CORPUS

Corpus

15 | 2016
Corpus de français parlé et français parlé des corpus

Henry TYNE, Virginie ANDRÉ, Christophe BENZITOUN, Alex BOULTON et Yan GREUB (éd.), *French through corpora : ecological and data-driven perspectives in French language studies.* Newcastle upon Tyne UK : Cambridge Scholars Publishing, 2014, 343 p.

Filip Verroens

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/corpus/3085
ISSN : 1765-3126

Éditeur
Bases ; corpus et langage - UMR 6039

Édition imprimée
Date de publication : 15 octobre 2016
ISSN : 1638-9808

Référence électronique
Filip Verroens, « Henry TYNE, Virginie ANDRÉ, Christophe BENZITOUN, Alex BOULTON et Yan GREUB (éd.), *French through corpora : ecological and data-driven perspectives in French language studies.* Newcastle upon Tyne UK : Cambridge Scholars Publishing, 2014, 343 p. », *Corpus* [En ligne], 15 | 2016, mis en ligne le 15 janvier 2017, consulté le 08 septembre 2020. URL : http://journals.openedition.org/corpus/3085

Ce document a été généré automatiquement le 8 septembre 2020.

© Tous droits réservés

Henry TYNE, Virginie ANDRÉ, Christophe BENZITOUN, Alex BOULTON et Yan GREUB (éd.), *French through corpora : ecological and data-driven perspectives in French language studies.* Newcastle upon Tyne UK : Cambridge Scholars Publishing, 2014, 343 p.

Filip Verroens

1 Cet ouvrage vise à montrer comment la méthodologie de corpus fait fructifier plusieurs domaines linguistiques. Il importe donc de démontrer les liens entre les outils, les méthodes et les analyses. Comme le titre l'indique clairement, l'approche adoptée se veut inductive (*corpus-driven*) et écologique. Autrement dit, le corpus y est utilisé comme point de départ pour élaborer une théorie linguistique et les données sont authentiques tant dans leur origine que dans leur traitement. Le volume se compose de seize chapitres regroupés en quatre parties respectivement consacrées à la diachronie, à la syntaxe, à la sociolinguistique et à l'apprentissage du français. Chaque partie contient plusieurs contributions empiriques et est introduite par une contribution générale où l'on établit la relation entre le domaine de recherche en question et la méthodologie de corpus.

2 L'introduction à la première partie du volume (Bernard Combettes) résume les avantages de l'emploi de corpus et caractérise les problèmes auxquels les études diachroniques sont confrontées. Parmi les points positifs, l'auteur note comment les corpus peuvent changer le regard du diachronicien et par conséquent la manière de

traiter les données linguistiques historiques. Si auparavant les études étaient surtout de nature synchronique, c.-à-d. une photographie instantanée d'un item grammatical dans une période bien délimitée, elles portent de nos jours surtout sur le changement linguistique proprement dit à travers plusieurs périodes. Il s'ensuit que la périodisation traditionnellement reconnue et la question de la variation seront remises en cause. Finalement, la possibilité de la lecture 'verticale' à l'aide de concordanciers permet de mettre en lumière plusieurs paramètres contextuels qui restaient inaperçus auparavant. Quant aux inconvénients, l'auteur observe d'abord que le mérite des corpus dépend trop souvent du degré d'annotation. Pour l'instant, les seuls domaines de recherche qui profitent des corpus sont ceux où les données sont facilement repérables par l'ordinateur, p. ex. l'étude des expressions figées. L'attention accordée aux expressions figées peut amener des conséquences théoriques importantes étant donné que la notion de syntaxe est de plus en plus considérée en rapport avec des constructions figées plutôt qu'avec des constructions libres. Ensuite, une certaine prudence dans l'annotation automatique des textes historiques s'impose afin d'éviter des anachronismes au niveau des catégories et des unités syntaxiques. Enfin, il y a la question largement discutée de la représentativité qui, due à l'absence de certains genres et registres, semble plus problématique qu'en synchronie vu l'inventaire donné et clos des textes disponibles. Une première étude (Paul Isambert) montre comment le peu de données historiques semble à première vue contredire l'hypothèse que l'adverbe de manière *autrement* se grammaticalise vers un connecteur d'hypothèse négative. Or, une analyse synchronique détaillée permet ensuite de reconstruire la grammaticalisation et de montrer que l'évolution a eu lieu à travers la construction incluant l'adverbe. La position syntaxique qu'occupe cette construction convient bien à une réanalyse et ouvre ainsi la voie au connecteur. Une deuxième étude (Richard Ingham) porte sur la naissance des constructions discursives elliptiques en anglais (p. ex. *Haven't you heard Peter ? - Yes, I have*). Contrairement à d'autres qui expliquent cette construction par l'influence du substrat celtique, l'hypothèse avancée ici est qu'elle résulte du contact avec l'anglo-normand. Les arguments en faveur sont, primo, que les questions et réponses elliptiques en anglo-normand préexistent à l'anglais, qui est d'ailleurs la seule langue germanique qui en dispose. Secundo, elles apparaissent dans le même genre (les farces) et registre (le dialogue informel). Tertio, le parallélisme structurel révèle une réduplication grammaticale et le prestige de l'anglo-normand en Angleterre a favorisé sa distribution. Le décalage entre le moment de contact présumé et la première apparition dans les textes est expliqué par l'usage de la construction dans un registre oral informel précédant le registre écrit. Cette étude montre entre autres que la linguistique diachronique nécessite un corpus de dialogues en français pré-moderne. Une troisième étude (Inka Wissner) pose la question de savoir ce que les corpus textuels peuvent contribuer à l'étude des expressions diatopiquement marquées en français moderne, en l'occurrence la collocation *tomber en amour*. Cette expression a le statut de marqueur diatopique, puisqu'elle est généralement considérée comme un calque de l'anglais (*to fall in love*) usité en français québécois. Or, une analyse lexicographique exhaustive et une analyse sur corpus montrent que cette expression apparaît déjà au XIII^e^ siècle en France ce qui rend l'hypothèse d'un calque moins probable. Sans que l'auteur ne le dise explicitement, on suppose alors que *tomber en amour* est un des archaïsmes qui a subsisté dans le Nouveau Monde. Enfin, l'auteur regrette qu'il n'y ait toujours pas de corpus global pour toute la francophonie pourvu entre autres d'annotations diatopiques.

