Linguistique et Traitement Automatique des Langues : une coopération nécessaire

Max SILBERZTEIN
Université de Franche-Comté

Résumé

Aujourd'hui, la plupart des applications logicielles du Traitement Automatique des Langues (analyse du discours, extraction d'information, moteurs de recherche, etc.) analysent les textes comme étant des séquences de formes graphiques. Mais les utilisateurs de ces logiciels cherchent typiquement des unités de sens : concepts, entités, relations dans leurs textes. Il faut donc établir une relation entre les formes graphiques apparaissant dans les textes et les unités de sens qu'elles représentent. Cette mise en relation nécessite des ressources et des méthodes de traitement linguistiques, que je présente ici.

Mots clé : Traitement Automatique des Langues, linguistique.

Abstract

Today, most software applications (discourse analysis, information extraction, search engines, etc.) process texts as sequences of word forms. However, users of these applications are looking for units of meanings: concepts, entities and relations in their texts. Therefore, one needs to link these word forms with the units of meanings that they represent. Linking these two types of objects requires the availability of linguistic resources and methods that I present here.

Keywords: Natural Language Processing, Linguistics.

Resumen

Hoy en día, la mayoría de programas de Procesamiento Automático del Lenguaje Natural (análisis del discurso, extracción de información, motores de búsqueda, etc.) analizan los textos como secuencias de formas gráficas. Ahora bien, les usuarios de dichos programas van en busca, normalmente, de unidades de sentido: conceptos, entidades, relaciones en los textos. Es preciso, pues, establecer un vínculo entre las formas gráficas que aparecen en los textos y las unidades de sentido a las que estas representan. Dicha puesta en relación requiere recursos y métodos de procesamiento de tipo lingüístico, que presento en el artículo.

Palabras clave: Procesamiento del Lenguaje Natural, lingüística.

1. Introduction

Aujourd'hui, la plupart des applications logicielles du Traitement Automatique des Langues (TAL) traitent les textes comme des séquences linéaires de formes graphiques : les moteurs de recherche indexent des formes graphiques et les retrouvent par une simple comparaison avec les requêtes des utilisateurs (elles aussi traitées comme des formes graphiques) ; les logiciels d'extraction d'information associent des mots-clés ou formes graphiques avec des annotations représentant des concepts, entités et relations ; les logiciels d'analyse du discours comptent les occurrences de formes graphiques pour détecter des fréquences anormalement élevées ou faibles, typiquement des fréquences qui n'obéissent pas à une loi normale de distribution[1]. Par exemple, il peut être intéressant de découvrir qu'une personnalité politique utilise avec une fréquence exceptionnellement élevée le mot *weak* [faible]. Inversement, découvrir dans un corpus d'entretiens dans une école qu'un enfant n'utilise jamais le mot « Maman » peut constituer un signe alarmant pour les psychologues[2].

Dans tous les cas, ces logiciels unifient « information » (concepts, entités et relations) et « formes graphiques », sans toujours le dire à leurs utilisateurs[3]. Mais les formes graphiques ne correspondent presque jamais aux unités de sens[4] qu'elles sont censées représenter.

En effet, tous les éléments de sens d'une langue sont portés par des unités linguistiques atomiques, c'est-à-dire indivisibles : les concepts (par

[1] La loi normale est présentée à la page http://fr.wikipedia.org/wiki/Loi_normale. Pour des applications linguistiques de l'analyse statistique, voir entre autres les contributions rassemblées par S. Loiseau (2015) ou par P. Blumenthal et D. Vigier (2017). Noter cependant que F. Rastier (2011) affirme : *« La pertinence n'émerge pas du quantitatif ».*

[2] F. Neveu (2016) montre ainsi que dans un texte, l'absence ou l'hapax peuvent être aussi révélateurs que la pléthore d'occurrences.

[3] Il existe des dizaines d'outils de traitement automatique des textes en langue naturelle (TAL), mais la plupart exigent des connaissances en programmation (ex. Gate, NLTK) et n'ont pas pour but de produire les analyses dont les chercheurs en sciences sociales ont besoin pour analyser leur corpus : évolution du vocabulaire, mesures de pertinence, de similarité, analyse factorielle, etc. Je m'intéresse ici aux logiciels conçus pour répondre aux besoins des chercheurs en sciences sociales : Hyperbase (Brunet, 2010), Iramuteq (Ratinaud, 2009 ; Loubère, 2014), Lexico (Lamalle *et alii*, 2002), Sketch Engine (Kilgarriff *et alii*, 2004) et TXM (Heiden *et alii,* 2010), entre autres.

[4] Ce que nous savons depuis Martinet (1966).

ex. *le socialisme*[5]), les entités (par ex. *un parti politique*), les prédicats et relations entre concepts ou entités (ex. *adhérer, aimer*), etc. Si l'on veut apprendre une langue, il est indispensable d'apprendre ses unités linguistiques atomiques : il ne viendrait à l'idée de personne par exemple d'essayer de deviner comment on dit « maison » en japonais à partir de la traduction japonaise des mots *mais* et *on* : le mot « maison » est une unité linguistique atomique ; personne ne tentera non plus de deviner comment on dit « tout de suite » en japonais à partir des mots *tout*, *de* et *suite*. De même, aucune chance qu'un logiciel d'analyse de textes retrouve des informations de façon fiable s'il confond les occurrences du mot *carte* dans les contextes *carte bancaire*, *carte blanche*, *carte postale ou carte routière*.

Pour fournir aux utilisateurs des outils de TAL les informations dont ils ont vraiment besoin, il faut donc construire une passerelle entre les formes graphiques présentes dans les textes et les unités de sens qui les intéressent. Dans cette présentation, nous montrons à partir d'exemples concrets comment les méthodes [6] et les ressources linguistiques peuvent être utilisées pour construire cette passerelle.

2. Orthographe

La première étape de toute analyse doit être d'établir une correspondance entre les formes graphiques – que l'on trouve dans ces textes écrits – et les unités de sens que l'on veut traiter. Mais l'orthographe n'a pas toujours été fixée. Ainsi par exemple, dans les textes en moyen-français, le mot *seigneur* est orthographié de plus de cinquante façons différentes. Pour pouvoir offrir aux historiens et aux médiévistes les moyens d'analyser leur corpus, il faut intégrer au logiciel une correspondance entre ce terme et toutes ses variantes orthographiques.

[5] Le terme *socialisme* a acquis une autonomie par rapport à l'adjectif *social* : il réfère à une idéologie et à une théorie économique et a sa propre histoire, ses partis politiques, etc. On ne peut donc pas réduire le sens de ce terme comme étant juste une dérivation de l'adjectif *social*. Il s'agit donc d'une unité linguistique atomique.

[6] Cf. Silberztein (2015). Le logiciel NooJ, son manuel et ses ressources linguistiques pour une vingtaine de langues sont disponibles sur le site www.nooj-association.org. Les fonctionnalités discutées ici sont nouvelles et ont pour but de rendre NooJ (qui est fondamentalement une plateforme linguistique) utilisable par les chercheurs en sciences humaines et sociales (non-linguistes) qui veulent exploiter leur corpus.

Dans le logiciel d'exploration PALM[7], cette mise en correspondance est décrite par la grammaire électronique suivante.

