Histoire Épistémologie Langage
38/2 (2016), p. 87-99

DOI : 10.1051/hel/2016380207

Disponible en ligne sur
www.hel-journal.org

Une linguistique outillée, pour quels objets ?

Marie-Paule Jacques
Université Grenoble Alpes, UJF / LIDILEM

Résumé

Nous prenons dans cet article le parti d'interroger non les apports des corpus en tant que tels à la recherche linguistique et à la constitution de nouveaux savoirs mais la nature des objets de recherche et le type de recherches induits par une linguistique de corpus, plus précisément une linguistique outillée. Nous défendons l'idée que l'outillage même de la linguistique oriente la recherche vers les formes, selon une approche sémasiologique. Toutefois, nombre de recherches en corpus visent à cerner les expressions linguistiques de certaines significations, selon une approche onomasiologique, tout en utilisant des outils et des apports du TAL pour ce faire. Nous en tirons des propositions méthodologiques pour accroitre la valeur scientifique des recherches.

Mots-clés

Approche onomasiologique, objets linguistiques, outils, méthodologie, TAL

Abstract

In this article, I question the relationship between corpus linguistics and the kind of research objects that a computer-based linguistics deals with. I try to show that using software to carry out linguistic investigations influences the kind of objects linguists are studying. Computers help to grab forms, so linguists tend to take forms as a point of departure and investigate their meaning(s). Another orientation exists, from meaning to the expressions that express this given meaning, and it may also use software tools. In this case, in order to limit subjectivity, I propose to adopt a methodology which combines a measure of judgments about the collected data and in-corpus annotation of the phenomena under study.

Keywords

Corpus linguistics, tools, methods, NLP

1. INTRODUCTION

Nous écrivions, il y a quelque dix ans (Jacques 2005), à la suite de Habert *et al.* (1997), qu'il n'y a pas *une* linguistique de corpus mais des linguistiques de corpus. Ce qui nous semble faire la diversité des linguistiques de corpus est la diversité d'approche et de conception tout à la fois du ou des corpus et des recherches qu'il supporte. Ainsi, interroger le rapport entre corpus et constitution des savoirs linguistiques suppose de préciser quelque peu la nature des corpus sur lesquels se fondent les recherches linguistiques et conjointement d'interroger l'objet de ces recherches.

Notre propos ici se focalisera sur le type de recherches que les corpus supportent. Nous voulons mettre en lumière le fait qu'évoquer le travail sur corpus emporte avec soi certains présupposés, plus ou moins explicites, sur les objets traités et sur l'approche adoptée.

Nous montrerons dans la section 2 que, si l'on considère l'école anglaise comme l'un des points de départ historiques de la linguistique de corpus, la perspective presque naturelle de celle-ci est l'élucidation du sens, selon une approche sémasiologique, c'est-à-dire qui part des formes pour aller vers leur signification. D'une démarche similaire – mais quelque peu différente – procède l'analyse de discours, surtout quand elle se fonde sur les métriques, lexicométrie et textométrie.

Or, c'est ce que nous mettrons en évidence, quoique l'outillage de la linguistique, c'est-à-dire tout à la fois la mise en œuvre de logiciels pour le traitement des corpus et l'accès de plus en plus facile à des corpus informatisés de plus en plus variés et volumineux, l'incline plus volontiers vers un travail sur les formes, ce même outillage peut être mobilisé pour un programme de recherche différent, partant du sens pour aller vers les formes, selon une approche onomasiologique. Pour appuyer cette perspective, nous examinerons section 3 divers travaux de recherche mettant en œuvre une telle approche, afin d'en abstraire dans la section 4 les lignes directrices et conditions méthodologiques, en montrant aussi comment les outils de TAL peuvent être mis à contribution pour ces recherches.

2. REGARD HISTORIQUE : CORPUS ET RECHERCHES LINGUISTIQUES

2.1. Du côté britannique

Associer au terme *linguistique* « de corpus » et « outillée » renvoie à la C*orpus Linguistics* et à l'école britannique qui s'est, d'après Léon (2008), confrontée dès son origine à la constitution et à l'exploitation de corpus informatisés. Linguistique empirique, elle rassemble des textes dans un corpus pour en faire le matériau de

base du travail linguistique – ce qui l'oppose à la linguistique introspective, évoquée avec humour sous les traits de l'*armchair linguist* par Fillmore (1992).

Quels objets vise-t-elle, sur quoi se focalisent ses recherches ?

