Corela

Cognition, représentation, langage

HS-21 | 2017

**Linguistique de corpus : vues sur la constitution,
l'analyse et l'outillage**

Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage

Iris Eshkol-Taravella et Anaïs Lefeuvre-Halftermeyer

Édition électronique

URL : <http://journals.openedition.org/corela/4800>

DOI : 10.4000/corela.4800

ISSN : 1638-573X

Éditeur

Cercle linguistique du Centre et de l'Ouest - CerLICO

Référence électronique

Iris Eshkol-Taravella et Anaïs Lefeuvre-Halftermeyer, « Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage », *Corela* [En ligne], HS-21 | 2017, mis en ligne le 27 janvier 2017, consulté le 30 avril 2019. URL : <http://journals.openedition.org/corela/4800> ; DOI : 10.4000/corela.4800

Ce document a été généré automatiquement le 30 avril 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.

Linguistique de corpus : vues sur la constitution, l'analyse et l'outillage

Iris Eshkol-Taravella et Anaïs Lefeuvre-Halftermeyer

- 1 Les possibilités offertes par l'évolution des nouvelles technologies et notamment les techniques d'exploration des documents numériques ont permis des développements théoriques fondés sur l'exploitation de corpus, mettant ceux-ci aux centres de la description et de l'analyse linguistiques. Elles sont devenues l'atout principal de la linguistique de corpus. En effet, les gros corpus de textes électroniques aujourd'hui disponibles, notamment sous leur forme brute ou annotée, permettent de faire des recherches et des analyses nouvelles.
- 2 Pour bien apprécier les enjeux contemporains de la linguistique de corpus, commençons par un bref aperçu historique. A partir des années soixante, l'avènement d'ordinateurs de plus en plus puissants commence à faciliter l'analyse linguistique des textes. L'année 1964 marque le véritable coup d'envoi des travaux sur corpus, avec la constitution du premier corpus informatisé par une équipe de chercheurs de l'université Brown aux Etats-Unis¹. En France, une dizaine d'années plus tard, on commence à rassembler la base textuelle *Frantext*². Avec l'arrivée du traitement de texte dans les années quatre-vingt, la plupart des données textuelles sont automatiquement saisies sur support électronique, ce qui permet une apparition des corpus écrits de taille variable. Les corpus de l'oral sont moins volumineux et demandent plus de temps de constitution car ils nécessitent une étape supplémentaire, une transcription manuelle des enregistrements sonores. Mais petit à petit ils apparaissent aussi sur la Toile et deviennent donc accessibles à l'analyse (CLAPI³, PFC⁴, CRFP⁵, ESLO⁶, etc.). Les questions de la place et du rôle des corpus dans la recherche en linguistique se posent (Williams 2005), et nous commençons tout juste à voir apparaître des corpus natifs du Web ou des corpus multimodaux.
- 3 Le terme de « linguistique(s) de corpus » est emprunté au courant britannique « corpus linguistics ». La linguistique de corpus remplace la linguistique qui utilise des exemples préconstruits fondés sur les intuitions des locuteurs natifs. Les exemples sont tirés maintenant directement des observations et des exploitations des corpus grâce à des

outils informatiques développés spécialement ou non pour cette tâche. La démarche est donc inductive, on construit des descriptions de formes à partir de leur usage. Les vastes corpus électroniques permettent d'obtenir des exemples attestés d'un phénomène donné, de découvrir de nouvelles observations et de prendre en compte des faits non accessibles manuellement, de développer des outils linguistiques comme des dictionnaires, des bases de données, des grammaires ou encore de construire des outils du TAL.