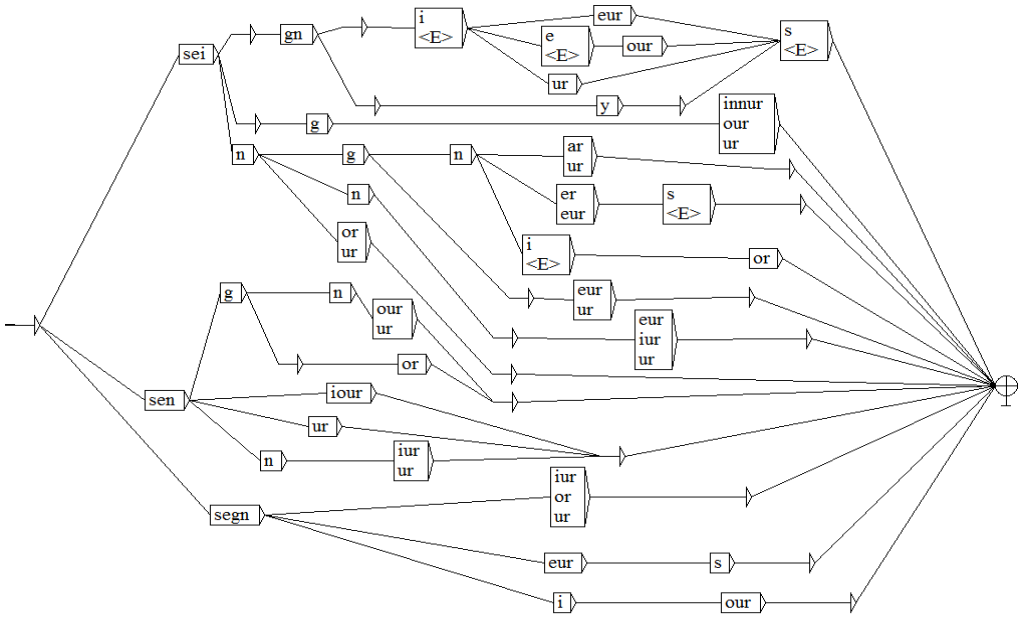


Figure 1. Grammaire qui représente les variantes graphiques du mot seigneur

Cette grammaire reconnaît toutes les formes graphiques que l'on peut épeler en partant du nœud initial (flèche horizontale à droite), en suivant les connections et en rejoignant le nœud terminal (la cible à droite). Grâce à cette grammaire, le logiciel PALM[8] donne aux médiévistes la possibilité de retrouver cette entité partout dans leur corpus, quelle que soit sa graphie, pour étudier ses contextes, caractériser sa fréquence dans des sous-corpus, retrouver des collocations intéressantes avec d'autres entités ou relations, mettre à jour des oppositions avec d'autres entités, etc.

Aujourd'hui, la norme orthographique est relativement fixée, même si elle n'est pas toujours appliquée dans les corpus utilisés par les chercheurs en SHS (voir les tweets et les messages SMS), mais même dans le meilleur des cas, les dictionnaires dits de référence contiennent de nombreuses

[7] Le logiciel PALM d'exploration de corpus médiéval est compatible avec les ressources linguistiques construites avec NooJ, cf. Aouini (2018).
[8] Cf. Aouini (2018).

imprécisions orthographiques[9]. Il existe ainsi des hésitations systématiques dans l'orthographe des mots d'origine étrangère (ex. *casher, cacher, kosher, koscher*), dans l'usage de la soudure, de l'espace et du trait d'union (ex. *audiovisuel, audio-visuel, audio visuel*), des majuscules (ex. *moyen-âge, Moyen-âge, Moyen-Âge*), des lettres accentuées (ex. *événement* vs. *évènement*), du pluriel (ex. *attrape-nigaud*, *attrape-nigauds, simulateur de vol, simulateur de vols*), etc.

Les logiciels d'analyse de textes qui ne rassemblent pas les variantes orthographiques sous-estiment donc systématiquement la fréquence des concepts ou entités qu'ils traitent. Par exemple, une simple recherche Google ne permet de trouver que 13 millions d'occurrences pour la forme « audiovisuel », alors que si l'on combine toutes les variantes orthographiques de ce terme, on trouve en fait 29 millions d'occurrences. Avec seulement 45 % de résultats trouvés, on est loin des performances typiquement mises en avant (95 % +) par les partisans des méthodes stochastiques de TAL !

Les logiciels qui veulent offrir aux chercheurs en SHS des analyses statistiques vraiment fiables doivent prendre en compte la variation orthographique ; ces variantes peuvent être décrites soit en utilisant des règles plus ou moins productives (à l'aide de grammaires comme celle de la figure 1), soit à l'aide de dictionnaires. Ainsi par exemple, le logiciel NooJ offre la possibilité de décrire les variantes du terme *tsar* à l'aide d'un dictionnaire comme le suivant :

tsar,N+Hum

csar,tsar,N+Hum

czar,tsar,N+Hum

tzar,tsar,N+Hum

Grâce à ces ressources linguistiques, toutes les variantes orthographiques peuvent être rassemblées avant de compter les occurrences, et les mesures statistiques qui en découlent sont alors bien plus précises.

[9] Mathieu-Colas (1990) avait déjà montré qu'environ 5 % des entrées des dictionnaires *Grand Dictionnaire Encyclopédique Larousse*, *Grand Robert* et *Trésor de la Langue Française* ne sont pas orthographiées de la même façon. Silberztein (2010a) montre par ailleurs que la réforme de l'orthographe de 1990 (dont les recommandations ne sont pas toujours suivies ni même connues) a introduit des règles sans définir précisément leur domaine de définition, ce qui ne va pas dans la direction d'une normalisation.

3. Morphologie

Les logiciels d'analyse statistique de textes qui traitent comme objets de base les formes graphiques sont condamnés à distinguer les variantes morphologiques comme par exemple :

manger, mange, mangera, mangeur, immangeable, remangerait, etc.

Les auteurs de certains de ces logiciels revendiquent ces distinctions ; ainsi par exemple, pour le logiciel Lexico, André Salem remarque que dans les textes de gauche des années 1970, on parle *des libertés* dans le cadre de la défense des droits au logement, à l'éducation, etc. tandis que le terme singulier *la liberté* apparaît dans des textes de droite pour désigner les concepts plus traditionnels de liberté démocratique, et de liberté d'entreprendre, cf. (Lamalle et al., 2002). Il s'agit donc de deux concepts différents, et il ne faut surtout pas fusionner leurs occurrences.

La remarque est pertinente car en effet, certains noms abstraits ne se mettent pas au pluriel de façon sémantiquement transparente, par exemple les termes *honneur* (vs. *les honneurs de la guerre*), *amour* (vs. *les amours de jeunesse*), etc. Mais la très grande majorité des noms abstraits, concrets, animés et humains se mettent au pluriel de façon totalement transparente, sans différence de sens autre que le nombre ; ainsi :

Une révolution → *deux révolutions*

Une table → *deux tables*

Une girafe → *deux girafes*

Un boulanger → *deux boulangers*

En vérité, la différence de sens entre *une révolution* et *deux révolutions* ne serait pas pertinente pour la plupart des politologues lorsqu'ils cherchent à évaluer l'état de la paix dans le monde, par exemple en analysant des archives de journaux. En ne rassemblant pas les occurrences singulières et plurielles des termes, on minimise leur fréquence, et donc leur importance potentielle. Une meilleure solution, proposée depuis longtemps par les linguistes, est de découpler les termes

singuliers de leurs formes plurielles lorsque leur sens est différent[10], ce qu'on peut représenter de la façon suivante :

liberté,N+Abst+f+s

libertés,N+Abst+f+p

révolution,N+Abst+f+FLX=TABLE

Ces trois entrées décrivent le fait que le nom « liberté » n'a pas de pluriel, le nom « libertés » n'a pas de singulier, tandis que le nom « révolution » a un singulier et un pluriel (paradigme flexionnel « FLX=TABLE », c'est-à-dire nom singulier qui prend un « s » au pluriel).

Un argument plus général pour ne pas rassembler les formes morphologiques consiste à dire qu'on ne veut pas perdre l'information morphologique lorsqu'on effectue des analyses. Un littéraire pourrait par exemple s'intéresser au changement de sens entre les formes *mangeons* et *mangez* par exemple, puisque la première forme correspond à la première personne (elle contient en quelque sorte le pronom « nous »), tandis que la seconde correspond à la deuxième personne (elle contient en quelque sorte le pronom « vous »). Mais dans ce cas, pourquoi alors fusionner les occurrences des cinq formes homographes *mange* :

> *(1) première personne du singulier, présent de l'indicatif, (2) première personne du singulier, présent du subjonctif, (3) deuxième personne du singulier, impératif, (4) troisième personne du singulier, présent de l'indicatif, (5) troisième personne du singulier, présent du subjonctif*

En fait, cela aurait plus de sens de rassembler les deux occurrences *mange* (4) et *mangeons* puisqu'elles sont souvent synonymes, par exemple dans la phrase :

On mange = Nous mangeons

et de distinguer la forme à l'impératif (ex. *Mange ta soupe !*) de la forme à la troisième personne du singulier (ex. *Elle mange ta soupe*) qui correspondent à des situations très différentes.