Il s'agit, d'une part, de construire théoriquement une linguistique de l'usage, c'est-à-dire un appareil qui permette l'analyse du corpus, donc qui définisse les phénomènes linguistiques pertinents et qui en procure une mise en ordre et une exploitation. Par exemple, l'appareil théorique de la *Systemic Functional Grammar* de M.A.K. Halliday (Matthiessen & Halliday 1997) met l'accent sur la *grammaire*[1] de la langue en tant que système de ressources, disponible pour le locuteur dans ses interactions sociales et lui permettant avant tout de construire du sens.

Le cœur du programme, originel et continué, de la *corpus linguistics* est ainsi le sens : « The focus of corpus linguistics is on meaning. Meaning is what is being verbally communicated between the members of a discourse community. » (Teubert 2005, p. 2). Teubert, après ce premier principe, poursuit son exposé de ce qu'est pour lui la linguistique de corpus en élaborant les rapports du sens et des formes : « The form is what represents the meaning, and there is no meaning without the form by which it is represented. » (p. 3). Plus loin dans l'article, il ancre ce sens dans le discours, dans une opposition à la fois à une vision référentielle de la langue – pour lui, il n'y a pas de lien direct entre le discours et le « monde réel » – et à une saisie décontextualisée des unités étudiées.

C'est là le fondement majeur de la *corpus linguistics* : quoique la nature des unités étudiées soit variable, selon le domaine d'étude – pour Leech (1992), la linguistique de corpus traite aussi bien de phonologie que de syntaxe ou de sociolinguistique –, leur fonctionnement est appréhendé dans le contexte qui leur donne sens.

Mais les faits à étudier ne surgissent pas spontanément du corpus, il est nécessaire – d'autre part – de construire les observables qui vont alimenter l'étude des phénomènes. Et dans la mesure où, depuis les années 1980, dans l'univers de la linguistique anglo-saxonne, la constitution de corpus est en fait la constitution de corpus *informatisés*, les observables correspondent à « observables par une machine ». Or ce que peut facilement faire l'ordinateur, c'est retrouver des formes, puis les assembler, les dénombrer, lister leur voisinage, calculer leurs affinités mutuelles...

Se dessine ainsi, sur le plan théorique aussi bien que sur le plan pratique, un cadre pour des recherches dont les objets sont des unités formelles, qu'il s'agisse d'unités « naturelles » de la langue écrite comme le mot ou le groupe de mots ou d'unités obtenues par transcription ou codage comme le phonème (dans le cas de

1 Qu'il nomme *grammar* par opposition à *grammatics* qu'il réserve à la théorie grammaticale, ce dernier sens présent dans l'expression *traditional grammar*.

transcriptions phonémiques) ou la catégorie grammaticale (dans le cas de codage), et dont le propos est d'élucider le sens de ces unités elles-mêmes ou la façon dont elles concourent à la signification.

On comprend ainsi à quel point l'outillage (logiciel) est déterminant pour la *corpus linguistics* : l'ordinateur rend possible extractions et comptages, calcul de cooccurrences, de collocations, qui permettent de récupérer et traiter les formes des textes puis de fonder l'élucidation de leur sens tout en prenant en compte les phénomènes de fréquence.

Du côté français, l'ordinateur a eu sa place aussi pour le traitement de corpus, en prenant là aussi en compte la quantification, mais à l'origine dans une des disciplines de la linguistique[2] : l'analyse de discours.

2.2. En France, analyse de discours et textométrie

Contrairement à la *corpus linguistics*, l'objet de l'analyse de discours « à la française » n'est pas le système linguistique mais les discours, en tant qu'ils sont situés socialement et en tant que l'on peut les rapporter à une énonciation et à des conditions socio-historiques de production. Pour Pêcheux (1984), fondateur de l'Analyse Automatique du Discours (AAD), elle est une discipline à la croisée de la linguistique, la sociologie, l'histoire, la politologie...

Malgré cette différence d'objets et de programme de recherche entre linguistique de corpus et analyse de discours, existent un certain nombre de convergences que nous explicitons ci-après pour montrer qu'il en ressort des effets tout à fait similaires.

Travailler sur les discours implique de les collecter, de façon raisonnée et argumentée. Ainsi, l'AAD est elle aussi nécessairement, de par son objet même, une discipline « de corpus », quoique la notion de corpus n'y soit pas problématisée dans les mêmes termes que pour des études linguistiques (Charaudeau 2009). C'est le corpus qui fixe le discours objet d'analyse, aussi les études en analyse du discours prennent-elles le soin de le décrire et le caractériser, en ce que c'est lui qui permet de fixer le cadre non-linguistique dans lequel prendront place les analyses du matériau linguistique, lesquelles en retour contribuent à éclairer le discours : « L'analyse de discours (AD), qui a pour objet les productions écrites ou orales, envisagées dans leur matérialité linguistique et dans leurs conditions de production historiques et politiques, offre un point de vue qui structure l'interprétation de ces textes. » (Cislaru et Sitri 2009, p. 85).

