Corela
Cognition, représentation, langage

HS-13 | 2013
Statut et utilisation des corpus en linguistique

Statut et utilisation des corpus en linguistique

Laurence Vincent-Durroux et Philip Carr

Édition électronique
URL : http://journals.openedition.org/corela/3004
DOI : 10.4000/corela.3004
ISSN : 1638-573X

Éditeur
Cercle linguistique du Centre et de l'Ouest - CerLICO

Référence électronique
Laurence Vincent-Durroux et Philip Carr, « Statut et utilisation des corpus en linguistique », *Corela* [En ligne], HS-13 | 2013, mis en ligne le 07 janvier 2014, consulté le 01 mai 2019. URL : http://journals.openedition.org/corela/3004 ; DOI : 10.4000/corela.3004

Ce document a été généré automatiquement le 1 mai 2019.

Corela – cognition, représentation, langage est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Partage dans les Mêmes Conditions 4.0 International.

Statut et utilisation des corpus en linguistique

Laurence Vincent-Durroux et Philip Carr

1 La recherche en linguistique s'appuie de plus en plus sur des corpus, et ce, dans des domaines variés de la linguistique, qu'il s'agisse de la sociolinguistique (Docherty & Foulkes, 2000), de la syntaxe (Sampson, 1996, 2001), de la phonologie (Durand & Eychenne, 2004) ou encore de l'apprentissage des langues étrangères (Gregg, 2003). Les contributions rassemblées ici abordent les questions cruciales qui se posent aux linguistes quant au statut et à l'utilisation de corpus dans leurs travaux, questions, d'ordre métathéorique, théorique ou méthodologique.

2 Ces contributions font suite à deux journées d'étude organisées par l'Équipe EMMA (Études Montpelliéraines du Monde Anglophone), Université Montpellier III, les 1er et 2 juin 2012, journées suscitées par les questionnements suivants.

3 L'objet de l'analyse linguistique est-il constitué d'énoncés regroupés ? Les données des corpus devraient-elles venir en complément des jugements intuitifs d'acceptabilité (Durand, 2009) ou bien avoir préséance sur eux, voire les remplacer ainsi que le suggèrent Sampson (2005) pour la syntaxe et Pierrehumbert *et al.* (2000) pour la phonologie ? Ou bien Itkonen (1978) a-t-il raison de poser que l'analyse grammaticale repose fondamentalement sur des jugements intuitifs d'acceptabilité, puisque, selon lui, ces jugements sont accessibles grâce à une connaissance des conventions sociales plutôt que grâce aux phénomènes observables (raison pour laquelle Itkonen affirme que le travail sur corpus pour l'analyse grammaticale ne sert à rien ; voir la discussion dans Riemer, 2009a, 2009b et Lopez-Serena, 2009).

4 L'utilisation de corpus est-elle une garantie du statut empirique et / ou scientifique de l'analyse menée ? Que signifient exactement les termes "empirique" et "scientifique" ? Si l'analyse menée sur corpus est empirique, une telle analyse conforte-t-elle une certaine version de l'empirisme en linguistique ?

5 Le travail sur corpus conduit-il à favoriser une approche théorique plutôt qu'une autre ? Par exemple, Arndt-Lappe (2011) affirme que l'analyse des composés nominaux en anglais

dans des corpus plaide en faveur de la théorie des instances (" exemplar theory ": voir Bybee, 2001, Pierrehumbert *et al.* 2001) plutôt que des approches génératives (Giegerich, 2004 et Liberman & Sproat, 1992). L'analyse sur corpus peut-elle éclairer, et si oui de quelle manière, le rôle joué par la fréquence des occurrences, rôle auquel font appel les approches linguistiques fondées sur la théorie des instances ?

6 Dans quel sens différents corpus (tels que le *Brown corpus*, le *LOB corpus*, le *British National Corpus*) peuvent-ils être considérés comme représentatifs ? Et représentatifs de quoi, précisément ? Dans quelle mesure des corpus peuvent-ils être utilisés par des chercheurs qui ne les ont pas constitués ? Les corpus nous donnent-ils accès à des données "objectives", non liées à une théorie et renvoyant à une forme de réalité ? Ou bien Scheer (2004) a-t-il raison d'affirmer qu'"il n'existe pas de corpus sans théorie" et que "le corpus ne représente pas *la* réalité: il représente la réalité de celui qui l'a construit" ?