[10] Ce découplage est systématique dans les dictionnaires éditoriaux traditionnels, mais aussi dans les dictionnaires électroniques, cf. par exemple Courtois, Silberztein, éds (1990).

Un système fondé exclusivement sur des formes graphiques n'est donc pas cohérent d'un point de vue sémantique, et ne peut par conséquent pas produire des résultats fiables pour des applications dont le but est justement d'aider les chercheurs à extraire du sens de leur corpus.

Certains analyseurs statistiques de texte utilisent des lemmatiseurs. Par exemple, on peut connecter le logiciel Iramuteq[11] à un dictionnaire de formes fléchies, tandis que le logiciel TXM[12] utilise le logiciel de lemmatisation TreeTagger[13] pour lemmatiser les formes. Mais ces logiciels d'étiquetage et de lemmatisation statistique produisent des résultats notoirement insuffisants et très peu fiables[14].

Mais si un utilisateur cherche à étudier l'état de la paix sociale en analysant un corpus d'archives d'un quotidien sur plusieurs années, la simple lemmatisation ne suffit pas : il devra aussi tenir compte des dérivations dans son corpus, sinon les trois phrases suivantes ne seront par décomptées ensemble :

> ... *Voici les grèves, les* manifestations *et les rassemblements qui se tiennent à Paris...*
>
> www.evous.fr/Les-Manifestations-a-Paris-la-semaine-1176044.html
>
> ...*Les* manifestantes *ont rendu hommage à ces femmes tuées...*
>
> www.leparisien.fr/societe/paris-rassemblement-au-trocadero-pour-denoncer-le-100e-feminicide-de-l-annee-01-09-2019-8143560.php
>
> ... *Brest : les lycéens et les étudiants* remanifesteront *demain, vendredi...*
>
> rennes.maville.com/actu/actudet_-Brest-les-lyceens-et-les-etudiants-remanifesteront-demain-vendredi_-1548674_actu.Htm

Plus généralement, toutes les occurrences des formes *immangeable, mange, mangeurs, remanger* et *remangerait* doivent être rassemblées si l'on veut associer les phrases suivantes avec la même unité de sens *manger*.

---

[11] Cf. Loubère, Ratinaud (2014).

[12] Cf. Heiden (2010).

[13] Cf. Schmid (1994).

[14] Ils ne distinguent pas les différents types de noms (humain, concret, abstrait) et de verbes (intransitifs, transitifs directs, auxiliaires modaux, aspectuels, etc.) et ne traitent pas l'homographie. Silberztein (2018) a étudié les corpus *Open American National Corpus* (OANC) et *Corpus of Contemporary American English* (COCA) étiquetés avec Gate et montre qu'on y trouve de nombreuses fautes, comme par exemple *anomaly* étiqueté comme adverbe, *bible* étiqueté comme adjectif, *many* étiqueté comme verbe, etc. La non-fiabilité des étiqueteurs statistiques est en fait bien connue et a fait l'objet de nombreuses publications, ex. Green *et alii* (2010), Volokh *et alii* (2011), Dickinson *et alii* (2012), entre autres.

Quelle part est immangeable ↔ *Ce qui se* mange *ou non...*

toogoodtogo.fr/fr/movement/knowledge/qu-est-ce-que-gaspillage-alimentaire

Qui mange *le plus de fromage* ↔ *Les dix plus grands* mangeurs *de fromage...*

www.lexpress.fr/styles/saveurs/qui-mange-le-plus-de-fromage-dans-le-monde_1622803.html

... est celui qui l'a faite remanger *de la viande* = *Il* mangerait *tout ce que le chef...*

www.vonjour.fr/anne-hathaway-a-revele-quelle-ressentait-un-redemarrage-dordinateur-lorsquelle-a-remange-de-la-viande-apres-avoir-ete-vegetalienne/

De même si l'on s'intéresse à la place de la France dans le monde, la seule recherche de la forme « France » ne permettra de trouver qu'une infime partie des occurrences de ce concept. Dans la plateforme NooJ, on décrit des familles morphologiques à l'aide de graphes comme le suivant :

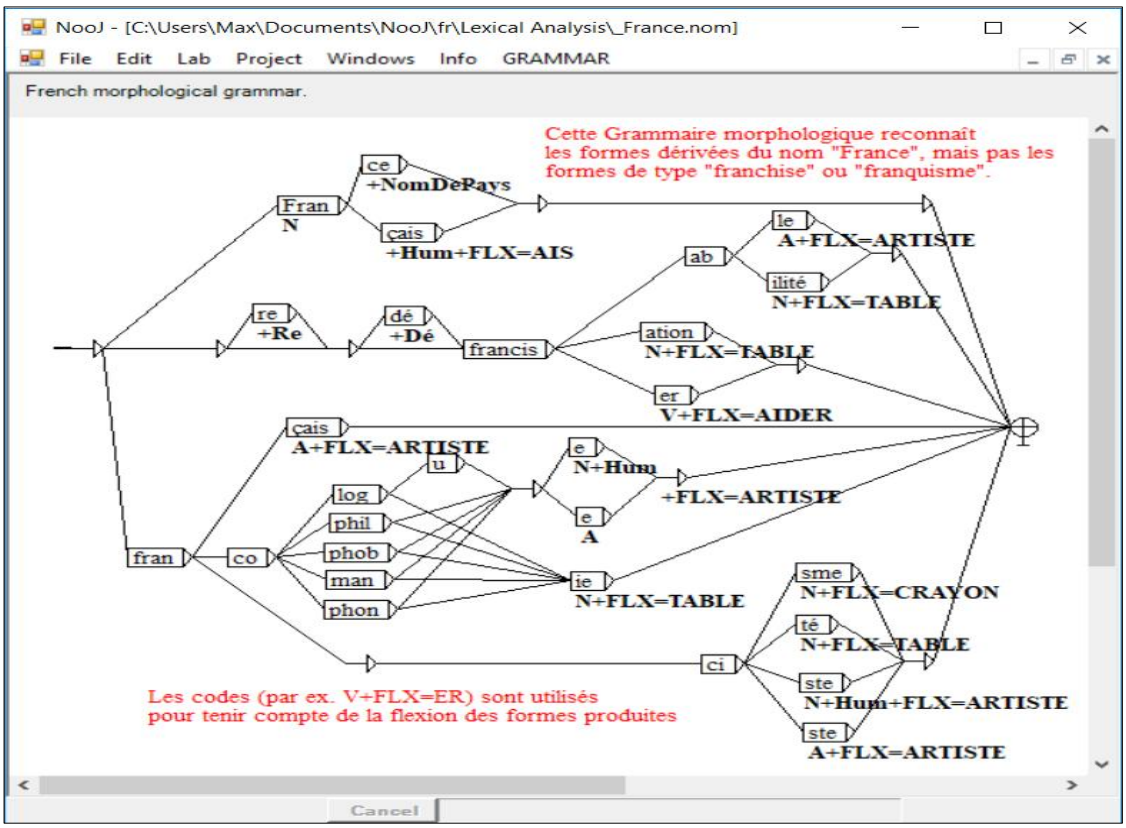


Figure 2. Grammaire qui représente les variantes graphiques du mot seigneur

Cette grammaire reconnaît une quarantaine de formes, dont *France*, *redéfranciser*, *francophone*, *francité*, etc. ainsi que toutes leurs formes fléchies (ex. *refranciserons*, *Françaises*), ce qui correspond à plus de trois cents formes

graphiques[15]. Les logiciels qui n'ont pas accès à ce type de description sont condamnés à diluer le concept *France* en comptant séparément les fréquences de plusieurs centaines de formes, ce qui le conduira immanquablement à sous-estimer la présence de ce concept dans les textes ; par exemple, dans le corpus du journal *Le Monde* (année 2006), la forme « France » apparaît 38.000 fois, tandis qu'elle apparaît 60.000 fois avec ses formes dérivées.

Notez que cette grammaire produit un résultat bien plus précis que ne le produirait un analyseur à base de racine[16] ; en effet, sans description linguistique précise, la recherche de l'affixe « fran » inclura parmi les occurrences du concept *France* des formes comme « franchir », « frangipane » ou « franchise ».