2 Certains considèrent l'analyse de discours française comme détachée de la linguistique. Sans prétention aucune concernant ce débat, nous nous contentons de noter que nombre d'articles de recherche en analyse de discours sont le fait de linguistes et qu'ils y mobilisent des concepts linguistiques.

Avec l'emploi de « interprétation », on pourrait penser retrouver la notion de *sens* qui était au cœur de la *corpus linguistics* et voir là un rapprochement supplémentaire entre les deux approches, mais le sens dont il est question en AD ou AAD est celui qui va se trouver déterminé par ce que Cislaru et Sitri (2009) appellent les « extérieurs » des textes. Par exemple, comment certaines expressions deviennent dans l'espace médiatico-politique des « formules » (Krieg-Planque 2009) cristallisant idéologie et dimension polémique (pour le dire rapidement). Ou encore, par quels cheminements textuels et sociaux un toponyme se charge des significations qui n'ont plus guère à voir avec le nom du lieu (Lecolle 2007).

Malgré cette différence de regard entre « sens linguistique » et « sens en discours », linguistique de corpus et analyse du discours abordent la détermination du sens dans des mouvements comparables. L'analyse de discours, surtout quand elle est explicitement automatique (AAD), use de procédures qui permettent de « décrire les structures, les régularités et les spécificités » (Sueur 1982, p. 145) des **formes** du texte. Nous insistons ici sur le mot « formes » qui revient de façon systématique sous la plume de Sueur (1982), notamment quand il critique la lexicométrie : « l'objet de la lexicométrie n'est pas le texte ou le discours ; il est le texte réduit à un dictionnaire de formes » (p. 145). Sueur déplore cette réduction qui gomme fonctionnements syntaxiques et faits d'énonciation et il préconise d'articuler l'étude des fréquences à une étude de séquences, qui réinsère les régularités et spécificités observées à l'intérieur des phrases et donc du déroulement linéaire du texte.

Il dessine ainsi une méthode exploratoire qui, trente ans plus tard, s'est développée parallèlement en analyse de discours et en linguistique de corpus. Fondée sur l'alternance du gros grain de mesures statistiques avec le grain fin de l'observation en contexte, elle est facilitée depuis quelques années par la floraison d'outils adaptés, tels que TXM[3] (Pincemin 2007 ; Heiden *et al.* 2010) ou Le Trameur[4] (Fleury et Zimina 2014), qui n'existaient évidemment pas quand Sueur formulait ses critiques et qui concilient approche statistique et prise en compte des niveaux textuel et discursif.

Ce parcours, nécessairement parcellaire et sommaire, met en évidence que, du côté britannique comme du côté français, des recherches fondées sur des corpus textuels et qui ont recours à l'ordinateur ont en commun de prendre pour départ les formes de ces textes pour arriver à l'interprétation et/ou à la signification.

Pour autant, nous voulons montrer maintenant que cette orientation 'forme → sens' n'est pas la seule possible et que les corpus aussi bien que l'ordinateur

3 http://textometrie.ens-lyon.fr/spip.php?article60
4 http://www.tal.univ-paris3.fr/trameur/

peuvent être mobilisés pour un programme de recherche linguistique différent, partant du sens ou de la signification pour aller vers les formes.

3. NON PLUS « QU'EST-CE QUE CELA VEUT DIRE ? » MAIS « COMMENT EST-CE QUE CELA SE DIT ? »…

Nous appuierons notre propos sur la présentation – là aussi réductrice – de plusieurs études pour tirer, section 4, des enseignements méthodologiques pour une approche onomasiologique des faits de langue.

Cette approche, renversant la perspective à 180 degrés, pourrait-on dire, consiste donc à passer d'un questionnement qui vise à expliciter « ce que ça veut dire » et « comment ça veut dire ce que ça veut dire », à un questionnement qui part d'une signification – ou d'un phénomène de nature sémantique ou discursive – et répond à la question « comment est-ce que cela se dit ? » en inventoriant les unités qui réaliseront la signification ou le phénomène étudiés.