- 4 Ce numéro réunit les communications sélectionnées parmi celles présentées lors des 8es Journées Internationales de Linguistique de Corpus (JLC2015) organisées par le Laboratoire Ligérien de Linguistique (LLL) qui ont eu lieu à Orléans les 3-5 septembre 2015 à l'Université d'Orléans (<http://jlc2015.sciencesconf.org/>). Les travaux exposés pendant ces journées illustrent bien la diversité des recherches menées généralement en linguistique de corpus. Ils se distinguent par types de corpus traités, par objectifs, par méthodes et outils de traitement aussi bien que par domaines des sciences du langage auxquels ils renvoient comme le montre la carte mentale ci-dessous.

- 5 Les corpus qui font l'objet de ce champ de recherche varient selon le type : corpus écrit vs oral, corpus multilingue aligné ou non, corpus d'apprenants, corpus collaboratif, corpus brut vs annoté, etc. La linguistique de corpus permet une diversité d'approches : extraction de patrons lexico-syntaxiques en utilisant les concordanciers ou en créant des grammaires locales, textométrie, TAL, annotation, etc. Les linguistes peuvent à présent tester leurs hypothèses ou vérifier leurs théories grâce aux logiciels (étiqueteur, concordancier ou logiciel d'aide à la transcription de l'oral) dédiés à un traitement automatique des données linguistiques que (Habert 2004, 2006) définit comme des « instruments ».
- 6 Le premier volet de cette sélection d'articles représente les travaux concernant l'exploration et la description des phénomènes linguistiques. C'est le cas de la contribution de Lotfi Abouda et Marie Skrovec qui s'intéressent au rapport micro-diachronique futur simple/futur périphrastique en français moderne dans le corpus oral ESLO⁷ (Enquêtes Sociolinguistique à Orléans). En se fondant sur une annotation manuelle, les auteurs relient des résultats quantitatifs à des hypothèses qualitatives en utilisant le logiciel TXM⁸. Olga Nádvorníková traite dans son travail de corpus parallèles. Ces corpus sont composés de textes originaux et de leurs traductions respectives. L'auteur pose des questions liées à la méthodologie de leur exploitation et les confronte au corpus InterCorp⁹. Sylvain Hatier et Rui Yan font une analyse comparative entre des scripteurs étudiants (apprenants du français et natifs) et des scripteurs experts (chercheurs) en ce qui concerne l'emploi de verbes fréquents dans l'écrit scientifique (*montrer, considérer, décrire, expliquer*). Pour procéder à cette analyse, les auteurs utilisent une méthodologie d'extraction semi-automatique de patrons verbaux. Suzanne Mpouli et Jean-Gabriel Ganascia proposent quant à eux la classification d'un lexique des adjectivaux contenant un nom de couleur à partir d'un corpus de romans britanniques du XIX^e et du XX^e siècles extraits depuis la base Gutenberg¹⁰ à l'aide de Treetagger¹¹. Cette étude diachronique sur

corpus offre une analyse du comportement du patron *nom+terme désignant une couleur* (comme *snow-white*) en comparaison avec le comportement des comparatifs explicites (*white as snow*).