3 La deuxième partie contient trois chapitres axés sur l'emploi de corpus en syntaxe. Dans l'introduction, Christophe Benzitoun souligne la révolution que l'emploi massif des corpus a déclenchée dans ce domaine. En revanche, le recours à des corpus arborés (*treebanks*) pour analyser les données n'est pas encore très fréquent en français. Il est vrai que leur emploi n'est ni neutre, à cause des choix théoriques adoptés, ni parfait, à cause d'éventuelles erreurs d'annotation et que l'exemple le plus connu, le *French Treebank*, se limite au corpus journalistique du *Monde*. Ces facteurs expliqueraient pourquoi certains préfèrent travailler à partir des données brutes. L'ampleur et la qualité de l'analyse syntaxique changent aussi. D'une part, en se basant sur des corpus pour définir une grammaire, on préfère donner des tendances descriptives plutôt que des règles absolues, ce qui amène une certaine fragmentation dans la description. D'autre part, la collaboration récente entre la linguistique de corpus et la linguistique expérimentale peut aboutir à des interprétations plus raffinées des données de corpus, notamment sur le plan des jugements de grammaticalité. L'avenir est à l'éclectisme, mais la diversité des corpus et des méthodes utilisés requiert une certaine prudence. La première étude syntaxique (Juliette Thuilier, Anne Abeillé et Benoît Crabbé) concerne les préférences d'ordre des compléments postverbaux en français. Plusieurs études ont déjà montré que la longueur de l'objet, le statut discursif et le sens verbal jouent un rôle. La conclusion générale de la présente analyse est que 70,4 % des données préfèrent l'ordre NP-PP, mais la longueur de l'objet et la sémantique du verbe sont susceptibles d'inverser cet ordre. De plus, une analyse multifactorielle montre que ces deux facteurs et le corpus sont significatifs, contrairement aux facteurs [±pronominal], [±défini] ou [±animé] du NP ou du PP. Une deuxième étude (Nathalie Rigaud et José Delofeu) porte sur l'ellipse modale et diffère de l'étude précédente par son caractère inductif. On y montre que le fragment de surface est dans 75 % des cas une construction idiomatique (p. ex. *comme il faut*) associée à une interprétation indépendante du contexte. Dans les autres cas, le fragment occupe une position VP sans contenu lexical et sans besoin de structure syntaxique. Habituellement, la reconstruction du VP se fait grâce à un antécédent (*trigger*) explicite dans une phrase adjacente. Or, l'antécédent, un verbe, est parfois assez éloigné et l'interprétation ne se fait pas toujours par un recours à l'élément explicite, mais par inférence sémantique de l'ensemble du contexte. Par conséquent, on présume que l'antécédent et le fragment ne forment pas pour autant une unité syntaxique.