Assez peu d'études linguistiques suivent cette voie, en raison de la difficulté de l'entreprise : comme ce que l'on veut trouver, ce sont des formes, on ne peut utiliser de la façon « classique » évoquée précédemment un ordinateur pour la récolte des occurrences – ce que soulignait fort justement Péry-Woodley (2001, p. 31) en insistant sur la difficulté de constitution à priori d'un corpus pour une telle étude.

Divers travaux ont tout de même affronté cette difficulté sur des objets divers.

Hearst (1992) pour l'anglais puis Borillo (1996) pour le français ont décrit les modalités d'expression de la relation lexicale d'hyperonymie telles qu'elles se manifestent dans des textes « spécialisés » du type encyclopédie ou textes scientifiques. Il s'agissait de produire une liste d'éléments lexicaux et de structures phrastiques en vue d'une utilisation automatique ultérieure. L'objectif pour Hearst était d'enrichir automatiquement ou semi-automatiquement une ressource telle que WordNet de couples hyperonyme-hyponyme. Borillo, pour le français, poursuivait un objectif similaire mais en fournissant une description linguistique bien plus fine des structures analysées.

C'est de description linguistique fine qu'il s'agit aussi dans le travail de Rebeyrolle (2000) sur la définition telle qu'elle peut apparaître spontanément dans les textes (et non pas la définition lexicographique, définition experte du dictionnaire). À partir de 4 corpus dans des disciplines différentes, elle a relevé manuellement tous les énoncés définitoires pour en extraire et en livrer les particularités linguistiques, sous la forme de « patrons » ensuite exploitables pour une recherche automatique, telle qu'elle est décrite dans (Rebeyrolle et Tanguy 2001).

C'est encore la description linguistique et la théorisation linguistique qui sont visées par Leroy (2001) dans son travail basé sur un corpus de textes de presse et

ambitionnant de « proposer une description et une analyse linguistiques du phénomène de l'antonomase du nom propre » (p. 11). Suivant une procédure bien établie en TAL, elle a partitionné le corpus en plusieurs sous-ensembles, les uns servant à construire « une grammaire de l'antonomase », d'autres à tester cette grammaire. La grammaire en question se situe dans la lignée des travaux précédemment évoqués, en ce qu'elle se présente sous la forme d'une modélisation-abstraction des énoncés antonomasiques analysés et se veut un moyen de repérer automatiquement les antonomases dans les textes. L'automatisation du repérage ambitionne de permettre aux chercheurs intéressés par le phénomène d'accéder plus facilement et plus massivement aux contextes des antonomases afin d'en poursuivre l'étude. Une particularité de ce travail, par rapport aux trois précédemment cités, est d'avoir testé la reconnaissance spontanée de l'antonomase auprès de 22 « informateurs », essentiellement dans la perspective de mesurer la faisabilité du repérage, en premier lieu par des êtres humains (Leroy 2004). Nous reviendrons plus loin sur la mobilisation d'informateurs dans le cadre d'une approche onomasiologique.

C'est un objectif similaire de description de structures et d'élaboration de « grammaires » destinées à un repérage automatique que poursuit Florez (2014) sur la citation qu'elle appelle « positionnée ». Dans le même esprit que Leroy, elle élabore des patrons de recherche afin de recueillir dans son corpus (thèses et articles scientifiques rendus disponibles par le projet Scientext[5]) les passages dans lesquels les auteurs citent les travaux d'autrui tout en se positionnant par rapport à ces travaux. Les patrons vont là aussi systématiser et accroître l'accès aux occurrences en vue d'une analyse linguistique de la citation : sa forme syntaxique et son type pragmatico-sémantique selon les disciplines scientifiques.

Toujours en mobilisant les outils de TAL et en poursuivant la confection de « grammaires », Jacques *et al.* (2013) ont étudié dans des articles scientifiques en anglais la formulation de leur objectif de recherche . Il s'agissait là moins de s'intéresser au phénomène lui-même dans sa dimension rhétorique et discursive que d'abstraire un petit nombre de schémas combinant des unités lexicales, grammaticales et des places syntaxiques pour rendre compte de la variété de réalisations de cette formulation. Ce sont donc tout à la fois l'inventaire des réalisations et la fabrication des grammaires qui étaient ciblées, pour une double exploitation ultérieure : dans d'éventuelles applications de TAL – du même type que le *argumentative zoning* de Teufel (1999) – mais surtout comme outil d'aide à la rédaction en anglais à destination de locuteurs non-natifs de cette langue, en situation de production d'un écrit scientifique.

