous les savoirs. »[Bou03]

Le poids des approches philosophiques (en partie fondés sur l'aristotélisme) reste important tout au long de plusieurs siècles (intérêt aux relations

8. En relation, les travaux sur la langue parfaite d'Umberto Eco, travaux de Leibniz, de l'archêveque Wilkins et aussi de Raimon Llull.

entre les mots, les idées, les choses, le tout établi par Dieu). P. Abélard, théologien et philosophe du début du XIIe siècle, élabore un postulat sémantique sur le caractère monosémique des mots qui suivrait un principe biblique [9]. Chaque mot serait ainsi institutionnalisé par l'Homme moyennant le pouvoir que Dieu lui a donné, dans le but de nommer les Choses [Pol14] :

> « Abélard's attitude towards word meaning is a prototypical illustration of philosophical, non-linguistic approaches to the study of language. Instead of being driven by systematic logical observation of linguistic facts, it is designed to abide by general axioms about God, Man and the relation between them. »

Plus tard, les philosophes de l'époque classique (XVIIe et XVIIIe) ont longuement réflechi sur le langage et les mots. Par exemple, J. Locke s'est intéressé à la relation entre le mot et son contenu, une relation influencée cette fois-ci par l'usage (empirisme). Locke est, par ailleurs, l'un des pionniers concernant la notion "d'associations d'idées".

> « L'usage des Mots consiste à être des marques sensibles des Idées : les Idées qu'on désigne par les Mots, sont ce qu'ils signifient proprement et immédiatement (...). L'homme habile & l'ignorant, le savant & l'idiot se servent des mots de la même manière, lorsqu'ils y attachent quelque signification. Je veux dire que les mots signifient dans la bouche de chaque homme les Idées qu'il a dans l'Esprit & qu'il voudroit exprimer par ces mots-là. »

J. Locke. *Essai philosophique concernant l'entendement humain.* IIIe Partie : Des Mots. Chapitre 2. Traduction de l'anglois par M. Coste. Amsterdam, 1729. (Dans [Rey70a].)

Si on regarde les définitions de l'entrée "mot" dans les dictionnaires de l'époque, on peut également constater cette vision sémantico-philosophique. Par exemple, dans le dictionnaire de la Real Academia Española de la Lengua, le *Diccionario de Autoridades* (1726-1739), le mot est la "voix articulée ou diction significative, qui comprend une ou plusieurs syllabes qui, unies les unes aux autres, forment des locutions. *Le mot explique les concepts de l'âme et appartient aux Hommes*" [10]. Et dans l'exemple d'usage de la même entrée, "le mot est la pensée prononcée par la bouche".

> « s. f. Voz articulada, o dicción significativa, que consta de una o muchas sylabas, y unida con otras, forma la locución, y explica los conceptos del ánimo, y es propria solo del hombre.

9. On retrouvera plus tard ce 'postulat monosémique' dans des approches comme le Lexique Génératif, pour des raisons principalement pratiques [Pol14], cf. 1.4.1.

10. C'est nous qui avons utilisé l'italique.

> Segun Covarr. y otros se dixo de Parábola, que en la baxa Latinidad significaba qualquiera locución. Latín. *Verbum. Vox, ocis.* HORTENS. Mar. f. 159. La palábra es el pensamiento pronunciado en la boca, y voz en rigor es lo mismo. »

Diccionario de Autoridades - Tomo V (1737) [11].

Les grammairiens de l'époque classique restent donc tributaires de la philosophie classique : analyse du contenu lexical en éléments conceptuels. Les aspects formels sont subordonnés aux aspects sémantiques (recherche du sens) et d'usage.

> « 2. La valeur des mots consiste dans la totalité des idées que l'usage a attachées à chaque mot. Les différentes espèces d'idées que les mots peuvent rassembler dans leur signification, donnent lieu à la Lexicologie de distinguer dans la valeur des mots trois sens différents ; le sens fondamental, le sens spécifique, & le sens accidentel. »

Douchet et Beauzée. Article 'Grammaire' dans Encyclopédie ou Dictionnaire raisonné des sciences. 1751 à 1772 sous la direction de Diderot et D'Alembert [12].

Dans le même article, on décrit le rôle d'une 'science du mot' :

> « L'office de la Lexicologie est donc d'expliquer tout ce qui concerne la connoissance des mots ; & pour y procéder avec méthode, elle en considère le matériel, la valeur, & l'étymologie. »

(*Ibid.*).

La Lexicologie est donc un domaine disciplinaire inséré dans la Grammaire (« science de la parole prononcée ou écrite », *ibid.*). Elle se consacre à l'étude des mots : leur matériel (sons et prosodie), leur valeur (les idées) et leur étymologie (leur source). La grammaire spécifie également les nouveaux mots formellement possibles.

En conclusion, bien que des préoccupations sur la forme émergent (« Les sons & les articulations sont les parties élémentaires des mots, & les syllabes qui résultent de leur combinaison, en sont les parties intégrantes & immédiates. », *ibid.*), pour la linguistique classique l'étude des mots se confondait avec celle de la signification : « la lexicologie était d'abord sémantique »[Rey70a].

11. `http://web.frl.es/DA.html`
12. `http://diderot.alembert.free.fr/G.html`

1.3 Approches contemporaines

On l'a vu, pendant une longue période, on étudie les unités lexicales du point de vue des réflexions philosophiques. Les connaissances étaient limitées, la description linguistique presque inexistante et principalement centrée sur un seul système linguistique.

À partir du XIXe siècle, les unités lexicales ne sont plus considérées uniquement comme des signes dotés de valeurs conceptuelles mais comme des formes pouvant être observées et décrites en termes phonétiques et morphologiques. Par ailleurs, les comparaisons des formes permettent de retracer l'histoire des langues (approche phylogénétique, travaux de linguistique comparative). L'étude du lexique prend de l'importance dans la comparaison de langues indo-européennes. Les grammairiens de l'époque sont, ainsi, éminemment lexicologues. Par exemple, F. Bopp utilise des critères morphologiques, avec des notions empruntées aux anciens grammairiens de l'Inde, pour comparer les racines de mots et interpréter leurs origines.

> « Il y a aussi des mots qui sont purement et simplement des mots-racines, c'est-à-dire que le thème présente la racine nue, sans suffixe dérivatif ni personnel ; dans la déclinaison, les syllabes représentent les rapports casuels viennent alors s'ajouter à la racine. Excepté à la fin des composés, les mots de cette sorte sont rares en sanskrit (...) ; en grec et en latin la racine pure est également la forme de mot la plus rare. »

F. Bopp (1885) *Grammaire comparée des langues indo-européennes* § 110-11. Traduit par M. Bréal. Paris. (Dans [Rey70a].)

À la fin du XIXe, la lexicologie historique s'intéresse à l'évolution des formes comparées (métaphore de l'évolution biologique : familles de langues, etc.), depuis un point de vue phonétique et morphologique. Les approches contemporaines (XXe siècle) sont, quant à elles, principalement intéressées par la morphologie et la syntaxe, mais des courants sémantiques coexistent également.

1.3.1 Structuralisme

Pour la linguistique moderne, la refléxion lexicologique est centrée sur le mot en tant que signe. En ce sens, Saussure rompt avec la tradition philosophique qui conçoit le signe en l'opposant à l'idée et au concept représenté. Pour Saussure, le mot-signe ne reproduit pas la réalité : l'idée est 'incorporée' au signe. Ainsi, le mot possède deux constituants indissociables : le signifiant, support matériel (image acoustique), et le signifié, idée contenue dans le signe. Le lien signifiant/signifié est radicalement arbitraire.

Cependant, l'étude du mot et du lexique en général n'est pas une priorité pour la linguistique moderne. En effet, la linguistique structurale, par exemple, conçoit la langue en tant que système où toutes les unités entretiennent des relations entre elles (axes paradigmatique et syntagmatique de Saussure) :

> « Les rapports et les différences entre termes linguistiques se déroulent dans deux sphères distinctes dont chacune est génératrice d'un certain ordre de valeurs ; l'opposition entre ces deux ordres fait mieux comprendre la nature de chacun d'eux. (...) Les mots se contractent entre eux, en vertu de leur enchaînement, des rapports fondés sur le caractère linéaire de la langue. (...) D'autre part, en dehors du discours, les mots offrant quelque chose de commun s'associent dans la mémoire, et il se forme ainsi des groupes au sein desquels règnent des rapports très divers. »

F. de Saussure (1916) *Cours de linguistique générale*. Paris : Payot, p.170.

La langue n'est plus considérée comme nomenclature (liste de mots, tel que cela avait était le cas pendant des siècles de lexicographie) mais comme structure : tous les éléments sont en relation [Gag65]. Les mots sont ainsi pris en compte dans un système, la description lexicologique se doit de tenir compte des oppositions et ne plus considérer les unités dans l'absolu (isolées). Par ailleurs, le structuralisme privilégie la synchronie (au détriment de la diachronie et donc de l'étymologie). Les analyses prennent souvent en compte des critères phonologiques et morphologiques :

> « L'entité linguistique n'est complètement déterminée que lorsqu'elle est délimitée, séparée de tout ce qui l'entoure sur la chaîne phonique. Ce sont des entités délimitées ou unités qui s'opposent dans le mécanisme de la langue. »

Ibid., p.145.

> « Ainsi, dès qu'on veut assimiler les unités concrètes à des mots, on se trouve en face d'un dilemme : ou bien ignorer la relation pourtant évidente, qui unit 'cheval' à 'chevaux', [mwa] à [mwaz], etc., dire que ce sont des mots différents, ou bien, au lieu d'unités concrètes, se contenter de l'abstraction qui réunit en diverses formes un même mot. »

Ibid., pp.147-148.

Certaines approches sémantiques se sont inspirées des méthodes d'analyse structuraliste dans leur description linguistique : ils analysent le contenu des mots en termes de traits pertinents et d'oppositions entre les traits : analyse componentielle ou analyse sémique (sèmes ou traits distinctifs), par exemple chez Pottier, sémantique structurale chez Greimas, etc.

1.3.2 Fonctionnalisme

Il y a eu différentes approches ayant en commun le fait de prendre la langue comme un ensemble complexe d'éléments dont les fonctions et les relations sont à déterminer. Par exemple, le mot représente "l'irrégularité face au système" d'après Bloomfield (on retrouve ici l'arbitraire du signe de Saussure) :

> « Nous avons vu que la fonction de certaines formes était déterminée par leurs constituants ou par leur construction. Toute fonction qui est ainsi déterminée est dite *régulière*, et une fonction qui ne l'est pas est dite *irrégulière*. (...) Le lexique est en réalité un appendice de la grammaire, une liste d'irrégularités fondamentales. Cela est d'autant plus évident si l'on prend les significations en considération, puisque la signification de chaque morphème lui est assignée par une tradition arbitraire. »

L. Bloomfield (1933) *Le langage (Language)*, traduction A. Rey, chapitre 16, p. 6-7. Paris : Editions Payot.

Par ailleurs, la linguistique américaine de l'époque se consacre principalement à l'analyse morphologique et phonologique des langues amérindiennes et à en faire un classement typologique sans tenir compte du sens.

A. Martinet, quant à lui, adopte une vision syntaxique en préférant le syntagme autonome au mot. Le syntagme constitue l'entité dans laquelle s'intègrent les mots (*Eléments de linguistique générale*, p.114-115).

Enfin, la notion de relation et d'association revient chez Bally :

> « Chaque mot est, dans notre mémoire, une maille d'un réseau aux fils ténus et innombrables ; dans chaque mot viennent aboutir, pour en repartir ensuite, mille associations diverses. Ainsi, d'une part les mots s'appelant les uns les autres, se retiennent plus facilement ; d'autre part, la variété de ces associations nous donne une grande liberté dans leur emploi, parce qu'elles offrent non pas une, mais de nombreuses possibilités dans la reproduction de ces mots. »

C. Bally (1921) *Traité de stylistique française*, T. I, § 79 (p. 67). Heidelberg : Carl Winters Universitätbuchhandlung [13].

1.3.3 Générativisme

Le générativisme est le courant théorique qui domine la linguistique entre les années 1950 et 1970. Il regroupe plusieurs théories et formalismes, principalement centrés sur la syntaxe, développés à l'origine par N. Chomsky. Le recours à des formalismes est lié au développement des outils informatiques et de traitement automatique des langues (TAL). Les formalismes devaient donc représenter l'information linguistique de façon explicite et systématique de façon à pouvoir être utilisés dans des applications de TAL comme la traduction automatique. Par ailleurs, ils devaient rester opératoires et calculables afin d'être implémentés pour analyser et prédire la grammaticalité des productions linguistiques.

Ce qui intéresse Chomsky et les générativistes est le développement d'une théorie linguistique générale des structures linguistiques (grammaire universelle), c'est-à-dire, proposer un ensemble de mécanismes qui permettent de décrire le langage de façon holistique. Il en résulte que le lexique, appelé souvent 'vocabulaire', est subordonné aux composantes principales et en aucun cas n'a un statut majeur (c'est un paramètre de plus de la théorie linguistique). Ainsi, selon Chomsky, la grammaire générative inclut un système de règles correspondant à trois composantes : les composantes syntaxique, phonologique et sémantique. Les unités lexicales sont alors analysées selon des règles appartenant à ces trois composantes.

> « I assume throughout that the syntactic component contains a lexicon, and that each lexical item is specified in the lexicon in terms of its intrinsic semantic features, whatever these may be. »

N. Chomsky (1965) *Aspects of the theory of Syntax*. MIT Press, p. 198.

Concrètement, dans les grammaires génératives, l'unité lexicale est décrite en tant qu'un ensemble de traits de différente nature :

> « Le lexique est un ensemble non ordonné d'*entrées lexicales*. Chaque entrée lexicale est simplement un ensemble de traits spécifiés. Ces traits qui constituent l'entrée lexicale peuvent être phonologiques (par ex. [± [14] voisé *n*] (...), sémantiques (par ex. [± objet fabriqué]), ou syntactiques (par ex. [± nom propre]). »

13. `https://ia700603.us.archive.org/23/items/traitdestylist01ball/traitdestylist01ball.pdf`

14. Le symbole ± indique 'plus' ou 'moins' par rapport à la variable qui le suit. Par exemple [g] est plus 'voisé', 'femme' plus 'humain' et moins 'objet fabriqué', etc.

N. Chomsky. Topics in the theory of generative grammar, III, p. 42-46. (Dans [Rey70a].)

La théorie standard de Chomsky a beaucoup évolué avec les années, avec toujours le même objectif de fournir une description des connaissances liées au langage (son acquisition, sa production, etc.) bâtie à partir de la syntaxe. Katz et Fodor [KF63] ont introduit une composante sémantique capable de rendre compte de phénomènes comme l'ambiguïté sémantique, la synonymie, les anomalies liées au sens, etc. [15]. Cette composante est, pour Katz et Fodor, composée d'un dictionnaire et d'un ensemble de règles de projection. Le dictionnaire décrit sous forme arborescente les significations des unités lexicales (nœuds terminaux avec des traits sémantiques).

> « On remarquera que le statut théorique des phénomènes d'ambiguïté, d'anomalie et de synonymie n'est pas étudié par lui-même : de façon circulaire, ceux-ci sont caractérisés par le traitement qu'ils reçoivent dans le cadre proposé. Tout repose donc, en dernier ressort, sur la structure des articles de dictionnaire ; or celle-ci est des plus traditionnelles : décomposition de la signification en atomes de sens, et absence de réflexion théorique sur la notion d'"unité lexicale". »

1.4 Études computationnelles et quantitatives

À partir des années 1950, l'utilisation des ordinateurs en linguistique ouvre la voie à de nouvelles perspectives d'étude du lexique.

D'une part, les approches mathématiques et computationnelles décrivent les langues en tant que systèmes formels dotés de grammaires et d'ensembles finis de mots et de symboles [16]. Or, l'aspect formel, c'est-à-dire, la formalisation en termes de structures de traits avec des contraintes syntagmatiques, empêche de prendre en compte tout ce qui ne peut pas être décrit dans ces termes (irrégularités, usage, etc.). C'est pourquoi Chomsky s'intéresse à peine à la composante lexicale. En revanche, pour des courants dérivés des premières théories génératives, l'hypothèse lexicaliste émerge (meilleure considération des éléments lexicaux car ils apportent beaucoup d'informations y compris des informations syntaxiques). Parmi ces courants, on peut citer la grammaire syntagmatique guidée par les têtes *Head-driven Phrase*

15. Chomsky n'a jamais adhéré aux hypothèses sémantiques de Katz et Fodor qu'il considère comme des "variantes notationnelles" dérivées des représentations syntaxiques (construction de la représentation sémantique à partir de l'interprétation de l'arbre syntaxique).

16. Cette idée fût déjà avancée par Humboldt (1836) : "le langage fait un usage infini de moyens finis".

Structure Grammar (HPSG) de Sag et Pollard, la *Lexical Functional Grammar* (LFG) de Bresnan et Kaplan, les grammaires d'arbres adjoints de Joshi, ainsi que des développements beaucoup plus récents autour du Lexique-Grammaire [Gro94] et du Lexique Génératif [Pus91] (1.4.1).

D'autre part, l'avènement de grands volumes de données sous format électronique, surtout depuis la généralisation du web dans les années 1990, favorise le développement des méthodes statistiques et probabilistes autant dans le traitement du lexique (statistique lexicale, 1.4.2) que, plus globalement, dans le domaine du traitement automatique des langues [17].

1.4.1 Descriptions formelles privilégiant le lexique

Lexique-Grammaire

Le Lexique-Grammaire [Gro94] est à la fois une théorie et une méthodologie d'analyse formelle des langues dans laquelle, à la différence des approches génératives, le lexique a ici une place primordiale : les règles de grammaire ne peuvent pas ignorer l'information lexicale. Pour ce faire, la notion de 'grammaire locale' est introduite. Il s'agit d'ensembles de règles qui capturent des informations d'ordre principalement lexical (expressions polylexicales, figements, entités nommées, etc.).

> « But grammarians operating at the level of sentences seem to be interested only in elaborating general rules and do so without performing any sort of systematic observation and without methodological accumulation of sentence forms to be used by further generations of scientists. (...) the model we advocate, and which we call it finite-state for short, is of a strictly local nature. In this perspective, the global nature of language results from the interaction of a multiplicity of local finite-state schemes which we call finite-state local automata. »[Gro97]

Sur la base des principes du distributionnalisme de Z. H. Harris (de qui M. Gross fut disciple, tout comme N. Chomsky), le Lexique-Grammaire est fondé sur une formalisation de la description lexicale, recueillie sous forme de matrices à double entrée, avec de l'information explicite sur les caractéristiques syntactico-sémantiques de chaque entrée. Les travaux issus de cette approche (de M. Gross et de ses collaborateurs) permettent de créer les fameuses tables du LADL [18], qui ont donné lieu à nombre de lexiques syntaxiques pour le TAL (nous y revenons plus loin dans ce mémoire : nous

17. Pour une introduction au domaine : Christopher D. Manning, Hinrich Schütze (1999) *Foundations of Statistical Natural Language Processing*, MIT Press.

18. Laboratoire d'Automatique Documentaire et Linguistique, nom du laboratoire de linguistique informatique qu'il créa en début des années 1970 et dirigea pendant trente ans.

avions travaillé à l'informatisation d'un lexique issu des tables du Lexique-Grammaire avec A. Valli [GV05], section 3.3.2).

D'un point de vue plus théorique, les travaux issus du Lexique-Grammaire s'attèlent à montrer la forte interaction entre la syntaxe et le lexique, à décrire les propriétés de valence des entrées verbales, à définir des classes lexicales et à identifier des sens en fonction du partage des mêmes propriétés syntaxiques. Il s'agit d'études systématiques des propriétés d'ordre combinatoire en syntaxe [BM11], sur le figement dans les années 1990 et sur les collocations dans les années 2000.

Lexique Génératif

Une des théories linguistiques récentes dans le domaine de la sémantique lexicale est le Lexique Génératif (*Generative Lexicon*, GL) [Pus91]. Il s'agit d'un modèle formel qui s'intéresse aux propriétés du lexique dans une optique générative. À la différence des approches centrées sur les verbes et ses propriétés syntaxiques (qui déterminent les propriétés sémantiques), cette théorie s'intéresse à la nature distributionnelle de la compositionnalité, c'est-à-dire, au traitement du sens dans une optique dynamique.

> « GL is concerned with explaining the creative use of language (...) It is the notion of constantly evolving lexicon that GL attempts to emulate. (...) GL was initially developed as a theoretical framework for encoding selectional knowledge in natural language »[Pus91]

Plutôt qu'une liste statique de sens associés à une unité lexicale, le GL s'emploie à déterminer un espace d'interprétations possibles pour chaque entrée lexicale à partir de quatre niveaux de représentation : *lexical typing structure, argument structure, event structure, qualia structure.* Par exemple, pour *book*[19], la figure 1.1 représente sa structure lexicale :

```
book
ARGSTR = ARG1 = y:information
         ARG2 = x:phys_obj
QUALIA = FORM = hold(x,y)
         TELIC = read(e,w.y)
         AGENT = write(e',v,x.y)
```

FIGURE 1.1 – Exemple d'entrée lexicale dans le Lexique Génératif.

Ce type de structure de traits spécifie différents niveaux de la représentation lexicale. Avec ces informations et l'application d'un nombre fini de

19. Pas d'*event structure* car cela ne s'applique qu'aux verbes.

mécanismes génératifs il doit être possible d'interpréter le sens de l'entrée *book* dans différents prédicats. Ce type de représentation est fondée sur la présupposition que les mots ont un seul sens 'noyau' (*core meaning*) et que tous les autres sens peuvent être dérivés par des règles d'inférence à partir du contexte [Pol14]. L'ambition de cette approche est de représenter le pouvoir sémantique expressif du langage ("the expressive semantic power of language") avec un nombre de mécanismes et contraintes limitées (on retrouve clairement la marque générative).

Lexicologie Explicative et Combinatoire

Enfin, la Lexicologie Explicative et Combinatoire [MCP95] constitue le volet consacré à la lexicologie (et à la lexicographie) de la Théorie Sens Texte (TST) [Mel81]. Les travaux de cette approche s'articulent autour de la description explicite et rigoureuse de la forme, le sens et les traits combinatoires des unités lexicales, de façon à les caractériser dans leur globalité. Dans ce sens, il s'agit d'une approche théorique qui a comme objectif la description des unités du lexique de façon le plus exhaustive possible. Une lexie est ainsi une « entité trilatérale avec un sens (le *signifié* saussurien), une forme phonique/graphique (le *signifiant* saussurien), et un ensemble de traits de combinatoire (le *syntactique* de la théorie Sens-Texte) »[MCP95].

Cette approche théorique et descriptive de la lexicologie a une visée pratique : l'élaboration de dictionnaires, dont le *Dictionnaire explicatif et combinatoire du français contemporain* (DEC) [Mel99]. Il s'agit d'un lexique théorique où toutes les information associées aux différentes lexies (vedettes) sont représentées de façon formelle et multilatérale dans des articles exhaustifs. Un article de dictionnaire comprend, ainsi, des informations sémantiques (définition), des informations syntaxiques (schémas de régime) et des informations lexico-combinatoires (fonctions lexicales).

Pour les auteurs de cette théorie, le lexique est fondamental dans la langue et sa description exhaustive doit se trouver, ainsi, au cœur de la linguistique théorique [20].

1.4.2 Approches statistiques sur grands corpus

La généralisation de l'informatique et la possibilité d'accès à de grands corpus a entraîné le développement d'approches quantitatives pour l'étude du lexique. Ces approches ont néanmoins des antécédents dans des travaux de la fin du XIXe siècle. Ainsi, F. Käding, avait utilisé un corpus d'environ 11 millions d'occurrences pour dénombrer la fréquence d'apparition des séquences de lettres, non pas à des fins linguistiques mais pour améliorer les performances des sténographes. Il montrait que les 15 mots les plus fréquents

20. Ceci n'est pas la tendance habituelle de la majorité des courants linguistiques de la fin du XXe siècle.

du texte représentaient 25% du total des occurrences. Sur le français, il y a des travaux similaires, ceux de Henmon et Vaun der Becke qui avec des corpus de 400.000 et 1.200.000 occurrences respectivement avaient constaté que 4.000 unités lexicales correspondaient à plus de 95 % du texte.

Les travaux quantitatifs sur les mots d'un texte ou d'une langue répondent à des besoins pratiques et relèvent de l'importance dans des domaines comme la linguistique de corpus, la lexicographie et l'enseignement de langues.

Bien que la statistique lexicale soit une méthodologie de traitement des données plus qu'une théorie, la problématique qu'elle aborde reste lexicologique dans la mesure où les unités lexicales sont au cœur des traitements automatiques. Ainsi, la fréquence d'une unité lexicale dans un corpus est considérée "comme une estimation de sa probabilité d'emploi dans les discours non analysés ou non encore produits" [Mul77]. Dans cette démarche, on met en rapport la notion de lexique dans l'absolu (le lexique d'une langue) par rapport à son actualisation dans le discours (vocabulaire). Une mise en garde est généralement faite dans ce sens : à la différence des caractéristiques morphologiques, syntaxiques et sémantiques, les probabilités d'emploi (les fréquences) des unités lexicales ne sont pas constantes. Elles sont tributaires de la nature, du genre et de la taille du corpus dépouillé. Il en résulte qu'il y a des différences considérables d'un corpus à l'autre, même si on constate des régularités statistiques (des lois), par exemple celle de G. K. Zipf, qui donne une vue plus synthétique de la structure globale d'un texte que celle d'un tableau de distribution de fréquences. Elle stipule que le produit du rang et la fréquence d'une unité lexicale d'un texte tend à être constant, c'est-à-dire, que le deuxième mot le plus fréquent d'un texte l'est à moitié du premier, le dixième deux fois plus fréquent que le vingtième, etc.

Les travaux en statistique lexicale sont donc plus liés aux corpus qu'à une vision théorique du lexique. Un autre exemple en est la notion de richesse lexicale, utilisée, par exemple, dans des études stylistiques :

> « Un texte peut faire appel à des parties très excentriques du lexique sans pour autant avoir un vocabulaire très étendu ; l'excentricité du vocabulaire est un fait de contenu, alors que la richesse lexicale, indépendante du contenu, est un fait de structure. Appliqué à un texte, le terme de richesse lexicale est donc défini par le nombre des vocables, et rien de plus. Cette façon de voir considère le texte comme un ensemble clos et achevé (même s'il s'agit d'un fragment ou d'une tranche), formé de n mots, et dont on mesure la richesse par le nombre des v vocables qui y figurent, sans référence extérieure et sans hypothèse sur le lexique dont ce vocabulaire est un échantillon. »[Mul77] (pp. 115-116).

La notion de richesse lexicale exprime le rapport entre le nombre de formes (lemmes) et leurs occurrences (apparitions dans le texte). Une autre

façon de la calculer est de prendre en compte le rapport hapax/occurrences : plus il y a de hapax par rapport au nombre total d'occurrences, plus on considère que le texte est riche en vocabulaire. D'autres mesures sur les corpus décrites dans la littérature et utilisées dans de nombreuses études sur le lexique sont l'accroissement lexical, la spécialisation lexicale, etc.

La linguistique de corpus, dont les pays anglo-saxons sont pionniers avec notamment les travaux sur le Brown Corpus (*University Standard Corpus of Present-Day Edited American English* [KF64]), a recours à la statistique lexicale pour l'analyse linguistique issue de données réelles (corpus). Par exemple, l'étude des collocations et d'expressions polylexicales occupe une place très importante en linguistique et en traitement automatique des langues, leur étude est intrinsèquement liée à l'usage et pour cela le recours à des corpus (et aux statistiques) est indispensable [Tut10].

Enfin, la statistique lexicale offre des possibilités nouvelles dans l'enseignement des langues, comme le montrent un certain nombre de ressources qui incluent les fréquences des mots comme information essentielle (représentative de l'usage). Parmi ces listes de fréquence de mots on peut signaler les premières, celle d'E. Thorndike pour l'anglais (*The Teacher's word book*, [Tho21]) ou, pour le français, le *Français Fondamental* de G. Gougenheim et ses collaborateurs [Gou58] (nous y reviendrons dans le chapitre 4 de ce mémoire).

1.5 Conclusion

Dans ce chapitre, nous avons voulu faire un bilan qui, loin d'être complet, présente différentes façons d'aborder les unités lexicales le long de l'Histoire. Le tableau suivant se veut un récapitulatif succinct de différentes approches (certaines se chevauchent dans le temps) :

Période	Courant	Domaine
Ve av J.-C.	Inde Antique	morphologie
IVe av J.-C.	Grèce Classique	sémantique
Ier av J.-C.	Rome Latine	morphologie
VIe - VIIe s.	Moyen Âge	étymologie, philosophie
XVIIe - XVIIIe s.	Lumières	sémantique, philosophie
XIXe	Liguistique comparée	morphologie, étymologie
Début XXe s.	Structuralisme	phonologie, morphologie
Années 50-60	Fonctionnalisme	syntaxe
Années 60-70	Générativisme	syntaxe
Années 60-80	Lexique-Grammaire	syntaxe
Années 90-00	Lexique Génératif	sémantique
Années 90-00	Lexicologie Explicative et Comb.	approche globale
Années 70-00	Statistique Lexicale	statistique

FIGURE 1.2 – Récapitulatif : prise en compte des unités lexicales à travers différents courants linguistiques.

La perception que l'on a eue des unités lexicales varie considérablement au fil des courants de pensée et des courants linguistiques du XXe siècle. La lexicologie, en tant que telle, longtemps subordonnée à la construction de dictionnaires (et donc, à la lexicographie), reste une science 'jeune' avec beaucoup de domaines de la description lexicale qui suscitent de l'intérêt pour le traitement automatique des langues. Avant d'aborder ces aspects dans le chapitre 3 de ce mémoire, nous consacrons dans le chapitre suivant quelques pages aux ressources lexicographiques imprimées et informatisées.

Chapitre 2

De la description du lexique à sa structuration dans des dictionnaires

« L'utilité des Dictionnaires est universellement reconnue. Tous ceux qui ont estudié les Langues Grecque et Latine, qui sont les sources de la nostre, n'ignorent pas les secours que l'on tire de ces sortes d'Ouvrages pour l'intelligence des Autheurs qui ont escrit en ces Langues, et pour se mettre soy-mesme en estat de les parler et de les escrire. C'est ce qui a engagé plusieurs sçavants hommes des derniers siècles à se faire une occupation serieuse de ranger sous un ordre methodique tous les mots et toutes les plus belles façons de parler de ces Langues »

Préface du *Dictionnaire de l'Académie françoise*, 1694.

Dans ce chapitre, nous nous intéressons au lexique via sa description et organisation dans les ressources lexicales "classiques", c'est-à-dire, les dictionnaires [1]. Nous aborderons dans un premier temps des aspects linguistiques et lexicographiques ; dans une deuxième partie, nous décrirons les ressources dictionnairiques imprimées et les changements qu'elles ont subis du fait de leur informatisation.

Les aspects liés à l'apparition des acteurs des technologies du langage dans la construction de ressources pour le traitement automatique des langues seront abordés dans le chapitre suivant (chapitre 3).

1. Nous nous intéressons, avant tout, aux ressources 'occidentales'.

2.1 Introduction

La construction de ressources lexicales est très ancienne et répond à des besoins humains. Nous nous étions intéressée à cette question dans [Gal13], avec l'édition d'un volume thématique consacré à différents aspects concernant les ressources lexicales au sens large.

Ainsi, pour J. C. Boulanger [Bou03], les inventeurs de dictionnaires seraient les Sumériens, avec les tablettes cunéiformes où des noms d'objets sont consignés dans des listes. Certes, répertorier le lexique constitue une première approche, mais on est encore loin d'une description linguistique : à cette époque lointaine, les besoins sont comptables, administratifs et bien entendu didactiques (garder une trace, enseigner les mots d'une langue qui sera amenée à évoluer, voire à disparaître).

L'étude empirique des mots et son application pratique (création de dictionnaires) n'est effective qu'à partir du moment où les mots sont listés et accompagnés d'informations les décrivant. En Europe occidentale, cela n'arrive qu'à la Renaissance, malgré quelques exemples significatifs au Moyen Âge qui témoignent de préoccupations méthodologiques qui sont déjà protolexicographiques. En sont des exemples des glossaires et des encyclopédies avec ordonnancement des mots et des informations sémantiques et étymologiques associées aux entrées : l'*Elementarium Doctrinae Erudimentum* de Papias (1050), le *Liber Derivationum* d'Hugutio de Pise (fin XIIe, début XIIIe), ou le lexique encyclopédie byzantin *Suidas* (IXe et Xe siècles).

Par ailleurs, c'est pendant la Renaissance que la différence entre dictionnaire (*Dictionarium* en latin) et encyclopédie (*en kuklos paideia* en grec [2]) se précise [Rey11], le premier avec un intérêt pour les formes ou "mots" (signifiants chez Saussure), le deuxième pour les "idées" qu'ils expriment (signifiés dans la terminologie saussurienne). Dans les deux, néanmoins, "l'alphabet instaure la suprématie du signifiant" (*ibid* p.47).

Cependant, ce n'est qu'au XIXe siècle que le dictionnaire moderne se forge, exhaustif, prescriptif ou descriptif, en diachronie ou en synchronie, par des institutions publiques ou des entreprises privées. Le XXe siècle verra également des bouleversements fondamentaux dus à la maturité de la discipline d'une part, et à la généralisation de l'informatique d'autre part : corpus électroniques, ressources informatisées en cédérom et/ou en ligne sur le Web.

2.1.1 Décrire les mots

Dans beaucoup de définitions, l'idée de dictionnaire correspond à un recueil de mots ordonnés alphabétiquement, dans un ouvrage imprimé, exhaustif et monolingue. J. Rey-Debove (1971, p. 20-27) citée par J. C. Boulanger

2. Littéralement "dans le cercle de l'enseignement", *kuklos* cercle et *paideia* enseignement.

(2003) donnait huit critères déterminants pour caractériser un dictionnaire, que nous résumons comme suit :

- suite de messages graphiques isolés
- la nomenclature [3] est un ensemble déterminé et structuré
- double structure, entrées (macrostructure) et informations associées (microstructure)
- données classées par la forme (sémasiologique) ou par le contenu (onomasiologique)
- caractère linguistique des informations associées aux entrées
- ouvrage de consultation à buts didactiques

Deux aspects nous semblent fondamentaux : la structure et le caractère linguistique des informations. Ainsi, "la mise en ordre de la langue dans le dictionnaire" [Dot12] distingue les dictionnaires des anciennes gloses (explications des mots en marge des textes). Cet ordre différencie aussi les ressources dictionnairiques "classiques" des nouvelles ressources informatiques où la notion d'ordre est peu significative tout au moins au niveau d'interface et de l'accès lexical. En revanche, le caractère linguistique des informations reste une caractéristique primordiale quel que soit le type de support [4].

2.1.2 Les 'mettre en ordre'

La notion d'ordre et de classement implique une organisation hiérarchique. Si les ressources classiques ordonnent leurs entrées principalement par ordre alphabétique, d'autres procédés formels existent, parmi lesquels : classements par familles morphologiques, classements phonétiques et classements thématiques (onomasiologiques) [5].

Classement alphabétique

> « Le dictionnaire classe les données par la forme ou par le contenu. Le classement le plus courant de nos jours est l'ordre alphabétique, simple et efficace. Depuis la Renaissance, il est

3. Ensemble d'entrées, classées méthodiquement par ordre alphabétique.

4. D'un point de vue informatique, la notion d'ordre alphabétique existe (appelé aussi 'ordre lexicographique') : c'est ce qui permet de comparer deux mots sans ambiguïté au moment de recherches et de tris de données. Cependant, cet ordre hiérarchique n'est pas utilisé directement pour la structuration (c'est-à-dire, le rangement) des données.

5. On pourrait encore ajouter, entre autres, les classements par catégories grammaticales ou par fréquences d'emploi (par exemple, le *Dictionnaire de fréquences. Vocabulaire littéraire des XIXe et XXe siècles* de P. Imbs (1971)).

conforme à l'image sociale communément admise du dictionnaire. »

De même, A. Rey (2011) souligne l'interdépendance entre la notion de 'dictionnaire' et 'ordre alphabétique' et justifie cela par l'aspect stable et mnémotechnique :

> « L'ordre formel de l'alphabet, qui l'emporte dans la plupart des langues ayant des glossaires et des dictionnaires (à l'exception des recueils arabes anciens, beaucoup plus subtils, beaucoup moins aisés), lorsqu'elles sont écrites en alphabets, a un avantage négatif important. Cet ordre en partie désordonné et stable, strict, mémorisé par toute personne sachant lire. Les autres ordres, théologiques, logiques, philosophiques, sont changeants selon la culture, discutables, idéologiques. Ils dominent dans les encyclopédies médiévales, dans les méthodes pédagogiques de vocabulaire, mais ont dû se restreindre et disparaître, au point qu'on a confondu 'dictionnaire' avec 'suite alphabétisée'. »[Rey11] (p. 742)

Pour ce qui est du caractère stable, force est de constater que l'alphabet latin a pratiquement conservé l'ordre de l'alphabet sémitique (seulement le 'x' et le 'z' ont été déplacés à la fin) et quelques lettres ont été ajoutées par les grecs. Il demeure donc inchangé depuis plus de deux mille ans. On peut imaginer le besoin d'un ordre fixe pour des raisons mnémotechniques, mais l'ordonnancement qui en résulte reste, malgré tout, également arbitraire.