En unifiant les occurrences de la famille morphologique *France*, on peut compter la fréquence de cette unité sémantique, et donc retrouver cette thématique, analyser sa spécificité par rapport à d'autres thèmes, détecter si elle est corrélée à des termes positifs ou négatifs dans la presse francophone étrangère par exemple, si elle est plus présente dans des rubriques politiques ou sportives, compter ses occurrences dans les discours d'une personnalité politique, etc. de façon bien plus complète que si l'on avait cherché la seule forme « France ».

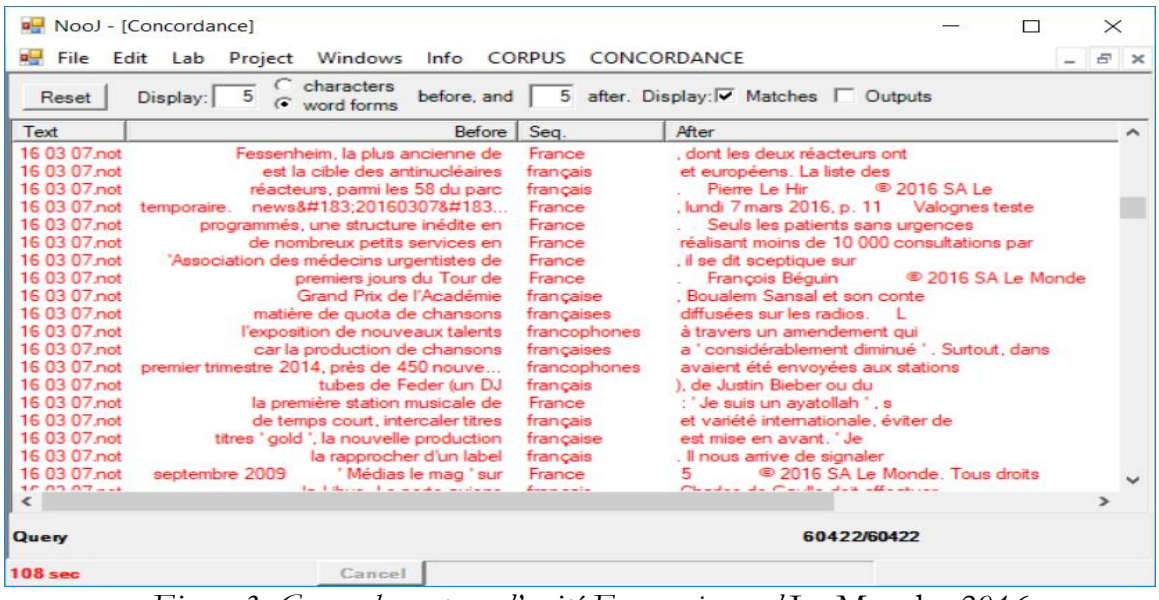


Figure 3. Concordance pour l'unité France, journal Le Monde*, 2016*

[15] Selon les besoins, on pourrait y ajouter le nom *franchouillard*.

[16] Le logiciel Alceste calcule la « forme réduite » (ou racine) des mots d'un texte à l'aide de simples algorithmes génériques, cf. Reinert (1999).

4. Champs lexico-sémantiques

Pour aider les utilisateurs à analyser leurs corpus de textes, la plupart des logiciels d'analyse automatique leur offrent la possibilité de calculer et/ou de construire des champs lexico-sémantiques, c'est-à-dire des familles de termes proches sémantiquement. Par exemple, Sketchengine[17] :

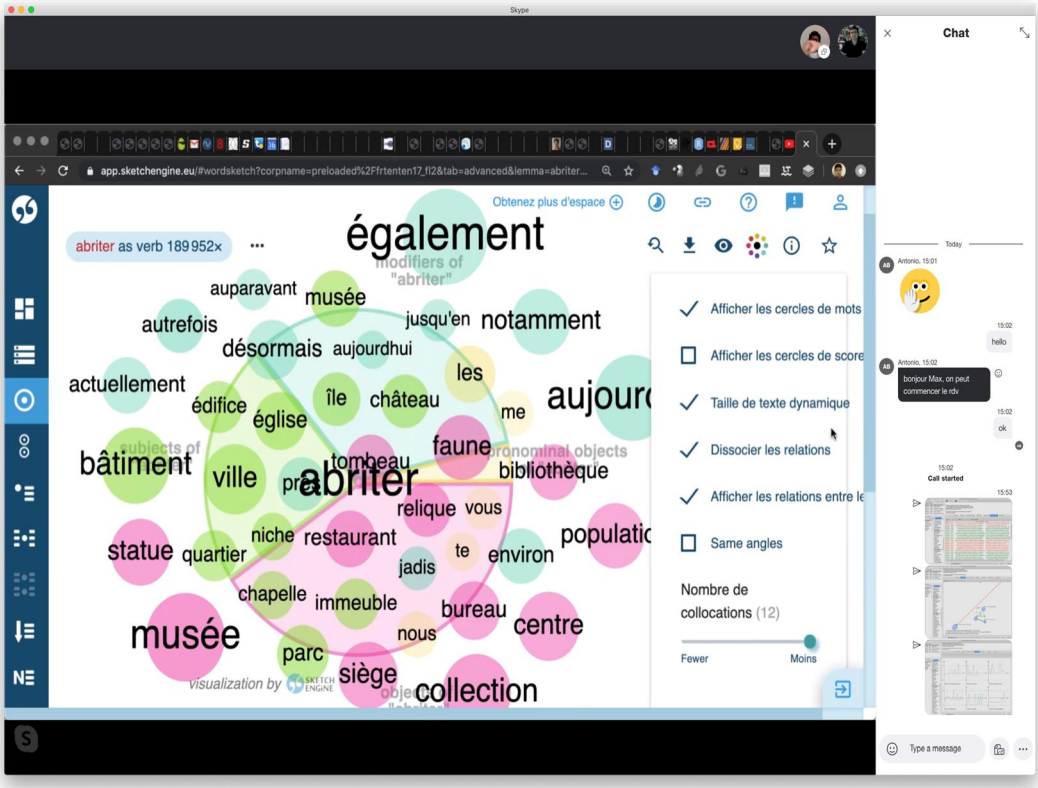


Figure 4. Un champ lexico-sémantique dans Sketch Engine

Mais les champs lexico-sémantiques calculés par ces logiciels ne sont pas fiables ; par exemple, dans la figure précédente, le logiciel nous propose les termes *jadis, environ, relique* et *siège* dans le champ « abriter ». Les champs lexicaux dépendent de chaque domaine, voire même des besoins spécifiques de chaque utilisateur ; par ex. on peut imaginer qu'un psychologue veuille lier les verbes *aimer* et *haïr* pour une analyse d'entretiens, tandis que ces deux verbes ne seraient pas rassemblés par un politologue qui analyserait les discours de personnalités politiques. Il est donc préférable de donner à chaque utilisateur la possibilité de construire ses propres champs lexico-sémantiques. Ainsi, le logiciel NooJ

[17] Cf. Kilgariff (2004).

donne aux utilisateurs la possibilité de construire des bibliothèques de requêtes comme la suivante[18] :

MORT = *<agoniser>* | *<assassiner>* | *<décéder>* | *<étrangler>* | *<massacrer>* | *<mourir>* | *<périr>* | *<succomber>* | *<suicider>* | *<tuer>* | *funérailles* | *<funéraire>* | *<cadavre>* | *<cimetière>* | *décès* | *<défunt>* | *<deuil>* | *<homicide>* | *<meurtre>* | *obsèques* | *<tombeau>* | *<veuf>* | *<orphelin>* | *<fatal>* | *<morbide>*.

En appliquant ce champ lexical à un corpus, on peut détecter des fréquences anormalement élevées de termes « morbides » dans des entretiens psychologiques ou dans les discours d'une personnalité politique, retrouver des cooccurrences ou oppositions intéressantes avec d'autres champs lexicaux (par ex. opposition amour/mort dans des poèmes), etc. La figure suivante montre la mesure du score standard[19] de ce champ lexical dans les romans de la série *Les Rougon-Macquart* d'Émile Zola (1871-1883).

