Fonctions lexicales et annotation de corpus : un outil pour l'apprentissage des collocations linguistiques

Les collocations jouent un rôle clé dans l'apprentissage des langues : leur maîtrise est un élément essentiel d'une compétence linguistique élevée, mais la nature idiosyncratique et imprévisible du lien entre bases et collocatifs pose des difficultés même aux étudiants de niveau avancé.

Depuis une vingtaine d'années, l'étude des expressions polylexicales a reçu une attention croissante en didactique des langues et en lexicographie (Primož, 2011). Malgré cela, l'apprentissage des collocations reste un point critique. Les ressources disponibles pour l'apprentissage des collocations présentent souvent des limites importantes. Par exemple, dans les dictionnaires de collocations, les collocatifs d'une lexie sont le plus souvent regroupés par partie du discours ou sont listés en ordre alphabétique, sans distinctions sémantiques, ni détails sur la fréquence d'usage. C'est sur ce dernier point que les corpus pourraient venir en aide, car ils permettent d'observer la fréquence et la distribution des mots. Les corpus sont désormais largement exploités en didactique des langues et dans la formation d'interprètes et de traducteurs, et leur utilité dans l'étude des collocations est reconnue (Bernardini et al., 2003 ; Sinclair, 2004 ; Gallego-Hernandez et Rodriguez-Inés, 2016). Toutefois les corpus (s'ils ne sont pas annotés sémantiquement) n'offrent pas d'informations explicites sur le lien sémantique entre une lexie et ses collocatifs, ni sur les différences de sens entre plusieurs collocatifs, donc les corpus n'aident pas toujours un étudiant étranger à s'orienter.

Pour toutes ces raisons, nous avons reconnu la nécessité d'un nouvel outil qui faciliterait le repérage, l'étude et l'apprentissage des collocations. L'outil que nous proposons consiste en une annotation des collocations en corpus basée sur les fonctions lexicales de la théorie Sens-Texte (Mel'čuk, 1996).

Les fonctions lexicales constituent une notation formelle permettant de représenter et de modéliser de façon systématique les collocations (en particulier le lien sémantique entre base et collocatif), et de classifier les collocatifs d'une lexie selon un critère sémantique. Les fonctions lexicales constitueraient donc une aide pour l'apprentissage des collocations.

L'annotation proposée consiste à assigner à chaque collocatif d'une lexie une étiquette avec la fonction lexicale qui lui correspond. De cette façon, plusieurs types d'informations seraient explicitées : information syntaxique (par exemple, la configuration des actants syntaxiques profonds d'un collocatif verbal) et information sémantique. Grâce à cette annotation, il serait possible de distinguer les collocatifs par leur sens.

Le corpus choisi pour l'annotation est EPTIC (*European Parliament Translation and Interpreting Corpus*), corpus trilingue, parallèle et intermodal réalisé par une équipe de chercheurs du Département d'Interprétation et Traduction de l'Université de Bologne (Ferraresi & Bernardini, sous presse). Le choix du corpus EPTIC est motivé par deux facteurs principaux : 1) les textes institutionnels européens sont une typologie textuelle très exploitée dans la formation de traducteurs et interprètes ; 2) notre annotation implique une série d'opérations à effectuer manuellement, donc les dimensions contenues du corpus EPTIC (330 000 mots environ) facilitent la tâche.

Notre étude consiste en une proposition méthodologique. L'annotation n'a pas été réalisée sur la totalité des collocations du corpus ; nous avons choisi de nous concentrer sur une seule langue (l'italien, c'est-à-dire les sous-corpus italiens d'EPTIC) et sur un seul lexème (DIRITTO 'droit', un substantif dont la fréquence et la variété de collocatifs sont intéressantes aux fins

de notre annotation). Notre étude de cas vise à présenter les bénéfices potentiels de ce type d'annotation si elle était réalisée sur l'ensemble des lexèmes d'un corpus (et, idéalement, sur un corpus de dimension plus importante). L'annotation a été réalisée manuellement, mais il serait possible de partiellement automatiser le processus en ayant recours au bootstrapping.

Le corpus EPTIC a été précédemment annoté avec TreeTagger,[1] dont le format de codage est un format vertical CWB-compliant (Ferraresi & Bernardini, sous presse). Notre annotation a donc été codée en format vertical, en ajoutant une colonne supplémentaire (nommée `lexf` pour *Lexical Function*) à la structure du corpus. Le tagset de l'annotation a été conçu selon les critères de Leech (2005), qui souligne l'importance d'utiliser des étiquettes *brèves*, *univoques* et *transparentes pour l'utilisateur humain* dans l'annotation de corpus. Nos étiquettes correspondent aux noms des différentes fonctions lexicales, qui possèdent les caractéristiques citées par Leech.

Une fois que l'annotation est indexée, le corpus peut être consulté à travers NoSketch Engine.[2] Différents modes d'exploration sont possibles : on peut choisir d'observer les concordances du lexème DIRITTO (où les collocatifs apparaîtront avec l'étiquette de la fonction lexicale correspondante) ou bien on peut effectuer des recherches ciblées à l'aide d'expressions régulières, par exemple en cherchant tous les collocatifs qui correspondent à une fonction lexicale donnée, ou trier les collocatifs par leur sens, etc. Pour ces raisons nous estimons que ce type d'annotation serait un outil efficace pour l'apprentissage des collocations (surtout pour les futurs traducteurs et interprètes) mais aussi pour la consultation.

