Champs linguistiques

RECHERCHES

Céline POUDAT
Frédéric LANDRAGIN

Explorer un corpus textuel

Méthodes – pratiques – outils

deboeck SUPÉRIEUR

Explorer un corpus textuel

Champs linguistiques

Collection dirigée par Marc Wilmet (Université libre de Bruxelles) et Dominique Willems (Universiteit Gent)

Recherches

Corminboeuf G., *L'expression de l'hypothèse en français. Entre hypotaxe et parataxe*
Demol A., *Les pronoms anaphoriques il et celui-ci*
Heyna F., *Étude morpho-syntaxique des parasynthétiques. Les dérivés en dé– et en anti–*
Horlacher A.-S., *La dislocation à droite revisitée. Une approche interactionniste*
Huyghe R., *Les noms généraux d'espace en français. Enquête linguistique sur la notion de lieu*
Jacquin J., *Débattre. L'argumentation et l'identité au cœur d'une pratique verbale*
Marchello-Nizia Ch., *Grammaticalisation et changement linguistique.*
Marengo S., *Les adjectifs jamais attributs. Syntaxe et sémantique des adjectifs constructeurs de la référence*
Martin F., *Les prédicats statifs. Étude sémantique et pragmatique*
Micheli R., *Les émotions dans les discours. Modèle d'analyse, perspectives empiriques*
Poudat C., Landragin Fr., *Explorer un corpus textuel. Méthodes – pratiques – outils*
Rézeau P., (études rassemblées par), *Richesses du français et géographie linguistique. Volume 1*
de Saussure L., *Temps et pertinence. Éléments de pragmatique cognitive du temps*
Schnedecker C., *De l'un à l'autre et réciproquement…Aspects sémantiques, discursifs et cognitifs des pronoms anaphoriques corrélés*
Thibault A. (sous la coordination de), *Richesses du français et géographie linguistique, Volume 2*
Van Goethem K., *L'emploi préverbal des prépositions en français. Typologie et grammaticalisation*

Manuels

Bal W., Germain J., Klein J., Swiggers P., *Bibliographie sélective de linguistique française et romane.* 2e édition
Bracops M., *Introduction à la pragmatique. Les théories fondatrices : actes de langage, pragmatique cognitive, pragmatique intégrée.* 2e édition
Chiss J.-L., Puech C., *Le langage et ses disciplines. XIXe -XXe siècles*
Delbecque N. (Éd.), *Linguistique cognitive. Comprendre comment fonctionne le langage*
Englebert A., *Introduction à la phonétique historique du français*
Gaudin Fr., *Socioterminologie. Une approche sociolinguistique de la terminologie*
Gross G., Prandi M., *La finalité. Fondements conceptuels et genèse linguistique*
Klinkenberg J.-M., *Des langues romanes. Introduction aux études de linguistique romane.* 2e édition
Kupferman L., *Le mot «de». Domaines prépositionnels et domaines quantificationnels*
Leeman D., *La phrase complexe. Les subordinations*
Mel'čuk I. A., Clas A., Polguère A., *Introduction à la lexicologie explicative et combinatoire.*
Coédition AUPELF-UREF. Collection Universités francophones
Mel'čuk I., Polguère A., *Lexique actif du français. L'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*
Revaz Fr., *Introduction à la narratologie. Action et narration*