> « Il faut noter que cet ordre n'a rien d'absolu. Les alphabets sémitiques anciens connaissaient d'ailleurs deux ordres, le sémitique occidental (par exemple le phénicien) et le sudarabique. (...) Si l'ordre sémitique du sud avait été retenu à la place de celui du nord, nous ne dirions donc pas *abécédaire* ni *alphabet* mais "*halahamédaire*" et "*halahama*". »[Tou14] (p. 82)

En espagnol, par exemple, les graphèmes 'ch' et 'll' ont longtemps fait partie de la liste alphabétique, après 'c' et 'l' respectivement. Ils n'ont été supprimés par la *Real Academia Española de la Lengua* (RAE) que tout récemment [6]. Les dictionnaires antérieurs à cette date contiennent des pages consacrées à ces digrammes, comme on peut le voir dans la figure 2.1 pour le 'ch' (correspondant à [tʃ]) :

Dans la définition [7] de 'ch' dans l'édition de 1917 (Alemany y Bolufer) on peut lire : "quatrième lettre de l'abécédaire castillan et troisième des

6. `http://www.rae.es/consultas/exclusion-de-ch-y-ll-del-abecedario`
7. `http://ntlle.rae.es/ntlle`

502 CHACA

CUYANO, NA. adj. *Amér.* En la República Argentina, natural de la antigua provincia de Cuyo. U. t. c. s. ‖ Perteneciente o relativo a esta provincia.

CUYO, YA. (del lat. *cujus*). pron. rel. con terminaciones distintas para los gén. m. y f., y con ambos números sing. y pl. De quien. U. también como posesivo, y entonces concierta, no con el poseedor, sino con la persona o cosa poseída. Precede inmediatamente al nombre y sólo puede anteponerse al verbo *ser*. No puede construirse con el artículo. ‖ m. fam. Galán o amante de una mujer. ‖ ant. Amo, dueño, señor.

CUYUCPA. f. *Amér.* En el Perú, se da este nombre a un ave pequeña, de la

CH. Cuarta letra del abecedario castellano y tercera de las consonantes. Su nombre es **che**. Considerada como signo gráfico, es una consonante compuesta, pero sencilla por su sonido, e indivisible en la escritura.

CHA. m. Nombre que dan los chinos al te. ‖ Tela de seda muy en uso en la China.

¡CHA! interj. *Amér.* En Chile, **¡char!** Es vulgarismo.

CHABABÉ. m. *Mús.* Especie de flauta originaria de Persia.

CHABACANADA. f. Chabacanería.

CHABACANAMENTE. adv. m. Con cha-

FIGURE 2.1 – 'Ch' dans le dictionnaire espagnol de la RAE.

consonnes. Son nom est ché. Considérée comme signe graphique, c'est une consonne composée, mais simple de par son son, et indivisible à l'écriture". En espagnol, on aurait donc dû parler d'"*abécétchaire*" ou "*abécétchédaire*"...

Historiquement, l'ordre alphabétique commence à se répandre dans des travaux (proto-)lexicographiques du Moyen Âge, étant peu rigoureux au début. Ainsi, le *Suidas*, encyclopédie byzantine du Xe siècle écrite en grec, est considérée comme l'un des premiers ouvrages avec agencement alphabétique des entrées. Ce type de classement se généralise en Europe avec l'*Elementarium Doctrinae Erudimentum* de Papias[8].

> « Celui qui voudra trouver dans notre ouvrage promptement quelque chose, saura que l'ordre en est alphabétique, non seulement dans les premières lettres, mais encore dans la seconde, la troisième, et quelquefois au delà. »(Papias cité par E. Littré, [Bou03], p. 276)

8. Premier ouvrage également ordonné selon le principe de la dérivation (familles morphologiques autour d'une tête)[Bou03].

D'autres ouvrages, comme le *Liber Floridus*[9] (XIIe) adoptent également le classement alphabétique (encore imparfait) au détriment de l'ordre thématique (ordre d'une certaine vision du monde, ordre de 'Dieu'), comme il était le cas pour la plupart des ouvrages encyclopédiques et réligieux de l'époque (*cf.* figure 2.2).

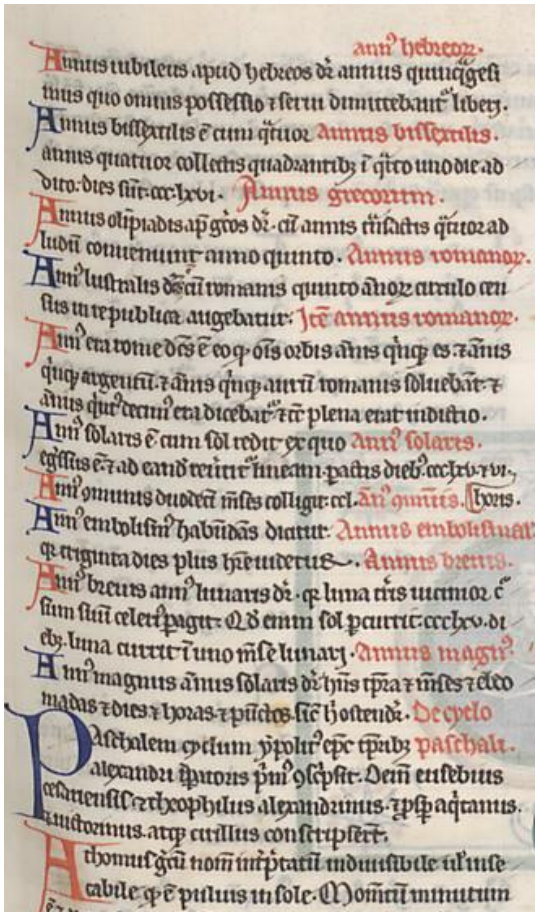


FIGURE 2.2 – *Liber Floridus* : exemple de classement alphabétique ancien.

> « Si l'ordre alphabétique absolu et complet est connu des Grecs aussi tôt qu'au IIe siècle, il n'a pas été vraiment utilisé dans les œuvres latines de l'Antiquité. L'alphabétisation totale des mots peut être regardée comme une pratique oubliée (...) la mise en ordre alphabétique des entrées de dictionnaires ne sera considérée comme une conquête du point de vue de la méthode que très tard au Moyen Âge. »[Bou03] (p. 276)

9. `http://gallica.bnf.fr/ark:/12148/btv1b6000541b` (p. 54r)

Ainsi, les premiers dictionnaires monolingues pour le français (et aussi pour l'anglais) font apparaître la notion d'ordre alphabétique dans les titres mêmes des ressources, certainement pour mettre en avant cette avancée méthodologique :

- *Les mots français selon l'ordre des lettres, ainsi que le faut écrire, tournés en latin, pour les enfants*[10] de R. Estienne (1544)
- *A Table Alphabeticall conteyning and teaching the true writing, and vndrestanding of hard usuall English wordes* de R. Cawdrey (1604)

Le *Thresor de la langue francoyse* de J. Nicot (1606) et la première édition du *Dictionnaire de l'Académie françoise* (1694) présentent également les entrées sous ordre alphabétique (en plus de morphologique). Il en va de même pour certains ouvrages philosophiques du siècle des Lumières. Pour les philosophes de l'époque, il y a un lien entre la raison et l'ordre alphabétique, même si cet ordre n'a rien de rationnel : "tel est le paradoxe des dictionnaires philosophiques du XVIIIe siècle. Bayle, Diderot, Voltaire ont su jouer avec art de cette contradiction"[11], par exemple : *La raison par l'alphabet* de Voltaire (1769).

Si aucun dictionnaire moderne imprimé n'échappe pas à la contrainte de l'ordre alphabétique, cette barrière est 'généralement' franchie avec les ressources électroniques, à l'exception de quelques dictionnaires informatisés pédagogiques où l'ordre alphabétique reste un outil d'apprentissage. Dans ce domaine, il existe, néanmoins, des dictionnaires 'dynamiques' où l'ordre alphabétique traditionnel n'est plus privilégié, par exemple le DAFLES [SVB03].

Classement morphologique

Pour les langues européennes, le classement par familles morphologiques apparaît pour la première fois avec Papias et 'cohabite' avec la méthode alphabétique pendant plusieurs siècles. Il s'agit de montrer les membres d'une même famille morphologique via l'accès par la base (racine ou étymon) à partir de laquelle différents mots de la famille ont été dérivés :

> « Papias fait tout de même progresser la méthode lexicographique sur les plans de la macrostructure et de l'ordonnancement de la nomenclature. (...) le premier, il introduit dans un dictionnaire le principe du rangement des unités qui s'appuie sur la méthode de la dérivation (la *derivatio*). (...) "*Aptus, -ta, -tum, -tior, -simus ; apte, -tius, -sime*, adverbia ; et *aptitudo, -inis* ; et

10. Au début de chaque entrée il y a, néanmoins, la liste des mots qui appartiennent à la même famille morphologique.

11. B. Didier (1996) *Alphabet et raison. Le paradoxe des dictionnaires au XVIIIe siècle.* PUF, « Écriture ».

apto, -tas, -vi, verbum activum ; quod componitur *adapto, coapto...*". »[Bou03] (p. 277)

D'autres ouvrages postérieurs, comme le *Thresor de la langue francoyse* de J. Nicot (1606) et le premier dictionnaire de l'Académie (1694) [12] font le choix d'ordonner la nomenclature par ordre alphabétique puis par racines autour des "mots primitifs" et des "chefs de famille". Un exemple beaucoup plus récent est celui du dictionnaire espagnol *Diccionario de uso del español* de María Moliner (1962). Malheureusement, des choix éditoriaux très controversés font que l'ordonnancement morphologique ait disparu des versions plus modernes, l'accès supposé 'difficile' aux entrées est parfois cité comme argument pour justifier ce choix [13]. Enfin, pour le français on peut citer le *Lexis* [Dub75], un ouvrage de soixante-dix mille termes où tous les dérivés sont regroupés autour du mot base correspondant à un sens précis (les homonymes apparaissent répartis dans différents articles avec leurs propres dérivés).

Classement phonologique

Le classement par la structure des mots, et plus précisément en tenant compte de la racine, est un classement généralisé dans les langues sémitiques. Bien qu'il s'agisse d'un procédé courant, il n'est pas unique, comme en fait preuve le fameux (et original) "dictionnaire de l'œil" ou "de la source" d'al-Khalîl ibn Ahmad (VIIIe siècle). Ce lexicographe arabe avait utilisé le critère phonétique pour classer les entrées à partir de la consonne la plus antérieure d'un point de vue articulatoire (ayn) jusqu'à la plus postérieure (consonne labiale mìm) :

> « (...) classer les mots de façon inattendue : non pas dans l'ordre de l'alphabet, mais en commençant par la consonne qui, selon lui, se prononçait au plus profond de la gorge, celle qui se trouve au début du mot *'ayn* "œil". C'est pourquoi on appelle ce dictionnaire *kitâb u l 'ayn*, "le livre de l'œil". Ce dictionnaire d'al-Khalîl est considéré comme le premier grand dictionnaire de la langue arabe, et de nombreux autres lexicographes anciens ont suivi cet ordre de classement phonétique : en partant de la consonne la plus profonde et en remontant le canal vocal, son par son, pour aboutir aux sons articulés avec les lèvres. »[WB06]

12. Cette première édition ne rencontre pas beaucoup de succès, mais on attribue cela à son élitisme et son décalage entre la langue décrite et son usage réel, plus qu'à l'ordonnancement morphologique.

13. Les conflits entre l'éditeur Gredos et les héritiers de la lexicographe, lors de la publication de la 2e édition du dictionnaire, sont de notoriété publique (voir M. Seco, *El País*, 29/05/1981, `http://www.mariamoliner.com/centenario_2.pdf`).

À notre connaissance, il n'existe pas de dictionnaire de ce type pour les langues européennes [14].

Classement onomasiologique

Il existe un certain nombre de ressources (dictionnaires onomasiologiques, thesaurus) [15] conçus pour aider à trouver les mots à partir d'une idée (démarche onomasiologique). "On part d'une *idée* ou d'un concept (...) qu'on traduit en un *concept lexical* (mot-sens) avant de le communiquer sous forme écrite ou orale, le *mot-forme*" [ZS13]. Dans ces ouvrages, c'est le contenu qui prime, c'est-à-dire que les mots partageant des notions sémantiques ou thématiques se retrouvent associés à une même entrée (organisation en classes, sous-classes, etc.). Cependant, ils exigent également un index alphabétique, qui reste le meilleur point d'entrée dans la ressource.

> « Les classements sémantiques sont construits en partant des sens des mots et des rapprochements en découlant. Il s'agit de permettre à l'usager de trouver ou retrouver un mot précis en fonction d'une idée, de mieux percevoir les liens établis entre des mots sémantiquement proches. La démarche est onomasiologique : on part d'un concept pour chercher les mots s'y rattachant. (...) On a une idée et on l'exprime avec les unités lexicales les mieux adaptées. (...) la grille sémantique offrant une série structurée de concepts permettant d'atteindre les signes linguistiques en relevant et pouvant accueillir tous les mots d'une langue. »[Pru06] (p. 110-111)

Par rapport à l'ordre alphabétique, le classement sémantique requiert une organisation méthodique du savoir. L'exemple le plus significatif est le *Roget's Thesaurus of English Words and Phrases* (1852) qui contient 6 classes, 39 sections et 990 têtes [16]. On retrouve des exemples de dictionnaires onomasiologiques à partir de la moitié du XIXe siècle et jusqu'à aujourd'hui :

- le *Dictionnaire analogique de la langue française. Répertoire complet des mots par les idées et des idées par les mots* (1852) de P. Boissière ;

14. En revanche, le classement phonologique est à la base de l'ordre alphabétique dans les écritures alphasyllabiques indiennes (bengali, tamoul, thaï, tibétain, etc.) : "C'est le classement phonologique, assorti à l'ordre des signes, qui constitue l'une des grandes originalités des systèmes dérivés de la brami. C'est probablement Panini, le grand grammairien indien, qui a systématisé ce classement adopté dans pratiquement toutes les écritures alphasyllabiques indiennes (...). Les consonnes apparaissent dans l'ordre suivant : vélaires, palatales affriquées, rétroflexes, dentales, bilabiales et sont donc classées en fonction de divers points d'articulation." [Tou14] (p. 88).

15. Nous avions fait un recueil de quelques-uns des ouvrages plus significatifs pour l'anglais, le français et l'espagnol dans [Gal13].

16. `http://en.wikipedia.org/wiki/Wikipedia:Outline_of_Roget%27s_Thesaurus`

- le *Diccionario ideológico de la lengua española. Desde la idea a la palabra, desde la palabra a la idea*[17] (1942) de J. Casares ;
- le *Dictionnaire alphabétique et analogique de la langue française. Les mots et les associations d'idées* (1958-1964) de P. Robert, repris par A. Rey, H. Cottez et J. Rey-Debove pour devenir, à partir de 1985, *Le Grand Robert de la langue française. Dictionnaire alphabétique et analogique de la langue française de Paul Robert* et, dans les années 2000, le *Petit Robert* ;
- le *Bernstein Reverse Dictionary* (1975) de Th. M. Bernstein.

Comme on peut le voir dans le titre du dictionnaire anglais, la notion de "dictionnaire à l'envers" est parfois utilisée (par défaut, le dictionnaire "à l'endroit" serait le dictionnaire sémasiologique, "la suprématie du signifiant" comme évoqué par [Rey11]). Cela donne une idée du caractère exceptionnel de ce type de dictionnaire[18].

2.1.3 Les interrelier

La grande majorité des ressources lexicales, jusqu'à très récemment, ont été conçues comme des objets textuels, statiques, avec des mots listés puis décrits et expliqués textuellement et linéairement :

> « Printed monolingual definition dictionaries are primarily structured as lists of lists : each entry is a list of items (...). These entries are themselves listed, e.g. alphabetically (macrostructure), and some of them are related by links. The links are usually only of few types : synonymy, antonymy, an unspecific "see also"-relation, etc. Electronic dictionaries which have been transformed from printed ones tend to conserve this linear structure, possibly with the inclusion of general links from each word form appearing in one of the items to that word's entry. The claim (...) is that electronic dictionaries could do better. »[Hei09] [19]

En effet, un des apports fondamentaux de l'informatique (outre le stockage de grandes quantités des données, l'aide aux lexicographes dans des tâches fastidieuses, etc., pour ne citer que quelques avantages) est la possibilité de modéliser les données lexicales autrement que de façon textuelle. Ainsi, de nouvelles approches ont émergé dans les dernières années avec la

17. "Dictionnaire ideologique de l'espagnol. De l'idée au mot, du mot à l'idée". On remarque la similitude avec le titre choisi par Boissière.

18. L'appellation 'reverse' est ambiguë car elle est aussi utilisée pour des dictionnaires de rimes, des dictionnaires qui classent leurs entrées en fonction de leur terminaison.

19. https ://www.uclouvain.be/en-271026.html

particularité de prendre en compte une dimension complètement dynamique du lexique : ce sont non seulement les unités lexicales mais aussi les informations linguistiques les concernant qui sont interreliées.

> « Le lexique est en effet, formellement, un graphe immensément riche de connexions entre entités lexicales alors que les dictionnaires, dans leur realité formelle, ne sont que des "textes". »[Pol12]

Cette vision novatrice (le lexique en tant qu'un réseau de connaissances lexicales interreliées) serait plus à même de représenter le dictionnaire mental, c'est-à-dire, la 'vraie' organisation du lexique au niveau cognitif [20].

On retrouve l'idée de dynamisme au niveau du lexique dans le *Lexique Génératif* de [Pus91] (section 1.4.1). A. Polguère, quant à lui, propose la notion de "système lexical" : "la véritable structure d'un modèle explicatif et combinatoire du lexique de la langue est un réseau lexical" [Pol06] (des graphes non hiérarchisés d'entités lexicales structurées grâce à un système de liens formalisés). Cette modélisation serait, par ailleurs, beaucoup plus apte à représenter des connexions interlingues [21].

Ainsi, d'après A. Polguère (*ibid.*), il y aurait deux approches pour représenter la structure lexique, celle des dictionnaires standard (vision *dictionary-like*, linéaire, textuelle) et celle de quelques ressources plus récentes conçues à la base comme des graphes (vision *net-like*). La première approche cible les unités lexicales et en décrit leur sens (d'autres informations peuvent être également présentes, mais c'est le sens qui guide la description lexicale et l'organisation des informations). Quant à la deuxième, elle n'est pas nécessairement centrée sur les mots mais plus généralement sur des informations lexicales, sur des connaissances permettant d'interrelier ces informations (au delà des renvois vers des synonymes ou antonymes), par exemple, des liens paradigmatiques, des relations lexicales variées, etc. Il s'agit d'une représentation multidimensionnelle, comme l'évoquait déjà G. Grefenstette dans son article 'visionnaire' *The Future of Linguistics and Lexicographers : Will there be Lexicographers in the year 3000 ?* :

> « (...) three and four and five dimensional in which information is stored about how each word is used with each other word, and how that pair of words is used with a third word, and that triple with a four word »[Gre98]

20. C'est dans les années 1960-1970 que les travaux sur les réseaux sémantiques et taxinomiques ont pris de l'essor (voir, par exemple, Quillian, M. R. (1967). *Word concepts : A theory and simulation of some basic semantic capabilities.* Behavioral Science, 12(5), 410-430. Allan M. Collins et Elizabeth F. Loftus (1975). *A spreading-activation theory of semantic processing.* Psychological Review 8).

21. Nous nous sommes penchée sur cette question lors de notre projet LexRom [Gal11] (section 3.3.1).

Les ressources lexicales conçues comme des graphes sont, ainsi, des réseaux associatifs très puissants qui permettraient un fonctionnement analogue à celui du cerveau humain (par exemple, pour l'accès lexical par des chemins multiples) [22]. Nous en montrerons quelques exemples dans la section 3.3.4. ; nous verrons également dans 2.3.3. quelques exemples de visualisation de données sous forme de graphe.

2.2 Ressources dictionnairiques imprimées

Après quelques considérations générales, notre intention n'est pas ici de dresser une typologie exhaustive de ces "objets protéiformes" [Rey13] que sont les dictionnaires, particulièrement les dictionnaires sous format papier. On pourra se référer pour cela à la myriade d'ouvrages dans la littérature (par exemple, [Rey70b]). L'idée est plutôt de donner une vision générale de l'existant afin de nous servir de base de comparaison avec les ressources informatisées et électroniques que nous décrirons plus tard (respectivement, section 2.3 et chapitre 3). De ce fait, nous avons identifié trois critères [23] pour lesquels l'informatique a apporté des changements significatifs au niveau du contenu ou de sa consultation, à savoir : le nombre de langues traitées, la nature des données d'un point de vue temporel, la mise à jour des données rapide (spécialement utile pour certains types de niveaux de langue, standard ou spécialisée).

2.2.1 Dictionnaires monolingues *vs* multilingues

Une première observation peut être faite par rapport au nombre de langues traitées.

Les premiers dictionnaires imprimés apparaissent dans la Renaissance, un moment historique qui foisonne de voyages et de découvertes. Les besoins de communication poussent à la création de dictionnaires bilingues, parfois rédigés par des moines ou missionnaires. Ces dictionnaires sont majoritairement bilingues ou trilingues (Nebrija 1492, Estienne 1539, de Molina 1555, Percyvale 1591) [24]. Le seule exception remarquable, quant au nombre de langues traitées, est le *Dictionarium, quanta maxima fide ac diligentia fieri*

22. Ce sujet prend de l'importance dans la lexicographie et le traitement automatique des langues, comme en font preuve des publications récentes ou des ateliers de conférences, notamment, plusieurs ateliers à Coling (Cogalex 2008, 2010, 2012, 2014 : `http://pageperso.lif.univ-mrs.fr/~michael.zock/CogALex-IV/cogalex-webpage/index.html`) et à TALN 2014 (*Réseaux Lexicaux pour le Traitement des Langues Naturelles*, RLTLN).

23. Il y en a d'autres, par exemple, nous n'aborderons pas ici des aspects concernant la construction des ressources (aspects méta-lexicographiques).

24. Nous nous étions intéressées à une perspective historique des ressources lexicographiques dans [Gal13].

potuit accurate emendatum multisque partibus cumulatum. Adjectae sunt latinis dictionibus, hebraea, graecae, gallicae, italicae, germanicae et hispanicae d'A. Calepino (1578), un ouvrage comprenant sept puis onze langues.

Parallèlement, les premiers dictionnaires monolingues exhaustifs, rigoureux et détaillés apparaissent (Cawdrey 1604, Nicot 1606, Covarrubias 1611, Richelet 1680). C'est également l'époque de la naissance des Académies des langues, à l'origine d'ouvrages normatifs à double finalité : stabiliser la langue, d'une part, et lui donner du prestige, de l'autre.

Dans les siècles qui suivent, la lexicographie est principalement monolingue, vouée à la description de plus en plus méthodique et détaillée du lexique d'une langue. Plus récemment, les progrès majoritaires quant au contenu des dictionnaires concernent les ouvrages monolingues (rajout d'information statistique -fréquences d'usage dans des corpus-, information relationnelle -collocations-, etc.), bien que les dictionnaires bilingues (ou multibilingues) aient continué à exister sous forme de listes plus ou moins enrichies (informations grammaticales, exemples d'usage sur corpus, etc.).

Avec l'informatique, et surtout avec le Web à partir des années 1990, ce sont les bases terminologiques multilingues qui font leur apparition de façon notoire, par exemple, IATE [25]. Il s'agit d'outils qui n'ont pas de version papier au préalable dû à la difficulté d'englober dans un seul volume plusieurs langues. La présence de lexiques terminologiques multilingues sur le Web, ainsi que le nombre de langues représentées, n'a cessé d'augmenter depuis. De plus en plus, la quantité de langues traitées est un 'plus' pour les systèmes de traduction en ligne comme c'est le cas de Reverso [26] ou Google Traduction, ce dernier avec quatre-vingt langues (février 2014).

Bien que ce soient des ressources lexicales multilingues, on ne peut pas appeler ces ressources des dictionnaires dans le sens 'classique', non pas à cause du contenu associé aux entrées (parfois elles proposent des définitions), mais plutôt du fait que, s'agissant de *termes d'un domaine de spécialité*, la polysémie est très ciblée [27]. Ce sont donc des bases de données terminologiques traduites dans plusieurs langues. Quant aux systèmes de traduction automatique statistique, ils ne peuvent pas non plus être considérés comme des dictionnaires multilingues.

2.2.2 Diachronie *vs* synchronie

L'axe du temps constitue un deuxième élément pour distinguer et caractériser le contenu linguistique des ressources. La diachronie concerne, dans

25. `www.iate.europa.eu`

26. `http://www.reverso.net/text_translation.aspx?lang=FR`

27. Un mot comme 'prise' a plusieurs équivalents classés par domaine de spécialité, par exemple dans IATE vers l'anglais, *leverage* en 'questions sociales', *seizing* en 'communication', *setting up* en 'bâtiment et travaux publics', etc. À titre de comparaison, ce même mot a deux entrées dans le TLFi, avec plusieurs acceptions chacune.

un premier temps, les listes lexicales bilingues : langue ancienne - langue parlée. Ainsi, les premiers recueils qui ont été retrouvés recensent des mots en sumérien (langue qui disparaît au IIIe millénaire av. J.-C.) et une langue vivante de l'époque (sumérien-éblaïte ou sumérien-akkadien) [Bou03].

De même, le Moyen Âge foisonne de dictionnaires bilingues latin - langues vernaculaires (le latin dévient de plus en plus éloigné des langues parlées, jusqu'à devenir inaccessible).

En ce qui concerne les dictionnaires monolingues, ce n'est qu'au XIXe siècle que la lexicographie s'intéresse à la création de grands ouvrages historiques, par exemple :

- *Deutsches Wörterbuch* de J. y W. Grimm, initié en 1838, publié entre 1852 et 1961 ;
- *Dictionnaire de la langue française* d'E. Littré, initié en 1844, publié entre 1863 et 1872 ;
- *Oxford English Dictonary* initié par J. Murray, publié entre 1884 et 1928 [28].

Un des apports de l'informatique dans ce domaine est la possibilité de consulter plusieurs dictionnaires en même temps et de voir, ainsi, l'évolution du traitement des entrées au fil du temps. Ceci est possible grâce à des plateformes multiressources, comme c'est le cas, par exemple, du *Nuevo Tesoro Lexicográfico de la Lengua Española* [29] qui recense tous les dictionnaires de RAE (figure 2.3) :

28. On peut constater qu'il n'y a pas eu de projet de cette envergure pour l'espagnol à cette époque. Le seul dictionnaire historique équivalent est actuellement en cours de construction par la RAE et est déjà un dictionnaire électronique (le *Nuevo diccionario històrico del español* `http://www.rae.es/recursos/diccionarios/nuevo-diccionario-historico`).

29. `http://buscon.rae.es/ntlle/SrvltGUILoginNtlle`

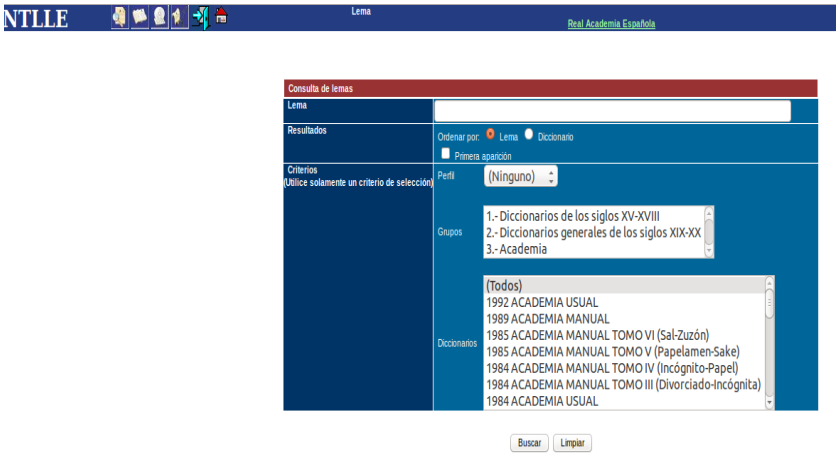


FIGURE 2.3 – NTLLE, plateforme multiressources

2.2.3 Dictionnaires de langue *vs* dictionnaires de spécialité

D'un point de vue de la 'portée', on peut caractériser les dictionnaires en deux grands groupes : les dictionnaires généraux (à visée exhaustive) et les dictionnaires de spécialité (restreints à des variantes de la langue ou à un champ d'expérience ou domaine particulier). Au sein des premiers, on distingue les dictionnaires normatifs (prescriptifs, issus d'institutions comme les Académies) et les descriptifs (d'usage, issus d'entreprises particulières ou privées). Selon la taille du dictionnaire, ainsi que le public visé (apprenants, public général, etc.), les dictionnaires de langue peuvent varier considérablement.

Quant aux dictionnaires de spécialité, on peut les classer en différents groupes, en fonction de la restriction choisie :

- **diatopique** : variantes liées à la géographie, variétés dialectales et/ou régionales, par exemple le *Diccionari català-valencià-balear*[30] (DCVB) d'A. M. Alcover et F. de B. Moll, initié fin du XIXe siècle et publié entre 1926-1935 et 1949-1962, un travail colossal qui réunit *"el conjunt de formes de parlar pròpies de totes les comarques indicades en el subtítol"*[31] ;

- **diastratique** : variantes liées à la couche sociale et au niveau de langue, principalement dictionnaires d'argot, par exemple *A New Dictionary of Terms, Ancient and Modern, of the Canting Crew*[32] d'au-

30. `http://dcvb.iecat.net/`
31. "L'ensemble des façons de parler des toutes les regions indiquées dans le titre" (le catalan compte environ 7 millions de locuteurs).
32. `http://www.bodleian.ox.ac.uk/news/2010-08-11`

teur anonyme (1699) (le mot 'slang' n'apparaît dans un dictionnaire anglais qu'en 1756). Cet ouvrage, parmi les premiers dans son genre, contient 4.000 entrées *"to educate the polite London classes in 'canting' – the language of thieves and ruffians"*, figure 2.4 ;

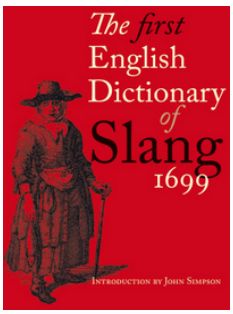


FIGURE 2.4 – Premier dictionnaire d'argot pour l'anglais britannique.

- **terminologique** : variantes liées à un domaine de spécialité, regroupant des termes propres à une aire d'activité, avec des degrés différents de technicité.

Un des progrès indéniables de l'informatique à ce niveau est la possibilité de mettre à jour les données très rapidement, ce qui reste un atout important pour des langues de spécialité ou pour des niveaux de langue soit soutenus soit familiers.

2.3 Ressources dictionnairiques informatisées

Après un aperçu rapide concernant les ouvrages imprimés, et une incursion inévitable dans des aspects liés à l'utilisation de l'informatique, dans cette section, nous nous intéressons à la lexicographie computationnelle. Il s'agit d'un vaste domaine qui peut être abordé sous des angles différents. Que ce soit sur des aspects méthodologiques (construction de ressources à partir de corpus électroniques, présentation des informations, typologie formelle de contenus, etc.), ou sur des aspects d'usage (accès aux données, technologie disponible : cédérom/web, pc/tablette/smartphone, etc.), les possibilités offertes grâce à l'utilisation des nouvelles technologies sont très vastes. Cette thématique a donné lieu à un volume de la revue tal en 2003 (*Les dictionnaires électroniques* [ZC03]), ainsi qu'à la naissance d'une conférence bi-annuelle depuis 2009, *Electronic lexicography* (*e-Lexicography*), qui vise à explorer des développements nouveaux où lexicographie et technologies du langage se rencontrent. Outre la mise en lumière des avantages et des inconvénients des versions papier et électronique des ressources, deux aspects

majeurs se sont dégagés lors des trois éditions de cette conférence (auxquelles nous avons assisté) : contenus et utilisateurs.

Ainsi, nous décrivons sommairement ici l'impact de l'utilisation des nouvelles technologies et de l'informatique en particulier, d'une part au niveau de la création et la mise à jour des dictionnaires et, d'autre part, au niveau de l'accès aux contenus par les utilisateurs. Dans cette section, nous considérons uniquement les dictionnaires informatisés, c'est-à-dire, les versions électroniques de ressources initialement imprimées, à usage principalement humain.

2.3.1 Informatisation de dictionnaires

La création d'un dictionnaire *ex nihilo* est très coûteuse. C'est la raison pour laquelle, actuellement, dans des situations socio-économiques défavorisées, on préfère informatiser des ouvrages imprimés existants. C'est le cas, par exemple, des langues peu dotées [ME13]. Pour des langues bien dotées comme l'anglais, informatiser les ressources existantes fut une activité généralisée il y a une trentaine d'années. Ainsi, les premiers dictionnaires informatisés virent le jour dans les années 1980, et ce ne fut que vers la fin des années 1990 que les premières versions en ligne apparurent, par exemple les *"Big Five"* (tous d'excellents dictionnaires monolingues pour des apprenants d'anglais [Béj10]) :

- le *Collins Birmingham University International Language Database* (COBUILD) [33])
- le *Oxford Advanced Learners Dictionary* (OALD) ([34])
- le *Longman Dictionary of Contemporary English* (LDOCE) [35])
- le *Collaborative International Dictionary of English* (CIDE), dérivé du *Merriam Webster's Dictionary*
- le *Macmillan Dictionary* ([36])

Du côté francophone, le *Trésor de la Langue Française* bénéficie d'une préparation informatique depuis ses débuts et donne lieu au *Trésor de Langue Française informatisé* (TLFi) [37] (1957, B. Quemada et P. Robert ; 1959 projet CNRS ; 1961 dynamique technologique, premières questions sur les aspects informatiques, accessible en ligne dès l'année 2000) [DP03]. D'autres

33. `http://www.collinslanguage.com/`
34. `https://oald8.oxfordlearnersdictionaries.com/`
35. `http://www.ldoceonline.com/`
36. `http://www.macmillandictionary.com/`
37. `www.atilf.fr/tlfi`

contributions institutionnelles à la lexicographie française ont également produit des versions informatisées dans les années 2000 [Pie13], par exemple, la 9e édition du *Dictionnaire de l'Académie Française*[38] (en 2005). Comme pour le TLFi, d'autres ressources sont complètement électroniques (élaboration et consultation), par exemple le *Dictionnaire du Moyen Français (1330-1500)*[39].

Du côté des grandes maisons d'édition privées, c'est aussi vers la fin des années 1990 - début 2000 que les versions en cédérom, puis en ligne voient le jour :

- Les dictionnaires *Robert*[40], en cédérom puis en ligne en 2005 (avec accès via abonnement)
- *Le Petit Larousse*, version de 1905 en ligne depuis 2010, élaboré à l'Université de Cergy-Pontoise [MBCH09]
- Les dictionnaires *Larousse*[41] sur le Web, avec récemment un onglet pour des applications mobiles (iPhone, iPad, Android et Windows Phone)

Quelle que soit leur langue et leur origine (institutionnel ou privé), en 2014 il est impensable que les grands dictionnaires n'aient pas des versions en ligne ou sur support électronique. À ce jour, certaines maisons d'édition en sont même à prendre des résolutions drastiques : il existe des dictionnaires qui ne seront plus jamais imprimés, comme le *Oxford-English-Dictionary* (principalement pour des raisons financières : *"increasing popularity of online alternatives"*)[42].

L'apparition et la généralisation d'Internet a eu des conséquences importantes pour la lexicographie : beaucoup plus que sur support cédérom, le nombre de ressources accessibles en ligne ne cesse d'augmenter quel que soit leur type (dictionnaires institutionnels, scolaires, de spécialité, etc.). Or, l'hétérogénéité de ressources va de pair avec l'hétérogénéité de sources. Si certains outils sont mis en ligne sous l'égide d'institutions reconnues (académies de langue, centres de recherche, universités, organismes officiels, etc.) d'autres existent avec paternité 'anonyme'. Dans ces cas, la véracité des informations, les sources des données, etc. sont alors mises en cause (mais ceci reste un problème plus général au niveau du Web, qui ne concerne pas seulement les ressources).