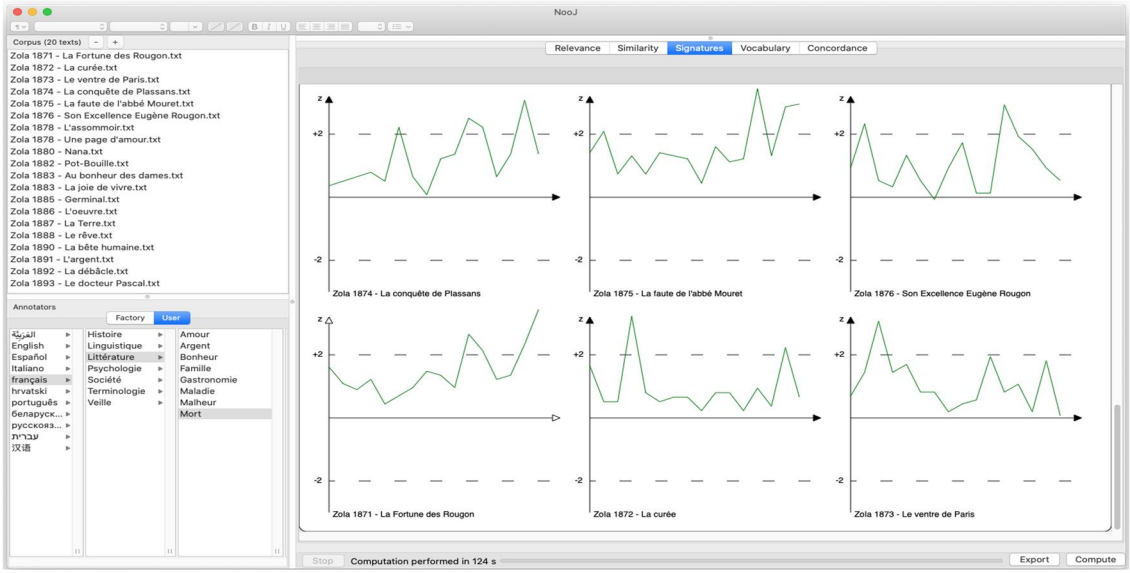


Figure 5. Analyse thématique de la série Les Rougon-Macquart

[18] Dans le logiciel NooJ, la notation entre crochets (ex. <mourir>) permet de représenter toutes les formes associées à une entrée lexicale : variantes orthographiques, flexionnelles (ex. *mourra*) et dérivationnelles (ex. *mourir* → *mort, mortel, mortellement*). Dans une requête NooJ, on peut décrire et exclure des contextes afin d'écarter les homographes, par exemple, lorsqu'on veut la forme « tombe » mais pas dans le contexte verbal « Il tombe ».

[19] Le score standard permet de détecter les différences « anormales » entre la fréquence attendue et la fréquence observée. Les limites en pointillé correspondent à deux écarts-types au-dessus ou en dessous de la fréquence attendue, qui correspondent à une probabilité de 2,5 % ; en d'autres termes, à une fréquence « intéressante ».

On voit ainsi qu'il est question de mort au début des romans *la Curée* et *Le ventre de Paris*, au contraire des autres romans qui se terminent mal.

5. Homographie

Dans les sections précédentes, nous avons vu qu'il faut souvent rassembler un plus ou moins grand nombre de formes graphiques pour traiter correctement certaines unités de sens. L'inverse est aussi vrai : la plupart des formes graphiques correspondent en fait à plusieurs unités de sens. Par exemple, la forme *volé* dans les trois phrases suivantes correspond à trois unités de sens différentes :

Sens 1 : ... *Mobile* volé. *Si vous perdez votre téléphone ou votre clé 3G/4G* ...

www.orange.ma/Assistance/Mobile-vole2

Sens 2 : ... *Malgré le temps, tous les avions ont* volé ...

www.leparisien.fr/essonne-91/malgre-le-temps-tous-les-avions-ont-vole-20-05-2013-2817219.php

Sens 3 : ... *on nous a* volé *de 100 euros* ...

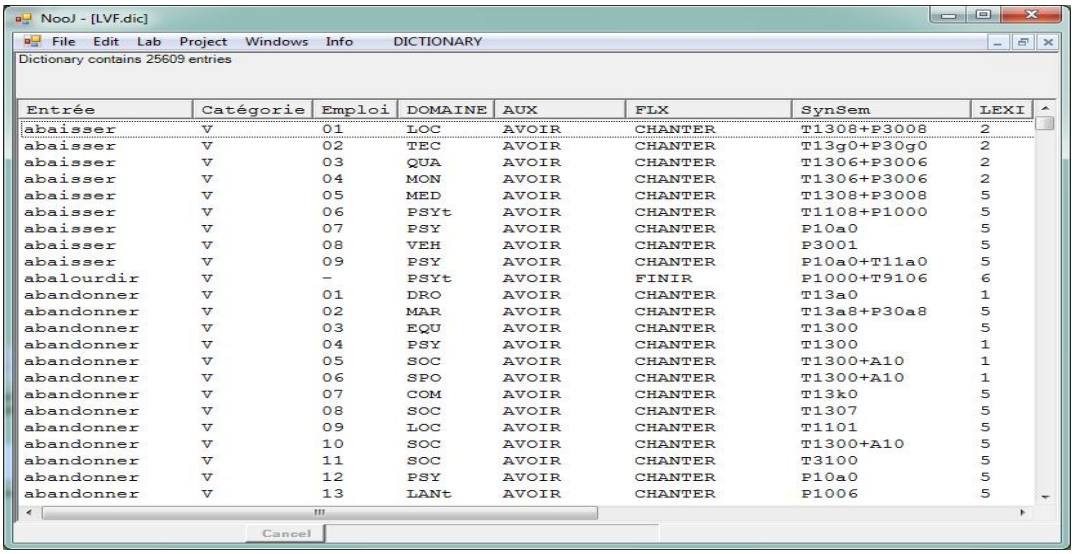
www.agoda.com/fr-fr/sawasdee-village-resort-spa/reviews/phuket-th.html?cid=-218

Dans quelques cas très rares, la morphologie permet de distinguer certaines unités de sens ; par exemple, la forme *volées* ne peut pas représenter le sens 2 puisque celui-ci correspond à une construction intransitive.

Pour pouvoir distinguer les différents sens que peut prendre une forme, il faut pouvoir accéder à un dictionnaire qui les recense. En effet, si un logiciel ne sait même pas que la forme *voler* peut correspondre à plusieurs sens, il ne pourra que produire des résultats d'analyse statistique totalement inutiles, par ex. « le verbe *voler* est plus utilisé par Georges Simenon que par Antoine de Saint-Exupéry » (?!).

Le choix de Jean Dubois et Françoise Dubois-Charlier dans leur dictionnaire *Les Verbes Français* (LVF) a été de séparer chaque homographe, afin que chaque entrée du dictionnaire *LVF* corresponde à un seul sens. Ce dictionnaire, qui répertorie 25.000 emplois verbaux, se présente comme une vaste classification dont chaque classe et sous-

classe est définie par un ensemble de propriétés formelles corrélé à un sens fondamental[20].