7 Les trois premières contributions (T. Scheer, P. ten Hacken et R. Panocová, N. Arbach et S. Ali) présentées dans ce volume abordent ces questions des points de vue métathéorique et théorique. Dans les quatre contributions suivantes (C. Brasart, L. David, J. Sauvage, C. Dodane, F. Hirsch et M. Barkat-Defradas, P. Artero et A. Şerban), les auteurs illustrent et débattent de ces questions en se fondant sur l'analyse de la méthodologie de leurs travaux respectifs.

8 La contribution de Tobias Scheer (Université de Nice, France) permet de situer les corpus en rapport avec l'engouement technologique de ces dernières décennies et dans le cadre de la confusion fréquente entre les données elles-mêmes et les connaissances que celles-ci sont susceptibles d'apporter. T. Scheer montre que les corpus véhiculent un *a apriori* d'objectivité, et ce, à tort, puisqu'en tant qu'outils, ils sont constitués dans un but précis et dans un cadre théorique. De plus, les corpus présentent de multiples limites puisqu'ils ne peuvent attester de ce qu'ils ne comportent pas et ne font état que de la performance. La réflexion de T. Scheer porte également sur la linguistique de corpus.

9 Pius ten Hacken (Swansea University) et Renáta Panocová (P.J. Šafárik University, Košice) abordent les corpus sous l'angle de leur nécessité ou de leur caractère non indispensable selon la recherche menée. Le contexte observé est celui de la formation des mots, et plus particulièrement l'étude de la productivité. En comparant trois approches de l'étude de la productivité, celle de Baayen, celle de la linguistique chomskyenne et celle de Štekauer qui distingue la productivité et la fréquence et définit celle-là de telle façon qu'on peut se passer de l'utilisation d'un corpus, les auteurs montrent qu'une utilisation systématique de corpus permettrait de renforcer les conclusions obtenues sur la productivité.

10 Najib Arbach et Saandia Ali (Université Rennes 2) consacrent leur article au critère de représentativité des corpus en illustrant ses différents aspects et en présentant des méthodologies susceptibles d'être utilisées pour rechercher à atteindre le caractère représentatif d'un corpus : catégorisations, échantillonnage, volume des données. Les auteurs portent un regard critique sur deux courants méthodologiques : la « stratification en amont » représenté par Biber et le courant des *monitor corpus* représenté par Sinclair.

11 Charles Brasart (Université Paris 4) démontre le caractère indispensable d'un travail croisé sur différents corpus dans le domaine de l'alternance codique chez les sujets bilingues. C. Brasart montre que l'alternance codique est un phénomène discursif qui ne peut être caractérisé par les jugements d'acceptabilité mais dont peut rendre compte une analyse croisée de corpus émanant de locuteurs bilingues avec différents couples de langues. L'étude présentée porte sur les couples de langues Français / Anglais et

Allemand / Anglais et fait apparaître que l'alternance codique se fait selon les mêmes modalités dans les deux groupes et concerne des éléments de la langue qui, jusque là, étaient souvent considérés comme étant non affectés par l'alternance codique.

12 Laurent David (Université Paris 3) met en relation d'une part, différentes approches théoriques sur le *Present Perfect* en anglais, d'autre part les études psycholinguistiques antérieures sur l'acquisition du *Present Perfect* , en vue de sélectionner les paramètres que doit avoir un corpus destiné à l'analyse de l'acquisition du *Present Perfect.* Cela a guidé son choix de corpus issus du projet CHILDES. Sur ces données, la mise en œuvre conjointe d'analyses quantitatives (distinction entre les formes reprises et les formes initiées par l'enfant) et qualitatives (étude du développement morpho-syntaxique, sémantique et pragmatique) permet de montrer le rapport entre le développement linguistique et l'émergence des capacités cognitives de l'enfant.

13 Jérémi Sauvage, Christelle Dodane, Fabrice Hirsch et Melissa Barkat-Defradas (Université Montpellier 3) abordent les difficultés de construire un corpus leur permettant d'allier des analyses en synchronie et une approche longitudinale dans le cadre d'une étude sur les liens entre le niveau de développement du langage de l'enfant et la structure acoustique du rire. Cela soulève des questionnements pour ce type d'étude quant à la pertinence des grands corpus mutualisés de données enfantines, quant au choix des indicateurs à extraire et à analyser, et quant à la mise en œuvre conjointe d'approches quantitatives et qualitatives.