On voit par ces exemples que la description ne s'en tient pas à la structure et aux traits linguistiques, elle est mise au service d'autres fins. L'outillage ouvre sur

5 http://scientext.msh-alpes.fr/scientext-site/spip.php?article1

une visée « applicative », récurrente dans ce type d'études. En effet, comprendre et décrire « comment ça se dit » ne semble pas une orientation très prisée des linguistes dans la seule perspective de description du système – à quelques exceptions près, par ex. Nazarenko (2000). Elle semble en revanche être une préoccupation recevable dans des études de linguistique appliquée, que l'application soit le TAL ou l'enseignement.

Sans doute est-ce dû au fait que l'approche onomasiologique ne limite pas l'effort de recherche à l'analyse des données et à la production d'un savoir nouveau rendant compte de ces données, comme lorsque l'on cherche par exemple à caractériser les emplois de tel ou tel connecteur, mais englobe la phase même de recueil de ces données. L'effort ne concerne pas seulement le temps que consomme l'analyse des occurrences mais aussi la mise au point et l'application d'une méthodologie adaptée.

4. CONDITIONS POUR UNE APPROCHE ONOMASIOLOGIQUE

4.1. Principes méthodologiques

Quel que soit l'angle retenu, sémasiologique ou onomasiologique, le travail sur corpus confronte à la nécessité d'une sélection et d'une mise en ordre des données. Si l'on étudie par exemple l'unité *cause*, il faudra sérier ses occurrences selon qu'il s'agit d'un nom en emploi libre, d'un nom au sein d'une construction figée (*à cause de*, *pour cause*, *pour la bonne cause*...), du verbe..., donc prendre des décisions : en retenir certaines, en écarter d'autres, et au final rendre compte des occurrences sélectionnées. L'approche onomasiologique rend encore plus crucial ce travail de sélection en ce qu'elle fait intervenir, à la racine même de la recherche, une part non négligeable d'interprétation et donc de subjectivité.

Dans Jacques & Poibeau (2010) sont exposées certaines propositions méthodologiques, que nous résumons et complétons en deux points : la mesure de l'accord entre juges et l'annotation du corpus.

4.1.1. Limiter la subjectivité en mesurant l'accord inter-juges

Décider, face à une occurrence donnée, si oui ou non elle est l'expression d'une relation d'hyperonymie, si elle constitue une définition ou un exemple d'antonomase pourrait paraitre trivial. Or, dès que l'on confronte plusieurs chercheurs, même avertis, à des données réelles donc bruitées (c'est-à-dire comportant aussi bien des occurrences de ce que l'on cherche que des occurrences qui n'ont rien à voir), le jugement peut différer.

La mesure de l'accord inter-juges, qui consiste à soumettre des données au jugement de plusieurs sujets puis vérifier dans quelle proportion les jugements concordent ou divergent, éclaire sur la part de variation individuelle dans l'appréhension des phénomènes de langue et dans la compréhension linguistique. Les résultats incitent à la plus grande circonspection face aux analyses qui s'appuient sur l'interprétation d'un seul chercheur : lorsque le sens est convoqué et que l'analyse met en œuvre une compréhension à grain fin, c'est-à-dire soucieuse de nuances et de discriminations subtiles, le désaccord est plus souvent la règle que l'accord parfait. Pour s'en convaincre, deux exemples issus de recherches en linguistique.

Véronis (2004) a soumis 600 mots en contexte (60 contextes pour chaque mot) à 6 juges (étudiants en sciences du langage, avertis des questions de sémantique), avec la question « Ce mot a-t-il *un* ou *plusieurs* sens dans les contextes ci-dessous ? ». Sur cette tâche qui ne demandait pas une identification fine du ou des sens mais seulement le repérage d'une quelconque polysémie, l'accord entre les 6 juges, mesuré avec un coefficient appelé *Kappa* (Carletta 1996), est de 0,49, ce qui correspond à un accord modéré.

Second exemple, pour une analyse de l'expression des procédures (Jacques & Poibeau 2010), nous avons soumis des portions de textes médicaux ou extraits de fiches de bricolage à plusieurs juges avec la question « cet extrait permet-il au lecteur d'inférer des actions à accomplir ? ». Là encore, sans plus de précision sur les caractéristiques à prendre en compte, l'accord inter-annotateurs est bon sans être excellent (*Kappa* = 0,72).