NooJ - [LVF.dic]

File Edit Lab Project Windows Info DICTIONARY

Dictionary contains 25609 entries

Entrée	Catégorie	Emploi	DOMAINE	AUX	FLX	SynSem	LEXI
abaisser	V	01	LOC	AVOIR	CHANTER	T1308+P3008	2
abaisser	V	02	TEC	AVOIR	CHANTER	T13g0+P30g0	2
abaisser	V	03	QUA	AVOIR	CHANTER	T1306+P3006	2
abaisser	V	04	MON	AVOIR	CHANTER	T1306+P3006	2
abaisser	V	05	MED	AVOIR	CHANTER	T1308+P3008	5
abaisser	V	06	PSYt	AVOIR	CHANTER	T1108+P1000	5
abaisser	V	07	PSY	AVOIR	CHANTER	P10a0	5
abaisser	V	08	VEH	AVOIR	CHANTER	P3001	5
abaisser	V	09	PSY	AVOIR	CHANTER	P10a0+T11a0	5
abalourdir	V	-	PSYt	AVOIR	FINIR	P1000+T9106	6
abandonner	V	01	DRO	AVOIR	CHANTER	T13a0	1
abandonner	V	02	MAR	AVOIR	CHANTER	T13a8+P30a8	5
abandonner	V	03	EQU	AVOIR	CHANTER	T1300	5
abandonner	V	04	PSY	AVOIR	CHANTER	T1300	1
abandonner	V	05	SOC	AVOIR	CHANTER	T1300+A10	1
abandonner	V	06	SPO	AVOIR	CHANTER	T1300+A10	1
abandonner	V	07	COM	AVOIR	CHANTER	T13k0	5
abandonner	V	08	SOC	AVOIR	CHANTER	T1307	5
abandonner	V	09	LOC	AVOIR	CHANTER	T1101	5
abandonner	V	10	SOC	AVOIR	CHANTER	T1300+A10	5
abandonner	V	11	SOC	AVOIR	CHANTER	T3100	5
abandonner	V	12	PSY	AVOIR	CHANTER	P10a0	5
abandonner	V	13	LANt	AVOIR	CHANTER	P1006	5

Cancel

Figure 6. Extrait du dictionnaire Les Verbes Français

Reprenons l'exemple du mot *abriter* de la figure 4. En fait, ce verbe correspond aux cinq entrées lexicales distinctes suivantes dans *LVF* :

- *abriter*$_1$: *Luc abrite Léa de la pluie avec un parapluie*
- *abriter*$_2$: *Luc abrite des réfugiés chez lui*
- *abriter*$_3$: *Cet immeuble abrite le service juridique*
- *abriter*$_4$: *Luc s'abrite des ennuis derrière son chef*
- *abriter*$_5$: On *abrite le port des vagues avec des digues*

On peut se demander l'intérêt de présenter à l'utilisateur le « champ lexico-sémantique » du mot *abriter* si ce dernier a des sens si différents, par exemple dans :

*... Bayrou s'*abrite *derrière l'indépendance des juges pour ne pas commenter...*

www.lefigaro.fr/politique/2017/06/01/01002-20170601ARTFIG00224-bayrou-s-abrite-derriere-l-independance-de-la-justice-pour-ne-pas-commenter-l-affaire-ferrand.php

et :

[20] Cf. Dubois, Dubois-Charlier (1997). Pour une présentation générale de ce dictionnaire, voir par exemple François *et alii* (2007) ou Leeman, Sabatier (2010). Le dictionnaire LVF a été formalisé sous la forme d'un dictionnaire NooJ, cf. Silberztein (2010b).

... Elle abrite *des personnes traquées dans son appartement...*

fr.wikipedia.org/wiki/Odette_Pilpoul

sous peine de produire des analyses imprécises, inutilisables, voire même à contresens.

Noms homographes ou polysémiques

L'homographie est aussi massive pour les noms. Par exemple, considérons les sens du nom *carte* :

- **carte de paiement :** carte d'abonnement, carte American Express, carte bancaire, carte bancaire prépayée, carte bleue, carte de cantine, carte sans contact, carte de crédit, carte à débit différé, carte de débit immédiat, carte Diners, carte de fidélité, carte JCB, carte jeune, carte MasterCard, carte de paiement virtuelle, carte de réduction, carte de rationnement, carte de restaurant, carte sénior, carte téléphonique, carte vermeil, carte Visa, e-carte bleue, mandat-carte ;

- **carte géographique :** carte aéronautique, carte astronomique, carte du ciel, carte d'état-major, carte fluviale, carte géologique, carte des marées, carte marine, carte routière, carte sanitaire, carte scolaire, carte sectionnelle ;

- **carte d'identité :** carte électorale, carte d'étudiant, carte du parti, carte de presse, carte professionnelle, carte de résident, carte de travail, carte de visite, carte vitale, carte grise, carte navigo, carte d'embarquement, carte orange ;

- **carte à jouer :** carte d'atout, carte de Tarot ;

- **divers :** carte d'accès, carte à puces, (avoir | donner) carte blanche, carte magnétique, carte mémoire, carte-mère, carte postale, carte des vins, carte de vœux.

Parmi les « cartes colorées » par exemple, une *carte bleue* est associée à un crédit auprès d'une banque, une *carte orange* était associée à un abonnement RATP, une *carte vermeil* est associée à un âge, une *carte grise* est associée à un véhicule, une *carte blanche* signifie faire confiance :

impossible de déduire la différence de sens entre ces termes à partir des propriétés optiques des couleurs *bleue*, *orange*, *gris, vermeil ou blanc*.

Parmi les « cartes de paiement », les cartes bancaires sont plus générales que les cartes restaurant ; les cartes American Express et Diners ne fonctionnent pas de la même façon ; ni, de leur côté, les cartes d'abonnement, de réduction et de fidélité qui ont toutes des fonctionnements différents ; les cartes de rationnement n'ont pas les mêmes contraintes d'utilisation que les cartes de restaurant.

Parmi les « cartes d'identification », une *carte d'identité* n'est pas du tout utilisée comme une *carte de visite* : impossible de déduire à partir des mots « identité » et « visite » laquelle des deux cartes a un rapport avec le monde des affaires. Une *carte vitale* n'est pas vraiment vitale.

Parmi les « cartes géographiques », une *carte routière* ne s'utilise pas de la même façon qu'une *carte sectionnelle* et ces deux cartes ne se ressemblent pas : les *cartes sectionnelles* sont utilisées pour le vol à vue (VFR) et comportent donc des points de repère codés, des altitudes à respecter, des classes d'espace aérien, etc. Les *cartes scolaires* ne sont pas utilisées pour se déplacer dans les écoles (elles sont utilisées pour l'affectation des élèves), et les *cartes sanitaires* ne localisent pas les sanitaires (elles situent les offres de soin de santé).

Parce que la forme graphique *carte* correspond à tant de sens si différents, la traiter comme une unité de traitement à part entière pour en compter ses occurrences, rechercher sa « spécificité » par rapport à un corpus ou à un domaine, en rechercher les termes co-occurrents ou opposés, proposer un nuage de mots proches, etc. n'a aucun intérêt : que pourrait en déduire un politologue si un logiciel « découvre » que le mot *carte* est utilisé plus fréquemment dans des journaux politiques de gauche (*carte du parti*) que dans des magazines d'informatique (*carte mère*) ?

De la même façon, un analyseur qui ne prend en compte que des formes simples risque de « détecter » que la forme *coup* a une fréquence anormalement élevée dans un quotidien, ce qui pourrait suggérer que le texte a un contenu violent, alors qu'en fait cette forme n'apparaîtrait que dans *tout à coup, coup de pouce* (à propos d'une aide au logement), *coup de soleil* (au sujet des vacances), *coups de cœur* (rubrique « cinéma »), *coup de pinceau* (une exposition) *coup du berger* (rubrique « Échecs »).

L'approche linguistique est bien sûr de considérer que ces séquences monosémiques sont toutes des unités linguistiques atomiques, éléments du vocabulaire qu'il faut répertorier dans un dictionnaire[21].

Si l'on se base sur les dictionnaires du système DELA, le vocabulaire français standard contient environ 35.000 noms simples pour plus de 200.000 mots composés[22]. Les logiciels qui ne traitent pas les mots composés comme des unités atomiques doivent alors traiter chaque nom simple comme ayant en moyenne cinq sens différents, sous peine d'unifier dans leurs calculs des occurrences d'unités de sens qui n'ont rien à voir entre elles.

Désambiguïsation grâce au contexte

Un contre-argument sur la nécessité de traiter les mots composés comme des unités atomiques serait de dire que les formes simples sont naturellement polysémiques, mais que leur contexte permet de les désambiguïser. De la même façon qu'on peut lever l'ambiguïté du verbe *voler* dans la phrase « On nous a volé de 100 euros » (puisque cette phrase est transitive directe avec un complément direct humain et un complément en *de*), on pourrait lever l'ambiguïté du nom *carte* simplement en examinant son contexte immédiat : s'il a bien une cinquantaine de sens potentiels, en revanche, suivi du mot « routière », il ne représenterait plus qu'une seule unité de sens.