Recueils

Albert L., Nicolas L. (sous la direction de), *Polémique et rhétorique de l'Antiquité à nos jours*
Bavoux C. (dir.), *Le français des dictionnaires. L'autre versant de la lexicographie française*
Bavoux C., *Le français de Madagascar. Contribution à un inventaire des particularités lexicales.*
Coédition AUF. Série Actualités linguistiques francophones
Berthoud A.-Cl., Burger M., *Repenser le rôle des pratiques langagières dans la constitution des espaces sociaux contemporains*
Bouchard D., Evrard I., Vocaj E., *Représentation du sens linguistique. Actes du colloque international de Montréal*
Conseil supérieur de la langue française et Service de la langue française de la Communauté française de Belgique (Eds), *Langue française et diversité linguistique. Actes du Séminaire de Bruxelles* (2005)
Corminboeuf G., Béguelin M.-J. (sous la direction de), *Du système linguistique aux actions langagières. Mélanges en l'honneur d'Alain Berrendonner*
Dendale P., Coltier D. (sous la direction de), *La prise en charge énonciative. Études théoriques et empiriques*
Evrard I., Pierrard M., Rosier L., Van Raemdonck D. (dir.), *Représentations du sens linguistique III. Actes du colloque international de Bruxelles* (2005)
Englebert A., Pierrard M., Rosier L., Van Raemdonck D. (Éds), *La ligne claire. De la linguistique à la grammaire. Mélanges offerts à Marc Wilmet à l'occasion de son 60e anniversaire*
Gradoux X., Jacquin J., Merminod G. (dir.), *Agir dans la diversité des langues. Mélanges en l'honneur d'Anne-Claude Berthoud*
Hadermann P., Van Slijcke A., Berré M. (Éds), *La syntaxe raisonnée. Mélanges de linguistique générale et française offerts à Annie Boone à l'occasion de son 60e anniversaire.* Préface de Marc Wilmet
Rézeau P. (sous la direction de), *Variétés géographiques du français de France aujourd'hui. Approche lexicographique*
Service de la langue française et Conseil de la langue française et de la politique linguistique (Eds), *La communication avec le citoyen : efficace et accessible ? Actes du colloque de Liège, Belgique, 27 et 28 novembre 2009*
Service de la langue française et Conseil de la langue française et de la politique linguistique (Eds), *S'approprier le français. Pour une langue conviviale. Actes du colloque de Bruxelles (2013)*
Simon A. C. (sous la direction de), *La variation prosodique régionale en français*

Céline POUDAT
Frédéric LANDRAGIN

Explorer un corpus textuel

Méthodes – pratiques – outils

Champs linguistiques

deboeck SUPÉRIEUR

© De Boeck Supérieur s.a., 2017 1re édition
Rue du Bosquet, 7 – B-1348 Louvain-la-Neuve

Imprimé en Belgique

Dépôt légal :
Bibliothèque nationale, Paris : février 2017 ISSN 1374-089X
Bibliothèque royale de Belgique, Bruxelles : 2017/13647/005 ISBN 978-2-8073-0563-2

SOMMAIRE

AVANT-PROPOS

Cet ouvrage est issu d'une part des réflexions du groupe de travail « Exploration de corpus » du consortium « Corpus écrits » de la TGIR Huma-Num, la « Très Grande Infrastructure de Recherche » des Humanités Numériques, et d'autre part d'une collaboration entre deux chercheurs membres de ce groupe de travail. Tous les deux travaillent dans le domaine de la linguistique de corpus, l'un plutôt dans ses aspects textométriques – Céline Poudat, laboratoire « Bases, Corpus, Langage » (UMR 7320, Nice), coordinatrice principale du groupe « Exploration de corpus » (avec Serge Heiden et Marie-Paule Jacques) – et l'autre dans la facette liée à l'annotation – Frédéric Landragin, laboratoire « Langues, Textes, Traitements Informatiques, Cognition » (UMR 8094, Montrouge). Cette collaboration a permis de confronter des méthodologies provenant de diverses communautés et de poser les bases d'une pratique commune pour l'analyse de données textuelles, annotées ou non.

Nous remercions tout d'abord la TGIR Huma-Num pour avoir rendu possible la préparation de cet ouvrage, ainsi que pour les encouragements institutionnels utiles à cette entreprise méthodologique participative. Merci aux collègues du comité de pilotage du consortium « Corpus écrits », et aux membres du GT10 « Exploration de corpus » pour leur enthousiasme et leur soutien.