4 Les cinq chapitres de la troisième partie portent sur des études sociolinguistiques. L'introduction (Virginie André et Henry Tyne) rappelle d'abord que dans la longue tradition en dialectologie française, les méthodes utilisées ont été, et continuent d'être, celles de l'interview et du protocole, bref celles orientées par la recherche même (*researcher-driven*). Récemment, une approche écologique (*speaker-driven*) partant plutôt des données de la situation communicative authentique s'instaure. Cependant, en attendant de grands corpus pourvus de métadonnées sociolinguistiques, on se retourne encore fréquemment à de petits corpus locaux. Une seconde observation concerne l'influence de la masse de données disponible. Cette quantité peut dévier le focus du domaine de recherche qui est plutôt la manière dont la langue est utilisée et non la langue utilisée. Enfin, on constate une grande variation dans les théories et méthodes utilisées, ce que les études suivantes démontrent. La première étude (Emmanuelle Guerin et Roberto Paternostro) examine les caractéristiques de la langue des jeunes (LDJ) et de ses locuteurs dans le corpus *Multicultural Paris French* (MPF). Il s'avère que les traits de LDJ, à savoir l'emploi du /r/ arabisé, l'affrication de plosives et la structure de

la prosodie finale ne se retrouvent pas que chez les jeunes immigrés de la banlieue défavorisée. Une analyse du discours rapporté établit une relation entre les stratégies discursives et prosodiques utilisées en LDJ et la proximité communicative. Plutôt qu'à une langue, LDJ réfère à une situation communicative qui manifeste une grande complicité entre les interlocuteurs. Une deuxième contribution (Heike Baldauf-Quilliatre, Sylvie Bruxelles, Sabine Diao-Klaeger, Emilie Jouin-Cardon, Sandra Teston-Bonnard et Véronique Traverso) traite de l'élément *oh là là* dans le *Corpus de LAngue Parlée en Interaction* (CLAPI). L'analyse du corpus montre que cette particule assume des fonctions évaluative et affective dans l'interaction. Mais, elle apparaît aussi comme élément autonome où elle sert à attirer l'attention de l'interlocuteur. Ce n'est que grâce à une analyse multimodale que le rôle de *oh là là* comme moyen de dramatisation devient très apparent. Une troisième étude (Kate Beeching) décrit les caractéristiques du marqueur discursif postposé *quoi* dans des corpus parallèles afin de vérifier si les différentes traductions sont susceptibles de dévoiler un changement sémantique diachronique. Trois corpus oraux (1968-2002) montrent d'abord que la fréquence de *quoi* postposé a nettement augmenté. Les corpus parallèles d'INTERSECT et d'OPUS à eux seuls ne sont pas en mesure de vérifier si cette hausse entraîne un enrichissement pragmatique. Cependant, ensemble avec une analyse historique et un inventaire de traductions équivalentes, on arrive à la conclusion que *quoi* postposé tend à perdre sa force emphatique d'interjection (fonction subjective) en faveur de sens plus larges (réflexif, interpersonnel et approximatif) à fonction intersubjective. Dans la quatrième contribution, Fabienne Baider et Evelyne Jacquey vont à la recherche de préjugés sexuels dans le discours socialiste de 2012 (Aubry versus Hollande). Une analyse du corpus journalistique montre la perception des deux candidats : les données soulignent le dynamisme d'Aubry et la faible personnalité de Hollande. Bien que ce soit Hollande qui l'emporte dans le second tour, il n'y a aucune indication dans les données qui aurait pu annoncer sa victoire. Aubry, comme Royal en 2007, sont estimées compétentes, mais, bien que le discours ne manifeste pas de stéréotypes sexuels, elles semblent exposées à une discrimination négative.

5 Les quatre chapitres de la dernière partie se concentrent sur l'application de corpus dans un contexte d'apprentissage. Dans l'introduction (Alex Boulton et Henry Tyne), les notions d'écologie et d'apprentissage sont mises en rapport. L'*input* ne devient *intake* qu'à condition qu'il y ait une relation pertinente entre l'apprenant et son environnement linguistique (*affordance*). Une manière de créer cette relation est en travaillant sur des corpus. Une première étude (Tom Cobb) décrit ce que l'implémentation de DDL (*data-driven learning,* approche inductive basée sur des données de corpus) en français requiert. L'avantage d'une telle approche est que l'apprenant s'aperçoit plus vite de certaines infos en L2 (p. ex. collocations fréquentes) lorsque les données sont explicitées par le logiciel. Lextutor est un exemple d'un outil qui s'inscrit dans la DDL permettant à l'apprenant d'entraîner sur corpus ses compétences et connaissances en L2. Cependant, le nombre d'outils français en DDL est encore très limité faute de corpus plus larges et d'une adaptation pédagogique. Une seconde étude (Elodie Vialleton et Tim Lewis) examine dans quelle mesure l'accroissement de nouveaux corpus oraux a influencé l'authenticité dans le matériel éducatif pour des débutants adultes. Cependant, il s'avère que la plupart des dialogues sont enregistrés en studio. De plus, il y a une nette différence entre les dialogues en interaction naturelle et ceux du manuel au niveau de l'hésitation, des tours de parole et de l'articulation. Les propriétés de la parole authentique ne se retrouvent pas (assez)