Cet argument ne serait valable que si la forme « routière » est elle-même monosémique. Or, cette forme est aussi polysémique comme on le voit en essayant de définir son sens dans les contextes suivants :

gare routière, infraction routière, moto routière, prévention routière, etc.

On peut expliquer ce qu'est une *gare routière* en disant que c'est une gare (implicitement ferroviaire) dans laquelle on a remplacé les trains par des

[21] Pour la définition des mots composés utilisés pour construire le dictionnaire DELAC, voir Silberztein (2015). Cette définition est basée sur trois critères : atomicité sémantique (*carte bancaire* ne peut pas désigner une *carte* qui situe les banques) ; usage (on dit *machine à laver*, et non *laveuse automatique de vêtements*) ; exception grammaticale (*voyage présidentiel* = le président voyage, mais *élection présidentielle* ≠ le président élit).

[22] Cf. Courtois, Silberztein (1990). Le vocabulaire standard correspond au vocabulaire partagé par les locuteurs francophones à l'exclusion des termes régionaux (par exemple du français du Jura, du Nord, du Québec, du Mali, etc.) et des millions de termes scientifiques (ex. *réaction chélotrope*) et techniques (ex. *dermite cortico-induite*).

autocars ; une *infraction routière* est une infraction au code de la route ; une *moto routière* s'oppose à une moto-cross par ses caractéristiques de tenue de route ; la *prévention routière* est liée aux accidents de la circulation des véhicules.

Le mot « carte » correspond donc à une cinquantaine de sens et le mot « routière » a lui-même une demi-douzaine de sens, donc la séquence « carte routière » a potentiellement $50 \times 6 = 300$ sens : aucune règle morphologique, syntaxique ni même sémantique ne s'opposerait en effet à ce que le terme *carte routière* représente :

- une carte qui situe les arrêts d'autocar (un peu comme les cartes de métro),
- ou alors une carte qui représente l'état de la circulation à un moment donné (comme les cartes présentées par « Bison futé » au journal télévisé),
- ou alors une sorte d'identifiant (comme une carte grise) pour habiliter certains véhicules à circuler sur les autoroutes (ex. les motos routières) par opposition aux véhicules (ex. les moto-cross) qui ne le sont pas,
- ou alors une carte de paiement à distance que l'on fixerait sur le parebrise de son automobile pour accéder rapidement à l'entrée des autoroutes à péage,
- ou alors une carte plastifiée que les agents de la circulation utiliseraient comme memento, dans lequel seraient recensées chaque infraction routière et sa contravention correspondante,
- etc.

La forme *carte* correspond donc à une cinquantaine d'unités de sens, la forme *routière* a elle aussi une demi-douzaine de sens, mais aucun locuteur francophone ne bute, ne serait-ce un instant, sur les sens multiples que cette combinaison pourrait représenter. Si je demande dans une banque à obtenir une carte bancaire, aucun employé ne pensera, ne serait-ce qu'une demi-seconde, à me donner le dépliant qui situe toutes les agences de la banque dans la ville.

La séquence *carte routière* n'a en vérité qu'un seul sens, et ce sens n'est jamais « calculé » par les locuteurs francophones, qui connaissent ce terme

« par cœur ». Il s'agit d'un élément du vocabulaire français. Les logiciels d'analyse automatique doivent traiter cette séquence comme une unité linguistique atomique, exactement comme le nom *maison*, surtout sans chercher à la découper.

En conclusion : un grand nombre de mots simples sont en fait des constituants de termes composés ; lors de l'analyse des textes, il ne faut pas ignorer ces termes composés (pour la plupart non-ambigus), sous peine d'avoir à gérer une polysémie massive des formes simples.

6. Règle d'identité

Gross (1988) nomme ainsi le fait que de nombreux noms composés acceptent l'effacement d'un ou de plusieurs de leurs constituants[23] :

une association caritative est une association ; le ministre de la culture est un ministre

Cette règle est systématiquement utilisée pour alléger les textes, comme on le voit dans un article du journal *Le Monde* (mars 2019) :

> *Venu dans la capitale lombarde à l'occasion de l'inauguration de la Triennale, le* ministre *français* de la culture*, Franck Riester, a rencontré son homologue italien, Alberto Bonisoli, pendant près de deux heures... Les deux* ministres *ont affiché une volonté renouvelée de collaborer...*

Si un analyseur automatique a accès à la liste des noms composés et qu'il sait que « ministre de la culture » correspond à un terme qui peut être abrégé en « ministre », alors une simple exploration du contexte lui permettra de relier automatiquement l'occurrence du mot simple « ministre » au terme explicite[24], et donc compter les occurrences du terme, même lorsqu'il est dans sa forme abrégée.

Malheureusement, les analyseurs qui se contentent de compter les occurrences de la forme simple « ministre » sans chercher à les relier aux termes plus explicites sont incapables de discriminer les différents référents du mot *ministre* dans les textes analysés (par exemple *ministre du culte* vs. *ministre de la culture*).

[23] Ce phénomène est un cas particulier des anaphores fidèles ; au sujet des anaphores, voir par exemple Kleiber (1988).

[24] Il faut bien sûr désactiver la règle d'explicitation *ministre* → *ministre de la culture* dès qu'un autre terme en *ministre* apparaît, par exemple *ministre des affaires étrangères*.

7. Collocations

Les spécialistes de TAL et auteurs de logiciels d'analyse de textes savent qu'il est nécessaire de traiter certains termes composés comme des unités en bloc. Mais, plutôt que de construire des dictionnaires ou même d'utiliser les ressources linguistiques déjà existantes[25], la plupart d'entre eux préfèrent tenter de détecter automatiquement ces mots composés en exploitant des corpus d'apprentissage grâce à des méthodes stochastiques (probabilistes, statistiques ou réseaux de neurones)[26].

L'idée de base de ces méthodes est d'assimiler les mots composés à des collocations, c'est-à-dire à des paires de mots dont les fréquences sont corrélées. Ainsi, en détectant automatiquement dans un corpus d'apprentissage que les mots *plant* et *nuclear* apparaissent ensemble de façon anormalement fréquente, on en déduit que *nuclear plant* est un terme du domaine[27].

L'idée de repérer des collocations en détectant des fréquences liées pour les constituants d'un mot composé va à l'encontre de l'utilisation de la règle d'identité vue dans la section précédente : si le terme « ministre de la culture » n'apparaît qu'une seule fois explicitement dans un texte (par ex. dans le titre d'un article), et qu'il est suivi par une dizaine d'occurrences de sa forme abrégée « ministre » (au sein de l'article), les méthodes automatiques ne pourront pas détecter que « ministre » et « culture » constituent une collocation, donc un terme composé.

En fait, il est illusoire de chercher à retrouver les mots composés dans des corpus de textes, et ce quel que soit leur taille. Par exemple, si un annotateur lisait minutieusement l'ensemble des textes du journal *Le Monde*, 2006 (qui contient environ vingt-trois millions de formes ; à 1 seconde par mot, cela représenterait trois ans de travail ininterrompu), il ne trouverait qu'environ 20 % des noms composés du vocabulaire standard. Des termes comme :

> *abandon de navire, accident de terrain, admission sur concours, aéronautique militaire, affaire d'honneur, agence d'assurance, aiguillage ferroviaire*, etc.

[25] Cf. par exemple Courtois & Silberztein (1990).

[26] Par exemple, l'atelier MWE 2008 (http://multiword.sf.net) utilisait des techniques d'intelligence artificielle pour calculer la probabilité qu'une séquence de mots correspondent à un mot composé.

[27] Cf. par exemple Gale, Church & Yarowsky (1986). Ce principe est toujours suivi, cf. par exemple Ramisch *et alii* (2018).

n'ont pas une seule occurrence dans ce corpus. Et les techniques semi-automatiques seraient de plus confrontées au fait que parmi les termes qui apparaissent, la moitié ne se présentent qu'une seule fois, ce qui les empêcherait de les détecter, par exemple :

> *bac scientifique, bec-de-lièvre, bien de consommation, bleu de cobalt, bol de lait*, etc.

De plus, la très grande majorité des noms composés qui se montrent plus d'une fois n'apparaissent en fait que dans une seule de leurs formes, au masculin mais pas au féminin, ou alors au féminin mais pas au masculin, au singulier mais pas au pluriel, ou alors au pluriel mais pas au singulier, par exemple :

> *cabinets comptables, centres de dépistage, chaînes de solidarité, circonscriptions territoriales, cloisons vitrées, collisions en chaînes*, etc.