38. www.academie-francaise.fr/dictionnaire/
39. www.atilf.fr/dmf
40. www.lerobert.fr
41. www.larousse.fr/dictionnaires/francais-monolingue
42. http://www.telegraph.co.uk/culture/books/booknews/7970391/Oxford-English-Dictionary-will-not-be-printed-again.html

Par ailleurs, le phénomène Wikipédia a entraîné la naissance des Wiktionnaires, dictionnaires libres et gratuits "que chacun peut construire" [43]. Cette philosophie contributive, qui suscite débat (cohérence de la ressource, qualité des contributions, processus de relecture, etc.), semble réussir pour ce projet qui affiche 2 485 387 articles décrivant en français les mots de plus de 3 500 langues et 1 293 390 d'entrées pour le français (janvier 2014) [44]. À titre de comparaison, et selon les mêmes sources :

- *Robert Junior* 1999 : 20 000 mots
- *Petit Larousse* 2009 : 59 000 articles (sans compter les noms propres : 28 000 articles)
- *Petit Robert* 2001 : 60 000 mots
- *Grand Robert* 2001 : 86 000 articles et 100 000 mots

L'avantage de disposer de données libres dans le cadre du développement d'outils de TAL est indéniable. Ainsi, les données de Wikipédia et du Wiktionnaire sont régulièrement utilisées dans de nombreux projets de création ou d'enrichissement de ressources. Un exemple significatif en est *BabelNet* [NP10] (section 3.3.4). Nous y avions eu recours aussi pour notre ressource Polymots [GRT09] (extraction semi-automatique du texte introductif de Wikipédia pour obtenir des "unités de sens", section 3.3.3).

2.3.2 Le poids du papier et au-delà

Le poids du papier a déjà été évoqué dans la littérature. En effet, malgré des possibilités nouvelles fournies par les moyens informatiques, trop souvent les versions informatisées des dictionnaires restent tributaires de leurs versions imprimées :

> « Still, we can observe that only a few existing e-dictionaries really use the technical possibilities of the technical medium in the conception and preparation of dictionaries, and in the access to and presentation of data in them. »[FO10]

Ainsi, le contenu de ces ressources reste le même que dans les versions papier (même structure textuelle). Quant à la recherche d'informations, bien que l'option fenêtre de recherche soit généralisée, très souvent il y a aussi une liste alphabétique, par exemple dans le cédérom du Petit Robert ou le TLFi. Cette présentation hiérarchique offre des avantages dans deux types de cas. Premièrement, lors que l'utilisateur ne maîtrise pas la langue et cherche un

43. `http://fr.wiktionary.org/wiki/`
44. Nous discutons des ressources construites de façon contributive dans la section 3.1.2.

mot dont il n'est pas sûr de son orthographe. Dans un deuxième cas, lors que l'utilisateur cherche des familles morphologiques (nous en avons fait l'expérience lors de l'enrichissement de Polymots).

Néanmoins, de nombreux lexicographes sont d'accord sur le fait que les dictionnaires électroniques de demain devraient aller plus loin et même être complètement repensés, par exemple :

> « Electronic dictionaries would be most effective if they were designed from scratch with computer capabilities and computer search mechanisms in mind »[Nes00]

Comment mieux utiliser le milieu informatique autant du point de vue du lexicographe que de l'utilisateur (*"To think outside the paper"*) était le thème central de la dernière édition d'*e-Lexicography* en 2013 [45] :

> « The focus from the actual task at hand – to find different ways on how to exploit the rich potential of electronic medium in order to respond (quickly) to the needs of the new types of users, as well as to the needs of modern lexicographers, to forget about conventional approaches and be innovative, to conceptualize the dictionary with an electronic format in mind ; in other words, "to think outside the paper". »

Nous avions fait un premier bilan de certains aspects liés à cette thématique dans [Gal13]. Les points suivants résument ces idées, que nous avons enrichies avec des réflexions formulées par des spécialistes du domaine. Ils concernent, d'une part, la présentation et la nature des contenus et, d'autre part, les utilisations possibles.

2.3.3 Présentation des contenus

Pour ce qui est de la façon d'accéder et de visualiser le contenu lexicographique, les dictionnaires en ligne, libérés de la "camisole papier" [ZC03], offrent des possibilités améliorées, aussi bien au niveau de l'accès lexical (a) que de la visualisation des informations sous forme textuelle (b), statistique (c) ou diagrammatique (d) .

a) Accès lexical

Outre les listes alphabétiques, l'accès lexical plus commun dans les dictionnaires sur support informatique est la fenêtre de recherche. Au niveau de l'interface, cette fenêtre se retrouve généralement en haut de la page :

a) centrée, par exemple dans le DAELE (*Diccionario de Aprendizaje del Español como Lengua Extranjera*) 2.5 :

45. `http://eki.ee/elex2013/`

FIGURE 2.5 – DAELE, fenêtre de texte centrée.

b) à droite, par exemple dans le TLFi

c) à gauche, par exemple dans le GDLC (*Gran Diccionari de la Llengua Catalana*) [46]

La plupart de dictionnaires sont suffisamment robustes comme pour accepter des erreurs orthographiques, des mots fléchis, omission de tiret, etc. Dans ces cas, le dictionnaire renvoie à une liste de lemmes se rapprochant du mot entré par l'utilisateur. Certains dictionnaires, par exemple le TLFi, proposent également d'autres moyens de recherche comme la saisie phonétique (qui peut s'avérer très utile pour des apprenants maîtrisant mal l'orthographe de la langue). De manière plus générale, on peut aussi définir des clés d'indexation sur n'importe quel type d'information puis effectuer une recherche dessus, voire combiner les critères de recherche (par exemple : phonétique, catégorie grammaticale, exemples, traduction, etc.) ou même des méta-données (auteur, date de création, etc.)

b) Visualisation textuelle

Les dictionnaires informatisés et/ou électroniques restent, on l'a vu, très liés à la textualisation des informations. La lecture sur écran se voit néanmoins facilitée grâce à des systèmes de mise en évidence de certains types d'information (le choix d'une couleur associée à une zone clé). Par exemple, dans le cas de la première version du *Collins On-Line* bilingue français-anglais (fin des années 1990) la couleur permet de mettre en évidence la langue cible (figure 2.6, [Man01]) :

46. `http://www.diccionari.cat/`

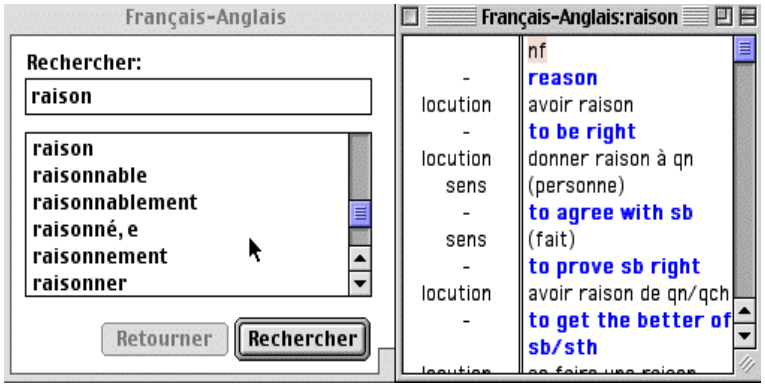


FIGURE 2.6 – Premier Collins on-line bilingue.

Plus récemment, dans le LDOCE (*Longman Dictionary of Contemporary English*) la définition apparaît soulignée automatiquement. Quant au TLFi (figure 2.7), c'est l'utilisateur qui peut choisir un champ particulier à mettre en valeur (définition, catégorie grammaticale, exemple, synonyme, etc.) :

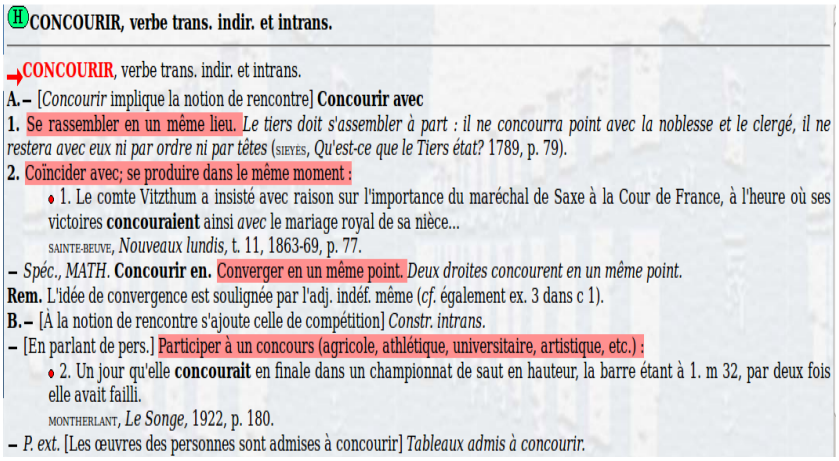


FIGURE 2.7 – TLFi, mise en relief par couleur

Par ailleurs, les hyperliens permettent de naviguer dans la nomenclature des dictionnaires. Certaines ressources sont à même de permettre l'accès à l'article d'un mot présent dans le texte d'une entrée, par exemple dans le Larousse [47] il faut cliquer pour qu'un mot devienne souligné, l'hyperlien apparaît alors sous la forme d'une loupe (figure 2.8) :

47. `http://www.larousse.fr/dictionnaires/francais`

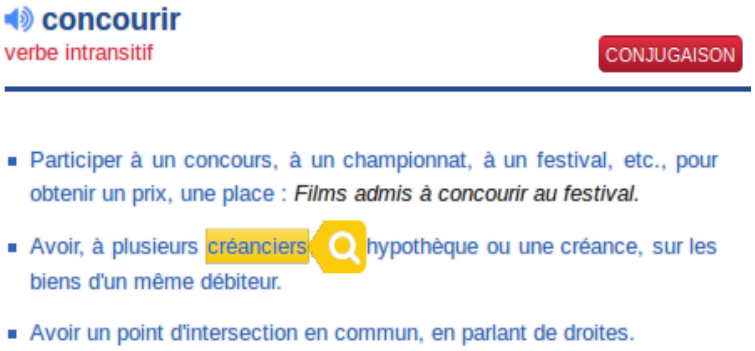


FIGURE 2.8 – Larousse, hypernavigation.

c) Visualisation statistique

Dans les nouveaux types de visualisation de contenu il y a les nuages de mots [48]. Ce sont des représentations visuelles permettant d'afficher les mots avec des polices de caractères d'autant plus grandes que le mot est significatif d'un point de vue statistique (plus fréquent dans un corpus). L'ajout des couleurs peut aussi être significatif, par exemple, pour différentier des catégories grammaticales. Le dictionnaire anglais LDOCE offre ce type de visualisation pour certaines entrées (chacune d'elles est aussi un hyperlien vers l'entrée textuelle) :

FIGURE 2.9 – LDOCE, nuage de mots.

48. Il s'agit d'une technologie très utilisée pour représenter le contenu sémantique d'un site web (on parle aussi de "nuage de tags" ou simplement de *"cloud"*).

d) Visualisation diagrammatique

Enfin, un des moyens de visualiser les liens entre les unités lexicales et les différentes informations les reliant (liens lexicaux) est la représentation par graphe. À partir d'un mot, on peut visualiser les mots en relation et on peut naviguer dans ces liens. Un exemple en est le *Visual Thesaurus*[49] ou le *GraphWords*[50] (figure 2.10, les couleurs sont également porteuses d'informations sur la catégorie grammaticale) :

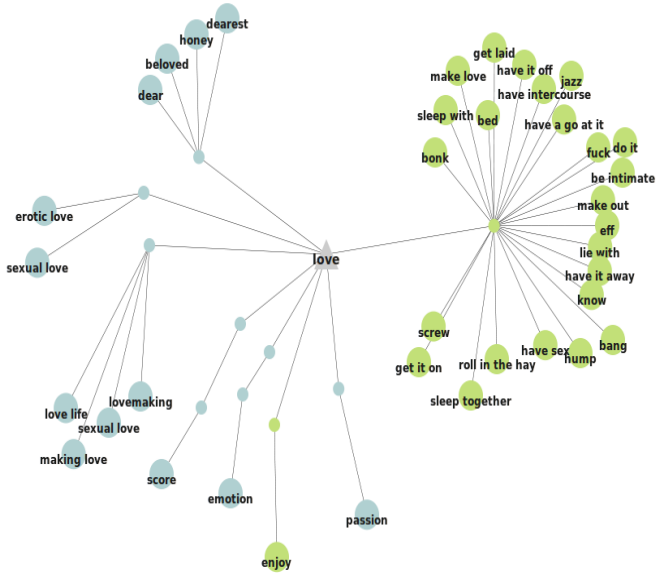


FIGURE 2.10 – GraphWords.

Comme nous l'avons vu dans la section 2.1.3, ce type de représentation *net-like* [Pol06] est très appropriée pour mettre en évidence les liens entre les différents objets lexicaux et les cliques sémantiques (regroupements de sens), que ce soit dans le cadre d'un thesaurus ou d'un autre type de ressource lexico-sémantique, monolingue et surtout interlingue.

2.3.4 Nature des informations

Les dictionnaires en version électronique ont la possibilité d'offrir des informations sur différents supports et de nature très différente :

- Information **multimédia** : outre les images, qui étaient déjà présentes surtout dans les dictionnaires encyclopédiques, le son est un apport

49. `https://www.visualthesaurus.com/`
50. `http://graphwords.com/`

important que ce soit dans les dictionnaires monolingues (spécialement pour apprenants) ou dans les bilingues, parfois avec des propositions multilingues [51] (figure 2.11) :

FIGURE 2.11 – Prononciations multilingues dans le Collins French-English.

- Information **multimodale** : en plus d'inclure du son, certaines ressources bénéficient d'images animées, adressées à des publics diversifiés. Par exemple, *ARASAAC*[52] propose des ressources graphiques pour des personnes présentant des difficultés de communication. Les dictionnaires pour les langues de signes ont également vu le jour, par exemple, le lexique LSF de l'Institut National de Jeunes Sourds de Metz [53], le *American Sign Language Dictionary* [54], *Auslan* pour la langue de signes australienne [55], etc. Sematos [56] est, quant à lui, un dictionnaire multilingue en cours de construction comprenant quatre langues de signes (figure 2.12) : espagnole (6.076 mots), française (3600 mots), anglaise (440 mots) et catalane (297 mots) :

51. http://www.collinsdictionary.com/dictionary/french-english
52. http://www.catedu.es/arasaac/
53. http://www.lsfdico-injsmetz.fr/
54. http://www.handspeak.com/
55. http://www.auslan.org.au/
56. http://www.sematos.eu/lsf.html

FIGURE 2.12 – Dictionnaire Sematos, version langue de signes catalane (LSC).

- Information **multi-ressource(s)** : l'accès en ligne permet la consultation de plusieurs ressources à partir de la même interface. Les dictionnaires deviennent alors des plateformes multi-ressources, avec accès passif (automatique) ou actif (choix de l'utilisateur) à différentes sources d'information. Par exemple, le résultat de la recherche d'un mot dans *Wordnik*[57] est un ensemble varié d'informations (définitions classiques, exemples d'usage, mots utilisés dans les mêmes contextes, images, etc.) issues de différentes ressources ainsi que de participations individuelles pour les aspects plus ludiques (*cf.* 'informations périphériques'). Un deuxième exemple est le *Nuevo Tesoro Lexicográfico de la Lengua Española* (NTLLE), avec différents choix possibles sur les éditions du dictionnaire de la RAE (ici c'est l'utilisateur qui choisit quelle édition du *Diccionario* il souhaite consulter, voir section 2.2.2).

- Information **méta-lexicographique** : la majorité des interfaces en ligne présentent tout un ensemble d'informations sur le dictionnaire même (nombre d'entrées, sources, etc.) et sur les informations décrivant les entrées (codes, domaines, constructions, emplois, exemples d'usage, etc.). Pour cela, l'utilisateur peut accéder à des menus d'aide, des boutons d'information, des guides d'utilisation, des adresses mail de contact, etc.

- Information **périphérique** : par information périphérique nous faisons allusion à des informations diverses. Par exemple, dans le cas de

57. `www.wordnik.com`

verbes on peut accéder à des paradigmes de conjugaison. Certaines ressources présentent des aspects ludiques (par exemple *Wordnik*) : le mot du jour, des listes de mots ayant une caractéristique spéciale (rime, nombre de caractères, etc.), listes pour le scrabble, quizz, possibilité de commenter le mot, etc. Parfois, on retrouve des informations plus techniques (possibilité de transférer une entrée sur Facebook ou Tweeter, amélioration de la visualisation des contenus, téléchargement de navigateurs, etc.) et même des demandes de contribution financière au projet.

2.3.5 Utilisateurs et scénarios d'usage

Des spécialistes du domaine de la lexicographie s'accordent sur le fait que les nouveaux dictionnaires doivent être adaptés aux utilisateurs. Ceci ne concerne pas seulement le support ou la technologie utilisée mais surtout le contenu. Ainsi, P. Battaner, professeur de philologie espagnole et présidente de l'Association Espagnole d'Études Lexicographiques, soulignait dans un entretien lors de la conférence de lexicographie hispanique en juin 2012 [58] l'importance d'adapter les dictionnaires aux besoins des usagers. De même, A. Klosa, lexicographe à l'*Institut für Deutsche Sprache* (Mannheim) dans son intervention en tant que conférencière invitée à elex-2013 (*Internet lexicography : requirements, concepts, research approaches*) [59], mettait également en lumière l'interdépendance des contenus avec les utilisateurs.

Hélas, dans ce domaine, les innovations plus récentes concernent beaucoup plus la technologie que le contenu même des ressources (consultation via des smartphones ou des tablettes), allant parfois jusqu'à la démesure dans la 'gadgétisation'. Dès lors, la plupart des grands dictionnaires proposent leurs versions pour appareils mobiles, par exemple le *Diccionari de la llengua catalana de l'Institut d'Estudis Catalans* [60] (DIEC), en ligne et téléchargeable pour iPhone et iPad (figure 2.13) :

58. `http://www.youtube.com/watch?v=4olLJFLMDmc`, consulté en janvier 2014.
59. `http://www.youtube.com/watch?v=O8BQ5Of1228`, consulté en janvier 2014
60. `http://www.iec.cat/mobil/diec2_promo.asp`, Dictionnaire de l'académie de la langue catalane, version 2.

FIGURE 2.13 – Le DIEC pour applications mobiles.

Si la prise en compte de publics particuliers est un critère nécessaire et généralement répandu lors qu'on construit un dictionnaire (par exemple, dictionnaires pour apprenants *vs* dictionnaire de langue générale *vs* lexique d'un domaine de spécialité), la considération de scénarios d'usage divers reste un objectif à atteindre et un des grands défis à relever. En effet, on ne consulte pas de la même façon une ressource en situation mobile qu'au bureau, dans la vie quotidienne lors qu'on déchiffre un manuel ou en vacances lorsqu'on essaye d'identifier un aliment dans une langue inconnue, lorsqu'on maîtrise une langue ou lorsqu'on souffre d'une pathologie qui affecte la parole d'une façon ou d'une autre, etc. Ainsi, les différents types d'utilisateurs et leurs besoins demeurent fondamentaux lors de la construction de ressources.

2.4 Conclusion

En parallèle à nos activités dans le domaine du traitement automatique des langues, nous nous intéressons depuis quelques années à la lexicographie computationnelle. La participation à la première conférence *e-Lexicography* à Louvain-la-Neuve en 2009 fût un déclic dans ce sens. Outre l'approfondissement d'aspects historiques et récents (comme dans [Gal13]), nous nous intéressons aux évolutions du domaine. Ainsi, nous restons convaincue que les technologies du langage et le traitement automatique des langues auront un rôle important dans ces évolutions. Le rapprochement entre ces disciplines n'a fait que commencer.

Chapitre 3

De la 'mécanisation' du lexique à la construction de ressources lexicales

« Soumettre les formes matérielles du langage aux méthodes d'analyse des machines capables d'opérations arithmétiques et logiques était une entreprise trop séduisante pour ne pas s'imposer. »

E. Delavenay. *La machine à traduire*. Que sais-je ? Paris : 1972.

« Natural Language Processing would be impossible without lexical databases that are both large and sophisticated. »

C. Fellbaum et G. A. Miller. Morphosemantic links in WordNet. Dans *Les dictionnaires électroniques*. TAL 44(2). 2003. [FM03]

Dans les dernières décennies du XXe siècle, l'intérêt pour le lexique s'est intensifié, fondamentalement parce que les acteurs du domaine des technologies du langage ont perçu l'importance de la composante lexicale dans les outils de traitement automatique des langues. Dans ce chapitre, nous nous intéressons spécifiquement à ce domaine et, concrètement, aux traitements automatiques du lexique et à la construction de ressources. Dans la première partie (section 3.1), nous décrivons le cadre et les motivations du domaine ainsi que quelques aboutissements méthodologiques d'ordre général (standards, recommandations, méthodes de construction et d'évaluation). Dans la deuxième partie, nous présentons de façon synthétique quelques techniques état-de-l'art pour le traitement des données lexicales (section 3.2). Enfin, dans la troisième partie (section 3.3), nous décrivons différentes ressources existantes, y compris celles auxquelles nous avons contribué dans la pé-

riode 2005 à 2012 : LexValf [GV05], Polymots ([GR08], [GRT09], [GRZ10], [RG11]), LexRom ([Gal11], [GBM13]) et Polarimots [GB12].

3.1 Généralités

Les travaux sur le lexique ont été favorisés par l'accessibilité des corpus électroniques et le besoin de ressources pour les outils de TAL. En France, les travaux en linguistique informatique de M. Gross et ceux de J. Dubois et F. Dubois-Charlier en lexicographie, ont sans doute eu aussi de l'influence dans la façon de décrire les données linguistiques en vue de leur traitement automatique. Ainsi, les premiers traitements automatiques (années 60) ont concerné le vocabulaire [Léo04] :

> « Le traitement automatique du langage commence par la mécanisation du lexique qui s'inscrit en droite ligne dans les préoccupations de nombre de linguistes de l'époque. »

Par ailleurs, lors du colloque *Lexicologie et lexicographie françaises et romanes* (Strasbourg, 1957) il a été évoqué « l'apport prometteur des machines mécanographiques et électroniques dans l'accélération des dépouillements et des classements du lexique »(B. Quemada et P. Robert commencent à travailler au *Trésor de la Langue Française* en 1957 et s'intéressent très rapidement à des aspects informatiques liés à la création et à la consultation du dictionnaire).

Ainsi, de la *mécanisation* du lexique dont parle J. Léon on est passés à son *automatisation* (termes que nous considérons abusifs, dans la mesure où ce n'est pas le lexique qui se mécanise ou s'automatise mais plutôt les méthodes pour construire les ressources ou pour accéder aux informations). *On Automating the Lexicon* est, par ailleurs, le titre d'un atelier qui eut lieu à Marina di Grosseto, en Italie, presque vingt ans plus tard (en 1986). Cet atelier de plusieurs chercheurs en traitement automatique des langues a posé les bases de toute une série de recommandations pour la création et l'enrichissement de ressources lexicales [1][Fra13] :

> « (...) a favorable climate for converging towards the common goal of demonstrating the feasability of large lexicons, which needed to be *reusable, polytheoretical* and *multifunctional*. This reflection has led to the definition of the concept of reusability of lexical resources as (1) the possibility of reusing the wealth of information contained in machine-readable dictionaries, by converting their data for incorporation into a variety of different

1. Ce workshop a, par ailleurs, débouché sur un livre de référence dans le domaine qui porte le même titre : *Automating the Lexicon*, D.E. Walter, A. Zampolli and N. Calzolari. Oxford : OUP. 1995.

NLP modules ; (2) the feasibility of building large-scale lexical resources that can be reused in different theoretical frameworks, for different types of application and by different users. »

Deux aspects liés à la réutilisation des ressources sont importants d'après [Fra13]. D'une part, l'utilisation d'informations déjà présentes dans des dictionnaires informatisés ou électroniques [2] et, d'autre part, la construction de ressources à large couverture, à visée étendue quant aux applications et aux utilisateurs.

Quant à la construction de ressources à large couverture, les acteurs du domaine, autant industriels que chercheurs d'organismes publics, ont convergé vers le développement de standards pour le lexique (section 3.1.1). Les méthodes de construction et d'évaluation (respectivement, sections 3.1.2 et 3.1.3) se sont aussi diversifiées dans le but d'atteindre le graal de la "large couverture" (*large-scale lexical resources*).

3.1.1 Normes, standards, consortiums

Dans les années 1990, on a vu la profusion de normes, recommandations et standards pour les ressources linguistiques en général. Pour les corpus citons la TEI (*Text Encoding Initiative*), EAGLES CES *Corpus Encoding Standard Initiative* (recommandations pour la création et l'évaluation de ressources au sens large, c'est-à-dire, aussi bien lexiques que corpus textuels et multimodaux) ou encore XCES (dérivé de EAGLES) pour baliser les corpus avec XML.

En terminologie, citons XLT (*XML representation of Lexicon and Terminology*) ou le TMF (*Terminological Markup Framework*, ISO 16642) qui permet de définir les structures et les mécanismes nécessaires pour représenter informatiquement les données terminologiques de façon indépendante à tout formalisme (méta-format d'annotation en XML) [Rom01]. Citons également OLIF2 (*Open Lexicon Interchange Format*), initiative similaire pour formaliser la structure et l'encodage des lexiques multilingues, l'échange des données lexicales et l'intégration dans des systèmes de traduction assistée par ordinateur.

Pour ce qui est du traitement des données lexicales et la création de lexiques, parmi les standards plus connus il y a eu la TEI P3 pour des dictionnaires ([IV95]), ACQUILEX, Eurotra-7, MULTILEX (lexiques multilingues), GENELEX (modèle générique) et le LMF (*Lexical Markup Framework* [3], ISO-24613 :2008) [Fra13]. Issu des recommandations EAGLES, ce

2. Le terme anglais *machine-readable* appliqué aux dictionnaires (MRD) fait allusion au fait que les données de ceux-ci sont analysables par un ordinateur. En français, il faut distinguer dictionnaires informatisés (chapitre 2) et dictionnaires électroniques -ceux construits expressement pour être lus par des machines et pour servir dans des applications en traitement automatique des langues.

3. `http://www.lexicalmarkupframework.org/`

```
<Lexicon>
  <feat att="language" val="fra"/>
  <LexicalEntry>
    <feat att="pos" val="noun"/>
    <Lemma>
      <feat att="writtenForm" val="treillis"/>
    </Lemma>
    <Sense id="s1">
      <feat att="definition" val="Entrecroisement de lattes"/>
      <SenseRelation target="s14">
         <feat att="label" val="synonym"/>
      </SenseRelation>
    </Sense>
    <Sense id="s2">
      <feat att="definition" val="Toile gommée"/>
    </Sense>
  </LexicalEntry>
  ...
  <LexicalEntry>
   <feat att="pos" val="noun"/>
    <Lemma>
      <feat att="writtenForm" val="treillage"/>
    </Lemma>
    <Sense id="s14">
    ...
  </LexicalEntry>
</Lexicon>
```

FIGURE 3.1 – Exemple d'encodage LMF pour l'entrée "treillis".

dernier est peut être le plus significatif et largement utilisé. Ce qui rend ce standard intéressant est le fait d'être un méta-modèle : il n'impose aucun élément XML et il propose une organisation modulaire pour encoder les différentes informations contenues dans un lexique, ce qui en fait un modèle générique [FMC+06].

Par exemple, la figure 3.1 présente un exemple d'entrée encodée selon le standard LMF. On peut distinguer des attributs liés au lexique et aux entrées lexicales encodées sous forme de traits (*feat* > *feature*, pour la langue (du lexique), pour les informations morphologiques (de l'entrée), pour le lemme et pour le sens). Le sens est noté à l'aide d'un identifiant : cet identifiant sert de pointeur pour des références croisées (le sens de la deuxième entrée lexicale `sense id = "s14"` est mis en rapport avec la première entrée `<SenseRelation target="s14">`).

Enfin, tous les efforts autour des normes et standards ont également donné lieu à des consortiums et organismes visant le partage de ressources, notamment ELRA-ELDA en Europe et le LDC (*Linguistic Data Consortium*) aux Etats Unis. Plus récemment, citons -en France- le CNRTL (Centre National de Ressources Textuelles et Lexicales, créé en 2005 à Nancy) et ORTOLANG (Outils et Ressources pour un Traitement Optimisé de la Langue, en 2012) avec comme objectifs : « proposer une infrastructure en réseau offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement clairement disponibles et documentés » [4]. Au niveau européen, citons le réseau CLARIN (*Common Language Resources and Technology Infrastructure* [5]), un consortium qui fédère différents organismes acteurs du domaine afin de gérer et de rendre accessibles des ressources et des technologies linguistiques variées, et la LRE Map [6].

3.1.2 Construction

Si pendant de longues années la construction de ressources lexicales était exclusivement réservée au domaine de la lexicographie (chapitre 2 de ce mémoire), il est évident que les acteurs des technologies du langage et du traitement automatique des langues ont radicalement changé la donne. Nous nous étions intéressée à ces aspects lors d'une publication pour la conférence *E-lexicography* en 2011, en collaboration avec M. Lafourcade [GL11].

En tant que composants essentiels des systèmes de traitement automatique des langues, le besoin de construire des ressources lexicales s'est très rapidement imposé. L'idée d'utiliser des dictionnaires informatisés (*machine-readable dictionaries*) comme source pour extraire des connaissances lexicales a été exploitée dans de nombreux travaux dans les années 1980 (quelques références dans [Bri91]). Cependant, les attentes initiales s'avérèrent assez vite très loin de pouvoir répondre aux attentes : « encouraging line of research »[VI90] mais « the information they contain is both too inconsistent and incomplete to provide ready-made source of comprehensive lexical knowledge »[IV94]. D'autres auteurs s'accordent sur ce point :

> « Conventional dictionaries tend naturally to be relatively informal and unsystematic, and, by the very nature of their organisation, to focus on the individual word rather than generalisations about classes of lexical items. »[Bri91]

Ainsi, le TAL s'est vu doter de ses propres ressources, méthodologiquement exploitables par les ordinateurs et adaptées aux différentes applications

4. `http://www.ortolang.fr/`
5. `http://www.mpi.nl/clarin/`
6. `http://www.resourcebook.eu/searchll.php`

du domaine [7] . Pour cela, on a construit des ressources structurées, explicites, dynamiques et multidimensionnelles (les informations sont souvent reparties dans des modules différents).

Plusieurs projets ont ainsi démontré la faisabilité des lexiques à large couverture, "réutilisables, polythéoriques et multifonctionnels" [Fra13]. Cependant, étant donnée la nature des langues, leur réalisation demeure un défi important [CCFH99]. Trois méthodes de construction peuvent être identifiées :

1. Manuelle (avec certaines tâches automatisées) : WordNet [Mil90], EuroWordNet [Vos00] ;

2. Automatique, acquisition automatique de données : (fusion de ressources) BabelNet [NP10], (adaptation de ressources pour les langues peu dotées) [HGN14] ;

3. Contributive : Papillon [MSL03], JeuxDeMots [Laf07], Wiktionnaire [8], etc.

Constructions manuelles

Il existe de nombreuses ressources construites de façon manuelle (WordNet parmi les plus notoires). En effet, certaines analyses ne peuvent pas se faire sur la base de régularités formelles, par exemple, des regroupements en familles morphophonologiques (comme dans Polymots [GR08] (3.3.1) ou dans [VORN96]) ou bien nécessitent des distinctions sémantiques fines :

> « The morphologically related word groups they find need to be further distinguished on the basis of meaning, and rules cannot account for all cases. »[FM03]

Il est évident que la création ou l'enrichissement manuel de ressources est extrêmement coûteux. D'où l'intérêt d'automatiser un maximum de tâches, ou de combiner des méthodes (méthodologie hybride : besoin de procédures automatiques pour viser une large-couverture, mais nécessité de validations manuelles pour certaines tâches).

---

7. Bien que destinées aux machines, certaines de ces ressources ont également été adaptées pour des usages humains (visualisations, ergonomie pour les recherches, etc.). Par exemple WordNet dans sa version consultable en ligne : `http://wordnetweb.princeton.edu/perl/webwn`.

8. Appellation francisée de Wiktionary, ressource lexicale créée en tant que “complément” du projet encyclopédique Wikipédia.

Acquisition automatique

La méthode plus largement exploitée en TAL est la construction automatique. Un bon nombre d'approches non supervisées ou semi-supervisées se sont développées pour l'extraction d'information linguistique à différents niveaux : morphologique [CSL04], syntaxique [BC97], sémantique [NVG03]. Grâce à l'accessibilité à de grands volumes de texte 'brut' (notamment le Web), les méthodes non supervisées permettent d'acquérir de l'information à grande échelle et construire, ainsi, des 'très grand lexiques' (*very large lexicons* [Gre99]).

Quant aux méthodes (semi-)supervisées, elles exploitent des annotations obtenues automatiquement, généralement par des outils de TAL, étiqueteurs (*taggers*), analyseurs (*parsers*), etc. (3.2.2).

L'adaptation de ressources d'une langue pour une autre, dans le cas des langues peu dotées, est également une option qui suscite de l'intérêt, par exemple dans le cas des dialectes de l'arabe (cf. 5.3.1).

Méthodes contributives

Le principe des méthodes contributives est le travail parcellisé : la ressource se construit au fur et à mesure des collaborations des participants. Les perspectives de cette approche sont intéressantes (contributions humaines, qualité en théorie assurée, on pallie au côté fastidieux des tâches répétitives par les collaborations multiples). Malgré tout, il est assez difficile d'obtenir des contributions régulières et de qualité [CFRZ08]. Si depuis quelques années les projets *wiki* se sont multipliés, ils restent néanmoins peu adaptés pour des développements à très grande échelle [FAC11].

L'alternative la plus largement répandue est celle des jeux sérieux (*game with a purpose*, GWAP). Les participants aux tâches sont motivés grâce à la compétition et/ou des récompenses à gagner. L'exemple le plus remarquable en est JeuxDeMots [Laf07] qui depuis 2007 a réussi à collecter 239 704 termes (termes avec au moins une relation de type idée associée).

Les différentes méthodes de construction (respectivement, contributive (C), automatique (A), manuelle (M)) ont des répercussions au niveau de la taille des ressources créées, comme on peut le voir dans la figure 3.2 :

Ressource	Nb lemmes	Auteurs	Version	Méthode
JeuxDeMots	239 704	[Laf07]	(mai 2014)	C
Lexique 3	157 920	[New06]	v3 (2006)	A
Morphalou	68 075	[RSAF04]	v1 (2004)	A
Lefff	58 647	[Sag10]	v3 (2010)	A
Manulex	23 900	[LSCC04]	v1 (2004)	A
Polymots	19 510	[GR08]	v3 (2014)	M (A)

FIGURE 3.2 – Taille de quelques lexiques pour le français.

3.1.3 Évaluation

La question de l'évaluation des ressources s'avère cruciale et en même temps très complexe. Deux méthodes sont généralement envisagées : des validations manuelles (évaluation qualitative intrinsèque) et des évaluations par rapport à une tâche ou application visée (évaluation extrinsèque).

L'évaluation qualitative intrinsèque implique une observation manuelle d'une partie des données de la ressource. Il s'agit de déterminer, d'après l'expertise de l'évaluateur, si les données ont été ou non correctement annotées. Le travail d'évaluation manuelle est coûteux et demande beaucoup de précision. Il est néanmoins indispensable pour assurer la qualité de la ressource. Nous avons procédé comme cela pour l'évaluation des familles morphologiques de Polymots (3.3.1).

Il existe une autre façon d'évaluer la qualité d'une ressource : il s'agit de concevoir une tâche dédiée, c'est-à-dire, identifier une application qui se serve de la ressource et observer si oui ou non les résultats du système s'améliorent grâce à l'introduction de la ressource. Par exemple, [JLSZ12] avaient évalué un réseau lexical via un outil pour retrouver le mot sur le bout de la langue. Également, dans [GB12], nous avions montré un exemple d'intégration d'un lexique de polarités dans un système d'analyse d'opinions. Nous avions pu démontrer que le lexique améliorait les résultats du système (3.3.3).

3.1.4 Vers des données ouvertes et liées

Une des dernières évolutions dans le domaine est celui de la constitution de données liées ouvertes (*Open Linked Data*). Ce sujet a été l'un des sujets clés (*hot topic*) de la conférence LREC 2014 (à laquelle nous avons assisté), conférence qui a également hébergé le troisième atelier sur les données liées en linguistique (*3rd Workshop on Linked Data in Linguistics : Multilingual Knowledge Resources and Natural Language Processing*). Le besoin de partage et d'accessibilité à des données (déjà amorcé avec les standards

et consortiums des années 1980-1990, cf. 3.1.1) est ainsi en train de prendre de l'ampleur avec l'approche "données liées" :

> « Modeling and publishing language resources as linked data offers crucial advantages as compared to existing formalisms. In particular, (...) this can enhance the interoperability and the integration of linguistic resources. Further benefits of this approach include unambiguous identifiability of elements of linguistic description, the creation of dynamic, but unambiguous links between different resources, the possibility to query across distributed resources, and the availability of a mature technological infrastructure »[CCMF13]

Cette approche permet de référencer les données sur le Web avec des URIs (*Uniform Resource Identifiers*)[9], et d'y accéder avec un langage de requêtes puissant (SPARQL, *Protocol and RDF Query Language*). Un des avantages majeurs est le fait de pouvoir sélectionner des données de différentes ressources, physiquement distribuées dans le *cloud*, et d'en obtenir différentes vues selon les besoins.