Nous remercions vivement tous les relecteurs que nous avons impliqués dans ce travail : ceux et celles qui ont bien voulu relire l'intégralité du manuscrit pour en détecter les incohérences et les faiblesses, et ceux et celles qui se sont attachés à des chapitres relevant de leurs compétences. Toute notre gratitude va ainsi, par ordre alphabétique, à Lucie Barque, Magali Guaresi, Ludovic Lebart, Muriel Marmurstein, Damon Mayaffre, Sylvie Mellet, Bénédicte Pincemin, Clément Plancq, Thierry Poibeau, Sophie Prévost.

Merci aux collègues qui ont corrigé et validé les nombreux encadrés illustrant page après page notre propos : Thierry Chanier, Georgeta Cislaru, Claire Doquet, Magali Guaresi, Céline Guillot-Barbance, Serge Heiden, Alexey Lavrentev, Giancarlo Luxardo, Véronique Magri, Ramon Martí Solano, Damon Mayaffre, Véronique Montémont, Émilie Née, Bénédicte Pincemin, Frédérique Sitri, Agnès Steuckardt, Agnès Tutin. Merci également aux concepteurs des outils qui ont bien voulu corriger et amender les encadrés logiciels les concernant : Serge Fleury, Ludovic Lebart, Bénédicte Pincemin, Pierre Ratinaud, Laurent Vanni.

En dehors de la rédaction de cet ouvrage mais pour l'inspiration générale et les échanges préalables – parfois anciens – qui influent indirectement sur le résultat, Céline Poudat remercie – en plus des personnes déjà citées – Étienne Brunet, Dominique Longrée et François Rastier. De même, Frédéric Landragin remercie (entre autres) Jean-Yves Antoine, Yann Mathet, Christophe Parisse, Bernard Victorri, Antoine Widlöcher.

CHAPITRE 1
INTRODUCTION

L'exploration de corpus fait appel à de nombreux procédés, pour lesquels des repères méthodologiques s'avèrent indispensables. De même qu'il existe une méthodologie pour constituer un corpus, c'est-à-dire pour rassembler plusieurs textes ou productions linguistiques dans un même ensemble, qui servira d'objet d'étude, il existe une méthodologie pour explorer un corpus. Cette méthodologie a pour objet de décrire comment appréhender les données textuelles regroupées dans le corpus. Elle inclut la recherche de mots, de portions de mots, d'expressions, avec des possibilités beaucoup plus nombreuses que celles offertes par un logiciel de traitement de texte. Elle intègre également la méthodologie de l'annotation – on explore un corpus en l'annotant – et celles de la linguistique de corpus et des statistiques textuelles, qui fournissent des indicateurs (numériques, graphiques, multidimensionnels) pour aider l'analyste, linguiste ou non, à mieux comprendre les données constituant son corpus de travail, à mieux caractériser celui-ci, à mieux en dégager les spécificités. La méthodologie d'exploration de corpus s'avère ainsi très complexe, et c'est l'objet de cet ouvrage que de la présenter.

Pour cela, nous nous appuyons sur un contexte institutionnel et des réflexions collectives. Depuis 2011, un groupe de travail réunissant une trentaine de chercheurs et d'utilisateurs de corpus s'est penché sur cette méthodologie, dans le cadre de la TGIR Huma-Num, la « Très Grande Infrastructure de Recherche » des Humanités Numériques, dont l'objectif général est de faciliter le tournant numérique de la recherche en sciences humaines et sociales. Ce groupe de travail, « Exploration de corpus », le numéro 10 du consortium « Corpus écrits » (intégré depuis dans le consortium « Corpus, Langues et Interactions »), a recensé diverses méthodes et pratiques d'exploration, se matérialisant avec divers outils de manipulation de corpus. Cet ouvrage en propose une synthèse, qui manquait jusqu'à

présent dans la littérature scientifique consacrée à la linguistique de corpus et à ses méthodologies. Son originalité repose sur la prise en compte simultanée de méthodes qui n'avaient pas été reliées auparavant : en nous intéressant aux pratiques de chaque utilisateur de corpus, nous avons rapproché méthodes et outils, et nous avons identifié des usages, c'est-à-dire des façons d'explorer un corpus, qui suivent par exemple un scénario précis. Pour donner des résultats pertinents, ces usages font appel à des méthodes variées, que nous avons alors approfondies.