Dans ces deux cas, les « juges » n'étaient pas des néophytes mais bien des experts d'analyse linguistique, au tout au moins des juges avertis dans le cas des étudiants. Ils se rapprochaient donc fortement de la situation du linguiste qui constitue son échantillon de données pour l'analyse et doit, comme nous l'indiquions plus haut, retenir ou éliminer telle ou telle construction en répondant – au moins intérieurement – à la question : cette construction est-elle une manifestation du phénomène que je veux étudier ? Dans chacun des deux cas évoqués, selon le juge impliqué, les données retenues auraient varié en qualité et en quantité dans une plus ou moins grande proportion.

Si l'on veut donc qu'une étude d'orientation onomasiologique fournisse des données et des analyses un tant soit peu fiables non au niveau d'un idiolecte mais à un niveau plus global, il est fortement souhaitable, sinon obligatoire, d'accompagner la phase de construction de l'objet d'étude d'une mesure du partage de la notion ou du phénomène en jeu.

Il n'est toutefois pas toujours aisé, les conditions de la recherche étant ce qu'elles sont, de mobiliser plusieurs collègues et étudiants pour se prononcer sur

des énoncés, d'autant plus lorsque le jugement nécessite un temps de réflexion et/ou de consultation de consignes, de guides et autres références. Une seconde méthode, non exclusive mais plutôt complémentaire, consiste à annoter, dans les corpus eux-mêmes, les données qui servent de matière à la réflexion.

4.1.2. Annotation des occurrences en corpus

S'il est bien un aspect par lequel l'appui sur les corpus peut changer l'angle de vue de la recherche linguistique, c'est l'enrichissement que permettent tout à la fois la constitution de corpus informatisés et le développement d'outils logiciels idoines. Nous synthétiserons dans la section suivante les bénéfices de l'outillage de la linguistique, nous nous focalisons ici sur le point relatif aux données.

Nous avons mentionné dans les sections précédentes divers travaux sur des phénomènes linguistiques non réductibles à une forme. Peu d'entre eux ont rendu publique la totalité des données qui ont alimenté leur réflexion[6] – y compris les nôtres... –, ce qui ne peut guère leur être reproché dans la mesure où une telle publication contreviendrait parfois au droit d'auteur. L'essor de corpus librement accessibles, non seulement consultables mais modifiables, peut changer cette relative opacité des données en permettant que les publications d'une recherche linguistique sur corpus ne soient pas seulement les analyses scientifiques produites, mais aussi l'annotation dans les corpus eux-mêmes des phénomènes observés.

Cette voie, exigeante conceptuellement et techniquement, en ce qu'elle nécessite l'élaboration d'un schéma d'annotation ainsi que sa réalisation concrète, constitue un prolongement logique de l'outillage des linguistes et augmente la portée des études sur corpus. Un projet comme Annodis[7] est à cet égard exemplaire : il cumule la constitution d'un corpus informatisé, l'analyse en corpus de phénomènes discursifs (relations rhétoriques, énumérations, chaines topicales...), la production de savoir sur ces phénomènes et la mise à disposition du corpus dans lequel les phénomènes sont annotés avec le guide d'annotation utilisé. Ainsi, tout chercheur peut reproduire la recherche et/ou juger par lui-même des choix faits sur les données, sans parler de la possibilité de réutiliser le schéma d'annotation proposé sur un autre corpus. De la même manière, le corpus annoté en « zones rhétoriques » utilisé par Teufel (1999) est rendu disponible avec l'ensemble du matériel permettant l'annotation automatique[8].

C'est selon nous la voie à suivre pour accroitre la scientificité de la linguistique de corpus. Tout comme les sciences expérimentales qui décrivent de façon très précise les conditions d'obtention des résultats qui sont ensuite discutés et peuvent

6 A l'inverse, Leroy (2001) fournit dans ses annexes la totalité des phrases comportant une antonomase sur lesquelles elle fonde sa thèse.
7 http://redac.univ-tlse2.fr/corpus/annodis/
8 http://wing.comp.nus.edu.sg/raz/

servir à faire évoluer les théories, la linguistique de corpus, si elle fait l'effort de rendre disponible son matériau de travail, ouvre de nouveaux modes d'investigation des faits linguistiques.

Corpus et outillage semblent ainsi faire très bon ménage, revenons maintenant sur le rapport complexe entre le matériau et l'outil qui le travaille.