3.2 Traitement des données lexicales

Par traitement des données, nous faisons allusion aux procédés qui permettent d'enrichir automatiquement les données lexicales. Les informations obtenues se retrouvent alors associées aux entrées dans le cadre de la construction (et l'enrichissement) de ressources lexicales ou bien sont utilisées pour le calcul d'autres informations à des niveaux de traitement supérieurs (par exemple, l'étiquetage morphologique est indispensable dans les traitements syntaxiques).

3.2.1 Traitements de 'bas niveau'

Les traitements de surface (ou de 'bas niveau') sont les premiers à intervenir dans la chaîne de traitement. Lorsqu'on construit une ressource à partir de corpus, la première étape est celle de la segmentation (*tokenization*) qui permet d'identifier les différentes unités du texte. C'est une étape cruciale qui permet de déterminer les unités lexicales (et aussi phrastiques) [GT94]. Un des enjeux de cette étape est l'identification des mots-forme inconnus, leur prise en compte est essentielle pour garantir la robustesse et la précision de la lemmatisation [Nam05]. L'analyse morphologique des unités est ainsi fondamentale pour toutes les applications du TAL, et ce pour deux

9. Au niveau du recensement de resources, ELRA-ELDA est également en train de mettre en place un standard d'identification de resources (ISLRI, *Internet Standard for Language Resources Identifier*).

raisons principales : le classement des données lexicales en catégories et la reconnaissance des formes inconnues.

Une fois les unités identifiées, l'analyse morphologique se poursuit avec le calcul automatique du lemme le plus vraisemblable pour une entrée donnée et par le calcul des catégories grammaticales possibles. Selon les outils, la liste de catégories contient toutes les possibilités (pondérées ou non) ou bien l'analyseur calcule la catégorie la plus probable. Le système le plus largement répandu pour cette tâche est le TreeTagger [Sch94], bien que d'autres outils existent.

L'annotation en catégories grammaticales peut également se faire à partir d'une ressource qui contient déjà ces informations, par exemple, nous avons utilisé les informations du TLFi pour Polymots (3.3.1), du Lexique 3 pour ReSyf (4.2.3), de NovLex pour Polymarmots (4.2.2), etc.

3.2.2 Apprentissage d'informations lexicales de 'haut niveau'

Dans la perspective d'obtention automatique de grands volumes de données annotées, la méthodologie la plus généralisée en TAL est l'acquisition d'informations à partir de corpus bruts (non supervisée) ou bien à partir de corpus annotés ou des lexiques (semi-supervisée). Par exemple, pour l'identification de familles morphologiques : [Gau99] ou [Ber07] (analyse morphologique non supervisée), [Hat09a] (à partir d'un dictionnaire).

Les informations de haut niveau concernent, très souvent, la structure argumentale des verbes et l'étiquetage de rôles sémantiques. Un autre type d'information qu'on peut acquérir par apprentissage sur des données est de l'information pour lever des ambiguïtés syntaxiques. Nous avions travaillé dans ce sens déjà lors de nos travaux de thèse [Gal03] et avions poursuivi en collaboration avec M. Lafourcade (LIRMM) [GL05] et [GL06]. Nous avions ainsi proposée une méthode qui combinait des fréquences lexicales à des informations sémantiques (signatures lexicales) dans le but de résoudre les ambiguïtés de rattachement prépositionnel produites par un analyseur (XIP [AMC02]) [10].

3.2.3 Traitements statistiques

Une des évolutions plus marquantes des dernières décades dans le domaine du traitement automatique des langues est l'utilisation de méthodes statistiques pour obtenir des informations sur les données. Initialement utilisés dans les systèmes de reconnaissance automatique de la parole, les modèles de langage sont largement employés aussi au niveau des corpus écrits et des listes de vocabulaire pour identifier des formes lexicales spécifiques (par exemple, les expressions polylexicales) ou pour classer des phénomènes

10. Nous avions envisagé un travail de création de lexique de collocations avec les informations obtenues avec cette méthode, mais cela en est resté à un projet 'en chantier'.

linguistiques au sein de catégories définies (par exemple, des catégories morphosyntaxiques, des niveaux de difficulté, etc.). Avec les modèles n-grammes, par exemple, on estime la probabilité d'occurrence d'un symbole conditionné à l'occurrence préalable d'autres symboles (principalement, n-1 pour des bigrammes, n-2 pour des trigrammes, etc.). Le but est d'obtenir, à partir de grandes quantités de données, des probabilités pour des séquences d'une unité donnée (phonème, lettre, syllabe, lemme, etc.), c'est-à-dire, des informations sur le lexique à partir de régularités observées. Les applications pour la construction de ressources lexicales sont donc intéressantes.

Les travaux de ce type se rapprochent de la lexicométrie, domaine d'étude qui s'intéresse aux propriétés statistiques des unités lexicales dans des corpus (cf. 1.4.2). Les résultats obtenus s'intègrent ainsi à des ressources, par exemple, Lexique 3 [NPFM01] propose des fréquences brutes obtenues dans deux types de corpus (livres et sous-titres de films). Cependant, comme discuté dans [Fra11] (pp. 222-223), une des principales limitations de l'approche lexicale fréquentielle est la qualité de l'estimation de ces fréquences, qui se dégrade sensiblement pour les mots peu fréquents. De même, les résultats obtenus sont interdépendants du type et de la taille des corpus utilisés. De ce fait, les informations statistiques associées aux entrées lexicales dans les ressources tendent à intégrer des paramètres plus variés (différentes tailles et/ou échantillonnages de corpus,etc.). Par exemple, Manulex [LSCC04] présente des informations statistiques plus complexes (indice de dispersion, index de fréquence) afin de prendre en compte la variabilité des corpus et la distribution du lexique au travers des différents textes.

L'utilisation de mesures (indices de dispersion) ou de techniques statistiques plus complexes (catégorisation ou tri de formes via des modèles de classification) vise, ainsi, à dépasser les limites d'une approche fréquentielle brute et est devenue une tendance très importante dans les travaux de TAL appliqués au lexique. Nous avons travaillé dans cette perspective, en collaboration avec T. François et C. Fairon (CENTAL) pour la construction de deux lexiques où les mots sont gradués en fonction de leur difficulté [GFF13] et [FGWF14] (cf. 4.2.3).

3.3 Ressources pour le TAL

Comme nous l'avons décrit plus haut, les ressources pour le TAL requièrent une description fine et explicite des informations linguistiques. Dans certaines ressources, cette description est souvent accompagnée d'informations statistiques de différente complexité. L'information est multidimensionnelle et structurée selon différents modèles informatiques (bases de données relationnelles, graphes et/ou réseaux, etc.).

Dans cette section, nous présentons quatre types de lexiques dans lesquels nous avons travaillé, classés en fonction du type d'information linguis-

tique [11].

3.3.1 Lexiques morphologiques (Polymots)

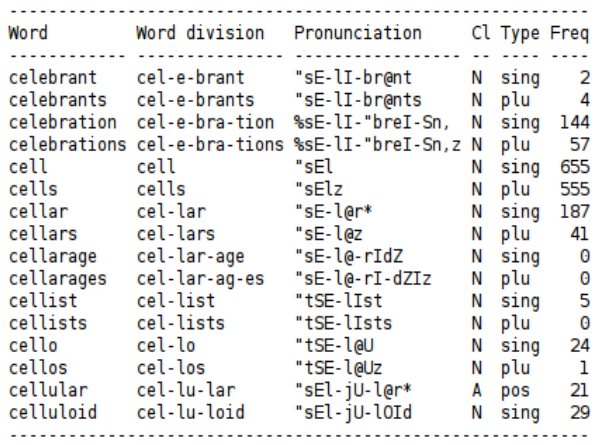
Les lexiques morphologiques sont des ressources lexicales répertoriant, pour chaque entrée, un ensemble d'informations variées d'ordre morpho-phonologique ou morphosyntaxique. Généralement, associer un lemme et une étiquette morphosyntaxique à chaque forme fléchie est fait par des analyseurs morphologiques (traitements de surface ou de 'bas niveau', 3.2.1). Il existe néanmoins des ressources accessibles qui contiennent ces informations. Pour le français, par exemple, Morphalou [RSAF04], totalise 524 725 formes fléchies correspondant à 95 810 lemmes issus de la base Frantext. À chaque entrée sont associées : la catégorie grammaticale, le genre, le nombre et un lien vers l'entrée d'origine dans le TLFi.

D'autres ressources morphologiques s'intéressent plus spécifiquement à la morphologie dérivationnelle. Par exemple, Morphonnette ([Hat08] et [Hat09b]) est un réseau morphologique visant à faire émerger la structure morphologique du lexique à partir des propriétés sémantiques et formelles des mots. Elle a été conçue à partir des représentations phonologiques des entrées du TLF (83 082) et utilise une mesure de similarité morphologique pour déterminer les analogies entre les formes au niveau des familles de mots et des séries (*fructificateur :fructification, modificateur :modification, rectificateur :rectification* et aussi *fructificateur :rectificateur*, etc.).

La ressource morphologique la plus connue est la base CELEX [12] [BPvR95]. Elle est constitué de trois bases de données lexicales (124 136 formes en néerlandais, 52 447 formes en anglais et 51 728 formes en allemand), la figure 3.3 est une capture d'écran pour l'anglais :

11. Nous aborderons les lexiques génériques utilisés en TAL (par exemple, Lexique 3 [NPFM01]) et les lexiques scolaires utilisés pour des études en psycholinguistique (Brulex [CMR90], NovLex [LC01], Manulex [LSCC04]) dans le chapitre suivant de ce mémoire.

12. Il existe une version récente de consultation sur le web `http://celex.mpi.nl/`.



Word	Word division	Pronunciation	Cl	Type	Freq
celebrant	cel-e-brant	"sE-lI-br@nt	N	sing	2
celebrants	cel-e-brants	"sE-lI-br@nts	N	plu	4
celebration	cel-e-bra-tion	%sE-lI-"breI-Sn,	N	sing	144
celebrations	cel-e-bra-tions	%sE-lI-"breI-Sn,z	N	plu	57
cell	cell	"sEl	N	sing	655
cells	cells	"sElz	N	plu	555
cellar	cel-lar	"sE-l@r*	N	sing	187
cellars	cel-lars	"sE-l@z	N	plu	41
cellarage	cel-lar-age	"sE-l@-rIdZ	N	sing	0
cellarages	cel-lar-ag-es	"sE-l@-rI-dZIz	N	plu	0
cellist	cel-list	"tSE-lIst	N	sing	5
cellists	cel-lists	"tSE-lIsts	N	plu	0
cello	cel-lo	"tSE-l@U	N	sing	24
cellos	cel-los	"tSE-l@Uz	N	plu	1
cellular	cel-lu-lar	"sEl-jU-l@r*	A	pos	21
celluloid	cel-lu-loid	"sEl-jU-lOId	N	sing	29

FIGURE 3.3 – CELEX.

À chaque entrée sont associées : la structure morphologique dérivationnelle (*word division*), une transcription phonétique SAMPA [13], la catégorie morphosyntaxique et un indice de fréquence.

Polymots

Polymots [14] [GR08], [GRZ10] et [RG11] est une ressource lexicale pour le français qui présente des mots regroupés en familles. Elle a été construite de façon semi-automatique [15] à partir d'une liste de mots extraits du dictionnaire Larousse 2000 et du TLFi. A l'origine, l'objectif d'une telle ressource était fondamentalement pédagogique : apprentissage du vocabulaire et de l'orthographe du français en milieu scolaire ou clinique (orthophoniste). Il s'agit d'un travail réalisé en collaboration avec V. Rey (CREDO) pour les aspects linguistiques, M. Zock (LIF) pour les aspects sémantiques et L. Tichit (IML) pour les aspects informatiques.

L'approche proposée, bien que liée à la morphologie dérivationnelle, s'en éloigne dans la mesure où les formes de base, que nous avons appelées "radicaux phonologiques", ne sont pas forcément des lemmes correspondants à des mots existants dans la langue, mais plutôt des formes phonologiques communes à un ensemble de mots de la même famille, ayant ou non un sens établi. Le parti pris était le suivant : la productivité morphologique dans la construction lexicale dépend non seulement de la structure de surface des mots (forme sonore immédiate des mots) mais aussi de la structure

13. (*Speech Assessment Methods Phonetic Alphabet*, jeu de caractères phonétiques utilisant les caractères ASCII 7-bits imprimables.

14. `http://polymots.lif.univ-mrs.fr/v2/`

15. Découpage morphologique manuel, acquisition d'information sémantique semi-supervisée.

profonde (règle sous-jacente rendant compte des constructions irrégulières). Nous rejoignions ainsi les travaux de Corbin [Cor87] sur la morphologie dérivationnelle :

> « Toute personne qui a acquis la connaissance d'une langue a intériorisé un système de règles qui détermine des connexions son-sens pour une infinité de mots construits . (...) C'est ce système de règles qui le rend capable de produire et d'interpréter des mots construits qu'il n'a jamais rencontrés auparavant.(...) Chacun comprend et produit des mots construits nouveaux sans aucune conscience de leur nouveauté (...). »[Cor87] (pp. 47-48)

L'approche se voulait synchronique. Cependant, Corbin montre que l'analyse morphologique ne doit pas séparer les deux dimensions : des informations historiques peuvent fournir certains renseignements parfois nécessaires et/ou compléter une défaillance dans la compétence du morphologue. Notre démarche reposait donc sur l'analyse morphologique des mots, à partir des radicaux, en tenant compte des contraintes morpho-phonologiques. Chaque famille de mots ainsi construite a été contrôlée avec un dictionnaire historique [16] afin de valider les rapprochements : deux formes présentées dans la même famille mais n'ayant pas une origine étymologique commune avaient été séparées. L'histoire des mots de la langue n'est pas donc à l'origine de la segmentation des mots mais sert à contrôler la validité du découpage.

À l'origine (2008), le travail réalisé reposait sur une liste de vingt mille mots communs extraits des 59 000 entrées du dictionnaire Larousse (2000). Ces mots désignaient soit un *objet* (concret ou abstrait), soit une *activité*, soit une *qualité* (il s'agissait donc d'unités de dénomination). Les mots grammaticaux, les noms scientifiques et les noms propres n'ont pas été intégrés dans cette base. La première version contenait vingt mille mots découpés en unités morphologiques : l'objectif était d'isoler une forme commune aux membres d'une même famille et de lister les affixes composant une entrée donnée.

Une fois résolu le traitement morpho-phonologique des unités, la question de la variation lexicale au sein d'une même famille s'est alors posée : avions-nous le droit de rassembler dans une même famille, des mots partageant un même radical morpho-phonologique mais appartenant à des champs sémantiques apparemment différents (comme "ride" et "rideau") ? Si oui, pourquoi ? Si le concept de famille de mots reposait sur un radical commun et une signification commune, alors il s'agissait, d'après nous, d'une construction très réductrice. En effet, comme unité de désignation, le mot peut comporter plusieurs traits sémantiques (dans une perspective structuraliste) générant des mots très différents. Ce phénomène est très bien documenté d'un point de vue historique (c.f. note 12). Notre hypothèse était qu'il en était

16. *Dictionnaire historique de langue française* d'A. Rey.

de même dans l'actualité de la langue. Entre les mots "boule", "boulotte" et "boulon", le trait sémantique de 'rondeur' est commun ; cela est donc relativement transparent pour le lecteur. Cependant, le mot "bouleverser" pose question : il y a bien la forme "boule" ; il y a également la forme "verse" qui indique un mouvement de retournement (on ne peut verser que vers le bas). La construction de ces deux mots, si on accepte cette analyse, conduit à la désignation d'un état d'âme. Le regroupement de mots sur un principe morpho-phonologique interroge donc la continuité sémantique.

Tel était le propos du travail que nous avons mené pour essayer de déterminer le continuité ou la dispersion sémantique au sein des familles [GRT09]. Pour ce faire, nous avons intégré de l'information sémantique : chaque mot de Polymots s'est vu associer des 'unités de sens' extraites automatiquement à partir de différentes ressources. La figure 3.4 montre les résultats sur l'interface de consultation pour l'entrée 'athlète' :

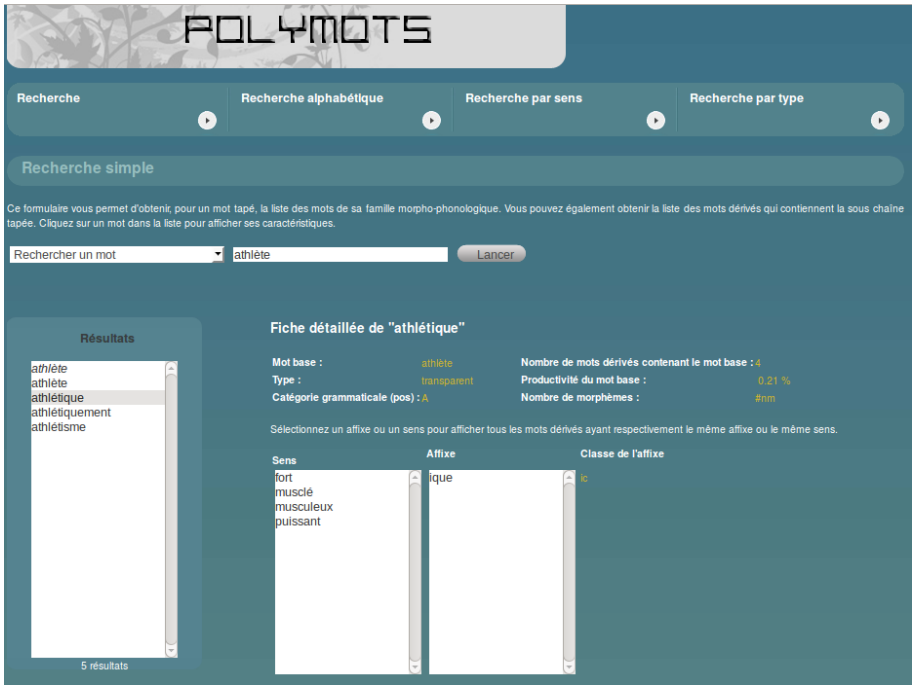


FIGURE 3.4 – Polymots.

La colonne de gauche montre la liste de mots dans la famille morphologique, la colonne de droite montre l'affixe pour le mot de la famille sélectionné (athlétique), la colonne centrale donne les 'unités de sens' pour ce même mot (fort, musclé, musculeux, puissant).

Dans un souci de diversification de corpus, mais confrontés à la difficulté d'obtenir facilement des ressources structurées, pour obtenir les 'uni-

tés de sens' nous avions choisi d'utiliser des sources lexicographiques et encyclopédiques accessibles librement, à savoir, Wiktionnaire et Wikipédia. Nous avions aussi à notre disposition le dictionnaire Hachette-XML. La méthodologie reposait sur l'extraction semi-automatique des définitions (hors exemples) dans le cas des dictionnaires, et du texte introductif (avant la table de matières) dans le cas de Wikipédia. Nous avions fait l'hypothèse de la présence de termes significatifs dans ces extraits, c'est-à-dire des termes ('unités de sens'), caractérisant sémantiquement chaque mot m donné. Par exemple, à partir de la première définition de "bras" du Wiktionnaire nous avions obtenu "partie, membre, supérieur, humain, bipède" etc. (lemmatisation avec TreeTagger [Sch94]) :

« (Anatomie) Partie du membre supérieur des humains (et des bipèdes en général) qui s'étend depuis l'épaule jusqu'au coude »

Par la suite, nous avions attribué à chaque unité de sens un poids $w(u_i)$ en fonction de la distance de chaque unité u_i par rapport au mot m, en tenant compte du nombre total de mots n dans chaque définition (avec $0 \leq i < n$). Nous avions enfin harmonisé pour chaque mot m, les poids calculés en divisant l'ensemble des valeurs par le poids maximal. Nous avions ainsi obtenu des espaces sémantiques comme par exemple, pour 'embrasser' (figure 3.5) :

[serrer 1] [contenir 0.666] [saisir 0.666] [bras 0.585] [attacher 0.444] [entourer 0.444] [étendre 0.314] [regard 0.314] [adopter 0.296] [baiser 0.248] [englober 0.166] [étreindre 0.148] ...

FIGURE 3.5 – Polymots : espace sémantique pour le mot 'embrasser'.

Une première version de Polymots avec des informations morphologiques et sémantiques accessibles sur le Web a été présentée à la conférence LREC (*Language Ressources Evaluation Conference*) en 2010 [GRZ10]. Le développement informatique (modèle relationnel, SGBD, interfaces web) a été fait par deux groupes d'étudiants de L3 informatique de Luminy, que L. Tichit et nous-même avons encadrés au premier semestre 2008 et 2009.

Le travail sur Polymots s'est poursuivi (2010-2012) avec une amélioration des informations morphologiques et un enrichissement des données. Par ailleurs, des découpages plus fins des familles ont été faits avec des méthodes semi-supervisées (création de clusters sémantiques) [GHN+11], en collaboration avec N. Hathout (CLLE-ERSS), A. Nasr (LIF) et S. Seppälä (LIF). Nous avons également collaboré avec P. Gambette pour la construction d'arbres de mots [GGNG11] et [GGN12].

La version 2 (2012), version actuellement mise en ligne, contient 19 193 lemmes étiquetés, regroupés en 2 289 familles. Les améliorations proposées

pour cette deuxième version ont trait aux bases, aux affixes et aux unités de sens :

- Bases
  - majorité de bases transparentes (2 157, 94%) à cause des rapprochements sémantiques en synchronie (la distinction transparent/opaque disparaît dans la version 3)
  - prise en compte de l'homonymie pour certaines bases : 'botte1' (agric.) / 'botte2' (chaussure) ; 'canon1' (arme) / 'canon2' (religion) ; 'casse1' (bris) / 'casse2' (boîte) / 'casse3' (poêle), etc.
  - suppression de mots fléchis (principalement de genre féminin)
  - ajout de la catégorie morphosyntaxique (N, V, A, Adv) à partir du TLFi
  - ajout du nombre de morphèmes (distinction 1-n)
- Affixes
  - classification des affixes et éléments formants en classes (98 préfixes, 126 suffixes), afin de prendre en compte des allomorphies et des variantes phonologiques, par exemple :
    - la classe i- regroupe les occurrences [i-, il-, in-, im-, ir-]
    - la classe -ais regroupe ['-ais', '-aise', '-ois', '-oise']
    - la classe -crate regroupe ['- crate', '-crat', '-ocrat', '-ocrate']
  - distinction d'affixes homographes, par exemple :
    - -is(V) dans des bases verbales (ex. 'actualiser', etc.) par rapport à -is(N) en fin de mot ('éboulis', etc.)
    - -esse(f) ('abesse', 'poetesse') par rapport à -esse ('allégresse', 'ivresse')
    - -ement(Adv) ('allègrement', 'pareillement') par rapport à -ement(N) ('apparentement', 'questionnement')
  - les affixes de flexion ont été enlevés, seules certaines marques ont été gardées :
    - infinitifs : -er(V), -re(V), etc. regroupés dans une classe VerbFin(flex)
    - nombre : -es(flex_pl_f) dans 'fiançailles', -s(flex_pl_m) dans 'vivres'
  - les affixes provenant d'autres langues (latin, anglais) ont également été marqués : -um(lat) dans 'référendum', -us(lat) dans 'terminus', -y(angl) dans 'sexy', 'jazzy', etc.

  - les formants participant à une construction par redoublement ont été marqués : bon(redoublm) dans 'bonbon', chou(redoublm) dans 'chouchou', dou(redoublm) dans 'doudou', 'doudoune', etc.
  - les affixes et formants hapax ont été enlevés
  - les quelques mots créés par composition en latin ont été gardés au sein d'une famille morphologique, mais le formant latin de droite n'a pas été considéré comme faisant partie de la liste d'affixes/formants (il est absent dans le découpage) : armistice, florilège, hélicoptère, lieutenant, majordome, orfèvre, piédestal, sacrilège, sangsue, septentrion, torticolis, vignoble.

- Sens
  - extraction automatique de synonymes et de mots thématiquement liés provenant du réseau de JeuxDeMots
  - 11 117 mots de Polymots v2 ont des synonymes (58%) et 8 076 n'en ont pas

La version 3, en cours d'enrichissement au niveau de la couverture, contient à ce jour (Février 2014) : 19 510 mots et 2 364 familles.

3.3.2 Lexiques syntaxiques (LexValF)

Les lexiques syntaxiques décrivent les propriétés syntaxiques des mots, par exemple, des cadres de sous-catégorisation, des restrictions de sélection, des classes d'argument, etc. L'intégration de ces lexiques à des analyseurs syntaxiques améliore significativement les résultats de l'analyse. En français, il existe un nombre important de ces ressources, dont nombre d'elles sont issues des tables du Lexique Grammaire (le Lefff [CSL04], SynLex [GGPF06], etc.). Les lexiques tabulaires du LADL ont également subi des transformations en automates à états finis ou en structures de traits [CT08].

Par ailleurs, DicoValence est un lexique de valences verbales pour le français, successeur du lexique PROTON [VM03]. Il a été développé dans le cadre méthodologique de l'Approche Pronominale. Pour identifier la valence d'un prédicat (ses dépendants et leurs caractéristiques), l'Approche Pronominale exploite la relation qui existe entre les dépendants dits lexicalisés (réalisés sous forme de syntagmes) et les pronoms qui couvrent en "intention" ces lexicalisations possibles. DicoValence se présente comme une liste d'entrées correspondant chacune à un emploi d'un lemme verbal. A chaque entrée lui est associée un type (prédicateur simple, verbe adjoint, verbe auxiliaire, verbe copule, etc.). Suivent alors les différents paradigmes qui dépendent du prédicateur (les termes de valences), avec pour chacun d'eux la liste des pronoms qui peuvent en être la réalisation. Sont enfin indiquées certaines propriétés complémentaires, dont les passivisations possibles.

LexValF

LexValF [17], quant à lui, est un lexique syntaxique qui a comme objectif l'encodage explicite des patrons de la complémentation verbale spécifique (les valences) des verbes du français. À chaque patron sont associés l'ensemble de restrictions lexicales et grammaticales et des indices sur la fréquence d'usage tirée d'informations issues du Web. Nous avions participé à ce projet dans la période 2005-2007.

Les données du lexique avaient été obtenues manuellement à partir d'une liste de patrons de valence (objets) établie par A. Valli [Val80] et sur la base desquels sont étudiés et interprétés les tables du Lexique-Grammaire des verbes du français développées au LADL par M. Gross et son équipe [18]. Ce travail systématique avait conduit à compléter la liste des patrons de bon nombre de verbes des tables, à formuler rigoureusement les restrictions lexicales et grammaticales, à les illustrer au moyen d'exemples (tirés des tables du Lexique-Grammaire, du *Petit Robert*, du *Grand Robert*, du *Trésor de la Langue Française* et du Web) et à indiquer leur fréquence d'usage sur le Web.

À chaque patron de valence d'un verbe sont associés différents types d'informations linguistiques :

- morphosyntaxiques (des types des syntagmes)
- restrictions grammaticales (des emplois passifs, impersonnels, etc.)
- sélections lexicales (des traits sémantiques ou classes d'objets)
- exemples d'emploi extraits de différentes sources

Nous avions structuré et formalisé ces informations dans le but de créer une ressource lexicale exploitable en ligne. Pour ce faire, nous avions créé un modèle relationnel encodant les différentes types d'informations [GV05] : pour une entrée donnée (lemme), il existe un cadre (frame) qui inclut l'ensemble de patrons de sous-catégorisation possibles avec les différentes restrictions de sélection et des exemples extraits de corpus [19].

17. `http://lexvalf.lif.univ-mrs.fr`

18. Tables mises à la disposition des chercheurs par l'Institut Gaspard Monge, Université Marne-la-Vallée, `http://ladl.univ-mlv.fr/`

19. Par exemple, le verbe 'distinguer' accepte un patron de sous-catégorisation de type SN Prép SN, uniquement avec les prépositions suivantes : 'à' (on le distingue *à* une cicatrice), 'de' (on distingue le vin *des* autres boissons), 'en' (on les distingue *en* deux classes), 'par' (la France se distingue *par* ses mathématiciens), 'pour' (le comité vient de le distinguer *pour* la découverte d'une molécule), 'd'avec' (il sait distinguer le vrai *d'avec* le faux), 'entre' (saurais-tu distinguer *entre* les effets et les causes ?), 'par rapport à' (distinguer le harcèlement *par rapport* à d'autres problématiques), etc. Tous ces exemples montrent que les classes de noms acceptées après l'une de ces préposition sont assez restreintes. Par ailleurs, d'autres prépositions comme 'dans' et 'parmi' n'imposent pas de restriction

Nous avions participé à la modélisation complète des données linguistiques (à l'origine, les données du lexique étaient non structurées dans un fichier réalisé avec un traitement de texte) et à une première version de l'interface de consultation, en collaboration avec C. Zaoui et L. Brieussel (membres de la jeune équipe DELIC). Ainsi, les informations fournies par LexValF sur les verbes français et leurs patrons syntaxiques (de valence) ont été rendues disponibles via une interface d'accès à la base de données, mises en ligne et récupérables pour des traitements automatiques. L'exemple de la figure 3.6 est un fragment du résultat obtenu pour le verbe 'développer' (45 résultats, 10 patrons de complémentation différents).

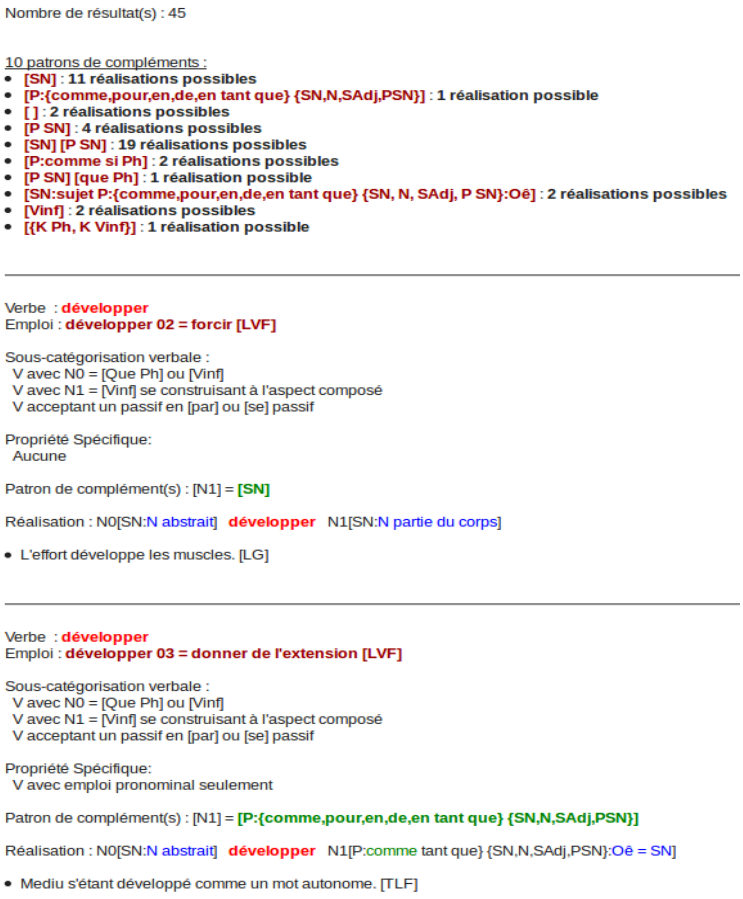


FIGURE 3.6 – LexValF.

À ce jour (mai 2014), la base contient 413 formes verbales représentées

(classe de noms ouverte). Dans LexValf ces différences sont cruciales : le type de syntagme qui suit le verbe est considéré comme argument dans le premier cas et comme modifieur pour le deuxième. Cette information est encodée dans le lexique.

par 10 031 réalisations. Le travail d'enrichissement au niveau linguistique se poursuit [20] : outre l'augmentation de la couverture, cela concerne principalement les équivalences entre l'emploi d'un patron syntaxique et un sens particulier du verbe.

3.3.3 Lexiques de polarités (Polarimots)

Depuis quelques années, l'analyse de sentiments suscite de l'intérêt dans la communauté du traitement automatique des langues (TAL) [21], comme conséquence d'un réel besoin dans le traitement de grandes masses de données : services web pour le tourisme ou la culture, discours politiques, etc. Par analyse de sentiments, on entend la détection de la polarité d'un texte, c'est-à-dire, l'obtention automatique de la tendance ou de l'opinion qui s'en dégage.

Deux approches ressortent dans la littérature. Les approches statistiques supervisées, fondées sur les co-occurrences de mots dans des corpus, et les approches plus linguistiques qui s'appuient, elles, sur des ressources lexicales. Nous avions travaillé dans cette dernière perspective en 2011-2012, en collaboration avec C. Brun (Xerox Research Centre Europe).

Si l'estimation de la polarité d'un texte passe par des phénomènes contextuels (intensificateurs, négation, etc.) et syntaxiques [Bru11], la qualité du lexique à la base du système reste cruciale. La construction d'un tel lexique demeure donc un aspect important.

Polarimots

À partir d'une liste initiale de 3 882 adjectifs annotés manuellement par trois annotateurs (C. Brun, P.-P. Hay-Napoleone [22] et nous même), nous avons voulu observer l'impact de la morphologie dérivationnelle dans le maintien ou non de la polarité. C'est-à-dire, nous avions voulu tester l'hypothèse selon laquelle la polarité intrinsèque d'un adjectif est la même que celle des unités lexicales de sa famille morphologique. L'idée était de voir :

(i) si on pouvait construire une ressource qui capitalisait sur les liens morphologiques pour propager des informations sémantiques,

(ii) si une telle ressource améliorerait les résultats d'un système d'analyse d'opinions [GB12].

20. Par A. Valli.

21. Par exemple, l'atelier *Sentiment and Subjectivity in Text* a COLING - ACL 2006, l'école d'été Eurolan 2007 *Semantics, Opinion and Sentiment in Text* , SemEval 2007 *Task 14 : Affective Text*, SemEval 2013 *Task 14 : Sentiment Analysis on Twitter*, etc.

22. Nous avions encadré cet étudiant de master Sciences du Langage, spécialité TAL, lors de son stage de master de 1e année (février-juin 2012).

Pour constituer ce lexique, nous avions alors utilisé la deuxième version de Polymots. Comme décrit plus haut, cette dernière version, outre une description plus fine de quelques familles de mots en clusters sémantiques [GHN+11], contient des étiquettes grammaticales, ceci nous avait permis d'extraire les 3 785 adjectifs et de les annoter manuellement avec trois valeurs (positif, négatif, neutre). Cette liste initiale d'adjectifs avait été complétée avec une centaine d'adjectifs supplémentaires provenant d'un lexique de l'analyseur XIP [AMC02]. Pour chacun des 3 882 adjectifs annotés, nous avions étendu automatiquement sa polarité vers les mots de sa famille morphologique.

Nous avions développé une interface de consultation (figure 3.7) et mis en ligne nos données [23] pour des recherches via l'interface ou bien pour des téléchargements des données (format csv ou xml).

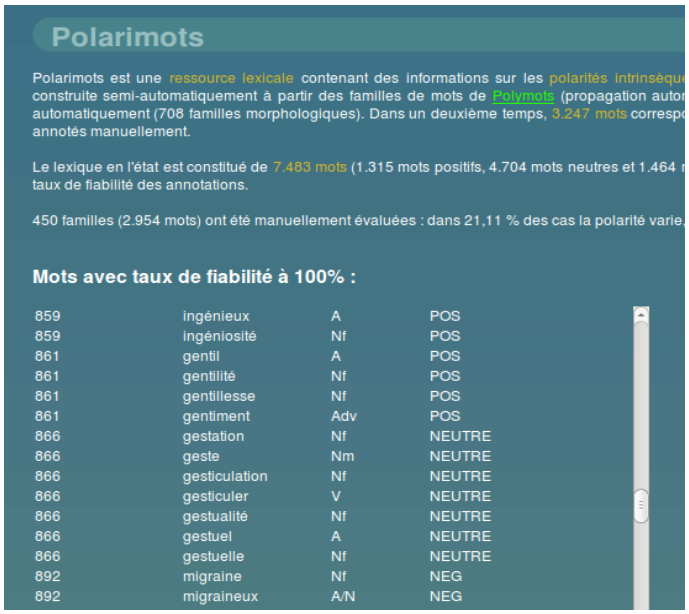


FIGURE 3.7 – Polarimots.

Pour ce qui est de l'évaluation de la ressource, nous avions procédé à une évaluation qualitative intrinsèque et à une évaluation extrinsèque.

Pour la première, une évaluation manuelle avait été faite pour 2 954 mots correspondant à 450 familles annotées automatiquement par propagation de la polarité à partir d'un adjectif (environ 70% du lexique). Les résultats de cette évaluation sont les suivants : sur 450 familles, 355 maintiennent la polarité de l'adjectif (78,89%) et 95 ne la maintiennent pas (21,11%). L'impact de la taille des familles morphologiques est un facteur essentiel dans

23. `http://polarimots.lif.univ-mrs.fr`

le maintien d'une polarité. Ainsi, plus celle-ci est réduite, plus la polarité reste identique, étant donné une cohésion sémantique plus forte.