- 7 Le second volet de cette sélection d'articles présente les travaux concernant la constitution des corpus annotés. Nathalie Gasiglia et Anne-Pascale Pouey-Mounou proposent un projet d'édition d'un corpus du XVI^e siècle, *Le dictionnaire d'Épithètes* de Maurice La Porte. Le dictionnaire, sous son format électronique, sera enrichi d'une analyse d'envergure : dans la continuité des travaux de La Porte, est proposée une classification des épithètes attestés au sein du corpus des textes de la Pléiade. Les auteurs mettent en évidence plusieurs écueils à éviter dans les diverses étapes de l'édition d'un tel corpus, tenant autant à la difficulté d'accès aux formes anciennes qu'au caractère insaisissable des interprétations linguistiques en diachronie. Les contraintes techniques de l'édition numérique sont abordées ainsi que les apports méthodologiques d'un travail d'exploration de textes anciens. Karën Fort met au centre de ses préoccupations l'annotation manuelle des corpus par la méthode collaborative appelée « myriadisation » produite par une masse de gens, via le Web et pose des questions quant au niveau d'expertise des annotateurs. Pour que ces annotateurs deviennent des experts de la tâche qui leur est attribuée, l'auteur propose l'utilisation de jeux bien conçus, comportant une solide phase de formation et un protocole d'évaluation strict.
- 8 L'article proposé par Audrey Bonvin et Amelia Lambellet explore et éprouve des métriques utilisées en didactique concernant la richesse lexicale des productions langagières d'apprenants L2 Français ou Allemand. Les auteurs discutent les intérêts et limites de ces métriques, subjectives ou algébriques, pour l'étude de la compétence langagière sous-jacente chez de jeunes enfants, particulièrement lorsque les productions sont courtes.
- 9 Enfin, le dernier volet d'articles sélectionnés porte sur l'outillage des corpus. Jorge Mauricio Molina Mejia présente un système d'Apprentissage des Langues Assisté par Ordinateur (ALAO) comprenant des ressources annotées par Cordial Analyseur¹² et manuellement en coréférence. Ces ressources sont assorties d'une série d'activités permettant aux étudiants de s'approprier progressivement divers phénomènes linguistiques en construisant leur expertise sur une expérience en corpus. George Christodoulides et Giulia Barreca proposent l'outil d'annotation multiniveau DisMo. Cet outil conçu pour l'oral est testé et évalué sur le corpus PFC¹³. DisMo est capable d'annoter différents phénomènes (informations morpho-syntaxiques, disfluences de l'oral, chunks, etc.) avec une précision comparable à celle des systèmes développés pour l'écrit.
- 10 Les articles présentés dans ce numéro montrent bien toute la diversité, l'hétérogénéité et l'interdisciplinarité mobilisés par le champ de la linguistique de corpus qui fait intervenir des didacticiens, syntacticiens, sociolinguistes, lexicologues et chercheurs en analyse du discours ou encore en traitement automatique du langage.
- 11 Nous tenons à remercier les auteurs pour la qualité de leurs contributions, le comité de lecture composé de Jean-Yves Antoine, Karen Fort, Nathalie Gasiglia, Marie-Hélène Lay, Julien Longhi, Denis Maurel, François Nemo, Marie Skrovec, Dominique Ulma pour leur précieux travail, ainsi que l'équipe de la revue Corela et Gilles Col pour avoir accueilli ces pages.

BIBLIOGRAPHIE

Cori M., David S., Léon J. (2008). Présentation : éléments de réflexion sur la place des corpus en linguistique. *Langages*, 3, n° 171, pp. 5-11.

Geoffrey W. (dir.) (2005). *La linguistique de corpus*. Presses Universitaires de Rennes.

Habert B. (2004). Outiller la linguistique : de l'emprunt de techniques aux rencontres de savoirs. *Revue Française de Linguistique Appliquée*, vol. 9, n° 1, pp. 5-24.

Habert B. (2006). TAL sur corpus : histoire, acquis, défis. In. G. Sabah (ed.) *Compréhension des langues et interaction*, Lavoisier, Hermes, Paris, pp. 249-271.

NOTES

1. <https://www.sketchengine.co.uk/brown-corpus/>
 2. <http://www.frantext.fr/>
 3. <http://clapi.univ-lyon2.fr/>
 4. <http://www.projet-pfc.net/?accueil:intro>
 5. <http://www.up.univ-mrs.fr/delic/crfp>
 6. <http://eslo.huma-num.fr/>
 7. <http://eslo.huma-num.fr/>
 8. <http://textometrie.ens-lyon.fr/>
 9. <https://ucnk.ff.cuni.cz/intercorp/?lang=fr>
 10. <https://www.gutenberg.org/>
 11. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/>
 12. http://www.cordial.fr/Cordial_Analyseur/Presentation_Cordial_Analyseur.htm
 13. <http://www.projet-pfc.net/>
-

AUTEURS

IRIS ESHKOL-TARAVELLA

ANAÏS LEFEUVRE-HALFTERMEYER