Le corpus ainsi annoté constituerait une ressource pour l'apprentissage qui présente les caractéristiques suivantes :

- en tant que corpus, il s'agit d'une ressource fondée sur des données textuelles réelles ; on peut donc l'utiliser pour effectuer des analyses statistiques sur la fréquence et la distribution des collocatifs ;
- l'annotation par fonctions lexicales fournit des informations sémantiques sur les collocatifs et des informations sur la nature du lien entre base et collocatif ;
- toutes les informations saillantes sont concentrées dans une seule ressource ;
- ce type de ressource encourage l'étudiant à analyser les données de façon autonome et à réfléchir sur les résultats, au lieu d'accepter passivement les informations fournies par des outils comme les dictionnaires.

Il est évident que, afin de pouvoir exploiter efficacement les informations contenues dans l'annotation, l'utilisateur devra posséder des notions de base relatives aux fonctions lexicales de Mel'čuk. Toutefois on estime que, avec le support de notre tagset, qui explique et vulgarise la notation formelle, l'utilisateur pourrait se servir de l'annotation sans connaître en détail le système des fonctions lexicales. Néanmoins, en accord avec Mel'čuk (1998, 2003), Polguère (2000, 2003), L'Homme (2009, 2010) et Primož (2011), nous estimons que l'introduction des fonctions lexicales et de notions de linguistique Sens-Texte dans la didactique des langues apporterait d'importants bénéfices, en particulier dans la formation des futurs traducteurs et interprètes, car "applied systematically and consistently within a given lexical field, the system of LFs can help translation students get a better grasp of the elusive collocability of lexemes" (Primoz, 2011 :129).

[1] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[2] https://nlp.fi.muni.cz/trac/noske

RÉFÉRENCES BIBLIOGRAPHIQUES

Bernardini, S. (2004). “Corpora in the classroom. An overview and some reflections on future developments”. Dans Sinclair J. dir. (2004) *How to Use Corpora in Language Teaching*. Amsterdam/Philadelphia: John Benjamins, 15-36.

Bernardini, S., Stewart, D., et Zanettin, F., dir. (2003). *Corpora in translator education*. Manchester: St. Jerome.

Ferraresi, A. et Bernardini, S. (sous presse). "Building EPTIC: A many-sided, multi-purpose corpus of EU Parliament proceedings”. Dans M. Sánchez Nieto and I. Doval (eds.) *Parallel Corpora: Creation and Application*. Amsterdam/Philadelphia: John Benjamins.

Fontenelle, T. (1998). “Discovering significant lexical functions in dictionary entries”. Dans Cowie A. P. (1998), 189-207.

Gallego-Hernandez, D. et Rodriguez-Inés, P., dir. (2016). “Special Issue: Corpus Use and Learning to Translate, almost 20 years on”. *Cadernos de Tradução*, 36(1).

Grossmann, F. et Tutin, A., dir. (2003). “Les Collocations. Analyse et traitement”. *Travaux et Recherches en Linguistique Appliquée*, E:1. Amsterdam: De Werelt.

Leech, G. (2005). “Adding Linguistic Annotation”. Dans Wynne M. ed. (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbrow Books., 17-29.

L’Homme, M.C. (2009). “A Methodology for Describing Collocations in a Specialized Dictionary”. Dans Nielsen, S. and S. Tarp, dir. (2010). *Lexicography in the 21st Century In honour of Henning Bergenholtz*. Amsterdam/Philadelphia: John Benjamins.

L’Homme, M.C. (2010). “Designing Terminological Dictionaries for Learners based on Lexical Semantics: The representation of actants”. Dans Fuertes-Olivera, P., dir. (2010) *Specialised Dictionaries for Learners*, Berlin/New York: De Gruyter, pp. 141-153.

Mel’čuk, I. A. (1996). “Lexical functions: a tool for the description of lexical relations in a lexicon”. Dans Wanner L., dir. (1996), 37-102.

Mel’čuk, I. A. (1998). “Collocations and Lexical Functions”. Dans Cowie P., dir. (1998). *Phraseology: theory, analysis and applications*. New York: Oxford University Press, 23-53.

Mel’čuk, I. A. (2003). “Les collocations: définition, rôle et utilité”. Dans Grossmann F. et Tutin A., dir. (2003), 23-31.

Mel’čuk, I. A. (2013). *Semantics: from meaning to text (Vol.II)*. Amsterdam/Philadelphia: John Benjamins.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S. et Polguère, A., (1984, 1988, 1992, 1999). *Dictionnaire explicatif et combinatoire du français contemporain: Recherches lexico-sémantiques I, II, III, IV*. Montréal: Presses de l'Université de Montréal.

Polguère, A. (2000). “Une base de données lexicales du français et ses applications possibles en didactique”. *Revue de linguistique et de didactique des langues* (Lidil), 21: 75-97

Polguère, A. (2003). “Collocations et fonctions lexicales: pour un modèle d’apprentissage”. Dans Grossmann F. et Tutin A., dir. (2003), 117-133.

Primož, J. (2011). "Meaning-Text Theory in the translator's classroom". *Rivista internazionale di tecnica della traduzione - International Journal of Translation*, 13:129-138.