Pour ce qui est de l'évaluation extrinsèque, nous avions intégré le lexique à un système d'extraction d'opinions afin de mesurer l'impact de cette intégration sur la capacité de ce système à classer correctement des revues en ligne en fonction de l'opinion globale de l'utilisateur [24]. Comme prévu, nous avions pu constater que l'intégration du lexique permettait d'améliorer les résultats du système, la configuration optimale étant obtenue lorsque les accords inter-annotateurs étaient de 100% [GB12].

3.3.4 Réseaux lexico-sémantiques (LexRom)

Les réseaux lexico-sémantiques sont des ressources qui mettent en évidence les relations qui s'établissent entre les unités lexicales, relations d'ordre principalement ontologique ou sémantique. La ressource la plus connue et la plus utilisée en TAL est WordNet [25] [Mil90]. D'autres exemples sont : WOLF [26] (wordnet libre du français) [SF06], JeuxDeMots [27] [Laf07], WordAssociation [28], BabelNet [29] [NP10], etc.

Pour le français, le *Dictionnaire de Combinatoire* (*DiCo*) encode les relations syntagmatiques et paradigmatiques. Il est conçu sur les principes de la Lexicologie et la Lexicographie Explicative et Combinatoire, dont la structure n'est pas textuelle mais de type réseau. Il en va de même pour le *Réseau Lexical du Français* [30] (*RLF*) [LPP11] et [Pol12]. Ce dernier est constitué en tant qu'un système lexical [Pol09], c'est-à-dire, un réseau constitué d'unités lexicales associées à des informations, celles-ci stockées dans différentes ressources interreliées (hiérarchie d'étiquettes sémantiques, base de fonctions lexicales, corpus d'exemples, etc.) [LP14]. Une autre ressource structurée comme un système lexical est FrameNet, constituée de trois composants (un lexique, une base de données de cadres de sous-catégorisation et un corpus de phrases annotées) [BFL98].

LexRom

La structuration du lexique des langues sous forme de familles, déjà décrite dans la littérature notamment par [Byb85], nous avait semblé une propriété intéressante à explorer d'un point de vue multilingue dans le cadre de quelques langues romanes (français, roumain, catalan et espagnol ; quelques

24. Travail réalisé par C. Brun.
25. `http://wordnet.princeton.edu/`
26. `http://alpage.inria.fr/~sagot/wolf.html`
27. `http://www.jeuxdemots.org`
28. `http://www.wordassociation.org/about/`
29. `http://babelnet.org/`
30. `http://www.atilf.fr/spip.php?article908`

expériences avaient aussi été menées pour l'italien, le portugais et le corse [31]). Nous nous étions ainsi proposée de créer un réseau lexical offrant des possibilités de recherche élargies (morphologie, sémantique) pour des langues typologiquement proches (figure 3.8).

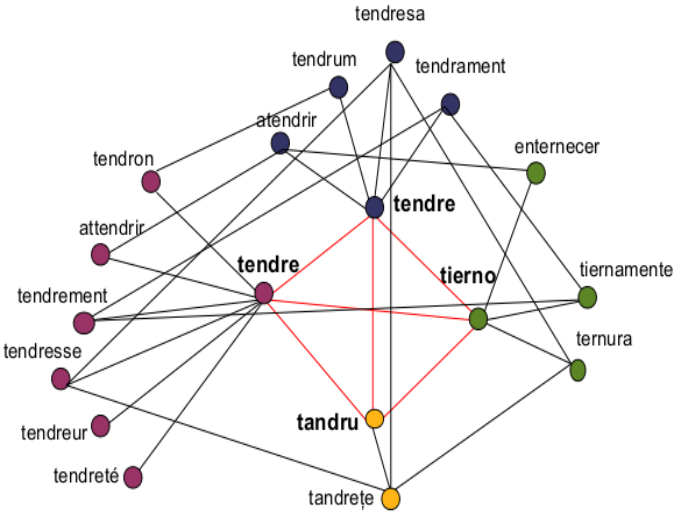


FIGURE 3.8 – LexRom : réseau pour la famille morphologique 'tendre'.

Exploiter les ressemblances formelles au niveau interlangue présente un intérêt aussi bien dans des applications humaines (études contrastives, apprentissage du vocabulaire ou aide à l'écriture dans une langue étrangère) qu'en traitement automatique des langues [32]. Les idées embryonnaires de ce travail avaient émergé en collaboration avec V. Rey (CREDO) et N. Hathout (CLLE-ERSS) [GRH10] dans une perspective d'aide à l'enseignement des langues.

Nous avions donc mené quelques expériences dans le but de créer une ressource fondée sur la notion de famille de mots interlangue [Gal11]. Notre objectif à terme était, ainsi, de proposer une ressource offrant des possibilités élargies pour des études contrastives, c'est-à-dire, des études portant sur des similarités ou des divergences dans les paradigmes (trous lexicaux), autant d'un point de vue morphologique (base commune ou non, affixes similaires ou non), sémantique (même sens ou non, évolution du sens vers un sens figuré ou non, etc.) que quantitatives (taille de la famille similaire ou divergences

31. Travail d'étudiants de master 1 TAL.

32. Il existe des travaux en didactique des langues romanes fondés sur l'intercompréhension, par exemple la plateforme Galanet `http://www.observatoireplurilinguisme.eu`. Le corpus C-ORAL-ROM, un corpus multilingue de parole spontanée pour les principales langues romanes composé d'environ 1 200 000 mots, est également un projet qui visait à faciliter l'étude comparative entre structures et vocabulaire des langues romanes.

importantes).

Par ailleurs, exploiter les ressemblances formelles au niveau interlangue présente un double intérêt, aussi bien dans des applications humaines (apprentissage du vocabulaire ou aide à l'écriture dans une langue étrangère) qu'en TAL (désambiguïsation sémantique, traduction, etc.).

Ce travail s'est déroulé principalement via l'encadrement de deux groupes d'étudiants de master Sciences du Langage, spécialité TAL (2010 et 2011), ainsi que le stage de recherche (été 2012) de V. Barbu-Mititelu (RACAI [33]). Ce dernier stage avait comme but de confronter les données de LexRom avec le Wordnet roumain [TBM14] afin (a) d'élargir la couverture de LexRom et (b) de l'enrichir avec des informations sémantiques issues du réseau. Il a donné lieu à une présentation au *Congrès International de Linguistique et Philologie Romanes* (CILPR) [GBM13]. Dans cette communication sans publication, nous faisions écho de l'absence de ressources pour des familles morphologiques dans des langues typologiquement proches et nous présentions les grandes lignes de notre ressource. En l'état actuel, le travail sur LexRom a été abandonné principalement par faute de financement et de temps.

3.4 Conclusion

Dans ce chapitre, nous avons présenté un ensemble de principes et méthodes en relation avec la construction, l'enrichissement et l'évaluation de ressources lexicales pour le domaine du traitement automatique des langues. Ayant participé à différents projets dans le domaine, nous avons décrit les ressources et les méthodes mises en œuvre pour les construire.

Dans le chapitre suivant, nous décrirons une problématique nouvelle (la complexité linguistique et, plus spécifiquement, lexicale) ainsi que des ressources adaptées visant des applications concrètes (enseignement du vocabulaire, apprentissage de la lecture) et des publics ciblés (enfants dyslexiques).

33. Institutul de Cercetari Pentru Inteligenta Artificiala, Academia Româna.

Chapitre 4

Vers des ressources intégrant la notion de complexité

« What is involved in knowing a word ? (...) What does a person who knows a word know ? A ready answer is that a person who knows a word must know its meaning(s). (...) The assumption that words have meanings presupposes an assumption (...) that words are doubly entered in lexical memory, once phonologically and once semantically, with associations between them. »

G. A. Miller. On knowing a word. *Annual Review of Psychology*, 50(1) : 1–19. 1999. [Mil99]

« Any claim about “complexity” is inherently about process, including an implicit description of the underlying cognitive machinery. By comparing different measures, one may better understand human language processing and similarly, understanding psycholinguistics may drive better measures. »

P. Juola. Assessing linguistic complexity. Dans M. Miestamo, K. Sinnemäki and F. Karlsson (éds.) *Language Complexity : Typology, contact, change*. 2008. [Juo08]

Dans ce chapitre, nous décrivons nos recherches autour de la notion de "complexité linguistique" et plus concrètement "lexicale". Ces notions ne peuvent pas être abordées sans tenir compte d'aspects liés à la compréhension et à la perception des unités lexicales par des publics très particuliers. Nos travaux autour de la construction de ressources graduées, avec des applications très concrètes, sont le cœur de nos recherches depuis deux ans.

4.1 Introduction

Depuis quelques années, la notion de "complexité" suscite de l'intérêt dans différents domaines de la linguistique (typologie, études créoles, ap-

prentissage des langues première et seconde, psycholinguistique, etc.). En font preuve un nombre significatif de manifestations scientifiques récentes : ateliers, conférences, ouvrages et publications diverses [Mie08].

Par ailleurs, le postulat selon lequel toutes les langues seraient équivalentes en complexité (ALEC, *All Languages are Equally Complex*), que l'on connaît également sous le nom d'"équi-complexité" (*linguistic equi-complexity dogma*), apparaît comme l'une des idées importantes de la linguistique moderne : la complexité d'une langue serait la somme de la complexité inhérente aux différents sous-domaines (phonologie, morphologie, syntaxe, sémantique, pragmatique). La simplicité d'un des sous-domaines serait compensée par la complexité de l'autre :

> « Objective measurement is difficult, but impressionistically it would seem that the total grammatical complexity of any language, counting both morphology and syntax, is about the same as that of any other. This is not surprising, since all languages have about equally complex jobs to do, and what is not done morphologically has to be done sytactically »(Charles Hocket 1958, cité par [SK12])

ALEC reste une notion intuitive et peu vérifiable d'un point de vue objectif : comment comparer la complexité entre deux langues ? Au sein d'une même langue, comment évaluer la complexité ? À notre sens, la question qui doit se poser en premier est la définition de la complexité même : qu'est-ce que la complexité linguistique ? Ce sujet, qui nous permet de nous immiscer dans des disciplines nouvelles, attire notre attention depuis un certain temps.

4.1.1 Quelle complexité ?

Pour répondre à cette question, il nous paraît intéressant de mettre en lumière quatre perspectives : linguistique (monolingue), typologique (multilingue), psycholinguistique (cognitive) et computationnelle (informatique).

Linguistique

M. Miestamo [Mie08] distingue, au sein d'un même système linguistique, entre la complexité globale et la complexité locale (celle d'un sous-domaine). Il considère que la complexité locale est plus facile à évaluer et donc à comparer entre langues (par exemple, inventaire de phonèmes, nombre de cas, nombre de nœuds dans un arbre syntaxique sémantiquement équivalent, etc.). De nombreux linguistes se sont ainsi consacrés à définir des critères pour mesurer la complexité au niveau des sous domaines de la linguistique, par exemple [SK12] :

- Phonologique : nombre de phonèmes, distinction tonales, restrictions phonotactiques, clusters consonantiques, etc.
- Morphologique : flexion, allomorphies, etc.
- Syntaxe : nombre de règles de la grammaire, ordre des mots, nombre d'emboîtements (*embeddings*) et de récursion autorisé par la grammaire, nombre de nœuds des arbres, etc.
- Sémantique : nombre d'unités homonymiques et/ou polysémiques, richesse lexicale, etc.
- Pragmatique : degré d'inférence pragmatique, etc.

Or, cette liste présente des problèmes. D'un point de vue méthodologique, et spécialement dans une perspective monolingue, est-ce possible de quantifier le nombre de règles de la grammaire d'une langue ou, de surcroît, le nombre d'unités ayant plusieurs sens ? De même, on peut mesurer la richesse lexicale d'un corpus (restreint à un nombre de mots) mais pas d'une langue dans sa globalité.

Enfin, quant à la récursion et les emboîtements, on ne peut pas parler d'une langue de façon générale car cela reste abstrait (« grammar is grammar, usage is usage », Frederick J. Newmeyer cité par [Deu13]). Les structures syntaxiques varient entre l'oral et l'écrit, et ce qui semble impossible à l'écrit reste parfaitement compréhensible à l'oral (emboîtements, enchâssements, concaténations d'éléments, listes à plusieurs niveaux, etc.) [1]. Par exemple, l'extrait suivant montre une concaténation d'éléments, c'est-à-dire un ensemble de syntagmes sans aucune relation grammaticale entre eux [Deu13] :

> « mais bon honnêtement / moi / bord de mer comme ça / Cannes / tout ça / c'est pas / c'est une ville de vieux quoi ».

Une telle construction non-canonique (par rapport à l'écrit) est, malgré sa structure, compréhensible par un locuteur du français.

Typologique

Dans une perspective multilingue (comparative, typologique), identifier la notion de complexité entraîne des problèmes méthodologiques : on ne peut comparer que ce qui est comparable, c'est-à-dire, des langues des mêmes familles ou des langues présentant les mêmes phénomènes. Par exemple, le nombre, bien que marqué morphologiquement ou grammaticalement et

1. Voir à ce propos les travaux du GARS (Groupe Aixois de Recherches en Syntaxe) et, notamment, de C. Blanche-Benveniste et J. Deulofeu.

ce avec des valeurs différentes selon les langues (singulier/pluriel, singulier/dual/pluriel, etc.) reste un élément comparable entre langues car 'universel'.

Cependant, ce n'est pas le cas, par exemple, pour l'expression d'expériences sensorielles différentes. En effet, seulement un quart des langues du monde possèdent ce type de marquage 'endopathique' qui permet au locuteur de donner une information précise de la source de l'information communiquée (expression du point de vue par rapport à des sensations physiques). Ainsi par exemple, en tibétain, l'expression "il y a quelqu'un" sera marquée avec des suffixes grammaticaux différents selon si (a) le locuteur exprime une connaissance personnelle de ce qu'il affirme, c'est-à-dire qu'il le *sait* (il n'a pas besoin de préciser l'accès à cette information), si (b) il *voit* (expérience sensorielle visuelle) ou bien si (c) il *perçoit* autrement (expérience non visuelle : ouïe, toucher, odorat) [2]. Il serait inadapté de dire que le tibétain est plus difficile que le français parce qu'il possède ce type de marquage (pour un locuteur natif de tibétain cela s'apprend). On ne peut pas comparer les langues tibétiques aux langues indo-européennes sur ce marquage car cela n'existe pas pour les dernières. Il en va de même pour les tons des langues d'Asie du sud-est. Se pose alors la question : quels critères utiliser pour comparer la complexité entre langues éloignées ? [3]

Psycholinguistique

Une autre façon de caractériser la notion de complexité est selon une perspective cognitive, c'est-à-dire, en tenant compte de paramètres psycholinguistiques liés au processus de compréhension. En réalité, cela faire écho ici à une question qui est fondamentale en sociolinguistique : « Who speaks which language to whom and when ? »[Dav03]. En effet, la notion de complexité ne peut pas être définie dans l'absolu mais en tenant compte des utilisateurs et des destinataires. Pour nous, ce n'est pas tant 'quelle langue' (standard) ni 'quand' (en synchronie) mais surtout 'pour qui'. P. Blache [Bla11] rajoute ainsi un troisième point par rapport à la complexité globale (du système linguistique) et locale (structurelle). Il s'agit de la 'difficulté'. Cette notion, liée à la subjectivité, a trait à des aspects de traitement : comment un individu particulier perçoit la langue.

Dans cette perspective, le profil du destinataire est essentiel : en fonction de son profil (enfant/adulte/âgé, instruit/peu d'instruction, normolecteur/faible-lecteur/dyslexique, entendant/sourd, etc.) certaines constructions seront plus difficiles que d'autres. Là aussi, une distinction entre pro-

2. Respectivement, 'mi-yod', 'mi-dug', 'mi-grag', 'mi' exprimant la notion de 'personne'. Plus de détails sur ce sujet dans : Tournadre, N. and LaPolla, R. J. (2014). Towards a new approach to evidentiality : Issues and directions for research. *Linguistics of the Tibeto-Burman Area*, 37(2).

3. Voir N. Tournadre [Tou14].

duction (parole/écriture) et analyse (écoute/lecture) entraînera des difficultés différentes. La problématique, dans cette perspective, peut être paraphrasée avec la question suivante : qu'est-ce qui rend une langue difficile à apprendre/comprendre pour un public donné ? Apprentissage (du vocabulaire, des structures grammaticales) et compréhension (du sens) sont les deux mots clefs dans ce domaine. Nous y reviendrons dans le cadre de notre projet de recherche SILK (cf. section 5.2.2).

Computationnelle

Enfin, on peut s'intéresser à la complexité en linguistique computationnelle et en informatique. Dans cette perspective, on s'intéresserait à identifier ce qui rend une production linguistique difficile à traiter computationnellement. Par exemple, P. Blache [Bla11] propose un modèle formel, la grammaire basée sur les contraintes, pour donner une vision précise et quantifiable de la complexité locale (incluant des facteurs liés à la difficulté). Il implémente différents paramètres cités dans la littérature dans le cadre d'un modèle formel basé sur l'analyse syntaxique : chacune de ces informations est représentée par une contrainte dans le formalisme. Au lieu de prendre en compte un arbre syntaxique, on tient ici compte des contraintes qui ont été satisfaites ou non, par exemple, précédence linéaire, dépendance, exclusion, obligation, etc. Il en découle que plus le nombre de contraintes non satisfaites est important (par rapport à la grammaire), plus la complexité est importante. La complexité computationnelle est ici liée à une théorie ou formalisme.

4.1.2 Parole pédagogique *vs* pathologique

Dans nos travaux en linguistique et en TAL, nous avons identifié deux publics distincts (a) pour lesquels la notion de complexité est pertinente et (b) pour lesquels il existe de réels besoins en ressources lexicales et en applications linguistiques (par exemple, d'aide à la lecture). Dans la suite de ce mémoire, nous utilisons le terme 'parole' dans le sens de Saussure, c'est-à-dire, en tant qu'*usage concret* de la langue fait par une classe d'individus identifiés.

Parole pédagogique

Dans le processus d'apprentissage pédagogique, la parole est intrinsèquement liée à l'acte d'apprentissage lui-même. Pour l'apprenant, la parole est le 'matériau' à acquérir (en plus des règles de combinaison etc.) [4], le moyen d'exprimer ses acquis. En même temps, pour l'enseignant, la parole est structurante et aide à l'acquisition de la langue par l'apprenant. Cette

4. Lexique, vocabulaire, mais aussi les règles de grammaire.

dichotomie est importante et se reflète dans la façon d'aborder les ressources utilisées pour l'apprentissage du vocabulaire, que ce soit en L1 comme en L2 (cf. 4.2).

Par 'parole pédagogique' nous faisons ainsi allusion à la langue utilisée dans un contexte d'apprentissage. La difficulté à acquérir et à maîtriser la parole dépend de différents facteurs cités dans la littérature [5], notamment, des facteurs liés à une source sociale (le niveau socioculturel et l'entourage de l'élève) et des facteurs liés à une source individuelle (le fait d'être atteint ou non d'une déficience intellectuelle, cognitive, auditive, etc.). Notre objectif n'est pas d'analyser ces facteurs (cela dépasse largement le cadre de nos recherches) mais plutôt d'apporter une contribution dans la caractérisation et la création des ressources utilisées dans un cadre pédagogique pour l'apprentissage du vocabulaire.

Dans ce contexte, nous nous sommes intéressée à des lexiques utilisés en milieu scolaire (français L1) et dans une perspective psycholinguistique, et à des lexiques pour l'apprentissage du français L2. L'étude des caractéristiques des unités lexicales dans ces outils nous a permis de proposer la notion de 'lexique gradué' (cf. 4.2.3).

Depuis 2012, et dans une perspective de lisibilité et de simplification automatique de textes, nous travaillons sur ce sujet en collaboration avec Thomas François et Cédrick Fairon au CENTAL (Centre de Traitement Automatique du Langage, à Louvain-La-Neuve).

Parole pathologique

La formulation 'parole pathologique' est un raccourci utilisé pour désigner la parole produite par des locuteurs atteints de dysfonctionnements de la voix et/ou de la parole [6]. Dans le cadre de notre travail, nous faisons allusion non seulement à la parole produite par des locuteurs mais aussi à la parole produite 'automatiquement' pour être lue par des individus atteints d'une difficulté de lecture. De nouveau, nous envisageons deux perspectives : celle des locuteurs eux-mêmes et celle des professionnels travaillant à la remédiation et à l'étude de certaines pathologies (orthophonistes, linguistes).

Depuis 2013, nous avons entamé deux collaborations, une avec le LPL (Laboratoire Parole et Langage), l'autre avec le LPC (Laboratoire de Psychologie Cognitive), favorisées par le labex BLRI (*Brain and Language Research Institute*) dont notre laboratoire (le LIF, Laboratoire d'Informatique Fondamentale) fait partie.

Nous avons ainsi travaillé sur des corpus de productions de locuteurs âgés atteints de la maladie de Parkinson, grâce à Serge Pinto (LPL). Notre

5. `http://www.cahiers-pedagogiques.com/Les-difficultes-ordinaires-d-apprentissage`

6. A. Ghio et al. (2006) Corpus de 'parole pathologique' : état d'avancement et enjeux méthodologiques. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence* (TIPA), 25, 109-126.

objectif initial était d'analyser la parole de patients en état 'off' (sans médicament qui aurait pu inhiber les effets de la maladie) et de confronter ces résultats, au niveau du lexique, à nos hypothèses en matière de lexique 'simple'.

Bien que de façon générale le public reconnaît la maladie de Parkinson par les troubles moteurs (tremblements, akinésie, rigidité), cette maladie entraîne des dysfonctionnements dans la parole. La dysarthrie [PGTV10] comprend une hypophonie (volume réduit de la voix), une parole monotone, des disfluences, et des difficultés au niveau de l'articulation de certains sons et de syllabes. La structure des phrases est aussi plus simple (plus courte) et le nombre d'unités lexicales 'pleines' (noms, verbes, adjectifs et adverbes) est plus important que le nombre d'unités grammaticales (pronoms, conjonctions, prépositions, etc.).

Lors de nos premiers travaux d'analyse de corpus de parole pathologique, nous avons émis l'hypothèse que les productions linguistiques de ce type de locuteurs parkinsoniens étaient simplifiées par rapport à des productions de locuteurs sains [GFF13]. Il s'agissait d'un premier travail exploratoire qui avait confirmé cette hypothèse. Malheureusement, le corpus utilisé est très restreint et ciblé à une tâche de description d'une image. Nous comptons poursuivre la collaboration avec Serge Pinto (et Elisabeth Godbert au LIF pour les aspects syntaxiques) pour l'analyse de corpus plus conséquents, comparés à des productions issues de personnes du même âge saines, dans le but de caractériser plus précisément la parole pathologique des parkinsoniens (cf. 5.2.4).

Une deuxième collaboration dans le cadre de la parole pathologique au sens large vient d'être initiée avec Johannes Ziegler (LPC). Elle donne lieu à un encadrement en cours de deux étudiantes en école d'orthophonie pour leur mémoire de fin d'études (soutenance prévue juin 2015). Cette collaboration est plus aboutie dans la mesure où elle a également donné lieu à la proposition d'un projet ANR [7], dont nous assurons la coordination scientifique, et au co-encadrement d'un doctorant (cf. 5.2.1). Globalement, le but de cette collaboration est la création d'un outil de simplification automatique de textes visant un public d'enfants dyslexiques ou faibles lecteurs. L'idée est d'analyser les résultats issus de quelques premières expériences avec des textes simplifiés à la main. Les variables qui influent d'une manière décisive dans la lenteur de la lecture seront prises en compte pour développer un système de simplification permettant aux enfants d'améliorer leurs performances de lecture (cf. 5.2.2).

7. SILK (*SImplification de textes pour faciliter leur Lisibilité et leur Compréhension*), projet finalement non sélectionné lors de la campagne 2014, soumis à nouveau -avec des modifications- en 2015. Le LPC à Marseille, le LIMSI à Orsay et le LiLPa à Strasbourg sont également partenaires, ainsi que le CENTAL en tant que partenaire non français.

4.1.3 Méthodes pour quantifier la complexité lexicale

La complexité lexicale n'est pas une notion qui puisse être définie dans l'absolu. En effet, un terme est perçu différemment en fonction du public qui y est confronté (apprenants de langue maternelle, apprenants de langue seconde, personnes avec une difficulté ou une pathologie liée au langage, etc.), d'où le terme de 'difficulté' (complexité subjective, [Bla11]). De même, s'appuyer sur le seul critère de la fréquence pour appréhender la complexité du lexique semble réducteur : bien que ce critère se soit avéré très efficace dans la littérature, cette variable ne peut seule expliquer l'ensemble des problèmes rencontrés par différentes catégories de lecteurs. La notion de 'complexité' est, ainsi, multidimensionnelle (vitesse d'accès au lexique mental, compréhension, mémorisation, prononciation, activation du sens, orthographe, etc.), difficilement saisissable à partir de critères uniquement statistiques et très liée aux caractéristiques du public envisagé.

En tenant compte de plusieurs ressources existantes, nous avons mené un travail en collaboration avec Thomas François et Cédrick Fairon (CENTAL) ainsi que Delphine Bernhard (LiLPa) dans le but d'identifier un ensemble de variables que nous avons intégrées dans un modèle cherchant à prédire le degré de complexité d'un mot [GFBF14]. Notre hypothèse est que seule la combinaison de plusieurs variables intralexicales fines, associées à des informations plus statistiques, pourra donner des indications précises sur le degré de complexité d'un mot.

La méthodologie que nous avons proposée pour évaluer la difficulté des mots du français se base essentiellement sur une approche par apprentissage automatisé. Il s'agit, à partir d'une liste de mots dont le niveau de difficulté est connu, d'apprendre les régularités existantes entre cette mesure de la difficulté et un large ensemble de variables issues de la littérature.

Critère de la fréquence

Analyser la complexité lexicale est une tâche qui, depuis toujours, a principalement intéressé les psycholinguistes et les pédagogues. En effet, de nombreux travaux sont décrits dans la littérature et se basent, par exemple, sur des tâches telles que la décision lexicale, la catégorisation sémantique, etc. pour explorer les propriétés du lexique. Ainsi, l'un des critères majeurs pour considérer qu'un mot est simple ou complexe est celui de la fréquence : de nombreux travaux démontrent la corrélation étroite entre la haute fréquence d'un terme et le fait que celui-ci soit perçu comme plus 'simple' [HS51, Mon91].

La fréquence est, d'ailleurs, le critère que plusieurs auteurs avaient utilisé dans la première moitié du 20e siècle pour construire les premières ressources de lexique 'simplifié', par exemple la liste de Thorndike [Tho21], le *Teachers' Book of Words*, qui reprend les 20 000 mots les plus courants de la langue

anglaise assortis de leur fréquence d'usage, ou encore le *Français fondamental* de Goughenheim et collaborateurs [Gou58] qui comprend 1 500 mots usuels pour l'apprentissage du français, aussi bien en tant que langue étrangère que maternelle. La liste de Thorndike reste une référence dans le domaine de la lisibilité (avant l'apparition des listes obtenues par traitement informatisé). Elle s'avère un instrument de mesure objectif de la difficulté lexicale des textes et ce malgré quelques faiblesses, comme la mauvaise estimation des fréquences des mots appelés *disponibles* (mots avec fréquence variée selon les corpus mais usuels et utiles [8]).

Autres critères avancés dans la littérature

D'autres critères avancés dans la littérature pour identifier des mots 'simples' concernent plutôt la familiarité d'un terme [Ger84] ou encore son âge d'acquisition [BLV00]. La familiarité lexicale a été utilisée pour la constitution d'une liste de mots simples par [Dal31]. Dans l'expérience menée par Dale et ses collègues, la mesure de familiarité a été définie comme suit : dans une liste de 10 000 mots, n'avaient été retenus que les termes connus par au moins 80% des élèves de quatrième primaire (CM1), ce qui avait réduit la liste à 3 000 mots.

Le nombre de voisins orthographiques (nombre d'unités de même longueur ne se différenciant que par une seule lettre) a aussi été envisagé par [CDJB77] comme une mesure discriminante de la difficulté d'accès au lexique mental, même si les résultats dans des tâches de décision lexicale semblent varier selon les langues. Enfin, la longueur (en nombre de syllabes et/ou caractères) apparaît aussi comme un facteur déterminant dans la façon de percevoir les unités lexicales (au niveau de la lecture), en particulier parce qu'un mot plus long augmente la probabilité de fixer la fovéa (zone de la rétine où la vision des détails est la plus précise) sur un point de position non optimal, ce qui augmente le temps de lecture [VOM90]. Plus récemment, [SB97] démontrent que le nombre de morphèmes et la taille de la famille morphologique jouent un rôle dans la décision lexicale visuelle (reconnaissance de mots parmi une série de mots et non-mots). [Lau97], pour sa part, identifie une série de facteurs linguistiques influençant l'acquisition du lexique, parmi lesquels : la familiarité des phonèmes, la régularité dans la prononciation, la cohérence graphème-phonème, la transparence morphologique ou la polysémie. Potentiellement, ces facteurs contribuent tous à la façon dont les mots sont perçus.

Répercussions

Les répercussions de tous ces travaux sont d'abord théoriques, aidant, par exemple, à comprendre l'organisation du lexique mental et comment il

8. Par exemple "fourchette", "coude", etc.

se distribue dans les différentes zones du cerveau. D'un point de vue plus pratique, certaines de ces études ont cependant débouché sur la construction de listes utilisées pour l'enseignement des langues. Plus récemment, la question de l'évaluation de la difficulté lexicale a fait l'objet d'un intérêt grandissant dans le domaine du traitement automatique des langues et, en particulier, en simplification automatique de textes. Dans ce domaine, le but reste d'identifier des termes et des structures difficiles à comprendre par un public cible et de proposer des outils de simplification automatisée de ces contenus. Bien que la plupart des travaux en simplification de textes se focalisent sur des aspects syntaxiques (par exemple [CDS96]), certains auteurs ont mis en œuvre des systèmes qui visent le traitement du lexique. Dans ce cas, différents aspects doivent être pris en compte : (i) la détection des mots ou termes complexes à remplacer, (ii) l'identification de substituts et (iii) l'adéquation au contexte. Ces trois aspects ne sont pas toujours pris en compte de manière conjointe. Sous sa forme la plus simple, la substitution lexicale se fait en fonction de la fréquence des synonymes extraits d'une ressource comme WordNet, sans prise en compte du contexte [CMC+98]. Récemment, des travaux ont fait appel à des corpus comparables comme Wikipedia et sa version simplifiée pour l'anglais (*Simple English Wikipedia*) pour acquérir des ressources utiles pour la simplification lexicale : ainsi, [BBE11] proposent une mesure de la complexité d'un mot en fonction de sa fréquence dans les deux versions de Wikipedia et de sa longueur. D'une manière générale, les critères utilisés pour sélectionner le meilleur substitut restent relativement simples. Pour la tâche de simplification lexicale organisée lors de la campagne SemEval 2012 [SJM12], la *baseline* correspondant à une simple mesure de fréquence dans un gros corpus n'a été battue que par un seul système. Ce résultat rend compte de la difficulté de la tâche : même si les travaux en psycholinguistique ont mis en évidence des facteurs complexes, leur intégration dans des systèmes automatisés n'est pas encore résolue.

Critères intralexicaux et statistiques

Nous travaux récents se situent dans cette direction : notre objectif à terme étant de créer un système de simplification automatique, nous sous sommes intéressée aux aspects liés au lexique. Ainsi, en tenant compte de plusieurs ressources existantes, nous avons identifié un ensemble de variables psycholinguistiques et intralexicales que nous avons intégrées dans un modèle cherchant à prédire le degré de complexité d'un mot [GFF13, GFBF14]. Notre hypothèse est que seule la combinaison de plusieurs variables intralexicales fines, associées à des informations statistiques, pourra donner des indications précises sur le degré de complexité d'un mot. La liste suivante présente la liste de variables obtenues à partir de données de différentes

ressources [9].

Critères orthographiques

1. *Nombre de lettres* : nombre de caractères alphabétiques dans un mot ;

2. *Nombre de phonèmes* : pour calculer le nombre de phonèmes dans un mot, un système mixte a été mis en place. Pour les mots présents dans *Lexique 3*, nous avons simplement récupéré l'information issue de cette ressource. Pour les mots absents de *Lexique 3*, nous avons généré leur représentation phonétique au vol via *eSpeak* [10] ;

3. *Nombre de syllabes* : comme pour le nombre de phonèmes, le nombre de syllabes d'un mot a été récupéré directement dans *Lexique 3*, quand l'information était disponible, sinon il a été calculé automatiquement en deux étapes. Tout d'abord, la forme phonétique a été générée (comme au point précédent), avant d'y appliquer l'outil de syllabification de [Pal99] ;

4. *Voisinage orthographique* : les informations concernant le nombre ou la fréquence des voisins orthographiques proviennent également de *Lexique 3* et nous les avons déclinées en 3 variables : (4a) nombre de voisins, (4b) fréquence cumulée de tous les voisins, (4c) nombre des voisins les plus fréquents ;

5. *Cohérence phonème-graphie* : le nombre de phonèmes et de lettres dans un mot ont été comparés sur la base de la classification suivante : 0 pour l'absence de différence (c'est-à-dire, une transparence parfaite), par exemple *abruti* [abRyti] ; 1 pour une différence de 1 ou 2 caractères, par exemple *abriter* [abRite] ; 2 pour une différence supérieure à 2 caractères, par exemple dans *lentement* [l@t-m@] [11] ;

6. *Patrons orthographiques* : 5 variables ont été définies autour de la présence de graphèmes complexes dans les mots, à savoir (6a) des voyelles orales (par ex. 'au' [o]), (6b) des voyelles nasales (par ex ; 'in' [ɛ̃]), (6c) des doubles consonnes (par ex. 'pp'), (6d) des doubles voyelles (par ex. 'éé'), (6e) ou encore des digrammes (par ex. 'ch' [ʃ]) ;

7. *Structure syllabique* : trois niveaux de complexité pour les structures syllabiques présentes dans les mots ont été définies sur la base des fré-

9. Pour plus de détails voir [GFBF14]. Les calculs statistiques (corrélations, SVM, etc.) ont été réalisés par T. François, les variables morphologiques ont été obtenues par D. Bernhard (apprentissage non supervisé).

10. http ://espeak.sourceforge.net

11. La transcription est celle de *Lexique 3* qui utilise l'alphabet SAMPA (*Speech Assessment Methods Phonetic Alphabet*).

quences de ces structures dans le corpus de parole « simple » Parkinson : (7a) les structures les plus fréquentes [12] (CYV, V, CVC, CV), (7b) les structures relativement fréquentes (CCVC, VCC, VC, YV, CVY, CYVC, CVCC, CCV), (7c) et les structures peu fréquentes (combinaisons de plusieurs consonnes, par exemple CCCVC) ;

Critères morphologiques

8. *Nombre de morphèmes* : nombre total de préfixes, suffixes et de bases dans le mot ;

9. *Fréquence minimale des affixes (préfixes et suffixes)* : nombre de mots différents (types) dans lesquels apparaît le préfixe / suffixe le moins fréquent ;

10. *Fréquence moyenne des affixes (préfixes et suffixes)* : moyenne des fréquences absolues des préfixes / suffixes ;

11. *Préfixation* : attestation ou non de la présence de préfixes ;

12. *Suffixation* : attestation ou non de la présence de suffixes ;

13. *Composition* : attestation ou non de la présence de deux bases ou plus ;

14. *Taille de la famille morphologique* : tous les mots qui contiennent la même base sont regroupés dans la même famille. Pour les mots composés qui appartiennent à plusieurs familles, seule la taille de la famille la plus petite a été prise en compte ;

Critères sémantiques

15. *Polysémie selon JeuxdeMots* : booléen indiquant si le mot est polysémique ou non ;

16. *Polysémie selon BabelNet* : nombre de synsets répertoriés dans BabelNet ;

Critères statistiques

17. *Fréquence dans Lexique 3* : logarithme des fréquences extraites de *Lexique 3* (calculées à partir d'un corpus de sous-titres de films).

12. La notation utilisée est la suivante : C pour consonne, V pour voyelle, Y pour les glides [j], [w] et [ɥ].

18. *Présence/absence dans la liste de Gougenheim* : pour chaque mot, un booléen indique s'il appartient ou non à la liste du *Français Fondamental* dans sa version longue (qui comprend 8 875 lemmes). Comme il est bien connu en lisibilité que la taille de la liste de mots simples utilisée comme variable influe sur la capacité de discrimination de celle-ci, nous avons expérimenté avec diverses tailles de liste, par tranche de 1 000 mots, de 1 000 à 8 875 mots.

Dans les travaux à venir nous comptons investiguer le poids d'autres variables dans la perception de la complexité lexicale, par exemple : le voisinage phonologique et le caractère abstrait ou concret des unités lexicales.