Notre objectif est d'expliquer de manière progressive et pédagogique ces méthodes de la linguistique outillée qui sont orientées vers l'exploration de corpus, et donc vers l'accès aux données, leur visualisation, le calcul de statistiques. Notre objectif est aussi d'informer l'utilisateur d'outils sur les grands principes de fonctionnement de ceux-ci, de manière à ce qu'il ne choisisse plus une fonctionnalité au hasard, mais – au moins pour les fonctionnalités courantes – en toute connaissance : (i) des prérequis indispensables à une exploitation pertinente de la fonctionnalité en question, (ii) des méthodes mises en œuvre par la fonctionnalité, et (iii) des interprétations qu'il peut tirer (ou non) des résultats obtenus avec cette fonctionnalité. Le but est que cette connaissance permette à chacun de choisir les familles d'outils qui lui seront utiles, de sélectionner et d'exploiter correctement les fonctionnalités qui seront adaptées à son corpus et à ses préoccupations de recherche.

Notre propos ne relève pas de la technique : il est méthodologique avant tout, et vise à fournir des repères sur le fonctionnement des outils d'exploration en général, et non sur l'utilisation de tel ou tel outil particulier. Quand rentrer dans la technique devient malgré tout nécessaire, nous renvoyons soit aux annexes de cet ouvrage, soit à des ouvrages complémentaires, cités dans la bibliographie, comme celui désormais classique de Ludovic Lebart et André Salem, *Statistique textuelle* (1994), ou le tout récent ouvrage de Ludovic Lebart, Bénédicte Pincemin et Céline Poudat, *Statistique exploratoire pour les textes* (2017). Ainsi, les aspects statistiques, qui peuvent nécessiter des explications mathématiques relativement longues, seront traités d'une manière accessible, orientée vers leur intérêt et leur usage concret pour explorer un corpus. Contrairement aux psychologues et aux sociologues, les linguistes français sont peu formés aux statistiques ; aussi les études linguistiques françaises ne recourent-elles souvent qu'aux statistiques descriptives (décomptes, moyennes, écart-types), beaucoup de linguistes éprouvant des difficultés à se saisir des méthodes de l'analyse des données et à exploiter les méthodes textométriques, pourtant développées en France depuis plusieurs décennies. Nous espérons combler ici ce manque, sans doute pas du point de vue mathématique sous-jacent, mais en tout cas du point de vue des pratiques que la connaissance de telles méthodes permet de mettre en œuvre.

Compte tenu de ces orientations, cet ouvrage se veut accessible à un large public. Il s'adresse prioritairement à des étudiants en sciences du langage, de

niveau master. Dans la mesure où nous procédons à un inventaire des méthodes et des pratiques d'exploration de corpus, il s'adresse également à tout étudiant et à tout chercheur intéressé par la manipulation et l'exploitation de données textuelles. Comme les méthodes impliquées relèvent de plusieurs disciplines – la linguistique de corpus, l'Analyse (statistique) de Données Textuelles (ADT), le Traitement Automatique des Langues (TAL) – il s'adresse d'une manière générale aux trois communautés correspondantes.

1.1 L'exploration de corpus : principes et définitions

Nous commençons ce chapitre introductif par quelques définitions, car la description des méthodes nécessite un vocabulaire spécialisé. Nous enchaînerons avec un rapide aperçu historique de l'exploration de corpus, avant de présenter les deux grandes méthodologies actuelles et d'expliquer l'organisation de l'ouvrage.

1.1.1 *Corpus : de la collection de textes au corpus organisé*

Un corpus est une collection de **productions langagières attestées**, écrites ou orales. Ces productions peuvent être des textes, des extraits de textes, des enregistrements audio ou vidéo. Plusieurs productions sont donc réunies dans un corpus, ce qui amène à parler de **division en sous-corpus** (figure 1.1). Les tailles de ces sous-corpus et du corpus complet peuvent varier fortement, et tous les sous-corpus d'un corpus n'ont pas forcément la même taille.