4.2. Contraintes et bénéfices de l'outillage

Nous avons montré dans la section 2 que la facilité qu'offre un outillage logiciel pour la recherche, le tri, le dénombrement de formes ou de combinaisons de formes, induit une focalisation sur les formes. Nous avons toutefois évoqué section 3 diverses études qui mobilisent cet outillage pour des investigations partant des significations. Ces études font face à une contrainte que nous avons soulignée : récolter les données du corpus. Il pourrait donc sembler paradoxal qu'elles soient à ce point liées, dans leur mise en œuvre comme dans leurs objectifs, à l'ordinateur mais cet outil autorise en fait de mettre des techniques de traitement automatique des langues (TAL) au service de la récolte d'occurrences. Les textes analysés subissent des traitements de type étiquetage morpho-syntaxique et/ou lemmatisation afin de permettre l'utilisation de catégories grammaticales ou de formes non fléchies des unités lexicales par des outils d'exploration textuelle (tels que ceux que nous avons cités : TXM ou le Trameur). Borillo (1996) et Hearst (1992) mobilisent le TAL dès la phase de recherche des expressions, en « projetant » sur les textes des couples déjà repérés afin d'atteindre des contextes dans lesquels s'exprime leur relation d'hyperonyme à hyponyme, pour découvrir des formes non encore répertoriées d'expression de cette relation. La phase ultérieure analyse les formulations recueillies et en abstrait des patrons lexico-syntaxiques, voie très exactement suivie par Rebeyrolle (2000), de même que par Jacques *et al.* (2013). La grammaire de l'antonomase de Leroy (2001) comme la grammaire des citations de Florez (2014) sont un résultat intermédiaire du travail en même temps qu'un des moyens d'accroitre automatiquement le nombre d'occurrences étudiées. Pour toutes ces recherches, une automatisation du repérage du phénomène étudié constitue un des objectifs à atteindre. Elle induit une modélisation du phénomène sous la forme de paramètres à fournir à un logiciel pour ce repérage.

Leroy (2001, p. 148) a souligné les bénéfices de l'utilisation du TAL – et nous dirons plus largement d'un outil : systématicité du recueil d'occurrences, qui repose alors sur des patrons précis, extension du nombre d'occurrences atteintes, prise en compte de la fréquence... Nous lui ajouterons un bénéfice apparu avec de nouveaux outils permettant l'annotation en même temps que le repérage (par ex.

Unitex[9] ou Glozz[10]) et que nous avons évoqué dans la section précédente, qui est de fournir un enrichissement des corpus sous la forme d'un balisage dans les textes mêmes des passages repérés et des analyses qui leur sont associées, sous forme de traits ou de types.

5. CONCLUSION

Nous avons souligné dans cet article que l'outillage de la linguistique oriente volontiers ses recherches vers les formes mais peut, à certaines conditions, servir des recherches onomasiologiques. Celles-ci, qui paraissent plus « naturelles » dans le champ de la linguistique appliquée, parce qu'elles y répondent à des besoins concrets tels que la rédaction dans une langue seconde ou le traitement de grandes masses textuelles, impliquent un double niveau de description des phénomènes : le niveau linguistique, qui permet de cerner leurs caractéristiques pertinentes, et un niveau plus formel, qui consiste à déterminer quels indices manipulables par un ordinateur permettent d'identifier dans le corpus les manifestations du phénomène à l'étude.

Cette double description, si elle s'accompagne par la suite d'une réinscription, dans les corpus mêmes, des analyses produites, ouvre plusieurs voies.

Elle rend possible un accès direct et contextualisé aux passages étudiés lors de la recherche initiale. Le format d'un article ou d'une communication scientifiques n'autorisent généralement qu'un petit nombre d'exemples, saisis dans un contexte de taille limitée, ce qui est peu favorable à l'exposé de phénomènes discursifs jouant sur des empans larges de textes. Les corpus enrichis par une recherche autoriseraient ainsi la perception de phénomènes jouant sur une dynamique textuelle (les chaines anaphoriques, les chaines topicales, la structuration du discours...).

Elle permet une sélection et une observation des contextes par le sens et permet à de nouvelles études de s'intéresser à d'autres interactions entre le sens et divers autres phénomènes. Elle participe donc d'une cumulativité de la recherche qui justifie les efforts engagés.

BIBLIOGRAPHIE

Borillo, Andrée, 1996. « Exploration automatisée de textes de spécialité : repérage et identification de la relation lexicale d'hyperonymie », *LINX* 34-35, 113-124.

Carletta, Jean, 1996. "Assessing Agreement on Classification Tasks: The *Kappa* Statistic", *Computational Linguistics* 22(2), 249-254.

Charaudeau, Patrick, 2009. « Dis-moi quel est ton corpus, je te dirai quelle est ta problématique », *Corpus* 8, 37-66.