On ne voit pas comment un analyseur automatique pourrait utiliser ces occurrences pour en déduire quels termes acceptent les deux genres ou sont obligatoirement masculins ou féminins, quels termes acceptent les deux nombres ou sont obligatoirement singuliers, ou obligatoirement pluriels[28].

Mesurer des fréquences de séquences ne peut pas être un critère direct pour détecter la présence de collocations puisqu'un grand nombre de séquences (par ex. « *est dans la* ») sont bien plus fréquentes dans les textes que la grande majorité des mots composés.

D'un point de vue linguistique, le principe même de la notion de collocations est contestable. Après tout, personne n'a besoin de calculer les fréquences relatives des syllabes [mɛ] et [zɔ̃] dans un corpus oral pour en déduire que le mot *maison* fait partie du vocabulaire français : et si cette séquence de deux syllabes n'apparaissait qu'une seule fois dans un corpus oral de taille considérable, ce ne serait pas un critère pour rejeter le mot « maison » du vocabulaire français. En vérité, les éléments qu'il faut traiter sont des unités de sens qui ne sont pas définies par des critères statistiques mais par des critères linguistiques.

[28] Ce problème est plus sérieux pour les langues germaniques, sémitiques ou slaves qui ont un système flexionnel bien plus riche que le français : on rencontrera beaucoup plus d'hapax dans un corpus allemand ou polonais que dans un corpus français.

8. Conclusions

Les logiciels de traitement automatique de la langue naturelle ne peuvent accéder aux textes qu'en lisant des séquences de caractères rangés dans un fichier informatique : ils ne « voient » que des formes graphiques. Or l'orthographe n'a un lien que très indirect avec le sens porté par les énoncés, et elle n'est même pas fiable.

Établir le lien entre les formes graphiques apparaissant dans les fichiers-textes et les unités de sens (concepts, entités, prédicats et relations) qui constituent les informations qui intéressent les utilisateurs de ces logiciels n'est pas simple et nécessite des méthodes et ressources linguistiques à plusieurs niveaux : orthographique, morphologique et syntaxique.

La plateforme d'analyse linguistique NooJ a été conçue pour répondre à ces besoins.

Références bibliographiques

AOUINI, M., *Outils d'analyse de corpus du moyen-français*, Thèse de doctorat, Université de Franche-Comté, 2018.

BLUMENTHAL, P., VIGIER, D., Du quantitatif au qualitatif en diachronie : prépositions françaises, *LANGAGES*, 2017, **206**, Armand Colin.

BRUNET, É., *HYPERBASE : Manuel de référence*, ⟨hal-01362721⟩, 2010.

COURTOIS, Bl., SILBERZTEIN M., éds., *Dictionnaires électroniques du français : le système DELA*, *LANGUE FRANÇAISE*, 1990, **87**.

DICKINSON, M., LEDBETTER, S., Annotating Errors in a Hungarian Learner Corpus, LREC. 2012.

DUBOIS J., DUBOIS-CHARLIER, Fr., *Les Verbes Français,* Paris, Larousse-Bordas. Téléchargeable à l'adresse : www.modyco.fr/fr/15-modyco/ressources.html, 1997.

FRANÇOIS, J., LE PESANT, D., LEEMAN, D., Présentation de la classification des Verbes Français de Jean Dubois et Françoise Dubois-Charlier. *LANGUE FRANÇAISE*, 2007, **153**, 3-19.

GALE, W., CHURCH, K., YAROWSKY D., Work on Statistical Methods for Word Sense Disambiguation, In AAAI Technical Report FS-92-04. Disponible à l'adresse : www.aaai.org/Papers/Symposia/Fall/1992/FS-92-04/FS92-04-008.pdf, 1992.

GREEN, S., MANNING, Ch. D., Better Arabic parsing: Baselines, evaluations, and analysis, *Proceedings of the 23rd International Conference on Computational Linguistics.* Association for Computational Linguistics, 2010.

GROSS, G., Degré de figement des noms composés, *LANGAGES Les expressions figées*. 1988, **90**, 57-72.

HEIDEN, S., MAGUE, J.-P., PINCEMIN B. TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement, *10th International Conference on the Statistical Analysis of Textual Data* - JADT 2010, June 2010, Rome, Italie. 1021-1032, ⟨halshs-00549779⟩, 2010.

KILGARRIFF A. RYCHLY, P., SMRZ, P., TUGWELL, D., The sketch engine, *INFORMATION TECHNOLOGY*, 2004, **105**, 116.

KLEIBER, G., Peut-on définir une catégorie générale de l'anaphore ? *VOX ROMANICA*, 1988, **47-1**.

LAMALLE, C., MARTINEZ, W., FLEURY, S., SALEM, A., FRACCHIOLLA, B., KUNCOVA, A., MAISONDIEU, A., *Lexico 3, Outils de statistique textuelle. Manuel d'utilisation,* Paris, Université de la Sorbonne Nouvelle, 2002.

LEEMAN, D., SABATIER P., *Empirie, Théorie, Exploitation : le travail de Jean Dubois sur les verbes français*, *LANGAGES,* 2010, **179-180**.

LOISEAU, S., *La fréquence textuelle : bilan et perspectives*, *LANGAGES,* 2015, **197**.

LOUBERE, L., RATINAUD, P., *Documentation IRaMuTeQ. Manuel d'utilisation* disponible à l'adresse :
http://iramuteq.org/documentation/html, 2014.

MARTINET, A., Le mot, *Problèmes du langage*, Paris, Gallimard, coll. "Diogène", 39-53, 1966.

MATHIEU-COLAS, M., Orthographe et informatique : établissement d'un dictionnaire électronique des variantes orthographiques, *LANGUE FRANÇAISE,* 1990, **87**, pp 104-111.

NEVEU, Fr., *Structures de la phrase en français moderne,* Paris, Université Paris-Sorbonne éds., 2011.

RAMISCH, C., RICARDO CORDEIRO, S., SAVARY, A., VINCZE, V., BARBU MITITELU, V., BHATIA, A., BULJAN, M., CANDIT, M., GANTA, P., GIOULI, V., GÜNGÖR, T., HAWWARI, A., IÑURRIETA, U., KOVALEVSKAITE, J., KREK, S., LICHTE, T., LIEBESKIND, C., MONTI, J., PARRA ESCARTIN, C., QASEMIZADEH, B., SCHNEIDER, N., STOYANOVA, I., VAIDYA, A., WALSH, A., *Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions*. LAW-MWE-CxG@COLING, 222-240, 2018.

RASTIER, Fr., *La mesure et le grain : sémantique de corpus*, Paris, Champion ; diff. Slatkine, 2011.

RATINAUD, P., IRAMUTEQ : Interface de R pour les analyses multidimensionnelles de textes et de questionnaires. Computer Software). Acesso em vol.15, 2009.

REINERT, M., Quelques interrogations à propos de l' « objet » d'une analyse de discours de type statistique et de la réponse "Alceste", *LANGAGE & SOCIETE*, 1999, **90**, 57-79.

SCHMID, H., *TreeTagger: a language independent part-of-speech tagger*. www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger. Stuttgard, Allemagne, 1994.

SILBERZTEIN, M., Orthographe compliquée ou orthographe fidèle ? *LE FRANCAIS AUJOURD'HUI*, 2010a, **3**, 83-98.

SILBERZTEIN, M., La formalisation du dictionnaire *LVF* avec NooJ et ses applications pour l'analyse automatique de corpus, *LANGAGES* 2010b, **179-180**, 221-241.

SILBERZTEIN, M., *La formalisation des langues : l'approche de NooJ*, Londres, ISTE (425 pages), 2015.

SILBERZTEIN, M., Using linguistic resources to evaluate the quality of annotated corpora, In *Proceedings of the LR4NLP Workshop at COLING2018*. www.aclweb.org/anthology/W18-38, 2018.

VOLOKH A., NEUMANN G., Automatic detection and correction of errors in dependency tree-banks, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* 2011, vol. 2, 346-350.