Prédicteurs de la complexité

Après avoir identifié un ensemble de variables nous avons effectué différentes expériences dans le but de déterminer quels sont, parmi nos prédicteurs, ceux qui apportent le plus d'information sur la difficulté des mots [GFBF14]. Ainsi, les meilleures variables sont celles basées sur la fréquence (17) et sur la présence dans une liste de mots simples (18) de taille moyenne (entre 4 000 et 5 000 mots). Une seconde constatation d'intérêt est l'efficacité et la robustesse des variables classiques telles que le nombre de lettres (01) et de syllabes (03). Moins utilisé dans la littérature, le nombre de phonèmes dans un mot (02) apparaît tout aussi efficace. Enfin, les informations relatives au statut polysémique des mots apportent également de l'information qui semble utile pour expliquer la difficulté du lexique, que ce soit via une information binaire sur le statut polysémique des mots issue de JeuxDeMots (15) ou via le nombre de synsets repris dans BabelNet (16).

Dans [GFBF14], nous détaillons les résultats d'une analyse correlationnelle que nous avons menée pour estimer l'efficacité de chacune des variables dans l'estimation de la difficulté lexicale. Les retombées de cette expérience nous permettront de construire un modèle pour graduer les mots que nous comptons utiliser, entre autres, dans un système de simplification de textes (cf. 5.2).

4.2 Bases de données générales, vocabulaires fondamentaux et simplifiés

Les travaux que nous menons actuellement sur le lexique, et que nous venons de décrire dans la section précédente, ne pourraient avoir lieu sans l'utilisation de ressources qui décrivent explicitement la structure intralexicale des mots, leur présence dans des corpus, etc. (celle-ci est également capitale pour permettre l'accès lexical). Ce type de ressources, qui dépasse le cadre 'classique' de la lexicographie, est très pertinent dans des disciplines comme

l'enseignement du vocabulaire des langues et de la psycholinguistique. Le TAL ne peut qu'en bénéficier également.

4.2.1 Des listes aux bases structurées

La création de listes pour l'apprentissage du vocabulaire est extrêmement ancienne (les premières ressources lexicales étaient des listes [Bou03], créées à des fins pédagogiques). Cette pratique perdure pendant très longtemps car il existe un réel besoin dans l'enseignement du vocabulaire mais aussi dans la préservation d'une langue, etc.

C'est au XXe siècle que l'on commence à introduire des informations quantitatives associées aux mots du fait de l'utilisation des corpus (cf. 1.4.2). Par exemple, *The Teacher's Word Book* [Tho21] est l'une des ressources pionnières. Il s'agit d'une liste de 10 000 mots ordonnés selon leur fréquence d'occurrence dans un corpus de 4 500 000 mots extraits de livres pour enfants, journaux, etc. Thorndike posa les bases de l'utilisation de l'information statistique à des fins pédagogiques avec l'idée que plus un mot est fréquent plus il est 'adéquat' pour les jeunes lecteurs.

La liste de Thorndike fut utilisée également dans de nombreuses formules de lisibilité. De même, elle inspira des travaux similaires pour d'autres langues comme *The Spanish Word Book* [Buc27] et *The French Word Book* [Van32], le *Français Fondamental* [Gou58], etc. Elle fut aussi élargie à 30 000 mots quelques années plus tard [TL44].

a to acacia

	G	T	L	J	S		G	T	L	J	S
a	AA	M	M	M	M	aborigines	1	7	8	5	12
Aaron	2	28	6	5	14	abortive	1	11	1	3	15
aback	2	10	15	11	12	abound	12	90	32	39	59
abandon	38	119	150	130	285	about	AA	M	M	M	M
abandoned (adj.)	3	11	14	12	27	above	AA	M	941	M*	?
abandonment	3	10	16	3	39	Abraham	11	115	47	26	22
abase	1	14	2	0	5	Abram	1	7	0	0	14
abash	3	16	14	24	13	abreast	4	16	17	23	20
abate	7	57	20	20	33	abridge	2	18	0	6	13
abatement	1	10	5	2	4	abridgment	1	11	1	0	9
abbé	3	7	18	0	44	abroad	48	200	198	200*	268
abbess	1	14	3	9	1	abrogate	1	10	0	2	9
abbey	11	57	19	51	83	abrupt	6	27*	43	20	26

FIGURE 4.1 – Extrait de la liste de 30 000 mots de Thorndike et Lorge.

Dans l'extrait, les colonnes correspondent à différentes mesures statistiques [13] : la première colonne indique le nombre d'occurrences par million

13. http://catalog.hathitrust.org/Record/000987642

de mots (AA indique plus de cent par million), les quatre colonnes suivantes donnent le nombre d'occurrences par rapport à d'autres corpus de Thorndike (1931) (T), de Lorge (L), etc.

Ce type de liste n'est pas exempt de critiques. Parmi celles citées dans la littérature, on peut signaler, par exemple, le problème de la représentativité des corpus (pour obtenir une estimation robuste de la fréquence d'apparition d'un mot il faut compter avec de gros corpus hétérogènes, ce qui n'était surement pas le cas lors des premiers travaux avant l'arrivée de l'informatique). Par ailleurs, il existe un certain type de mots ('mots disponibles', [Mic53]) qui sont très fréquents dans les langues mais qui ne sont pas forcément présents dans les corpus (mots de la vie quotidienne). Enfin, les mots utilisés dans les corpus pour l'apprentissage du vocabulaire (que ce soit en L1 ou en L2) sont choisis par les professionnels de façon pragmatique et plus ou moins aléatoire. Ainsi, qu'est-ce qui justifie qu'une leçon parle de princes et de princesses, qui sont dans l'imaginaire enfantin [14], ou d'un rat vert qui aime les olives (d'où sa couleur) [15] ? Dans les deux cas, le vocabulaire utilisé sera significativement différent, les fréquences obtenues sur corpus le seront aussi. On retrouve, ainsi, 'artificiellement', les mots qu'on veut faire apprendre aux enfants, et non pas les mots que les enfants apprennent 'naturellement'.

Avec le développement de la linguistique de corpus et du TAL les approches quantitatives du lexique ont connu d'énormes avancées (par exemple [CGHH91], cf. section 1.4.2). Sur la base du *Brown Corpus*, H. Kucera et W. N. Francis [KF67] proposèrent une liste de fréquences pour l'anglais américain. Ils remarquèrent que la fréquence des distributions dépend du type de document utilisé ainsi que du sujet abordé (si un mot est utilisé fréquemment dans un sujet, sa fréquence sera alors surestimée). Pour pallier ce problème, ils proposèrent de nouvelles mesures statistiques comme la dispersion, l'index de fréquence standard, etc. (cf. 3.2.3), qui sont devenues des mesures généralement utilisées. Dans des travaux plus récents, et grâce à l'application d'outils de TAL (analyseurs morphologiques et syntaxiques), les ressources incluent, également, des informations morphologiques et collocationnelles, par exemple *Les Voisins de Le Monde* [16].

Outre la sophistication dans le calcul de propriétés statistiques, les listes lexicales ont évolué, au niveau de la structuration des contenus associés aux entrées, en bases de données relationnelles. Les données associées aux entrées sont décomposées en tables (matrices ou relations) pouvant contenir des informations très explicites (non textuelles) tout en évitant toute redondance des informations. L'accès aux contenus se fait généralement par des opérations d'algèbre relationnelle (intersection, jointure, produit cartésien, etc.), ce qui facilite le traitement de grands volumes de données.

14. *La princesse au petit pois*, un conte de H. C. Andersen lu en CP `http://www.lagedeclasse.fr/la-princesse-au-petit-pois-cp-a42372845`

15. *Ratus et ses amis*. Méthode de lecture CP. (1994) J. et J. Guion. Editions Hatier.

16. `http://redac.univ-tlse2.fr/voisinsdelemonde/`

Du point de vue de leur contenu, ces ressources structurées en matrices sont des lexiques de langue générale, des lexiques scolaires et des lexiques gradués.

4.2.2 Lexiques de langue générale

Pour le français, la première base de données lexicales informatisée mise à disposition des psycholinguistes fut Brulex [CMR90], qui regroupait les 35 746 entrées lexicales du *Petit Robert* et leurs fréquences selon le *Trésor de la Langue Française* [Imb71]. Ces fréquences étaient estimées sur un corpus de textes littéraires datant de 1919 à 1964 et comprenant 26 millions de mots (l'ensemble de corpus est devenu la base FRANTEXT [17]).

Un deuxième exemple de ce type de ressource est MHATLex [PdC00], une base payante qui contient 81 000 lemmes et 854 000 formes fléchies ainsi leurs représentations phonologiques, des informations morphosyntaxiques et fréquentielles.

Lexique 3 [18] [NPFM01] reste, néanmoins, la plus connue en TAL. Pour les premières versions, les fréquences de Lexique 1 2 furent constituées à partir d'une sélection de textes publiés après 1950 du corpus de textes Frantext. Le Lexique 2 comprenait ainsi 130 000 formes fléchies ainsi que leur fréquence. Si Lexique 2 apportait un certain nombre d'innovations comparativement aux bases de données existantes, il subsistait encore quelques limitations (par exemple, les mots composés n'étaient pas présents dans la base, les homographes n'étaient pas pris en compte, etc.). Le *Lexique 3* pallie à ces limitations. Il s'agit d'une ressource librement accessible contenant un grand nombre d'informations intralexicales et statistiques (extraites à partir de corpus de textes et de sous-titres de films) pour 142 728 mots correspondant à 47 342 lemmes.

4.2.3 Lexiques scolaires (PolyMarmots)

Pour l'enseignement du vocabulaire, la question de définir un sous-ensemble "noyau" ou "élémentaire" a été abordée dans la littérature selon deux grandes approches : celle des vocabulaires de base "logiques" et celle des vocabulaires simplifiés obtenus grâce à des études de fréquence [Sus06].

Le premier type, appelé aussi dans la littérature "approche rationnelle", concerne des listes où les unités lexicales ont été choisies par des critères humains. On retrouve des exemples pour l'enseignement de l'anglais (ou du français) langue étrangère, plus tard adaptés aussi à l'apprentissage de ces langues en tant que langues maternelles. Parmi les exemples les plus significatifs de ces vocabulaires minimaux pour l'anglais on retrouve le *British American Scientific International Commercial* (BASIC) de C. K. Ogden et

17. `http://www.frantext.fr/`
18. `http://www.lexique.org`

I. R. Richard en 1928, ou le *Defining Vocabulary* de M. West en 1929. Cette approche a subi de nombreuses critiques sur l'apauvrissement du lexique. Cependant, force est de constater que l'utilisation de telles ressources se révela « un moyen efficace de diffusion de la langue anglaise »pendant les années 1930 et 1940 [Sus06].

Une deuxième approche très vite généralisée pour créer des lexiques "élémentaires" est celle qui utilise des ressources construites sur la base de statistiques lexicales (distributions de fréquence des mots dans des corpus), par exemple, comme nous l'avons mentionné plus haut, *The Teacher's Word Book* de [Tho21] ou, pour le français, le *Français Élémentaire* (1954) et *Français Fondamental* de Gougenheim [Gou58] destiné à l'enseignement du FLE.

Novlex [LC01] est une base lexicale destinée aux apprenants de français L1 [19]. Elle a été créée d'après l'analyse de différents corpus scolaires de niveau CE2 (8-9 ans) totalisant 417 000 occurrences (sans noms propres, ni prénoms, ni noms de ville, ni onomatopées). L'objectif poursuivi par les auteurs au moment de la construction de cette ressource était de mettre à la disposition des chercheurs psycholinguistes et didacticiens un outil susceptible d'estimer la fréquence lexicale pour les enfants.

Nous avions développé un projet de ressource lexicale dans cette lignée en 2010, PolyMarmots, à partir de deux ressources existantes :

a) Polymots (version initiale), 20 000 mots regroupés en 2 004 familles morpho-phonologiques ;

b) Novlex (accessible en ligne), 20 600 mots correspondant à 9 300 lemmes.

Les applications visées étaient fondamentalement pédagogiques : apprentissage du vocabulaire et de l'orthographe du français en milieu scolaire.

La ressource proposée contient proposée 4 221 mots (intersection de Polymots et Novlex) ainsi que leurs descriptions associées. Les unités lexicales sont également visibles en contexte : nous avions collecté (via le Web) un corpus de 92 802 mots, constitué de contes de Grimm et de fables de La Fontaine. Les informations extraites de Polymots comprennent la décomposition morphématique ainsi que des alternances vocaliques dans la famille. Le tableau suivant montre les informations pour le mot 'volontaire' :

base	forme	préfixe(s)	suffixe(s)	alternance phon.
voul-/vol-	volontaire	-	ont-aire	[u/o]

Table 4.1 – Exemples d'entrées de Polymots.

Les informations de Novlex pour ce même mot sont les suivantes :

19. http ://www2.mshs.univ-poitiers.fr/novlex/

fréq	nblet	phon.	nbphon	syll	nbsyll	struct	gram
5 712	10	[volôtèR]	7	vo-lô-tèR	3	CV-CV-CVC	N

TABLE 4.2 – Exemples d'entrées de Novlex.

Ainsi, grâce à Novlex nous avions pu acquérir des informations concernant : la fréquence d'apparition [20], le nombre de lettres, la représentation phonétique, le nombre de phonèmes, la décomposition syllabique, le nombre de syllabes, la structure vocalique (voyelle/consonne/semi-consonne), la catégorie morpho-syntaxique (étiquette grammaticale) et le genre (masculin, féminin, pas de différence de genre).

Dans Polymarmots, et en tenant compte que la ressource devait être destinée à des enfants, la fréquence avait été transformée en une étiquette explicite en fonction de différents seuils suivants : 'rare', 'peu fréquent', 'fréquent', 'très fréquent'.

Le développement informatique de Polymarmots avait été fait par des étudiants de L3 Informatique de Luminy (2e semestre 2010) que L. Tichit et nous-même avions encadrés (figure 4.2) [21].

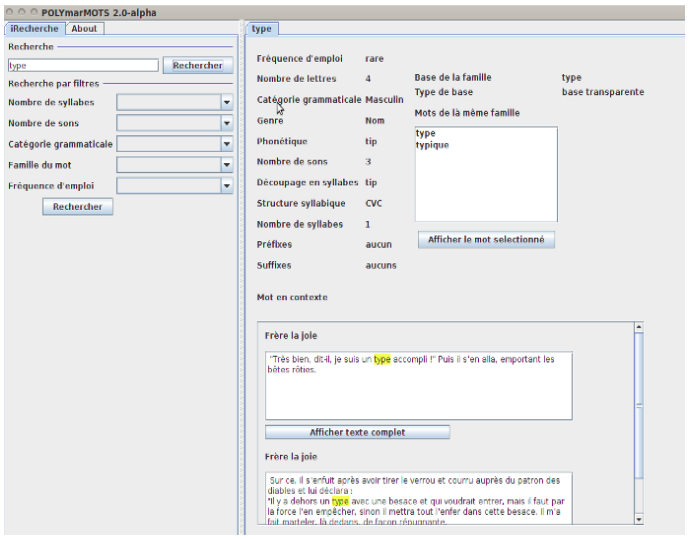


FIGURE 4.2 – Polymarmots.

Pour un mot donné, l'utilisateur peut accéder à toutes les informations le

20. La fréquence dans Novlex est calculée par rapport à un corpus de 417 000 mots provenant de manuels scolaires, [LC01].

21. Le projet en est resté à un stade trop embryonnaire pour pouvoir proposer une publication.

concernant : des informations phonologiques, statistiques et grammaticales de Novlex ; des informations morpho-phonologiques provenant de Polymots. Grâce à cette dernière ressource, pour un mot donné, l'utilisateur peut identifier un ensemble de mots appartenant à la même famille.

La recherche dans Polymarmots peut également se faire en sélectionnant des filtres, par exemple, des mots avec 3 syllabes, mots avec 3 syllabes et rares, etc. Le résultat de la requête est une nouvelle fenêtre avec tous les mots satisfaisant les critères de recherche. En sélectionnant un de ces mots, l'utilisateur accède aux différentes informations le concernant. Enfin, tout mot peut être visualisé en contexte (contes de Grimm et fables de La Fontaine).

4.2.4 Lexiques gradués (FLELex, ReSyF)

Si la notion de 'lexique scolaire' ou 'de langue générale' est repandue, à notre connaissance, la notion de 'lexique gradué' est inexistante dans la littérature. Deux seules ressources (avant les nôtres) correspondent à cette idée où les mots sont dotés d'un niveau lié à leur difficulté d'acquisition par un public donné. Il s'agit de *The Educator's Word Frequency Guide* [ZIMD95] et ses listes dérivées pour l'anglais, et Manulex [22] [LSCC04] pour le français.

Educator's Word Frequency Guide

The Educator's Word Frequency Guide [ZIMD95] est une ressource pour l'anglais avec des mots classés selon leur fréquence dans 60 527 textes pour apprenants.

Il existe une deuxième ressource dérivée de celle-ci avec une couverture plus restreinte (380 mots) mais avec une organisation favorisant les liens sémantiques entre les mots. Ainsi, il y a un niveau initial (*Early Vocabulary Connections : First Words to Know and Decode*) pour les primo-lecteurs et un deuxième niveau (*Early Vocabulary Connections : Important Words to Know and Spell*[23]) pour les enfants qui consolident leurs compétences de lecture. C'est donc une ressource destinée à l'apprentissage de la lecture de l'anglais.

Ce qui rend cette ressource intéressante est, premièrement, le fait qu'il y ait un classement en deux niveaux. Plus intéressant encore, les mots sont organisés en champs sémantiques grossiers (à la manière d'un thésaurus -très simple-) : actions et expériences, vie courante, êtres vivants, etc. Enfin, les mots apparaissent par paires selon différentes relations sémantiques pour favoriser un meilleur apprentissage (rivière/pont, guerre/paix, coût/prix,

22. www.manulex.org

23. http://assets.soprislearning.com/newsletters/NVSR/images/EVC_Level_2_Instructor%27s_Manual.pdf

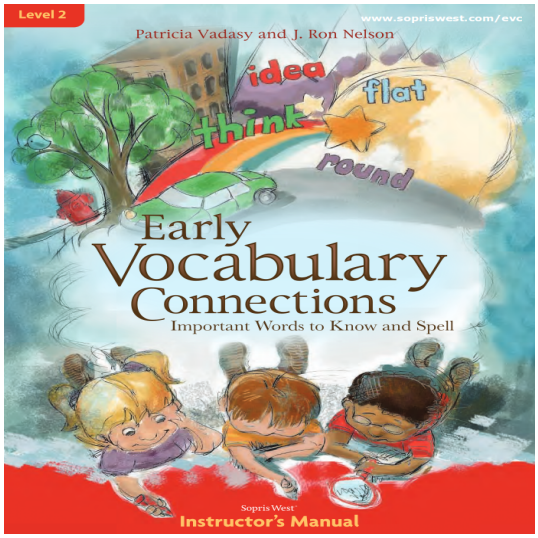


FIGURE 4.3 – Ressource graduée pour l'apprentissage de la lecture en anglais.

etc.). Il y a donc une structuration en niveaux de difficulté et en sémantique susceptible d'aider les enfants lors de l'apprentissage de la lecture.

Manulex

Manulex [24] est la première ressource de type 'graduée' créée pour le français. Les unités lexicales ont été classées en trois niveaux selon leur apparition dans 54 manuels scolaires de la première année de primaire, de la deuxième ou des trois suivantes. Ce choix se justifie en termes de volume d'acquisition de vocabulaire :

« le CP (6 ans) où se construit le lexique de l'enfant sur la base de la médiation phonologique, le CE1 (7 ans) où se construit le lexique orthographique par automatisation progressive de la reconnaissance du mot écrit et le cycle 3 (CE2-CM2, 8-11 ans) où se consolide et s'enrichit le stock lexical par exposition répétée à l'écrit. » [25].

La ressource, librement disponible, totalise 1,9 millions d'occurrences et 23 812 lemmes. Les fréquences associées à chaque entrée ne correspondent pas aux valeurs absolues observées dans les manuels, mais à des valeurs adaptées en fonction d'un indice de dispersion qui augmente l'importance des termes en fonction du nombre de documents dans lesquels ils sont apparus. La Figure 4.3 présente un exemple d'entrées issues de Manulex (N1 correspond au niveau CP, N2 au CE1, N3 à CE2-CM2).

24. `http://www.manulex.org`
25. `http://leadserv.u-bourgogne.fr/bases/manulex/manulexbase/indexFR.htm`

lemme	POS	Fréq. N1	Fréq N2	Fréq. N3
pomme	N	724	306	224
vieillard	N	-	13	68
patriarche	N	-	-	1

TABLE 4.3 – Exemples d'entrées de Manulex.

À la base, Manulex n'est pas une ressource explicitement graduée. Nous l'avons considérée ainsi pour nos expériences ([GFF13] et [GFBF14]) : nous avons considéré qu'un mot appartenait au premier niveau où il avait été vu (d'après l'exemple, 'pomme' serait un mot de niveau 1, 'vieillard' de niveau 2 et 'patriarche' de niveau 3.

ReSyF

Avec l'objectif de créer un lexique gradué, pour la simplification de textes et pour l'aide à l'apprentissage du vocabulaire, nous avons travaillé au développement d'un modèle permettant de classer les mots selon leur niveau de difficulté. Pour tester la validité de ce modèle, nous avons pris comme *baseline* les mots de Manulex selon l'attribution de niveaux telle que mentionnée plus haut. L'idée à moyen terme (travaux en cours) est de créer une ressource de synonymes gradués avec notre modèle.

Pour tester la validité d'un tel modèle, nous avons cherché, pour tous les mots de Manulex, une liste de synonymes, auxquels nous avons également attribué un niveau de Manulex. L'idée est de pouvoir attribuer un niveau de difficulté grâce à notre modèle, quel que soit le mot (appartenant ou non à Manulex).

Ce premier travail à donné lieu à une première version de ReSyF [GFF13]. À partir de la liste de 19 037 mots de Manulex (mots pleins) nous avons cherché leurs synonymes dans JeuxDeMots [Laf07] (un réseau sémantique construit de façon collaborative). Au moment de mener cette expérience (février 2013) nous avions obtenu 17 870 mots cibles de Manulex présents dans JdM, dont 12 687 avec des synonymes (le mot peut être présent dans le réseau mais avoir une relation de synonymie vide). Le tableau 4.4 montre la répartition en niveaux des mots de Manulex présents dans JdM.

Niveau 1	Niveau 2	Niveau 3
30.1%	21%	48.9%

TABLE 4.4 – Mots de Manulex dans JeuxDeMots

Le résultat est un lexique de synonymes où tous les mots sont gradués [26]. La figure 4.4 en montre quelques exemples :

armure(1) = protection(1), cuirasse(2), harnais(3)
piétiner(2) = marcher(1), fouler(3), piaffer(3), trépigner(3)
patriarche(3) = chef(1), père(1), vieillard(2)

FIGURE 4.4 – Exemple d'entrées dans ReSyF.

Nos travaux en cours visent à enrichir et à améliorer cette ressource : application du modèle pour graduer les mots, désambiguïsation sémantique, graduation de mots par rapport au groupe de synonymes (cf. 5.2.1). Les applications visées sont multiples, dans le domaine de la simplification automatique pour l'aide à la lecture, mais aussi pour l'aide à la rédaction, l'apprentissage du vocabulaire, etc. [GBFB15].

FLELex

Nous avons construit FLELex [27] [FGWF14] à partir d'un corpus de 28 manuels de français langue étrangère (FLE) et de 29 livres simplifiés également destinés à des lecteurs en FLE. Ces ouvrages étaient classés selon l'échelle de difficulté proposée par le cadre européen commun de référence pour les langues [Con01] ou CECR, qui définit six niveaux de maîtrise communicationnelle : A1 (niveau introductif ou de survie) ; A2 (niveau intermédiaire) ; B1 (niveau seuil) ; B2 (niveau avancé ou utilisateur indépendant) ; C1 (niveau autonome ou de compétence opérationnelle effective) et C2 (maîtrise). À ce jour (mai 2014), la ressource totalise 16 833 lemmes lexicaux et 1 038 lemmes grammaticaux, dont les fréquences ont été estimées sur 777 835 occurrences pour ces corpus. Comme pour Manulex, les valeurs rapportées ont été adaptées en fonction de leur plus ou moins grande dispersion dans les manuels du corpus.

26. Une version améliorée de ReSyf sera mise en ligne pendant l'été 2015 à l'adresse :`http://resyf.lif.univ-mrs.fr`

27. `http://cental.uclouvain.be/flelex/`

lemme	A1	A2	B1	B2	C1	C2
voiture	633.3	598.5	482.7	202.7	271.9	25.9
abandonner	35.5	62.3	104.8	79.8	73.6	28.5
justice	3.9	17.3	79.1	13.2	106.3	72.9
kilo	40.3	29.9	10.2	0.0	1.6	0.0
piétiner	0.0	0.39	0.0	0.53	15.7	0.0
logique	0.0	0.0	6.8	18.6	36.3	9.6
absurdité	0.0	0.0	0.34	4.55	3.29	67.36
en bas	34.9	28.5	13	32.8	1.6	0.0
en clair	0.0	0.0	0.0	0.0	8.2	19.5
de surcroît	0.0	0.0	0.0	0.0	15.67	0.0
donner rendez-vous	0.53	0.69	1.89	0.0	0.0	0.0
donner naissance	0.0	0.25	0.0	0.0	0.0	4.12

FIGURE 4.5 – Echantillon de données de FLELex.

D'après les exemples, on constate que certains mots concrets et quotidiens apparaissent dans les premiers niveaux, alors que d'autres (plus abstraits) n'apparaissent qu'à partir des niveaux intermédiaires. Le fait d'avoir extrait les mots à partir d'un corpus, et le fait d'avoir utilisé un traitement morphologique (segmenteur et étiqueteur), nous a permis de répérer des locutions et des expressions polylexicales (locutions et constructions à verbe support). À la différence de Manulex, FLELex intègre donc des unités lexicales variées.

4.3 Conclusion

Dans ce chapitre, nous nous sommes intéressée à la notion de complexité lexicale et à des ressources nouvelles pouvant intégrer cette notion. Il s'agit de travaux très récents que nous comptons développer par la suite. Les besoins pour des ressources de ce type nous semblent importants, autant dans un cadre pédagogique (enseignement du vocabulaire, apprentissage de la lecture, etc.) que dans un cadre pathologique (amélioration des compétences de lecture, meilleure accessibilité à des contenus). L'intégration de ces ressources dans des applications de TAL comme la simplification de textes reste, pour nous, un objectif majeur à ce jour, tel que nous le développons dans le chapitre suivant.

Deuxième partie

Travaux en cours et perspectives

Chapitre 5

Simplification Lexicale et Construction de Ressources

« Le vocabulaire ne s'enseigne pas comme une autre matière. L'apprentissage des mots a besoin d'une "base affective". »

J. Picoche. Dialogue autour de l'enseignement du vocabulaire *Études de Linguistique Appliquée* 116, pp. 421-434, 1999.

« Although the causes of dyslexia are still debated, all researchers agree that the main challenge is to find ways that allow a child with dyslexia to read more words in less time, because reading more is undisputedly the most efficient intervention for dyslexia. »

Zorzi, M., Barbiero, C., Facoetti, A., Lonciari, L., Carrozzi, M., Montico, M., Bravar, L., George, F., Pech-Georgel, C. et Ziegler, J. C. Extra-large letter spacing improves reading in dyslexia. *Proceedings of the National Academy of Sciences of the United States of Americ (PNAS)*, 2012. [ZBF$^+$12]

Dans ce chapitre, nous décrivons les grandes lignes du travail mené actuellement et les perspectives envisagées à court-moyen terme. Notre intérêt fait écho à des besoins réels dans trois domaines d'application où le lexique et la description lexicale sont au cœur des préoccupations : (a) la complexité linguistique, (b) la lisibilité et la simplification lexicale et (c) la construction de ressources pour les langues peu dotées. Le premier aspect, plus théorique, est abordé sous le prisme de la typologie linguistique et surtout sous celui de l'apprentissage des langues (5.1). Les deuxième et troisième aspects, plus appliqués, sont abordés dans la perspective de débouchés concrets autour de la notion de simplification de textes pour des publics particuliers (5.2), et autour de la construction de ressources lexicales pour des langues peu dotées (5.3).

5.1 Complexité - Difficulté

Nous avons évoqué dans le chapitre 4 la fluctuation terminologique entre 'complexité' et 'difficulté' linguistique. Dans la littérature (par exemple, [Mie08], [Bla11]), on distingue une complexité 'absolue' (intrinsèque à la langue, globale et locale selon s'il s'agit du système ou des structures concrètes), d'une complexité plus 'empirique' (relative à un individu ou groupe d'individus). C'est cette dernière que nous abordons sous le terme de 'difficulté'.

Outre la fluctuation terminologique, ces deux notions nous interpellent. C'est pourquoi nous comptons y réfléchir d'un point de vue théorique depuis deux perspectives : une perspective linguistique (aspects typologiques et multilingues) et une perspective d'apprentissage de langues (aspects cognitifs), que nous abordons dans les sections suivantes.

5.1.1 Aspects typologiques et multilingues

D'un point de vue objectif, la complexité est une propriété observable au sein d'un système, de façon intrinsèque (dans le système lui même) mais aussi par comparaison avec d'autres systèmes. Dans le cas des langues, N. Tournadre distingue la complexité *structurale* (structure du lexique, du système phonologique, morphologique ou syntaxique), *catégorielle* (nombre de catégories lexicales et grammaticales) ou *lexicale* (taille du vocabulaire, typologie lexicale) [Tou14].

Observer et comparer différentes langues en terme de leur complexité, dans une optique de modélisation informatique, est un projet qui nous intéresse et qui nous permettra de nouer des collaborations avec les membres de l'équipe *Langues en Contact et Typologie* d'Aix Marseille Université.

5.1.2 Aspects cognitifs

La relation entre complexité linguistique et complexité d'apprentissage d'une langue (par un type de public donné) est loin d'être triviale. La difficulté d'apprentissage peut être mesurée en termes de 'coûts', c'est-à-dire, de temps et de ressources mis en œuvre pour aboutir à l'acquisition d'un objectif (dans le cas linguistique : l'acquisition d'une liste de vocabulaire, d'une structure grammaticale, du décodage lors de la lecture, etc.). Nous mentionnons la notion d'apprentissage de façon générale, mais bien évidemment des différences existent si on s'en tient à l'apprentissage oral (système phonologique) ou bien écrit (système d'écriture) ou encore aux deux (écart graphème-phonème). Il en va de même s'il s'agit de l'apprentissage d'une langue étrangère (L2) ou d'une langue maternelle [1] (L1).

1. Le cas de l'apprentissage simultané de plusieurs langues est aussi intéressant, qu'il s'agisse dans le cadre d'un bilinguisme familial, d'une diglossie ou bien d'un apprentissage de différentes langues d'une même famille linguistique par effet de similarité.

Pour ce qui est de l'acquisition d'une L2, la difficulté d'apprentissage dépend très fortement de l'écart génétique entre la langue source et la langue cible (par écart génétique, nous faisons allusion à la parenté au niveau des familles de langues). De même, l'écart typologique jouera un rôle important : dans l'apprentissage d'une langue de famille différente, il sera plus facile d'assimiler certaines structures si elles existent déjà dans la langue d'origine (par exemple les clitiques qui marquent les cas en allemand, en roumain, en basque) [Tou14].

Dans le cas de l'acquisition d'une L1, l'apprentissage dépend de nombreux facteurs, dont des facteurs cognitifs [2]. Ainsi, au niveau du lexique, l'apprentissage du vocabulaire est fortement lié aux concepts auxquels l'enfant est confronté et aux sens (significations) qu'il est stimulé à communiquer [Lit06]. Cela fait écho aux mots appelés 'disponibles' par [GMRS64] (mots plutôt de type concret, avec une fréquence variée selon les corpus mais usuels et utiles, par exemple 'fourchette', 'coude', etc.). Il y aurait donc un paramètre très important dans la perception des mots en tant que mots 'moins difficiles' dès lors qu'ils font partie de cette classe de mots 'disponibles'. Caractériser ces mots, afin de les intégrer dans nos ressources et outils pour la simplification lexicale, reste une perspective intéressante.

Enfin, pour aller plus loin dans la prise en compte de différents aspects cognitifs liés à la perception et à la compréhension des mots, il est nécessaire de mener des expériences auprès des publics ciblés. Nous souhaitons nous investir dans ces aspects par le biais de collaborations avec le Laboratoire de Psychologie Cognitive d'Aix Marseille Université, dans le cadre du BLRI, dont le projet d'aide à la lecture pour enfants dyslexiques (5.2.2) est un premier objectif précis. Une deuxième collaboration très récente avec un membre de ce même laboratoire, F. Vitu [3], concerne la construction d'un corpus annoté avec des données oculomotrices issues de la lecture (*The BLRI Book-Reading Corpus : Des données oculomotrices uniques et fondamentales pour une approche écologique et pluridisciplinaire de la lecture*). Les informations issues de ce corpus, combinées à des informations plus formelles sur le lexique et la structure des phrases, devraient permettre l'obtention de résultats plus précis dans des applications de simplification automatique de textes.

L'ensemble de ces thématiques trouvent également écho dans l'équipe *Langues, usages, cognition et apprentissages* du Laboratoire Parole et Langage d'Aix Marseille Université. Nous comptons par conséquent nous en rapprocher.

2. D'autres facteurs, notamment socio-culturels, dépassent le cadre de notre recherche.

3. Projet validé par le BLRI fin décembre 2014, passations de lecture mises en place au premier semestre 2015.

5.2 Lisibilité et simplification

La lisibilité des textes est un domaine de recherche longuement exploré dans la littérature pour l'apprentissage des langues (L1 et L2) [4]. La notion de lisibilité reste étroitement liée à la compréhension :

> « Par lisibilité, nous désignons le degré de difficulté éprouvé par un lecteur essayant de comprendre un texte »[Hen75]

Depuis le début du 20e siècle, on a cherché à mesurer le degré de lisibilité d'un texte moyennant des formules qui fournissent une prédiction objective, par exemple celles de R. Flesch [Fle48] dans le monde anglo-saxon, ou celle de G. Henry [Hen75] pour le français. Ces formules s'appuient principalement sur des paramètres formels quantifiables : par exemple, la longueur des mots ou des phrases, le nombre de syllabes par mot, etc. Avec l'informatique, les formules de lisibilité s'automatisent mais ce n'est qu'à partir des années 2000 que des techniques issues du TAL (principalement des modèles statistiques) permettent d'analyser plus finement les textes : les études en lisibilité computationnelle foisonnent depuis [Fra11].

En lien étroit avec la lisibilité, un deuxième domaine émerge en parallèle, celui de la simplification automatique de textes. Il s'agit de transformer automatiquement un texte en un équivalent plus compréhensible pour un groupe d'individus donné, partageant une même difficulté de lecture (cf. 4.1.2), par exemple : apprenants de L1 ou L2 [BBLF14], personnes avec peu d'instruction [CMC+98] ou [IFT+03], dyslexiques [RBYS13], aphasiques [WJU+09], etc. Dans une certaine mesure, les problématiques rencontrées au niveau de la simplification rejoignent celles d'autres applications classiques en TAL comme la traduction ou le résumé automatique : transformation d'un texte tout en gardant le même contenu (ou le contenu essentiel).

Globalement, l'objectif principal de la simplification est ainsi de rendre un texte plus accessible à des publics en difficulté de lecture. Parmi ces publics, les enfants dyslexiques ou faibles lecteurs représentent une cible très importante (d'après des rapports nationaux (MJENR 2003) ou internationaux (PISA 2009), 20% à 30% des élèves français ont des difficultés pour comprendre les textes écrits, 5 à 10% sont des enfants dyslexiques).

Nous poursuivons actuellement des travaux en simplification lexicale, dans l'optique de fournir des outils d'aide à la lecture à des enfants dyslexiques, fruit d'une collaboration avec plusieurs équipes (le LPC à Marseille, le CENTAL à Louvain-la-Neuve, le LiLPa à Strasbourg et le LIMSI à Orsay). Ces projets sont décrits dans les sections 5.2.1 et 5.2.2.

Un travail sur des productions orales de patients parkinsoniens est également décrit en section 5.2.3. Enfin, une étude sur l'écrit des sourds est brièvement évoquée en section 5.2.4. Pour ces derniers, il s'agit d'ouvertures

4. Voir l'excellente thèse de T. François à ce sujet [Fra11]

possibles au niveau théorique (complexité lexicale) et applicatif (outils d'aide à la lecture).

5.2.1 Méthodes de TAL pour la simplification

En collaboration avec les équipes mentionnées plus haut, nous comptons développer un outil de simplification de textes pour le français qui visera, dans un premier temps, les enfants dyslexiques. Il s'agira ainsi d'évaluer le degré de difficulté d'un texte et d'en proposer une version plus compréhensible.