Cislaru, Georgeta & Sitri, Frédérique, 2009. « TEXTE ET DISCOURS. Corpus, co-texte et analyse automatique du point de vue de l'analyse de discours », *Corpus* 8, 85-104.

9 http://www-igm.univ-mlv.fr/~unitex/

10 http://www.glozz.org/

Fillmore, Charles J., 1992. " 'Corpus linguistics' or 'Computer-aided arm-chair linguistics' ", *Directions in Corpus Linguistics*, Berlin / New-York, Mouton de Gruyter, 35-60.
Fleury, Serge & Zimina, Maria, 2014. "Trameur: A Framework for Annotated Text Corpora Exploration", *COLING 2014*, 57-61.
Florez, Magda, 2014. « La citation positionnée dans l'écrit scientifique », *L'écrit scientifique : du lexique au discours. Autour de Scientext*, Rennes, PUR, 67-84.
Habert, Benoît, Nazarenko, Adeline & Salem, André, 1997. *Les linguistiques de corpus*, Paris, Armand Colin.
Hearst, Marti, 1992. "Automatic Acquisition of Hyponyms from Large Text Corpora», *COLING 92*, 539-545.
Heiden, Serge, Magué, Jean-Philippe & Pincemin, Bénédicte, 2010. « Une plateforme logicielle open-source pour la textométrie – conception et développement », *JADT 2010*, 1021-1032.
Jacques, Marie-Paule & Poibeau, Thierry, 2010. « Étudier des structures de discours : préoccupations pratiques et méthodologiques », *CORELA (Cognition, Représentation, Langages)* 8(1).
Jacques, Marie-Paule, 2005. « Pourquoi une linguistique de corpus ? », *La linguistique de corpus*, Rennes, PUR, 21-30.
Jacques, Marie-Paule, Hartwell, Laura & Falaise, Achille, 2013. « Techniques de TAL et corpus pour faciliter les formulations en anglais scientifique écrit », *TALN'2013*, 146-159.
Krieg-Planque, Alice, 2009. *La notion de « formule » en analyse du discours. Cadre théorique et méthodologique*, Besançon, Presses universitaires de Franche-Comté.
Lecolle, Michelle, 2007. « Polysignifiance du toponyme, historicité du sens et interprétation en corpus. Le cas de Outreau », *Corpus* 6, 101-125.
Leech, Geoffrey 1992. "Corpora and theories of linguistic performance", *Directions in Corpus Linguistics*, Berlin / New-York, Mouton de Gruyter, 105-122.
Léon, Jacqueline 2008. « Aux sources de la "Corpus Linguistics": Firth et la London School », *Langage*s 171, 12-33.
Leroy, Sarah 2004. « Extraire sur patrons : allers et retours entre analyse linguistique et repérage automatique », *Revue française de linguistique appliquée* 9 (1), 25-43.
— 2001. *Entre identification et catégorisation, l'antonomase du nom propre en français*, thèse, Université Paul Valéry – Montpellier.
Matthiessen, Christian & Halliday, Michaël, 1997. *Systemic Functional Grammar: A First Step into the Theory*, Beijing, Higher Education Press.
Nazarenko, Adeline, 2000. *La cause et son expression en français*, Paris, Ophrys.
Pêcheux, Michel, 1984. « Sur les contextes épistémologiques de l'analyse de discours », *Mots* 9, 7-17.
Péry-Woodley, Marie-Paule, 2001. « Modes d'organisation et de signalisation dans des textes procéduraux », *Langages* 141, 28-46.
Pincemin, Bénédicte, 2007. « Concordances et concordanciers : de l'art du bon KWAC », *XXVIIe colloque d'Albi Langages et Signification*, 33-42.
Rebeyrolle, Josette, 2000. *Forme et fonction de la définition en discours,* thèse, Université Toulouse II Le Mirail, Équipe de Recherche en Syntaxe et Sémantique.
Rebeyrolle, Josette & Tanguy, Ludovic, 2001. « Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires », *Cahiers de Grammaire* 25, 153-174.
Sueur, Jean-Pierre, 1982. « Pour une grammaire du discours », *Mots* 5, 143-185.
Teubert, Wolfgang, 2005. "My version of corpus linguistics", *International Journal of Corpus Linguistics* 10:1, 1-13.
Teufel, Simone, 1999. *Argumentative zoning: Information extraction from scientific text.* Ph.D. Thesis, University of Edinburgh.
Véronis, Jean, 2004. « L'étiquetage sémantique des corpus », *Le Français Moderne* 2004/1, 27-38.