dans les manuels. Par conséquent, les apprenants sont privés de la complexité de la parole authentique ainsi que de stratégies pour l'acquérir. Une troisième étude (Maud Dubois, Alain Kamber et Carine Skupien Dekens) présente une analyse de l'accord de l'adjectif en L2 (niveau B1). Le corpus comprend des textes narratifs, argumentatifs et des résumés rédigés par des locuteurs de six langues distinctes. Le nombre d'erreurs est différent d'après la L1 mais connaît une répartition homogène sur l'ensemble des adjectifs attributifs et prédicatifs. L'accord est surtout problématique dans la position post-nominale de l'adjectif attributif et lorsqu'il s'agit d'un adjectif au pluriel. Beaucoup d'erreurs s'expliquent par une prononciation incorrecte. C'est pourquoi le lien représentation phonétique - code écrit est crucial et mérite plus d'attention en classe de langue.

6 Il est vrai que *French through corpora* plaît pour plus d'une raison. Primo, à cause de l'organisation générale du livre. L'ouvrage ne manque pas d'articles forts et prévoit chaque fois une introduction dans laquelle des représentants éminents du domaine de recherche en question proposent des réflexions courtes mais pertinentes et où ils établissent le lien entre le domaine et la méthodologie de corpus, ce qui rend le livre très accessible à des chercheurs d'autres disciplines. Secundo, de l'approche pronominale en syntaxe à l'analyse du discours (CDA, *Critical Discourse Analysis*), sa force se situe incontestablement dans la richesse des disciplines linguistiques et des cadres théoriques présentés. Compte tenu de cette diversité et du choix de publier en anglais le livre est susceptible d'intéresser un grand public. La publication en anglais est un signal international important et montre la progression dans le domaine de la linguistique de corpus en français. Bien que plusieurs projets soient en cours, on peut en effet (e.a. p. 134 et 287) regretter le retard d'un grand corpus de référence, équilibré au niveau du genre. Mais, comme cet ouvrage et des bases de données (cf. Clarin, UGent Corpus Finder) l'indiquent, pendant longtemps il n'a pas manqué de corpus, mais plutôt d'études entièrement basées sur corpus. C'est précisément au niveau méthodologique que nous aurions voulu que ce volume soit plus ambitieux. Le sous-titre annonce une approche inductive, qui est l'approche généralement liée à la linguistique de corpus (Tognini-Bonelli 2001 ; Teubert & Krishnamurthy 2007), impliquant un traitement quantitatif et statistique (Biber & Reppen 2015 : 50-51). Or, nous constatons qu'un tiers des articles n'est pas quantitatif, certaines études (Thuilier et al. et Beeching) sont déductives (*corpus based*) et seulement deux études (Thuilier et al. et Vialleton & Lewis) utilisent des techniques statistiques. Si l'on fait un effort pour analyser les fréquences des phénomènes qu'on étudie, on devrait aussi vérifier si les différences observées dans la fréquence sont également statistiquement significatives. Renoncer à une telle analyse n'est pas seulement une chance ratée, mais cela augmente aussi le risque de trouver une explication linguistique pour des résultats de corpus qui, en fait, sont dus au hasard. Est-ce que nous devons en conclure que les auteurs n'ont pas atteint leurs objectifs ? Non, si le contenu ne répond pas entièrement aux attentes créées par le sous-titre, les auteurs ne se limitent pas à souligner l'importance de l'utilisation de corpus. L'aspect innovateur de *French through corpora* consiste, d'une part, en une discussion sommaire mais critique de l'état de la question méthodologique dans plusieurs domaines de recherche. D'autre part, il contient plusieurs études de cas empiriques intéressants sur le plan descriptif, méthodologique et théorique. Ceci dit, ce travail se présente comme un véritable complément à des guides d'introduction à la linguistique de corpus. Nous le recommandons non seulement aux linguistes actifs dans une des disciplines discutées dans le livre, mais certainement aussi aux didacticiens.

BIBLIOGRAPHIE

Biber D. & Reppen R. (2015). *The Cambridge handbook of English corpus linguistics.* Cambridge : Cambridge University Press.

Teubert W. & Krishnamurthy R. (éd.) (2007). *Corpus Linguistics. Critical Concepts in Linguistics. Vol. 1.* London : Routledge.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam : John Benjamins.

AUTEUR

FILIP VERROENS

Université de Gand