14 Paola Artero et Adriana Şerban (Université Montpellier 3) situent leur réflexion dans le domaine de la traductologie où les corpus ont pu être utilisés pour décrire des normes et des protocoles se révélant indépendants des langues traduites et pour fonder la discipline sur des bases empiriques. Les analyses quantitatives permettant notamment de situer les fréquences d'emplois de mots ou d'expression constituent aussi une aide majeure en traduction. Toutefois, l'apport d'analyses contextualisées et qualitatives aux analyses quantitatives semble indiscutable pour les auteurs qui appliquent cette méthodologie à une analyse de la traduction de *Narnia* afin de faire apparaître des tendances chez les traducteurs.

15 Comité scientifique : Philip Carr, Laurence Vincent-Durroux (EMMA - Université Montpellier 3), Jacques Durand (Université de Toulouse Le Mirail)

BIBLIOGRAPHIE

ARNDT-LAPPE, S. (2011). ‘Towards an exemplar-based model of stress in English noun-noun compounds’. *Journal of Linguistics* 47 (3) : 549-585.

BYBEE, J. (2001). *Phonology and language use*. Cambridge : Cambridge University Press.

DOCHERTY, G. & P. FOULKES (2000). ‘Speaker, speech and knowledge of sounds’ In Burton-Roberts, N., P. Carr & G. Docherty (eds) *Phonological knowledge: conceptual and empirical issues*. Oxford : Oxford University Press. 105-130.

DURAND, J. (2009). ‘On the scope of linguistics: data, intuitions, corpora’. In Y. Kawaguchi, M. Minegishi & J. Durand (eds), *Corpus and Variation in Linguistic Description and Language Education*. Amsterdam/Philadelphia : John Benjamins. 25-52.

DURAND, J. & J. EYCHENNE (2004). ‘Le schwa en français: pourquoi des corpus?’ In Scheer, T. (ed.) *Usage des corpus en phonologie*. *Corpus* 3 : 311-356.

GIEGERICH, H. (2004). ‘Compound or phrase? English noun-plus-noun constructions and the stress criterion.’ *English Language and Linguistics* 8 (1) : 1-24.

GREGG, K.R. (2003). ‘SLA theory: construction and assessment.’ In Doughty & Long (eds) *The Handbook of Second Language Assessment*. Oxford : Blackwell. 831-865.

ITKONEN, E. (1978). *Grammatical theory and metascience*. Amsterdam : Benjamins.

LIBERMAN, M. & R. SPROAT (1992). ‘The stress and structure of modified noun phrases in English.’ In Sag, I. & Szabolcsi (eds) *Lexical matters.* Stanford : CSLI publications. 131-181.

LOPEZ-SERENA, A. (2009). ‘Intuition, acceptability and grammaticality: a reply to Riemer.’ *Language Sciences* 31 (5) : 634-648.

PIERREHUMBERT, J. (2001). ‘Exemplar dynamics: word frequency, lenition, and contrast.’ In BYBEE, J & P; HOPPER (eds.) *Frequency effects and the emergence of lexical structure*. Amsterdam : Benjamins. 137-157.

PIERREHUMBERT, J., M.E. BECKMAN & D.R. LADD (2000). ‘Conceptual foundations of phonology as laboratory science. In Burton-Roberts, N., P. Carr & G. Docherty (eds.) *Phonological knowledge: conceptual and empirical issues*. Oxford : Oxford University Press. 273-304

RIEMER, N. (20079). ‘Grammaticality as evidence and as prediction in a Galilean linguistics’. *Language Sciences* 31 (5) : 612-633.

RIEMER, N. (2009b). ‘On not having read Itkonen: empiricism and intuitions in the generative data debate.’ *Language Sciences* 31.5: 649-662.

SAMPSON, G. (1996). ‘From central embedding to corpus linguistics’. In Thomas, J. & M. Short (eds.) *Using corpora for language research*. Londres : Longman. Reproduit dans Sampson (2001).

SAMPSON, G. (2001). *Empirical linguistics*. Londres : Continuum.

SAMPSON, G. (2005). *The ‘language instinct’ debate*. Londres : Continuum.

SCHEER, T. (2004). ‘Le corpus heuristique: un outil qui montre mais ne démontre pas’. In SCHEER, T. (ed.) *Usage des corpus en phonologie*. *Corpus* 3 : 153-192.

AUTEURS

LAURENCE VINCENT-DURROUX

Université Montpellier 3 / EMMA EA 471

PHILIP CARR

Université Montpellier 3 / EMMA EA 471