Concernant la taille : on compte le nombre de mots d'un corpus écrit, ce qui amène certains à parler de **grand corpus** quand ce nombre dépasse un certain seuil, par exemple un million de mots, ou dix millions de mots, voire beaucoup plus depuis l'exploitation croissante de données massives issues du web. À l'oral, l'unité est plutôt la durée de parole enregistrée, et les corpus regroupant quelques dizaines d'heures d'enregistrement sont déjà de grands corpus. Dans les deux cas, la taille n'est pas un indicateur fiable de la quantité d'information contenue – et du travail préparatoire réalisé – notamment quand le corpus est annoté : le repérage et l'annotation des quelques occurrences d'une suite de mots particulière (par exemple toutes les apparitions de la locution « tout compte fait ») représente une densité d'information bien plus faible que l'annotation de tous les mots du texte avec des informations lexicales, morphosyntaxiques, syntaxiques et sémantiques diverses. Évaluer la taille d'un corpus doit tenir compte du nombre de mots et du nombre d'annotations. En général, on cherche à ce que les différents sous-corpus d'un corpus se caractérisent par les mêmes principes d'annotation, et donc par la même densité d'information. On parle alors d'**homogénéité** ou de **consistance interne**.

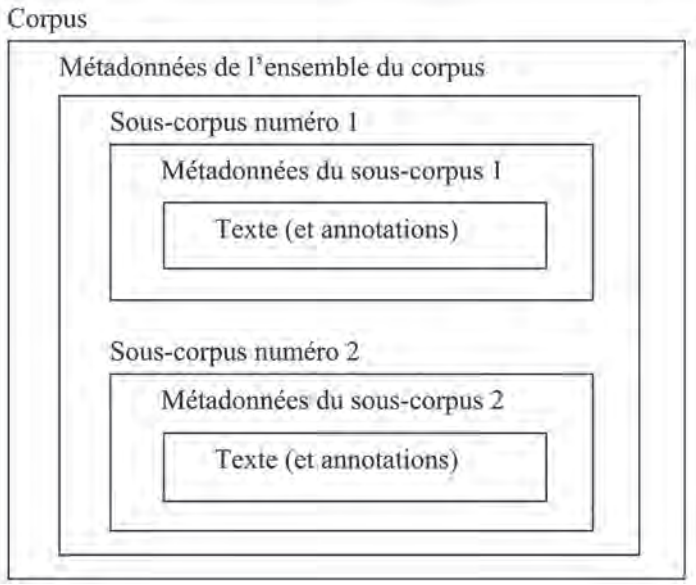

Figure 1.1
Organisation générale d'un corpus divisé en sous-corpus

Parmi les caractéristiques essentielles d'un corpus, notons que le format de représentation doit être **manipulable informatiquement**. Pour l'écrit, disposer du texte en format électronique – ASCII ou Unicode, par exemple – est indispensable, même si l'on garde par ailleurs les images des pages scannées correspondantes, notamment pour des textes anciens ou pour des documents multimédia. Pour l'oral, il s'avère difficile d'explorer des fichiers audio, et *a fortiori* des fichiers vidéo : représenter le langage est nécessaire, et ce sont le plus souvent des représentations (transcriptions) que l'on regroupe dans un corpus, que l'on annote et que l'on explore. Nous nous focaliserons dans cet ouvrage sur des exemples de corpus écrits, mais les principes décrits seront, compte tenu de cette représentation, souvent valables pour des corpus oraux.