Notre participation à ce projet, outre la coordination scientifique, a lieu au niveau de la simplification lexicale. Simplifier le lexique peut être mis en œuvre par l'ajout d'informations (reformulations, explications, définitions, etc.) ou par la réduction de la complexité linguistique. La plupart des systèmes de simplification existants se basent essentiellement sur cette dernière approche qui consiste à transformer le texte initial en un équivalent plus simple [Sha14]. Il s'agit alors de (1) repérer les formes qui posent problème (mots 'complexes'), (2) identifier une liste de substituts potentiels, (3) parmi ces derniers, choisir ceux qui véhiculent le même sens, (4) choisir le 'plus simple'.

La tâche est loin d'être triviale et se heurte à plusieurs écueils, notamment, les problèmes de substitution des formes désambiguïsées et le problème des transformations entraînant des modifications syntaxiques (figure 5.1). À titre de comparaison :

(a) *Un froid vif s'était installé* depuis une semaine. *À la veille des* départs en vacances de Noël, *chacun s'en réjouissait* : les *trajets* jusqu'aux stations de ski seraient agréables, la neige serait belle et douce. Mais ce matin *le froid a faibli* et *une pluie fine s'est mise à tomber*. Quels seront les effets de *ces conditions météorologiques* ?

(b) Il faisait très froid depuis une semaine. Avant les départs en vacances de Noël, tout le monde était content. Les voyages jusqu'aux stations de ski seraient agréables. La neige serait belle et douce. Mais ce matin il fait moins froid. Il pleut. Quels seront les effets de ce temps ?

FIGURE 5.1 – Simplification de textes en français.

Dans cet extrait (respectivement, texte original [5] et manuellement simplifié), on peut identifier, en rouge, les transformations lexicales directes ('trajet' > 'voyage', les unités lexicales peuvent aussi correspondre à des verbes pronominaux 'se réjouir de' > 'être content', des locutions 'avant' > 'à la veille de', des collocations 'temps' > 'conditions météorologiques',

5. `http://mon-cartable-du-net.perso.sfr.fr/lectures.html`

etc.) ; en bleu, les transformations lexicales entraînant des transformations syntaxiques ('froid' est sujet dans la première phrase mais objet dans la version simplifiée ; le nom 'pluie' devient le verbe 'pleuvoir', etc.). Dans ce deuxième type de transformation, il s'agira de paraphraser. Nous devrons avoir recours à des méthodes plus sophistiquées incluant des lexiques syntaxiques par exemple.

À ce jour (juillet 2014), nous avons proposé une ressource de synonymes gradués en trois niveaux selon Manulex (ReSyf, cf. 4.4). L'idée, à court terme, est de l'enrichir avec des mots automatiquement gradués. Pour ce faire, nous comptons améliorer notre modèle qui prend en compte des informations intra-lexicales et statistiques (4.1.3). Outre l'ajout de nouvelles variables (voisins phonologiques, imageabilité [6]), la piste que nous pensons plus prometteuse consiste à graduer les synonymes au sein de l'ensemble de synonymes et non pas sur l'ensemble de la liste de mots de la ressource (fonction de tri).

Nous rendrons disponible une version enrichie de ReSyf (plus large couverture) dans laquelle les mots seront associés à une liste de synonymes automatiquement gradués (été 2015). Dans un premier temps, nous comptons utiliser cette ressource dans notre système de simplification pour l'aide à la lecture des enfants dyslexiques, mais d'autres applications seront également envisagées, que ce soit en parole pathologique comme pédagogique [GBFB15].

5.2.2 Dyslexie et difficultés de lecture chez des enfants

La dyslexie est un trouble du développement qui entraîne des difficultés de lecture et de compréhension des textes écrits. Ces difficultés se traduisent par un décodage lent et laborieux : le temps de lecture augmente de façon exponentielle avec la longueur des mots [ZPMW+03]. Par conséquent, un enfant dyslexique lit en un an ce qu'un normo-lecteur lit en deux jours [CS98]. Il s'agit d'un cercle vicieux parce que lire avec fluidité implique beaucoup d'entraînement et d'exposition aux textes écrits [ZPZ14]. Par ailleurs, le manque de fluidité et la lenteur à lire nuisent à la compréhension des textes (les enfants sont plus concentrés au décodage des mots et des phrases qu'à la compréhension/interprétation du contenu) : cela a des conséquences au niveau du succès scolaire de l'enfant car la lecture est fondamentale pour tout ce que l'enfant doit apprendre pendant son parcours scolaire.

Ces dernières années, un nombre important de technologies ont été créées pour venir en aide aux personnes en difficulté [7]. Concrètement, pour le public

6. Terme des sciences cognitives correspondant à la propriété qu'aurait un concept à être plus ou moins difficile à visualiser ou imaginer (`http://www.cognitiveatlas.org/concept/imageability`).

7. Pour la recherche d'information [SBB10], pour la lecture chez les aphasiques [CMC+98], pour la lecture chez les adultes avec déficiences intellectuelles [HFE09], etc.

dyslexique, les systèmes proposés intègrent :

- des technologies de la parole : lecture «à voix haute», par exemple le système Gothit [8] (figure 5.2) ;
- des aides visuelles : paramétrage et/ou mise en couleur des polices [RSBY14], augmentation de l'espacement des lettres d'un mot et des mots d'un texte [ZBF+12], etc.

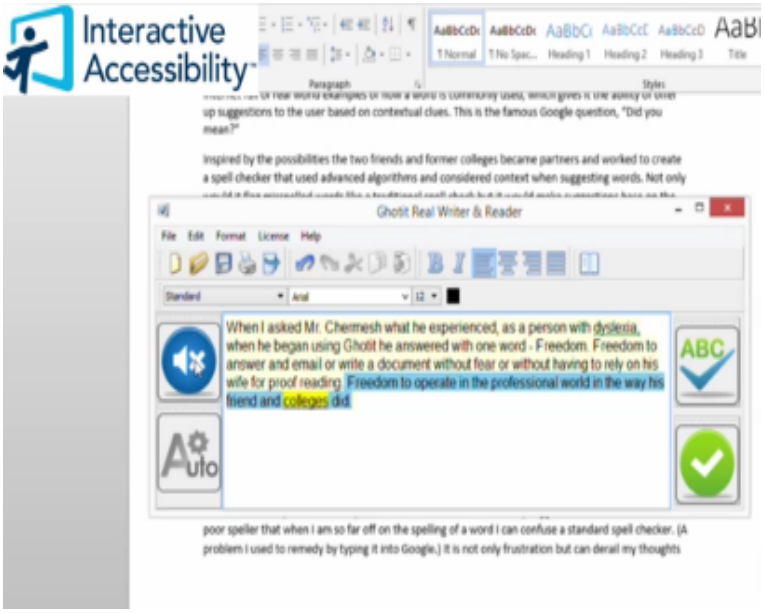


FIGURE 5.2 – Gothit : outil d'aide à la lecture et à l'écriture en anglais pour public dyslexique et dysgraphique.

Peu de systèmes proposent des transformations au niveau du contenu et encore moins des transformations adaptées aux besoins des dyslexiques. Par exemple, [SB08] ont proposé une méthode d'adaptation des requêtes dans un système de recherche d'informations basée sur la phonétique (l'inconsistance graphème-phonème est cruciale dans des langues comme le français où la différence entre langue orale et langue écrite est importante [Bru10]). Rello et ses collaborateurs ont, quant à eux, mené plusieurs expériences pour l'espagnol qui montrent que la longueur des mots, ainsi que leur fréquence, sont décisives pour la lisibilité et la compréhension des textes chez des dyslexiques [RBYS13]. Simplifier les mots des textes, tout en conservant le sens, s'avère ainsi une approche prometteuse à laquelle nous nous proposons de contribuer.

8. *Ghotit Real Writer Reader software* (`http://www.ghotit.com/`).

Dans le cadre de nos travaux futurs, nous nous proposons de concevoir des outils et des ressources venant en aide aux enfants et aux adultes avec des besoins spécifiques. Notre finalité première est ainsi de définir un modèle permettant d'évaluer la lisibilité des contenus textuels, que ce soit au niveau lexical, syntaxique ou discursif. Le deuxième objectif que nous nous fixons est celui de la génération automatique de contenus simplifiés et adaptés aux différents types de lecteurs avec difficultés de lecture et/ou compréhension. Enfin, le troisième objectif est l'évaluation de notre système auprès d'enfants ayant des difficultés de lecture et/compréhension. L'ouverture à des nouvelles technologies pour l'éducation (e-learning), et plus particulièrement orientées vers certains publics en difficulté, est un défi sociétal majeur. Nous nous proposons ainsi de relever un tel défi, d'autant plus qu'il n'existe pas à l'heure actuelle d'applications réelles pour les locuteurs du français. En revanche, il en existe, respectivement, pour des publics de dyslexiques, d'autistes et d'illettrés, dans les langues suivantes :

- pour l'espagnol (Simplext [9] [SGMA+11] et DysWebxia [Rel14]),
- pour l'anglais (First [10], *Flexible Interactive Reading Support Tool* [SEOM12])
- pour le portugais (PorSimples [11], *Simplificação Textual do Português para Inclusão e Acessibilidade Digital* [AG10])

Une partie du travail envisagé dans le domaine de la simplification afin de faciliter la lecture aux enfants dyslexiques se fera via l'encadrement d'étudiants.

Tests de lecture avec des textes simplifiés

A. Brunel et M. Combès (2014-2015), co-encadrement de mémoire de fin d'études d'orthophonie avec J. Ziegler (LPC-CNRS).

Améliorer la lecture chez des enfants dyslexiques ou en difficulté peut être abordé sous deux angles. D'une part, il peut y avoir un travail sur des mécanismes au niveau de la conscience phonologique (remédiation faite, notamment, par des orthophonistes). D'autre part, il peut y avoir un travail sur les matériaux de lecture (les rendre plus accessibles). Nous faisons l'hypothèse que la simplification des contenus aura des répercussions importantes dans la lecture : compréhension égale, meilleure vitesse, moins de fautes de lecture (meilleur décodage).

Pour tester cette hypothèse, nous allons mettre en place des tests auprès d'un public de 7 à 9 ans (CE1, CE2, CM1), en collaboration avec le Centre

9. `http://www.simplext.es/`
10. `http://www.first-asd.eu/`
11. `http://www.nilc.icmc.usp.br/porsimples/simplifica/`

de Référence pour les Troubles des Apprentissages CERTA, à Marseille (St. Salvador). Des expériences de test de lecture auprès d'enfants dyslexiques seront mises en place fin 2014 (20 enfants dyslexiques, 20 faibles lecteurs et 20 normo-lecteurs liront 30 textes simplifiés manuellement). Il s'agira de collecter les temps de lecture et les résultats des tests de compréhension (réponses à des questions sur le contenu des textes).

Au niveau des corpus, il sera question de textes provenant de textes scolaires librement accessibles sur le Web (comme celui de la figure 5.1). D'autres types de textes pour enfants seront également collectés (textes encyclopédiques [12], fables, contes, etc.). Ces corpus seront annotés et utilisés comme données source pour les simplifications (manuelles dans le cadre de ce mémoire, puis automatiques dans le cadre de notre collaboration avec les équipes de TAL mentionnées plus haut). Ils serviront également à l'évaluation du système de simplification.

À l'issue de ce travail, nous disposerons d'un premier corpus de textes parallèles (originaux-manuellement simplifiés) ainsi qu'une caractérisation de la typologie de phénomènes entraînant des difficultés pour le public d'enfants dyslexiques [13].

Désambiguïsation sémantique dans le cadre de la simplification lexicale pour enfants dyslexiques

M. B. Billami (2014-2017), co-encadrement de thèse avec J. Ziegler (LPC-CNRS).

En plus des problèmes de repérage et de délimitation des unités lexicales à simplifier mentionnés plus haut (cf. 5.2.1), la simplification lexicale se heurte au problème de la désambiguïsation sémantique (indispensable à d'autres applications en TAL comme la traduction automatique, etc.). À cet égard, il s'agit de sélectionner automatiquement le sens le plus approprié d'un mot en contexte.

La table suivante (5.3) est un exemple de l'entrée 'glacial' dans la ressource ReSyf. On peut constater une liste de synonymes gradués, sans aucune désambiguïsation sémantique (les couleurs ont été ajoutées pour cet exemple). Il est clair qu'il serait souhaitable de distinguer le sens propre ('basse température' en bleu) du sens figuré qualifiant une personne ou un comportement (en gris), ou un lieu, ambiance, etc. (en rouge) [14].

12. `http://fr.wikimini.org`

13. Les résultats de ce travail seront publiés dans un article de revue en psycholinguistique, courant 2016.

14. Les synonymes peuvent être à leur tour polysémiques, par exemple 'froid' (température et comportement).

glacial(n2)	impassible(n3), imperturbable(n3), froid(n1), rigoureux(n2), inhospitalier(n3), sec(n1), insensible(n3), glacé(n1), polaire(n2)

FIGURE 5.3 – Exemple d'entrée désambiguïsée dans ReSyf.

Le travail de thèse de M. Billami apportera une contribution dans l'étape la plus cruciale consistant à lever les ambiguïtés, puis simplifier (étapes 3 et 4). Cette tâche devra prendre en compte des besoins linguistiques (le remplacement lexical pourra entraîner des transformations syntaxiques) et surtout des besoins liés au public ciblé (identifier, puis formaliser, ce qui rend un mot difficile à comprendre pour un enfant dyslexique ou faible lecteur). Il apportera également une contribution à l'enrichissement de ressources lexicales intégrant des informations sur le degré de difficulté d'un mot-sens.

5.2.3 Parkinson et production de parole

Nous avons entamé, en 2013, une collaboration avec S. Pinto (neuro-linguiste au LPL) portant sur l'analyse de la production de la parole de personnes atteintes de Parkinson. Nous avons eu à notre disposition vingt fichiers audio correspondant à des enregistrements de patients dans une tâche de description d'une image de la vie 'quotidienne' (figure 5.4).

FIGURE 5.4 – Image décrite par des patients atteints de Parkinson.

Il s'agissait de patients en état 'off', c'est-à-dire sans l'effet de médicaments qui auraient pu amenuiser les conséquences de leur maladie (13 hommes et 7 femmes). Nous avons transcrit les fichiers audio afin de produire des fichiers en format XML (figure 5.5, exemple de fichier transcrit, produit par l'une des patientes enregistrées).