Il est possible de constituer un corpus en regroupant tous les romans écrits par un auteur précis – corpus flaubertien, par exemple – ou en regroupant tous les discours prononcés par une personnalité politique lors d'un événement tel qu'une campagne électorale (Mayaffre 2012). Avec des contraintes telles que l'unicité d'auteur ou l'unicité du contexte de production, le regroupement garantit une certaine **cohérence** du corpus. Cette cohérence permettra de donner sens aux statistiques qui seront ensuite calculées. Si l'objet de la recherche est diachronique, avec par exemple l'étude de la variation dans le temps de l'usage des prépositions ou des adverbes, le regroupement visera à sélectionner des textes de chaque siècle, de manière à couvrir avec régularité une large période

de variation linguistique (Reppen *et al*. 2002). D'autres types de variation peuvent être étudiées, et on peut regrouper des textes en prose et en vers, des œuvres finales (publiées par un éditeur) et leurs brouillons, des productions écrites et orales, des productions d'enfants et d'adultes, de locuteurs natifs et d'apprenants, etc. D'une manière générale, les linguistes composent des corpus en suivant des critères qualitatifs précis, ce qui contraste avec l'approche informatique construisant de très grands corpus, directement à partir de pages web, en faisant intervenir très peu de critères qualitatifs.

Les possibilités sont multiples et les contraintes ou « règles » de constitution de corpus s'adaptent à chaque cas particulier, surtout si l'on recherche une certaine **représentativité**. Dans cet ouvrage, nous ne détaillerons pas ces règles que l'on peut retrouver, entre autres références, dans (Atkins *et al.* 1992 ; Biber 1993 ; Sinclair 1996 ; Habert *et al.* 1998 ; Vandelanote *et al.* 2014).

L'auteur, la date de parution ou d'enregistrement, la nature des annotations et la taille permettent de caractériser un corpus et chacun de ses sous-corpus. Ce sont des informations indispensables à une bonne exploitation du corpus, que le concepteur ajoute aux données langagières elles-mêmes, de manière à documenter celles-ci. On les appelle **métadonnées**, et on réserve une partie du corpus à leur description structurée. Selon l'orientation scientifique, par exemple le type de variation étudiée, de nombreuses métadonnées peuvent être renseignées : la description de la méthode de recueil des sous-corpus, éventuellement de l'**échantillonnage** (quand on coupe certains textes pour obtenir des sous-corpus de même taille), la description du contexte de production, ou encore celle des choix effectués lors du formatage informatique. Quand chaque modification apportée au corpus y est renseignée, les métadonnées permettent une gestion adéquate de la **traçabilité**. Les métadonnées sont ainsi fondamentales pour toute recherche empirique sur des corpus textuels : les linguistiques de corpus sont en effet des linguistiques de l'usage. Poursuivant un principe de variation, elles contrastent et interprètent les phénomènes linguistiques d'un usage et d'un contexte à l'autre. Sans métadonnées, l'analyste est démuni et peut difficilement interpréter les contrastes et les écarts que les méthodes lui suggèrent.

Nous ne creuserons pas dans cet ouvrage la nature et la structuration des métadonnées, facette essentielle de la constitution de corpus, mais nous renvoyons là aussi à la littérature, avec par exemple (Habert *et al.* 1998 ; Burnard 2005 ; Wynne 2005).

On peut considérer les métadonnées comme une première catégorie d'annotations. Historiquement, une **annotation** est une remarque écrite à la main dans la marge d'un texte, dans le but d'ajouter une explication aidant le lecteur, que ce lecteur soit l'auteur de l'annotation lui-même (en vue d'une future relecture) ou un tiers. Nous verrons dans le prochain chapitre de nombreux exemples d'annotations plus

linguistiques – des **explications linguistiques interprétatives** – mais ce premier exemple nous permet déjà de mettre en avant un aspect essentiel de l'approche suivie dans cet ouvrage : quand on explore un corpus, on cherche à explorer le texte en même temps que les annotations ; du moins, on cherche à rendre possible un tel mode d'exploration et à ne pas se limiter au texte seul. Pour ce qui est des métadonnées, l'exploration de corpus vise ainsi à rechercher des éléments d'un corpus en les mettant en parallèle avec les métadonnées du sous-corpus concerné. Selon le corpus, explorer les occurrences de « tout compte fait » entraîne par exemple des comparaisons de fréquences et de contextes d'apparition dans les sous-corpus écrits et oraux, ou dans les productions de tel auteur par rapport à tel autre, ou encore à telle ou telle époque pour un corpus diachronique. Les résultats de ces comparaisons pourront amener l'analyste à apporter des arguments confirmant ou infirmant telle ou telle hypothèse linguistique. Nous entrons ainsi dans le domaine de la linguistique de corpus.

1.1.2 *Outils : pour une linguistique de corpus informatisée*

On parle de **linguistique de corpus** (Habert *et al.* 1997 ; Biber *et al.* 1998 ; McEnery & Wilson 2001 ; Cheng 2011 ; McEnery & Hardie 2012) à partir du moment où ce sont des données langagières attestées – et pas seulement la réflexion et l'intuition du linguiste – qui permettent d'étayer des théories et des hypothèses linguistiques. Avec l'apparition des ordinateurs, la linguistique de corpus s'est outillée, c'est-à-dire que la méthodologie de gestion des données langagières s'est informatisée, et s'appuie donc sur des **outils informatiques**. Ces outils sont multiples : outils permettant l'annotation manuelle de textes ou d'enregistrements oraux ; outils pour la visualisation des données annotées ; outils d'interrogation de corpus, permettant l'expression de requêtes d'exploration et l'affichage des résultats obtenus de manière ergonomique ; outils de traitement automatique des langues, permettant d'ajouter automatiquement certaines annotations à un corpus ; outils d'importation et d'exportation de corpus dans des formats divers ; outils d'analyse de données ; outils de calculs statistiques, d'affichage des résultats et de génération d'indicateurs numériques sur le corpus ; outils de traitement de texte ; tableurs ; Systèmes de Gestion de Base de Données (SGBD), etc. Parmi tous ces outils, on peut distinguer ceux qui sont utilisés en dehors de la linguistique de corpus – tableurs et SGBD, notamment – de ceux qui sont spécifiques aux données langagières et que, à la suite de (Habert 2005a ; Habert 2005b), on peut appeler **instruments**. La linguistique de corpus outillée inclut ainsi la linguistique à l'instrument. Dans la suite de cet ouvrage, nous utiliserons le mot « outil » à la fois pour les outils et les instruments, sauf cas particulier dûment indiqué : d'une manière générale, notre présentation commencera avec des outils pour s'orienter ensuite vers des instruments.

La diversité des outils actuellement disponibles reflète la diversité des approches sur corpus. Nous avons mentionné plus haut trois communautés : celle de la linguistique de corpus, celle de l'ADT et celle du TAL. Or, bien souvent, les linguistes qui pratiquent la linguistique de corpus outillée se focalisent sur les outils d'annotation manuelle et de gestion de corpus, mais ne connaissent que de loin les outils d'ADT et ceux de TAL. De même, les spécialistes de TAL conçoivent et s'intéressent aux outils d'analyse permettant d'annoter automatiquement, et ne connaissent que de loin les outils d'ADT voire les outils d'annotation manuelle. Et ainsi de suite. Cette image un peu caricaturale a déjà évolué et est en train de laisser la place à plus d'**interopérabilité**. Par exemple, l'ADT cherche de plus en plus à s'appuyer non seulement sur le texte, mais aussi sur les annotations qui peuvent s'y ajouter, manuellement ou automatiquement.

En même temps, la diversité des outils reflète la forte spécialisation des approches. Les outils d'annotation montrent leurs limites quand il s'agit d'exploiter les données obtenues, c'est-à-dire d'en tirer des observations pertinentes, et s'utilisent ainsi en complément d'autres outils ou instruments. C'est aussi cette **complémentarité des outils** que nous voulons souligner dans cet ouvrage, complémentarité que l'on retrouve bien sûr au niveau des méthodes. Nous serons amenés à mettre celles-ci en perspective et à montrer que les besoins d'outils ne sont pas encore totalement satisfaits : comme cela apparaissait dans Bilger (2000 : 94), nous sommes toujours à la recherche d'un outil polyvalent pour la gestion d'un corpus annoté, qui en permette une exploration intelligente.