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE Trans SYSTEM "trans-14.dtd">
<Trans scribe="NG" audio_filename="A02388_off" version="2" version_date="130315">
<Episode>
<Section type="report" startTime="0" endTime="45.220">
<Turn startTime="0" endTime="45.220">
<Sync time="0"/>
cette dame elle lave la vaisselle
<Sync time="4.312"/>
l'eau coule
<Sync time="7.284"/>
sais pas le robinet est bouché peut être
<Sync time="11.065"/>
le leurs enfants rangent les rangent la vaisselle peut être ou bien prend quelque
prend tout ce un gâteau
<Sync time="20.807"/>
ils tombent si tôt (rires)
(...)
</Turn>
</Section>
</Episode>
</Trans>
```

FIGURE 5.5 – Exemple de fichier transcrit 'Parkinson'.

Nous avons extrait les phrases et les avons étiquetées avec TreeTagger (un total de 2 271 occurrences, avec une moyenne de 114 occurrences par patient, un minimum de 42 occurrences et un maximum de 233). Par la suite, nous avons analysé les 1 106 occurrences lexicales obtenues (372 lemmes noms, verbes, adjectifs et adverbes) [15] :

- longueur des occurrences (moyenne de 3,66 phonèmes)
- nombre de syllabes (moyenne de 1,61)
- nombre de morphèmes (moyenne de 1,06)
- structure syllabique des occurrences (majorité CV 0,46%, puis CVC 0,18%, puis V 0,16%)
- richesse du vocabulaire (nombre d'hapax sur le nombre de lemmes, 0,58%)
- correspondance phonème-graphème pour les lemmes (8% équivalence totale entre le nombre de phonèmes et le nombre de graphèmes, 70,16% différence $<$ à 2, 21,77% différence $>$ à 2 phonèmes)
- comparaisons des lemmes avec les niveaux de Manulex (94,7% des lemmes du corpus correspondent à des mots du niveau 1 de Manulex)

L'ensemble des résultats obtenus à ce jour est très préliminaire et est à prendre avec précaution. Premièrement, nous sommes consciente du biais de la tâche (les patients décrivent une image en noir et blanc, pour laquelle il n'est pas possible d'aller très loin au niveau de la quantité et de la variété du vocabulaire). De même, le nombre de sujets étudiés (vingt patients) est insuffisant pour obtenir des résultats généralisables. Enfin, analyser des patients uniquement en état 'off' ne permet pas de comparer leurs productions de parole dans d'autres situations (mêmes patients sous l'effet de médicaments) ou avec celles d'autres individus (personnes saines dans la même tranche d'âge, personnes saines d'une tranche d'âge inférieure (jeunes)).

De ce fait, nous continuons notre collaboration avec S. Pinto (projet de co-encadrement de post-doctorat BLRI pour 2015). L'idée sera d'obtenir un échantillon de corpus beaucoup plus large et varié afin d'étudier et caractériser les structures linguistiques. Ces expériences seront menées avec trois catégories de population (parkinsoniens sans médicament, parkinsoniens avec médicament, public sain du même âge) et des nouvelles images.

15. Un travail d'analyse syntaxique sur ce corpus a été réalisé en collaboration avec E. Godbert du LIF.

5.2.4 Écrit des sourds

Collaborations avec M. Hamm et L. Boutora du Laboratoire Parole et Langage.

Une des pistes possibles dans les applications de simplification pour l'aide à la lecture concerne les personnes ayant la langue des signes française (LSF) comme langue maternelle. Selon le rapport Gillot [16] [Gil98], 80% de la population sourde adulte entretient un rapport difficile au français écrit. Bien que ce public de sourds et malentendants bénéficie de rééducations et de prises en charge spécifiques, les difficultés de compréhension sont assez persistantes et freinent trop souvent leur intégration sociale et professionnelle. Par ailleurs, la loi de 2005 « pour l'égalité des droits et des chances, la participation et la citoyenneté des personnes handicapées »prône l'accessibilité de toutes les personnes à tous les documents (support papier et Web). La simplification automatique de textes reste donc une piste intéressante pour toutes ces populations, facilitant l'accessibilité des textes au cours et au-delà de la période d'apprentissage en milieu scolaire.

À ce jour, la caractérisation des écrits des personnes sourdes ou malentendantes ([Ham12] ou [VBD12]) n'a pas donné lieu à des travaux en lisibilité ni en simplification alors que des enjeux existent. Il nous tient à cœur d'envisager dans nos perspectives futures un travail de recherche dans cette direction.

5.3 Construction de ressources pour des langues peu dotées

La construction de ressources linguistiques pour des langues peu dotées répond à des besoins clairement identifiés, autant d'un point de vue humain que de traitement automatique des langues. Cependant, malgré les efforts réalisés, le nombre de langues peu dotées est encore énorme. Les enjeux sont d'autant plus importants que doter ces langues, ne serait-ce que de ressources informatisées de base (lexiques, corpus), garantit leur visibilité et même leur survie. Depuis quelques années, les travaux en traitement automatique des langues peu dotées se sont généralisés [17], autant pour des langues ou variétés régionales dans des pays où le poids de la langue officielle est très important, que pour des langues de pays en voie de développement.

16. `http://www.ladocumentationfrancaise.fr/var/storage/rapports-publics/984001595/0000.pdf`

17. En font preuve plusieurs ateliers dans des conférences comme TALN, COLING, etc.. Par exemple en 2014, respectivement, TALAf (Traitement automatique des langues africaines) `http://jibiki.univ-savoie.fr/~mangeot/TALAf/2014/` et VarDial (Applying NLP Tools to Similar Languages, Varieties and Dialects) `http://corporavm.uni-koeln.de/vardial/index.html`.

Dans ce domaine, « il y a lieu de prendre en compte les contraintes socio-économiques s'exerçant sur la population des locuteurs : les ressources économiques sont limitées, les ressources humaines qualifiées sont rares, les recherches sont sporadiques et isolées, les résultats confidentiels et parcellaires. Il est donc nécessaire de définir des méthodologies économes en coût d'achat de logiciels et en temps de travail qualifié visant à produire des résultats pérennes, partagés et faciles à enrichir. La constitution de ressources linguistiques de manière générale (...) devrait donc respecter un certain nombre de principes : utilisation d'outils en source ouverte, définition et utilisation de standards (ISO, Unicode), transfert de connaissances entre les collègues des pays du Nord et du Sud, disponibilité des ressources sous licence ouverte (*Creative Commons*), etc. [18] ». En effet, il s'agit d'un domaine de recherche qui va bien au delà des aspects linguistiques-informatiques, franchissant ainsi le seuil de domaines comme la sociolinguistique et l'anthropologie. Les enjeux sont toujours très importants pour les communautés linguistiques et, d'un point de vue strictement du traitement automatique des langues, les défis scientifiques et méthodologiques demeurent considérables.

Depuis quelques années, nous explorons ce domaine par le biais de l'encadrement d'étudiants, notamment, des mémoires de master 2 (Sciences du Langage, spécialité Traitement Automatique des Langues, 2011 et 2012) et surtout deux thèses en cours.

5.3.1 Analyse morphosyntaxique automatique du tunisien

A. Hamdi (2011-2015), co-encadrement de thèse avec A. Nasr (LIF-CNRS).

Le travail de thèse d'Ahmed Hamdi fait suite à son travail de master 2 (2011) et porte sur le traitement automatique de la langue arabe et de sa variante tunisienne. Ces travaux se font également en collaboration avec le *Center for Computational Learning Systems* de l'université de Columbia à New York.

Le traitement automatique de l'arabe est confronté à des problèmes spécifiques dus principalement à sa morphologie riche et surtout, à son système de voyellation : en arabe, une partie des voyelles sont représentées comme des signes diacritiques (à l'image des accents en français). Pour des raisons d'économie, les diacritiques sont souvent omis en arabe et c'est au lecteur de les restituer. Dans certains cas, cette restitution se fonde sur des connaissances grammaticales (pour déterminer les marques casuelles, par exemple), mais dans d'autres cas, ce sont des connaissances sémantiques qu'il faut mobiliser pour aboutir à une diacritisation complète. Le *Center for Computational Learning Systems* a développé le système de diacritisation de l'arabe

18. Page d'accueil à l'atelier TALAf 2014.

MADA. Ce système prend en entrée du texte dépourvu de diacritiques et tente de les restituer. Il repose sur des techniques d'apprentissage automatique à partir de corpus dans lesquels figurent les diacritiques. Le travail de master d'Ahmed Hamdi (2011) avait porté sur l'étude de l'apport d'une diacritisation partielle en entrée du système. L'idée était de tirer parti de la diacritisation partielle des textes arabes pour améliorer les performances du système MADA, mais aussi de générer, à partir d'un texte partiellement diacritisé, un texte complètement diacritisé qui sera utilisé pour réapprendre le système MADA (self-training).

Le sujet de thèse poursuit en partie le travail de master : il s'agit d'étudier la relation existante entre diacritisation et analyse syntaxique. Concrètement, il s'agit de développer une chaîne de traitement morpho-syntaxique qui prenne en compte les spécificités du dialecte tunisien. Pour ce faire, la méthodologie mise en œuvre consiste à réaliser une traduction sommaire (ou superficielle) du dialecte vers l'arabe moderne standard afin d'utiliser les outils de ce dernier et les adapter pour le tunisien.

La collaboration avec A. Hamdi et A. Nasr a donné lieu à une publication dans un atelier COLING (VardDial) sur la création de ressources pour des langues peu dotées et sur l'adaptation de ressources et outils d'une langue pour une variante. Concrètement, nous avons automatiquement créé un lexique de déverbaux pour le tunisien à partir de ressources existantes pour l'arabe standard [HGN14]. La méthode consiste à générer des paires de noms tunisien-arabe à partir d'une table de correspondances de patrons de verbes. Pour chaque patron en arabe, il est possible de générer plusieurs déverbaux. Pour générer le lexique de noms, des correspondances ont été faites à partir d'un lexique de verbes. Une étape importante de filtrage a été nécessaire pour éliminer les candidats. Ce filtrage a été fait intrinsèquement (validation manuelle) et extrinsèquement (via un corpus annoté en tunisien). Le lexique final a été évalué en termes de couverture et d'ambiguïté. Ce lexique sera inclus dans la chaîne de traitement pour la traduction du tunisien vers une forme 'approximative' d'arabe standard.

5.3.2 Étude lexicologique des pratiques langagières liées à l'expression de la nature

J. Aznar (2012-2016), co-encadrement de thèse avec V. Rey (CREDO) et L. Dousset (CREDO).

Le travail de Jocelyn Aznar a une dimension ethnolinguistique du fait de son appartenance au Centre de Recherches et de Documentation sur l'Océanie (CREDO) à Marseille. En effet, il s'agit d'une thèse pluridisciplinaire avec une forte composante typologique, linguistique de terrain et ethnologique, dans le cadre de la documentation et de la révitalisation d'une langue du Vanuatu (Océanie) : le nisvai.

Ce travail a lieu à la suite du travail de master 2 en traitement automatique des langues : les aspects liés à l'informatisation et à la modélisation des données linguistiques, ainsi qu'à la création de ressources lexicales, gardent ainsi toute leur place.

Dans le cadre de son master 2 (2012), Jocelyn Aznar avait fait un premier séjour sur le terrain qui avait débouché sur une première proposition de description phonologique du nisvai et de notations graphiques (afin de transcrire la parole des locuteurs). Le travail de terrain avait également donné lieu au recueil de données permettant la création de trois ressources lexicales : un inventaire de noms vernaculaires des plantes, un inventaire de noms vernaculaires d'oiseaux et un lexique multilingue d'environ 120 termes nisvai-bislama-anglais-français sous format électronique.

Pour le travail de thèse, les analyses linguistiques du système phonologique et morphologique se poursuivent suite à un deuxième séjour sur le terrain. Outre ce travail d'analyse typologique, la collecte de données se situe dans un registre anthropologique (recueil de récits traditionnels). Enfin, pour ce qui est de l'aspect traitement automatique des langues, il sera question de développer une plate-forme d'accès aux ressources créées (corpus oraux et lexiques spécialisés).

5.4 Conclusion

Dans ce chapitre, nous avons présenté nos projets de recherche à court et à moyen terme. Notre intérêt pour le lexique reste multidisciplinaire et répond autant à des questionnements théoriques (caractérisation de la complexité lexicale) qu'à des besoins sociétaux pratiques (aide à la lecture dans le cas de publics bien identifiés). Les apports du TAL et des ressources lexicales dans ces domaines nous semblent indiscutables, tout comme la prise en compte d'aspects cognitifs et multilingues. Tous ces sujets suscitent notre intérêt et restent, de ce fait, présents dans le cadre de nos recherches futures.

Troisième partie

Conclusions

Conclusions

Au terme de la rédaction de ce mémoire, il nous semble nécessaire de récapituler les principaux sujets que nous y avons abordés et de réaffirmer l'importance d'une approche multidisciplinaire dans les études sur le lexique. En effet, après cette mise en perspective de nos recherches de ces dix dernières années, et une ébauche des grandes lignes de ce que seront nos travaux futurs à court et à moyen terme, nous revendiquons le rôle central du lexique dans plusieurs disciplines interreliées (qui devraient 'se parler' davantage) et le rôle indispensable des ressources lexicales dans de nombreuses applications. À ce stade, nous pensons que la multidisciplinarité reste un atout majeur pour mener à bien des projets aussi bien théoriques que de développement et d'enrichissement de ressources. Au vu des besoins sociétaux pour des applications nouvelles, que ce soit au niveau pédagogique comme pathologique, nous restons convaincue que les perspectives dans le domaine du traitement automatique du lexique sont considérables. Pour nous, elles restent, somme toute, alléchantes.

Bilan

Les unités lexicales, en tant que briques fondamentales des langues, ont été appréhendées sous différents prismes tout au long de l'Histoire, sous l'influence des courants de pensée et des pratiques en vigueur. L'étude des mots, liée de longue date à la création de ressources lexicographiques, a ainsi été abordée par la philosophie, la philologie, la psychologie et la lexicographie. Ce n'est que très tardivement que la lexicologie s'est constituée en tant que domaine des sciences du langage à part entière, étroitement liée à d'autres domaines de la linguistique (sémantique et morphologie, mais aussi syntaxe). Les progrès technologiques, et notamment l'informatique, ont eu une influence capitale sur la façon d'étudier et de formaliser le lexique (corpus, statistiques, réseaux lexicaux, etc.). C'est pourquoi l'avenir de la lexicologie reste, à notre sens, étroitement lié au traitement automatique des langues.

Par ailleurs, l'informatique a apporté des changements importants au niveau de la construction de ressources lexicographiques, de leur contenu et de la façon d'y accéder. La lexicographie computationnelle ou *e-lexicographie*

offre des perspectives intéressantes au niveau de l'utilisation de technologies du langage au sens large (analyseurs, recherche d'information, partage des données, etc.). Inversement, nous l'avons vu, la construction de ressources lexicales, historiquement liée à la lexicographie, est essentielle dans le traitement automatique des langues. La construction de ressources lexicales s'est ainsi diversifiée, à la différence qu'en TAL, les données lexicales sont nécessairement décrites de façon explicite, avec des informations structurées sous des formats divers, dans un cadre de plus en plus dynamique. De même, les projets collaboratifs et de partage de données (*open linked data*) voient aussi de plus en plus le jour.

Enfin, quelle que soit la ressource, et en tenant compte des efforts nécessaires malgré l'aide de l'informatique, il est clair qu'elle doit être créée et enrichie dans la perspective d'une application concrète. Dans ce sens, une problématique est venue depuis quelques années questionner nos recherches : il s'agit de l'exploitation possible de lexiques dans des applications d'aide aux lecteurs en difficulté. Cette problématique, qui nous permet de nous rapprocher de disciplines comme la psycholinguistique, la psychologie cognitive ou encore l'apprentissage des langues, nous a conduit à établir de nouvelles collaborations et ouvre les portes à des défis que nous comptons relever.

En somme, nous espérons pouvoir continuer à apporter notre grain de sable aux recherches en lexicologie, en linguistique appliquée et en traitement automatique des langues, tout en bénéficiant d'un cadre scientifique et humain idéal comme il a été le cas jusqu'ici, grâce aux excellentes collaborations en cours et grâce à celles que nous espérons avoir l'occasion de nouer pour nos recherches futures.

Annexe A

Glossaire[19]

Article (de dictionnaire) : bloc de texte suivant un schéma rigoureux pour décrire une unité lexicale (ou mot-vedette).

Base lexicale : ensemble structuré de données lexicales. Les données peuvent être consultables grâce à plusieurs clés d'accès (mot-vedette, sous-chaîne de caractères, prononciation, catégorie grammaticale, liens de traduction, nombre de syllabes, fréquence d'apparition dans un corpus déterminé, niveau de difficulté, etc.).

Collocation : association privilégiée (c'est-à-dire, fréquente) entre deux unités lexicales, résultant d'une forte contrainte sémantique de sélection, entérinée par l'usage. Voir co-occurrence.

Concept : représentation mentale abstraite, invariant sémantique d'une forme (le signifié du signe linguistique). Ainsi le sens de piano est l'invariant sémantique de toutes les instances de *piano*. Voir sens.

Co-occurrence : apparition simultanée de deux ou plusieurs unités lexicales dans un corpus. Deux mots co-occurrents ont un degré de proximité sémantique élevé, par exemple, *piano* et *droit*, *piano* et *numérique*, etc.

Corpus : recueil de données écrites, orales ou multimodales, sélectionnées et organisées selon des critères linguistiques et extra-linguistiques pour servir d'échantillon d'emplois déterminés d'une langue.

Décision lexicale : méthode expérimentale dans laquelle il s'agit d'identifier si une chaîne de caractères lue ou entendue est ou non un mot de la langue.

Dictionnaire : ressource recueillant les mots d'une langue et des informations afférentes, par exemple, leur définition, des exemples d'usage, des

19. La plupart des définitions ont été extraites du glossaire de Gala, N. et Zock, M. (éds). *Ressources lexicales : contenu, construction, utilisation, évaluation.* Lingvisticæ Investigationes Supplementa 30, Amsterdam : John Benjamins Publishing, 2013. La liste d'entrées a également été élargie, le contenu amélioré.

traductions vers une autre langue, etc.

Dictionnaire informatisé : dictionnaire qui a subi des transformations afin d'être consultable par un humain au moyen d'un ordinateur. Version électronique d'un dictionnaire papier.

Dictionnaire électronique : dictionnaire conçu pour être consulté par un humain ou par un programme ou application informatique. De ce fait, les informations associées aux entrées sont structurées et explicites. Voir base lexicale.

Dictionnaire mental : terme concernant les connaissances qu'un être humain a sur les mots, leur représentation, organisation et accès. Contrairement aux dictionnaires conventionnels, les mots ne sont pas stockés dans le cerveau de manière holistique : le sens, la forme et les sons sont stockés à différents endroits du cerveau.

Dictionnairique : processus relatifs à l'élaboration d'un dictionnaire en tant que produit commercial.

Encyclopédie : ouvrage visant à synthétiser les connaissances sur un ensemble de domaines. En théorie, il s'oppose aux dictionnaires, dans le sens que ces derniers ne proposent que des informations relatives aux mots. Par exemple, un dictionnaire ne fournit pour le terme *piano* que des informations linguistiques (définition, prononciation, étymologie, usage, etc.), tandis qu'une encyclopédie donne des informations sur le concept (histoire, évolution, typologie, etc.). Dans la réalité, dans beaucoup de cas il n'y a pas de distinction aussi nette entre le contenu d'un dictionnaire, d'un dictionnaire encyclopédique et d'une encyclopédie.

Entrée lexicale : élément par lequel on accède à une ressource au moment de sa consultation. Par exemple, pour trouver des informations concernant le concept *piano*, on cherche sous l'entrée lexicale *piano*. Les informations associées à l'entrée varieront en fonction de la ressource, par exemple, sens et usage linguistique dans un dictionnaire de langue, propriétés combinatoires dans un lexique syntaxique, traduction dans un lexique spécialisé, etc.

Expression polylexicale : unité complexe du lexique, entité lexicale formée de deux ou plusieurs lexèmes avec des degrés variables de figement, par exemple *piano à queue*, *caisse claire*, etc.

Fréquence lexicale : pour une forme donnée, nombre d'occurrences de cette forme dans un corpus déterminé.

Graphème : correspondance écrite d'un phonème pouvant être équivalente à une ou plusieurs lettres, par exemple dans *piano à queue*, le graphème 'a' correspond au phonème /a/, le graphème 'qu' correspond au phonème

/k/, le graphème 'eue' au phonème /ϕ/, etc.

Indice de dispersion : mesure de la répartition des occurrences d'un mot-forme dans des corpus, il augmente l'importance des termes en fonction du nombre de documents dans lesquels ils sont apparus.

Lemme : forme de base d'un mot choisie conventionnellement (en français, infinitifs pour les verbes, forme au singulier pour les noms et au masculin singulier pour les adjectifs). À noter que ce terme possède un autre sens pour les psychologues travaillant sur le dictionnaire mental : mot formellement sous-spécifié, c'est-à-dire variable, ou terme dont les sens et la catégorie syntaxique sont spécifiés, mais pas sa forme écrite ou orale.

Lexème : unité simple du lexique dépourvue de ses variations morphosyntaxiques. Par exemple, *piano* (voir mot-forme).

Lexicométrie : ensemble de méthodes qui permettent de réaliser des analyses statistiques du vocabulaire d'un corpus.

Lexie : unité abstraite du lexique, hyperonyme regroupant les notions de lexème (*piano*) et de locution (*piano à queue*).

Lexique : entité théorique correspondant à l'ensemble de lexies d'une langue (voir aussi vocabulaire). Ressource qui contient des unités lexicales et des informations associées. Par abus du langage, on utilise le terme indistinctement dans les deux sens.

Lexique gradué : ressource lexicale dans laquelle les entrées ont été classées en fonction d'un niveau de difficulté, par exemple Manulex, ReSyF et FLELex.

Locution : entité polylexicale figée, par exemple, *à plus forte raison*, *depuis le berceau*, *avoir la pêche*, *cordon bleu*, etc.

Morphème : plus petite unité significative, correspondant à des lexèmes simples (non construits, sans dérivation) ou à des affixes et des marques de flexion.

Mot : terme appartenant au langage général. On lui préfère des termes comme 'lexie', 'lemme', 'lexème', 'mot-forme' ou 'locution' qui sont des termes précis et non ambigus dans le domaine de la lexicologie.

Mot disponible : unité lexicale usuelle dans la langue, avec fréquence faible et peu stable dans les corpus (par exemple, mots familiers comme *fourchette*, mots sporadiquement utilisés comme *jupe*, *mur*, etc.).

Mot-forme : unité du lexique pourvue des marques morphosyntaxiques nécessaires à son utilisation dans la langue (accords, flexion, etc.) et formant une unité sémantique. Signe linguistique doté d'une certaine autonomie de fonctionnement et de cohésion interne, par exemple *piano*, *jouons*, *pianos*

droits, etc.

Mot-vedette : entrée d'un dictionnaire (voir Lexème).

Nomenclature : liste complète des entrées d'un dictionnaire. Par exemple, le Petit Robert et le TLFi contiennent respectivement 60 000 et 100 000 entrées.

Nonmot : chaîne de caractères qui ne correspond pas à une unité lexicale car elle ne respecte pas les règles phonologiques/phonotactiques de la langue (pas exemple, *pnspn*). Les non-mots sont utilisés en psycholinguistique dans des tâches de décision lexicale. Voir pseudomot.

Nœud : composant élémentaire d'un réseau ou d'un graphe.

Occurrence : apparition d'une unité lexicale dans un corpus. Cette notion est liée à la fréquence d'apparition.

Onomasiologique : classement sémantique à partir des notions ou de domaines, le Thesaurus de Roget étant l'archétype. Dans le cas de WordNet, le classement se fait en termes de classes de synonymes.

Ontologie : modèle représentatif d'un ensemble de concepts d'un domaine. Il encode les objets de base (termes/concepts/mots, individus), leur classe, leurs attributs et relations, ainsi que des axiomes permettant de faire des inférences sur les informations.

Pseudomot : chaîne de caractères qui ne correspond pas à une unité lexicale de la langue et n'a donc pas de signification. Elle serait phonologiquement possible mais elle n'est pas 'instanciée' (par exemple, *piana*). Les pseudo-mots sont utilisés en psycholinguistique dans des tâches de décision lexicale.

Réseau lexical : ensemble d'entités lexicales connectées dans un graphe, chaque lien correspondant à une relation lexicale. Les liens peuvent être conceptuels ou sémantiques. Exemples : WordNet, MindNet, etc. À noter qu'il y a des graphes lexicaux dont les liens ne sont pas explicités.

Réseau sémantique : graphe représentant les connaissances d'un domaine, souvent été utilisés pour représenter le sens d'une phrase ou les liens (conceptuels et/ou linguistiques) entre les mots. Cette dernière fonctionnalité est utilisée par des graphes lexicaux comme WordNet.

Ressource lexicale : ouvrage -quel que soit son support- contenant des unités lexicales, lexies, etc. ou des concepts, associés à des informations de nature très différente : des traductions vers une autre langue (ou vers d'autres langues), des explications à caractère linguistique (origine, caractéristiques grammaticales, emploi, etc.) ou conceptuel (liens thématiques,

relations lexicales, etc.).

Richesse lexicale : rapport entre le nombre de lemmes et leurs occurrences en corpus (la richesse augmente lors que le rapport augmente aussi) ; également, rapport entre le nombre d'hapax et les occurrences (plus il y a de hapax par rapport au nombre total d'occurrences, plus on considère que le texte est riche en vocabulaire).

Sémasiologique : classement à partir de la forme des unités lexicales (lemmes), ordre courant dans les dictionnaires classiques.

Sens : signification. Signifié du signe linguistique, ou encore, interprétation donnée aux formes linguistiques (mots, phrases).

Signe linguistique : association entre une forme et un contenu.

Signifiant : partie formelle (acoustique ou graphique) du signe linguistique, par exemple [pjano] ou *piano*.

Signifié : partie conceptuelle du signe linguistique, par exemple *piano* renvoie à la notion de piano (voir Concept).

Thésaurus : dictionnaire onomasiologique où les termes sont organisés par thèmes, qui eux peuvent être organisés à leur tour en sous-thèmes. La première ressource de ce type, et la plus connue reste le thésaurus de Roget.

Trait sémantique : propriété sémantique généralement encodée sous forme binaire -/+, par exemple *piano [- humain], [+ concret]*, etc.

Unité lexicale : entité appartenant au lexique, regroupant des mots-forme ou des constructions linguistiques se distinguant uniquement par la flexion (*piano* qui regroupe *piano* et *pianos*, *jouer* qui regroupe tout son paradigme *joue*, *jouais* etc.). Voir lemme.

Unité infralexicale : éléments susceptibles de contribuer à la forme ou au sens d'une unité lexicale (primitives conceptuelles, traits morphologiques, etc.)

Unité supralexicale : terme d'une classe à laquelle appartient une unité lexicale (verbes de mouvement, nourriture, etc.).

Vecteur de mots : lorsqu'on modélise le lexique, on utilise le terme mathématique de 'vecteur' pour signifier un ensemble structuré d'unités lexicales, regroupées en fonction d'une caractéristique commune, par exemple, la synonymie (on a, dans ce cas, un vecteur de synonymes, un *synset* dans le terminologie WordNet).

Vocable : forme regroupant deux lexies homonymes. Par exemple, le vocable *avocat* contient les deux lexies : lexie^1 (fruit de l'avocatier), lexie^2

(personne intercédant pour une autre).

Vocabulaire : recueil des unités lexicales d'un locuteur, d'un domaine ou d'une langue donnée. Voir lexique.

Voisin phonologique : unité lexicale de longueur identique à un mot cible et phonologiquement proche (un seul phonème varie, par exemple *bon* et *ton*).

Bibliographie

[ABD+14] D. Amiot, G. Boyé, G. Dal, B. Fradin, F. Kerleroux, S. Lignon, F. Montermini, F. Namer, and F. Villoing. *Manuel de morphologie*, 2014.

[AG10] S. M. Aluisio and C. Gasperin. Fostering Digital Inclusion and Accessibility : The PorSimples project for Simplification of Portuguese Texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, Los Angeles, CA, 2010.

[AMC02] S. Aït-Mokthar and J. P. Chanod. Robustness beyond Shallowness : Incremental Dependency Parsing. *Special Issue of Natural Language Engineering Journal*, 8(3) :121–144, 2002.

[Aur89] S. Auroux. *Histoire des idées linguistiques*. Mardaga, Liège, 1989.

[BBE11] Or Biran, Samuel Brody, and Noemie Elhadad. Putting it simply : a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, page 496–501, 2011.

[BBLF14] L. Brouwers, D. Bernhard, A.-L. Ligozat, and T. François. Syntactic Sentence Simplification for French. In *roceedings of the 3rd International Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR 2014)*, Gothenburg, Sweden, 2014.

[BC97] T. Briscoe and J. Carrol. Automatic extraction of subcategorization from corpora. In *Applied Natural Language Processing*, Washington DC, 1997.

[Béj10] H. Béjoint. *The Lexicography of English*. OUP, Oxford, 2010.

[Ber07] D. Bernhard. Apprentissage non supervisé de familles morphologiques par classification ascendante hiérarchique. In F. Benarmara, N. Hatout, P. Muller, and S. Ozdowska, editors, *Actes deTALN 2007(" TALN ")*, pages 367–376, Toulouse, June 2007. ATALA, IRIT.

[BFL98] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *COLING-ACL 98 : Proceedings of the Conference*, pages 86–90, Montreal, Canada, 1998.

[Bla11] P. Blache. A computational model for language complexity. In *1st Conference on Linguistics, Biology and Computational Science*, Tarragona, Spain, 2011.

[BLV00] M. Brysbaert, M. Lange, and I. Van Wijnendaele. The effects of age-of-acquisition and frequency-of-occurrence in visual word recognition : Further evidence from the Dutch language. *European Journal of Cognitive Psychology*, 12(1) :65–85, 2000.

[BM11] A. Balibar-Mrabti. Lexicographie, grammaire et lexique : une mise en perspective historique. *Cahiers de lexicologie*, 99(2) :255–263, 2011.

[Bou03] J. C. Boulanger. *Les inventeurs de dictionnaires : de l'eduba des scribes mésopotamiens au scriptorium des moines médiévaux*. University of Ottawa Press, Ottawa, 2003.

[BPvR95] R. H. Baayen, R. Piepenbrock, and H. van Rijn. The Celex lexical database (Release 1) [CD-ROM], 1995.

[Bri91] T. Briscoe. Lexical Issues in Natural Language Processing. In *Natural Language and Speech*. Springer-Verlag, 1991.

[Bru10] N. Brunswick. Unimpaired reading development and dyslexia accross different languages. In *Learning to read and spell in different orthographies.*, pages 131–154. Psychology Press, 2010.

[Bru11] C. Brun. Detecting opinions using Deep Syntactic Analysis. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2011)*, Hissar, Bulgaria, 2011.

[Buc27] M. A. Buchanan. *A graded Spanish Word Book*. University of Toronto Press, Toronto, 1927.

[Byb85] J. L. Bybee. Morphology. A study of the relation between meaning and form. *Typological studies in Language*, 9, 1985.

[Cal96] L. J. Calvet. *Histoire de l'écriture*. Pluriel, 1996.

[CCFH99] N. Calzolari, K. Choukri, C. Fellbaum, and E. Hovy. Multilingual resources. In *Multilingual Information Management : current levels and future abilities*. European Commission, 1999.

[CCMF13] C. Chiarcos, P. Cimiano, J. McCrae, and C. Fellbaum. Towards Open Data for Linguistics : Linguistic Linked Data. In *New Trends of Research in Ontologies and Lexical Resources*. Springer-Verlag, Berlin, 2013.

[CDJB77] M. Coltheart, E. Davelaar, T. Jonasson, and D. Besner. Access to the internal lexicon. In *Attention and Performance VI*, pages 535–555, London, 1977. Academic Press.

[CDS96] R. Chandrasekar, C. Doran, and B. Srinivas. Motivations and methods for text simplification. In *16th conference on Computational linguistics*, pages 1041–1044, 1996.

[CFRZ08] D. Cristea, C. Forascu, M. Raschip, and M. Zock. How to evaluate and raise the quality in a colaborative lexicographic approach. In *LREC 2008, International conference on Language Resources and Evaluation*, Marrakesh, 2008.

[CGHH91] K. Church, W. Gale, P. Hanks, and D. Hindle. Using statistics in lexical analysis. In *In Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon*, pages 115–164. Erlbaum, 1991.

[CMC+98] J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, 1998.

[CMR90] A. Content, P. Mousty, and M. Radeaux. Brulex : une base de données lexicale informatisée pour le français écrit et parlé. *L'année Psychologique*, 90 :551–566, 1990.

[Con01] Conseil de l'Europe. *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Hatier, Paris, 2001.

[Cor87] D. Corbin. Morphologie dérivationnelle et structuration du lexique. *Linguistische Arbeiten*, 193, 1987.

[CS98] A. E. Cunningham and K. E. Stanovich. What reading does for the mind. *Am Educator*, 22 :8–15, 1998.

[CSL04] L. Clément, B. Sagot, and B. Lang. Morphology based automatic acquisition of large-coverage lexica. In *LREC 2004, International conference on Language Resources and Evaluation*, pages 1841–1844, Lisbonne, Portugal, 2004.

[CT08] M. Constant and E. Tolone. A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In *27e Colloque international sur le Lexique et la Grammaire (LGC'08)*, pages 11–18, L'Aquila, Italie, 2008.

[Dal31] E. Dale. A comparison of two word lists. *Educational Research Bulletin*, 18(10) :484–489, 1931.

[Dav03] A. Davies. *The Native Speaker : Myth and Reality.* Multilingual Matters Ltd, Clevedon UK, 2003.

[Deu13] J. Deulofeu. Limites dans la compétence et performance dans la construction des énoncés., 2013.

[Dot12] G. Dotoli. *La mise en ordre de la langue dans le dictionnaire.* Hermann editeurs, Paris, 2012.

[DP03] J. Dendien and J. M. Pierrel. Le Trésor de la Langue Française informatisé. *Les dictionnaires électroniques*, 44(2) :11–37, 2003.

[Dub75] J. Dubois. *Lexis. Dictionnaire de la langue française.* Larousse, Paris, 1975.

[FAC11] K. Fort, G. Adda, and K. B Cohen. Amazon Mechanical Turk : gold mine or coal mine ? *Computational Linguistics*, 37(2), 2011.

[FGWF14] T. François, N. Gala, P. Watrin, and C. Fairon. FLELex : a graded lexical resource for French foreign learners. In *Proceedings of International conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, 2014.

[Fle48] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, 31(3) :221–233, 1948.

[FM03] C. Fellbaum and G. A. Miller. Morphosemantic Links in WordNet. *TAL*, 44(2) :69–80, 2003.

[FMC+06] G. Francopoulo, G. Monte, N. Calzolari, M. Monachini, N. Bel, M. Pet, and C. Soria. Lexical Markup Framework. In *LREC 2006, International conference on Language Resources and Evaluation*, Gènes, Italie, 2006.

[FO10] P. Fuertes-Olivera. *E-lexicography. The INternet, Digital Initiatives and Lexicography.* Henning BergenHoltz, London, 2010.

[Fra11] T. François. *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère.* PhD thesis, Université Catholique de Louvain, Louvain la Neuve, 2011.

[Fra13] G. Francopoulo. *Lexical Markup Framework.* ISTE Ltd., 2013.

[Gag65] J. Gagnepain. Lexicologie et structuralisme. *Annales de Bretagne*, 72(4) :537–539, 1965.

[Gal03] N. Gala. *Un modèle d'analyseur syntaxique robuste fondé sur la modularité et la lexicalisation de ses grammaires.* PhD thesis, Université de Paris Sud, Orsay, France, March 2003.

[Gal11] N. Gala. Developing a lexicon of word families for closely-related languages. In *ESSLLI International Workshop on Lexical Resources (WoLeR)*, Ljubljana, 2011.

[Gal13] N. Gala. Ressources lexicales mono- et multilingues : une évolution historique au fil des pratiques et des usages. In *Ressources Lexicales. Contenu, construction, utilisation, évaluation.*, volume 30, pages 1–42. John Benjamins, Amsterdam, Gala, N. et Zock, M. edition, 2013.

[Gau99] E. Gaussier. Unsupervised learning of derivational morphology from inflectional lexicons. In *ACL Workshop on Unsupervised Learning in Natural Language Processing*, College Park, MD, 1999.

[GB12] N. Gala and C. Brun. Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions. In *TALN 2012, Conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, 2012.

[GBFB15] N. Gala, M. B. Billami, T. François, and D. Bernhard. Graded lexicons : new resources for educational purposes and much more. In *Actes d'EUROCALL - 2015 Critical CALL*, 2015.

[GBM13] N. Gala and V. Barbu-Mititelu. LexRom : un réseau lexical pour des familles morphologiques dans des langues romanes. (Présentation sans publication), 2013.

[Ger84] M.A. Gernsbacher. Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology : General*, 113(2) :256–281, 1984.

[GFBF14] N. Gala, T. François, D. Bernhard, and C. Fairon. Un modèle pour prédire la complexité lexicale et graduer les mots. In *Actes de TALN 2014*, Marseille, 2014.

[GFF13] N. Gala, T. François, and C. Fairon. Towards a French lexicon with difficulty measures : NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. In *E-lexicography in the 21st century : thinking outside the paper*, Tallin, 2013.

[GGN12] P. Gambette, N. Gala, and A. Nasr. Longueur de branches et arbres de mots. *Corpus*, 11 :129–146, 2012.

[GGNG11] P. Gambette, N. Gala, A. Nasr, and A. Guénoche. Longueur de branches et arbres de mots. In *La cooccurrence : du fait statistique au fait textuel*, Février 2011.

[GGPF06] C. Gardent, B. Guillaume, G. Perrier, and I. Falk. Extraction d'information de sous-catégorisation à partir des tables du LADL. In *Actes de TALN 2006*, Louvain la Neuve, Belgium, 2006.

[GHN+11] N. Gala, N. Hathout, A. Nasr, V. Rey, and Seppälä. Création de clusters sémantiques à partir du TLFi. In *Actes de TALN 2011*, Montpellier, 2011.

[Gil98] D. Gillot. Le droit des sourds : 115 propositions., 1998. Rapport parlementaire au premier ministre.

[GL05] N. Gala and M. Lafourcade. Combining corpus-based pattern distributions with lexical signatures for PP attachment ambiguity resolution. In *Symposium on Natural Language Processing (SNLP 2005)*, Chiang Rai, December 2005.

[GL06] N. Gala and M. Lafourcade. PP Attachment Ambiguity Resolution with Corpus-Based Pattern Distributions and Lexical Signatures. *ECTI Journal*, 2(2) :116–120, 2006.

[GL11] N. Gala and M. Lafourcade. NLP lexicons : innovative constructions and usages for machines and humans. In *E-lexicography*, Ljubljana, 2011.

[GMRS64] G. Gougenheim, R. Michéa, P. Rivenc, and A. Sauvageot. *L'élaboration du français fondamental (1r degré)*. Didier, Paris, 1964.

[Gou58] G. Gougenheim. *Dictionnaire fondamental de la langue française*. Didier, Paris, 1958.

[GR08] N. Gala and V. Rey. Polymots : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. In *TALN 2008, Conférence sur le Traitement Automatique des Langues Naturelles*, Avignon, France, 2008.

[Gre98] G. Grefenstette. 'The Future of Linguistics and Lexicographers : Will there be Lexicographers in the year 3000 ? In *EURALEX-1998*, pages 25–41, 1998.

[Gre99] G. Grefenstette. Multilingual corpus-based extraction and the Very Large Lexicon. In *Parallel corpora, parallel worlds.*, volume 43, pages 137–149. Lars Borin, Uppsala, Suède, 1999.

[GRH10] N. Gala, V. Rey, and N. Hathout. Apprentissage du lexique des langues romanes à l'aide d'une ressource lexicale fondée sur la notion de familles et séries de mots. In *Actes d'EUROCALL - 2010 Languages, cultures and virtual communities*, Bordeaux, 2010.

[Gro94] M. Gross. Constructing Lexicon-Grammars. In B. T. Atkins and A. Zampolli, editors, *Computational Approaches to the Lexicon*. Oxford University Press, Oxford, 1994.

[Gro97] M. Gross. The Construction of Local Grammars. In *Finite-State Language Processing*, pages 329–352. The MIT Press, 1997.

[GRT09] N. Gala, V. Rey, and L. Tichit. Dispersion sémantique dans des familles morpho-phonologiques : éléments théoriques et empiriques. In *Actes de TALN 2009*, Senlis, 2009.

[GRZ10] N. Gala, V. Rey, and M. Zock. A tool for linking stems and conceptual fragments to enhance word access. In *LREC 2010, Seventh international conference on Language Resources and Evaluation*, La Valetta, Malta, 2010.

[GT94] G. Grefenstette and P. Tapanainen. What is a Word ? What is a Sentence ? Problems of Tokenization. In *Conference Computational Linguistics (COLING)*, 1994.

[GV05] N. Gala and A. Valli. Building a computational lexicon of verbal syntactic constructions in French. In *Workshop on Multilingual Lexical Databases (Papillon 2005)*, Chiang Rai, December 2005.

[Ham12] M. Hamm. Écrire sans entendre. Une exploration de la pratique de l'écriture chez quelques sujets sourds, devenus sourds et malentendants. *Éducation et formation*, 297, 2012.

[Hat08] N. Hathout. Acquistion of the morphological structure of the lexicon based on lexical similarity and formal analogy. . In *COLING Workshop Textgraphs-3*, pages 1–8, Manchester, UK, 2008.

[Hat09a] N. Hathout. Acquisition of morphological families and derivational series from a machine readable dictionary. In *6e Décembrettes* , Bordeaux, 2009.

[Hat09b] N. Hathout. Contributions à la description de la structure morphologique du lexique et à l'approche extensive en morphologie. , 2009. Habilitation à diriger des recherches. Universités de Toulouse II-Le Mirail.

[Hei09] U. Heid. Aspects of lexical description for electronic dictionaries, 2009. Key note speaker Electronic lexicography in the 21st century (ELEX-2009).

[Hen75] G. Henry. *Comment mesurer la lisibilité.* Labor, Bruxelles, 1975.

[HFE09] M. Huenerfauth, L. Feng, and N. Elhadad. Comparing evaluation techniques for text readability software for adults with intellectual disabilities. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility* , pages 3–10, New York, 2009.

[HGN14] A. Hamdi, N. Gala, and A. Nasr. Automatically building a Tunisian Lexicon for Deverbal Nouns. In *Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial), COLING Workshop*, Dublin, 2014.

[HS51] D.H. Howes and R.L. Solomon. Visual duration threshold as a function of word probability. *Journal of Experimental Psychology*, 41(40) :1–4, 1951.

[IFT+03] K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. Text simplification for reading assistance : a project note. In *Proceedings of the 2nd International Workshop on Paraphrasing :*

Paraphrase Acquisition and Applications (IWP), pages 9–16, 2003.

[Imb71] P. Imbs. Trésor général de la Langue Française, préface., 1971.

[IV94] N. Ide and J. Véronis. Machine Readable Dictionaries : what have we learned, where do we go ? In *COLING Workshop on Directions of Lexical Research*, Beijing, China, 1994.

[IV95] N. Ide and J. Véronis. Encoding Dictionaries. *Computers and the Humanities*, 29(1-3), 1995.

[JLSZ12] A. Joubert, M. Lafourcade, D. Schwab, and M. Zock. Évaluation et consolidation d'un réseau lexical via un outil pour retrouver le mot sur le bout de la langue. In *TALN 2012, Conférence sur le Traitement Automatique des Langues Naturelles*, Grenoble, France, 2012.

[Juo08] P. Juola. Assessing linguistic complexity. In *Language Complexity : Typology, contact, change*, pages 89–108. John Benjamins, Amsterdam, 2008.

[KF63] J. J. Katz and A. Fodor. The Structure of a Semantic Theory. *Language*, 39 :170–201, 1963.

[KF64] H. Kucera and W. N. Francis. Brown Corpus Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with Digital Computers, 1964.

[KF67] H. Kucera and W. N. Francis. *Computational analysis of present-day American English* . Brown University Press, Providence, RI, 1967.

[Laf07] M. Lafourcade. Making people play for Lexical Acquisition. In *Proc. SNLP 2007, 7th Symposium on Natural Language Processing.*, Pattaya, Thailande, 2007.

[Lau97] B. Laufer. *What's in a word that makes it hard or easy : Some intralexical factors that affect the learning of words.* Cambridge University Press, 1997.

[LC01] E. Lambert and D. Chesnet. Novlex : une base de données lexicales pour les élèves de primaire. *L'Année Psychologique*, 101 :277–288, 2001.

[Léo04] J. Léon. Lexies, synapsies, synthèmes : le renouveau des études lexicales en France au début des années 1960. *History of Linguistics in Texts and Concepts*, pages 405–418, 2004.

[Lit06] W. Littlewood. *Foreign and Second Language Learning*. Cambridge University Press, Cambridge, 20 edition, 2006.

[LP14] V. Lux-Pogodalla. Le Réseau Lexical du Français et son corpus encapsulé d'exemples lexicographiques annotés. In *Actes de TALN 2014, Conférence du Traitement Automatique du Langage Naturel.*, 2014.

[LPP11] V. Lux-Pogodalla and A. Polguère. Construction of a French Lexical Network : methodological issues. In *Workshop on Lexical Resources (WoLeR).*, pages 1–5, Ljubljana, Slovénie., 2011.

[LSCC04] B. Lété, L. Sprenger-Charolles, and P. Colé. Manulex : A grade-level lexical database from French elementary-school readers. *Behavior Research Methods, Instruments and Computers*, 36 :156–166, 2004.

[Man01] M. Mangeot. *Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue*. PhD thesis, Université Joseph Fourier, Grenoble, France, 2001.

[MBCH09] H. Manuélian, A. Bruscand, N. Cholewka, and A-M. Hetzel. Le Petit Larousse Illustré de 1905 en ligne : Présentation et secrets de fabrication. *Informatique et description de la langue d'hier et d'aujourd'hui, Études de Linguistique Appliquée (ÉLA)*, 4, 2009.

[MCP95] I. Mel'cûk, A. Clas, and A. Polguère. *Introduction à la lexicologie explicative et combinatoire*. Duculot, Louvain la Neuve, 1995.

[ME13] M. Mangeot and C. Enguehard. *Des dictionnaires éditoriaux aux représentations XML standardisées*, volume 30, pages 255–289. John Benjamins Publishing, Gala, N. et Zock, M. edition, 2013.

[Mel81] I. Mel'cûk. Meaning-Text Models : A Recent Trend in Soviet Linguistics . *Annual Review of Anthropology*, 10 :27–62, 1981.

[Mel99] I. Mel'cûk. Dictionnaire explicatif et combinatoire du français contemporain. *Recherches lexico-sémantiques*, IV :347, 1999.

[Mic53] R. Michéa. Mots fréquents et mots disponibles. Un aspect nouveau de la statistique du langage. *Les Langues Modernes*, 474 :338–344, 1953.

[Mie08] M. Miestamo. Introduction. In M. Miestamo, K. Sinnemäki, and F. Karlsson, editors, *Language Complexity : Typology, contact, change*. John Benjamins Publishing, 2008.

[Mil90] G. A. Miller. WordNet : An On-Line Lexical Database. *International Journal of Lexicography*, 3(4), 1990.

[Mil99] G. A. Miller. On knowing a word. *Annual Review of Psychology*, 50(1) :1–19, 1999.

[Mon91] S. Monsell. The nature and locus of word frequency effects in reading. In D. Besner and G.W. Humphreys, editors, *Basic processes in reading : Visual word recognition*, pages 148–197. Lawrence Erlbaum Associates Inc., Hillsdale, NJ, 1991.

[MSL03] M. Mangeot, G. Sérasset, and M. Lafourcade. Construction collaborative de données lexicales multilingues, le projet Papillon. *TAL*, 44(2) :151–176, 2003.

[Mul77] Ch. Muller. *Principes et méthodes de statistique lexicale.* Champion, Paris, 1977.

[Nam05] F. Nammer. La morphologie constructionnelle du français et les propriétés sémantiques du lexique, 2005. Habilitation à diriger des recherches. Université de Nancy 2.

[Nes00] H. Nesi. *The Use and Abuse of EFL Dictionaries. How learners of English as a foreign language read and interpret dictionary entries .* Max Niemeyer Verlag, Tübingen, 2000.

[New06] B. New. Lexique3 : une nouvelle base de données lexicales. In *Actes de TALN 2006*, Louvain la Neuve, Belgium, 2006.

[NP10] R. Navigli and S. P. Ponzetto. BabelNet : building a very large multilingual semantic network. In *48th annual meeting of the Association for Computational Llinguistics.*, pages 216–225, Uppsala, Suède, 2010.

[NPFM01] G. A. New, C. Pallier, L. Ferrand, and R. Matos. Une base de données lexicales du français contemporain sur Internet : Lexique 3. *L'année psychologique*, 101 :447–462, 2001.

[NVG03] R. Navigli, P. Velardi, and A. Gangemi. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18(1) :22–31, 2003.

[Pal99] C. Pallier. Syllabation des représentations phonétiques de Brulex et de Lexique. Technical report, Laboratoire de Sciences Cognitives et Psycholinguistique, Paris, 1999.

[PdC00] G. Pérennou and M. de Calmès. MHATLex : Lexical Resources for Modelling the French Pronunciation. In *Second*

International Conference on Language Resources and Evaluation (LREC), Athènes, Grèce, 2000.

[PGTV10] S. Pinto, A. Ghio, B. Teston, and F. Viallet. La dysarthrie au cours de la Maladie de Parkinson. Histoire naturelle de ses composantes : dysphonie, dysprosodie et dysarthrie. *Revue Neurologique*, 166(10) :800–810, 2010.

[Pie13] J. M. Pierrel. Structuration et usage de ressources lexicales institutionnelles sur le français. In *Ressources Lexicales. Contenu, construction, utilisation, évaluation.*, volume 30. John Benjamins, Amsterdam, Gala, N. et Zock, M. edition, 2013.

[Pol02] A. Polguère. *Notions de base en lexicologie.* Presses de l'université de Montréal, Montréal, 2002.

[Pol06] A. Polguère. Structural properties of lexical systems : monolingual and multilingual perspectives. In *COLING Workshop on multilingual resources and interoperability*, pages 50–59, Sydney, 2006.

[Pol09] A. Polguère. Lexical systems : graph models of natural language lexicons. *Language resources and evaluation*, 43, 2009.

[Pol12] A. Polguère. Lexicographie des dictionnaires virtuels. In *Meanings, Texts, and Other Exciting Things. A Festschrift to Commemorate the 80th Anniversary of Professor Igor Alexandrovič Mel'čuk*, pages 509–523. Jazyki slavjanskoj kultury Publishers, Moscou, 2012.

[Pol14] A. Polguère. Lexical contextualism : the Abélard Syndrome. In *Language production, Cognition, and the Lexicon. Festschrift in honour to M. Zock.* Springer, Gala, N., Rapp,R. and Bel-EnguixG. edition, 2014.

[Pru06] J. Pruvost. *Les dictionnaires français : outils d'une langue et d'une culture.* Ophris, Paris, 2006.

[Pus91] J. Pustejovsky. The Generative Lexicon. *Computational Linguistics*, 17(4), 1991.

[RBYS13] L. Rello, R. Baeza-Yates, and H. Saggion. The impact of lexical simplification by verbal paraphrases for people with and without Dyslexia. *Computational Linguistics and Intelligent Text Processing. Lecture Notes in Computer Science*, 7817 :501–512, 2013.

[Rel14] L. Rello. *DysWebxia. A Text Accessibility Model for People with Dyslexia.* PhD thesis, Universitat Pompeu Fabra, Barcelone, 2014.

[Rey70a] A. Rey. *La lexicologie. Lectures.* Klincksieck, 1970.

[Rey70b] A. Rey. Typologie génétique des dictionnaires. *La lexicographie. Revue langages*, 19, 1970.

[Rey11] A. Rey. *Dictionnaire amoureux des Dictionnaires.* Plon, 2011.

[Rey13] C. Rey. Dictionnaires d'hier et d'aujourd'hui. Ressouces lexicales par excellence. In *Ressources Lexicales. Contenu, construction, utilisation, évaluation.*, pages 85–118. John Benjamins Publishing, Gala, N. et Zock, M. edition, 2013.

[RG11] V. Rey and N. Gala. *Les mots de bouche à oreille*, pages 279–293. L'Harmattan, Paris, 2011.

[Rom01] L. Romary. *TMF - a tutorial.* Nancy, 2001.

[RSAF04] L. Romary, S. Salmon-Alt, and G. Francopoulo. Standards going concrete : from LMF to Morphalou. In *Workshop on Electronic Dictionaries, COLING, Conference on Computational Linguistics*, Geneva, Suisse, 2004.

[RSBY14] L. Rello, H. Saggion, and R. Baeza-Yates. Keyword Highlighting Improves Comprehension for People with Dyslexia. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) at EACL 2014*, 2014.

[Sag10] B. Sagot. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. In *Actes de LREC 2010*, La Valetta, Malta, 2010.

[SB97] R. Schreuder and R. H. Baayen. How complex simplex words can be. *Journal of Memory and Language*, pages 118–139, 1997.

[SB08] L. Sitbon and P. Bellot. How to cope with questions typed by dyslexic users ? In *2nd ACM SIGIR Workshop on Analytics for noisy unstructured text data*, pages 1–8, Singapour, 2008.

[SBB10] L. Sitbon, P. Bellot, and P. Blache. Vers une recherche d'informations adaptée aux capacités de lecture des utilisateurs. *RSTI Document Numérique*, 13(1) :161–186, 2010.

[Sch94] H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on new methods in language processing*, Manchester, UK, 1994.

[SEOM12] S. Stajner, R. Evans, C. Orasan, and R. Mitkov. What can readability measures really tell us about text complexity ? In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, Istanbul, Turkey, 2012.

[SF06] B. Sagot and D. Fišer. Building a free French wordnet from multilingual resources. In *LREC 2006, International conference on Language Resources and Evaluation, Ontolex Workshop*, Marrakesh, 2006.

[SGMA+11] H. Saggion, E. Gomez-Martin, A. Anula, L. Bourg, and E. Etayo. Text simplification in SImplext : Making texts more accessible. *Procesamiento del Lenguaje Natural (SEPLN)*, 46, 2011.

[Sha14] M. Shardlow. A Survey of Automated Text Simplification. *International Journal of Advances Computer Science and Applications, Special Issue on Natural Language Processing*, pages 58–70, 2014.

[SJM12] Lucia Specia, Sujay K. Jauhar, and Rada Mihalcea. Semeval-2012 task 1 : English lexical simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation (SemEval 2012), Montreal, Canada*, 2012.

[SK12] B. Szmrecsanyi and B. Kortmann. Introduction. In *Linguistic complexity. Second Language Acquisition, indigenization, contact*, pages 6–34. de Gruyter, Berlin, 2012.

[Sus06] J. Suso Lopez. Vocabulaires logiques, vocabulaires simplifiés et *Français élémentaire. Documents pour l'histoire du français langue étrangère ou seconde*, 36 :97–118, 2006.

[SVB03] T. Selva, S. Verlinde, and J. Binon. Vers une deuxième génération de dictionnaires électroniques. *TAL*, 2003.

[TBM14] D. Tufis and V. Barbu-Mititelu. The Lexical Ontology for Romanian. In *Language production, Cognition, and the Lexicon. Festschrift in honour to M. Zock*. Springer, Gala, N., Rapp, R., Bel-Enguix, G. edition, 2014.

[Tho21] E. Thorndike. *The Teacher's Word Book*. Teachers College, Columbia University, New York, 1921.

[TL44] E. Thorndike and I. Lorge. *The Teacher's Word Book of 30,000 words*. Teachers College, Columbia University, New York, 1944.

[Tou14] N. Tournadre. *Le Prisme des Langues*. Asiathèque, Paris, 2014.

[Tut10] A. Tutin. Sens et combinatoire lexicale : de la langue au discours, 2010. Mémoire d'Habilitation à diriger des recherches. Université de Grenoble.

[Val80] A. Valli. *Etablissement d'un lexique automatique de verbes français.* PhD thesis, Université Paris 7 (LADL), Paris, 1980.

[Van32] G. E. Vander Beke. *The French Word Book*. Macmillan, New York, 1932.

[VBD12] T. Vanrullen, L. Boutora, and J. Dagron. Enjeux méthodologiques, linguistiques et informatiques pour le traitement du français écrit des sourds. In *Actes de TALN 2012, Conférence du Traitement Automatique du Langage Naturel*, Grenoble, France, 2012.

[VI90] J. Véronis and N. Ide. Word Sense Disambiguation with Very Large Neural Networks extracted from Machine Readable Dictionaries. In *COLING-90 International Conference on Computational Linguistics*, pages 389–394, Helsinki, 1990.

[VM03] K. Van den Eynde and P. Mertens. La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13 :63–104, 2003.

[VOM90] F. Vitu, J.K. O'Regan, and M. Mittau. Optimal landing position in reading isolated words and continuous text. *Perception & Psychophysics*, 47(6) :583–600, 1990.

[VORN96] E. Viegas, B. Onyshkevych, V. Raskin, and S. Nirenburg. From Submit to Submitted via SUbmission : On lexical rules on Large-Lexicon Acquisition. In *Proceedings of Association for Computational Linguistics (ACL-1996)*, Santa Cruz, CA., 1996.

[Vos00] P. Vossen. *Eurowordnet. A multilingual database with Lexical Semantic Networks*. Kluwer Academic Publishers, Paris, 2000.

[WB06] H. Walter and B. Baraké. *Arabesques. L'aventure de la langue arabe en Occident*. Robert Laffont, 2006.

[WJU+09] W. M. Watanabe, A. C. Junior, V. R. Uzêda, R. P. de Mattos Fortes, T. A. S. Pardo, and S. M. Aluisio. Facilita : reading assistance for low-literacy readers. In *SIGDOC '09 : Proceedings of the 27th ACM international conference on Design of communication*, pages 29–36, New York, NY, US, 2009.

[ZBF+12] M. Zorzi, C. Barbiero, A. Facoetti, L. Lonciari, M. Carrozzi, M. Montico, L. Bravar, F. George, C. Pech-Georgel, and J. C. Ziegler. Extra-large letter spacing improves reading in dyslexia. *Proceedings of National Academy of Sciences of the United States of America (PNAS)*, 2012.

[ZC03] M. Zock and J. Carrol. Les dictionnaires électroniques. *Traitement Automatique des Langues*, 44(2) :7–10, 2003.

[ZIMD95] S. M. Zeno, S. H. Ivens, R. T. Millard, and R. Duvvuri. *The Educator's Word Frequency Guide*. Touchstone Applied Science Associates, Brewster, NY, 1995.

[ZPMW+03] J. C. Ziegler, C. Perry, A. Ma-Wyatt, D. Ladner, and G. Schulte-Korne. Developmental dyslexia in different languages : language-spexific or universal ? *Journal of Experimental Child Psychology*, 86(3) :169–193, 2003.

[ZPZ14] J. C. Ziegler, C. Perry, and M. Zorzi. Modeling reading development through phonological decoding and self-teaching : implications for dyslexia. *Philosophical Transactions of the Royal Society*, 2014.

[ZS13] M. Zock and D. Schwab. L'index, une ressource pour trouver le mot bloqué sur le bout de la langue. In *Ressources Lexicales. Contenu, construction, utilisation, évaluation.*, pages 313–354. John Benjamins Publishing, Gala, N. et Zock, M. edition, 2013.