1.1.3 *Exploration : visualisation, interrogation, calcul de statistiques*

Explorer un corpus peut se faire de plusieurs manières prototypiques. Premièrement, **en parcourant le texte**, tout simplement page après page, et en faisant apparaître au besoin quelques annotations, si le corpus en comprend. C'est l'exploration manuelle, la plus immédiate et la plus spontanée, qu'on appelle aussi lecture assistée. Quand le corpus est annoté, elle nécessite un outil de visualisation dédié à l'affichage ergonomique – parfois interactif – des annotations, en surimpression ou en marge du texte. Plus le corpus comprend d'annotations, plus la visualisation est riche et peut s'accompagner de fonctionnalités de paramétrage, de manière à personnaliser l'affichage en fonction des préoccupations scientifiques ou des préférences graphiques de l'utilisateur. Pour visualiser le texte dans son ensemble, ou pour avoir une idée de la répartition des annotations – et de leur concentration – dans le corpus, certains outils proposent des équivalents des fonctions « zoom in » et « zoom out » classiques dans les logiciels de gestion d'image. La figure 1.2 montre ainsi un exemple permettant d'appréhender visuellement la concentration des annotations dans les différents paragraphes d'un document. Exploration et visualisation sont ainsi liées.

TABLE DES MATIÈRES

Champs linguistiques
RECHERCHES

« Champs linguistiques » crée un nouvel espace de réflexion sur tous les aspects du langage en éclairant la recherche contemporaine en linguistique française, sans a priori théorique et en ne négligeant aucune discipline.
Pour les linguistes professionnels : une occasion de donner libre champ à leurs recherches.
Pour les amoureux de la langue : une manière d'élargir le champ de leurs connaissances.
Pour les étudiants : un outil de travail et de réflexion.

Avec le virage numérique, les pratiques du linguiste ont sensiblement évolué. Décrire des discours et des usages ou mettre en évidence des phénomènes linguistiques particuliers passe de plus en plus par l'exploitation de corpus numériques pour mettre à l'épreuve ses hypothèses.

Cette pratique fait appel à de nombreux procédés, pour lesquels des repères méthodologiques s'avèrent indispensables : quelle méthode choisir pour quel objectif de recherche ? Pourquoi annoter un corpus ? Comment mettre au jour sa structure, ou dégager ses spécificités ? Quels sont les outils mobilisables ?

L'originalité de cet ouvrage est de proposer à l'analyste, de manière pratique et située, un ensemble de repères méthodologiques en lien avec les usages et les outils d'exploration de corpus les plus mobilisés dans le champ linguistique français. Il s'appuie sur un contexte institutionnel et des réflexions collectives menées dans le cadre d'un groupe de travail sur l'exploration de corpus et balise les méthodes présentées d'exemples concrets de recherches et d'outils exploitables.

Maître de conférences à l'Université de Nice Côte d'Azur, **Céline POUDAT** *travaille depuis une quinzaine d'années en analyse du discours, en mobilisant les méthodes quantitatives de l'analyse de données textuelles et de la linguistique de corpus. Des textes scientifiques aux narrations de soi, en passant par des formes de communication médiée par les réseaux, elle a ainsi parcouru différents genres et types de discours qui lui ont permis de développer des compétences méthodologiques solides en matière d'exploration de corpus. Elle a coordonné le groupe de travail « Exploration de corpus » du consortium « Corpus écrits » de la Très Grande Infrastructure de Recherche sur les Humanités Numériques.*

Docteur en informatique-linguistique, directeur de recherche en linguistique au CNRS, **Frédéric LANDRAGIN** *s'est spécialisé dans l'étude des expressions référentielles et des chaînes de coréférences au laboratoire Lattice (Langues, Textes, Traitements Informatiques, Cognition). Ses publications portent sur les phénomènes de référence et de saillance, ainsi que sur leur application dans divers domaines, notamment celui du traitement automatique des langues et celui de la linguistique de corpus outillée. Il dirige actuellement un projet sur l'annotation et l'analyse des chaînes de coréférences dans un corpus regroupant des textes issus de périodes et de genres textuels variés.*

ISBN 978-2-8073-0563-2

9 782807 305632

deboeck SUPÉRIEUR

www.deboecksuperieur.com