InterCorp – projet de corpus parallèle

Section 16 – Projets en cours ; ressources et outils nouveaux

Olga Nádvorníková

InterCorp est un corpus parallèle créé par *l'Institut du Corpus National Tchèque* (www.korpus.cz). A l'époque actuelle, le corpus contient des textes en 27 langues, y compris les langues romanes (le français, le portugais, l'espagnol, l'italien et le roumain)[1]. Le projet, lancé en 2005, est financé dans son intégralité par le Ministère de l'Éducation nationale tchèque et la taille du corpus ne cesse de croître (www.korpus.cz/intercorp).

Le noyau du corpus est formé par des textes littéraires (91 529 000 mots), mais l'utilisateur a également accès à deux autres ensembles de textes : d'une part, des textes juridiques de l'Acquis communautaire, d'autre part, des textes journalistiques tirés des serveurs multilingues Presseurop (http://www.presseurop.eu/en) et SYNDICATE (http://www.project-syndicate.org/). Ainsi, en septembre 2012, la totalité du corpus *InterCorp* contenait plus que 540 millions de mots. Toutefois, les différentes sections linguistiques ne sont par représentées de manière égale ; parmi les plus grandes, nous comptons par exemple la section espagnole (plus que 40 millions de mots dans sa totalité, dont 12 millions dans le noyau littéraire). L'italien et le français appartiennent parmi les langues à représentation moyenne (approx. 30 millions de mots, dont 4 millions du noyau littéraire). Pour les mois à venir, nous envisageons dans la section française l'ajout de plusieurs textes littéraires.

En tant que corpus parallèle, *InterCorp* se compose de textes originaux et de leurs traductions respectives, alignés au niveau de phrases. Les textes journalistiques et ceux de l'Acquis communautaire sont alignés de manière automatique, tandis que pour les textes littéraires, nous réalisons l'alignement semi-manuel et par conséquent plus fiable. La langue tchèque reste la langue pivot de tout le corpus (chaque texte doit avoir sa version tchèque), mais les coordinnateurs des différentes sections linguistiques sont encouragés à chercher des intersections. Pour cette raison, nous trouvons parmi les textes représentés dans plusieurs langues non seulement par exemple *L'Immortalité* de Milan Kundera, mais aussi *Le Nom de la rose* de Umberto Eco ou *Harry Potter* de J.K. Rowling. Le logiciel permet à l'utilisateur de consulter jusqu'à cinq versions linguistiques de textes à la fois.

Le plupart des textes contenus dans le corpus sont lemmatisés ainsi que dotés de l'annotation morphologique (*TreeTagger* pour le français, l'italien, le portugais et l'espagnol). Le logiciel permet aussi d'utiliser des expressions régulières avancées, trier les résultats en fonction de plusieurs critères ou constituer des échantillons aléatoires en cas de concordances plus larges. D'autres possibilités de tri et de traitement statistique des résultats sont envisagés ; pour le moment, l'utilisateur peut effectuer des recherches statistiques sur ses textes en version unilingue, grâce aux fonctions sophistiquées du logiciel Bonito, créé pour le corpus unilingue tchèque (plus que 1 300 millions de mots).

Le corpus sert en premier lieu à des fins scientifiques : l'Institut du Corpus national tchèque a déjà organisé deux colloques internationaux présentant des recherches effectuées sur cette base de données, et les chercheurs des différentes sections linguistiques publient systématiquement les résultats des analyses basées sur *InterCorp*. Le corpus est utilisé également par des étudiants pour leurs travaux de fin d'études. Néanmoins, étant donné que

[1] Le liste complète des langues présentes dans le corpus *InterCorp* est la suivante : l'allemand, l'anglais, le biélorusse, le bulgare, le croate, l'espagnol, l'estonien, le finnois, le français, le grec, le hongrois, l'italien, le letton et le lithuanien, le macédonien, le maltais, le néerlandais, le norvégien, le polonais, le portugais, le roumain, le russe, le serbe, le slovaque, le slovène, le suédois et le tchèque.

l'accès au corpus est libre et gratuit, ses utilisateurs sont très variés : enseignants, traducteurs, etc.

Face à cette multitude d'utilisations, les créateurs d'*InterCorp* se posent aussi d'importantes questions concernant les limites de cet outil de recherche : limite de taille (le corpus s'avère pertinent surtout pour des recherches contrastives grammaticales, moins pour les analyses lexicales) et en particulier des spécificités de la langue de traduction (voir par exemple Olohan 2004 : 91–104 ou Baker 1996 : 176–177).

Bibliographie

Baker, Mona (1996). Corpus-based translation studies: The challenges that lie ahead. In: Somers, H. (ed.) *Terminology, LSP and Translation: Studies in languagae engineering, in honour of Juan C. Sager*. Amsterdam: John Benjamins, 1996, pp. 175–186.

Ballard, Michel (2007). *Les Corpus en linguistique et en traductologie*. Arras: Artois Presses Université, 2007.

Čermák, František – Corness, Patrick – Klégr, Aleš (eds). *InterCorp: Exploring a Multilingual Corpus*. Praha: Nakladatelství Lidové noviny/Ústav Českého národního korpusu, 2010.

Čermák, František (2010). Parallel Corpora: the case of InterCorp. In: Čermák, Fr. – Corness, P. – Klégr, A. (eds). *InterCorp: Exploring a Multilingual Corpus*. Praha: Nakladatelství Lidové noviny/Ústav Českého národního korpusu, 2010, pp. 10–18.

Granger, Sylviane – Lerot, Jacques – Petch-Tyson, Stephanie (eds). (2003). *Corpus-based Approaches to Contrastive Linguistics and Translations Studies*. Amsterdam – Atlanta: Rodopi, 2003.

Laviosa, Sara (2002). *Corpus-based Translation Studies*. *Theory, findings, applications*. Amsterdam: Rodopi, 2002.

Nádvorníková, Olga (2010a). Les corpus parallèles : L'Espace pour l'analyse contrastive. *Études Romanes de Brno*, 2010, 31, 1, pp. 7–27.

Nádvorníková, Olga (2010b). *The French* gérondif *and its Czech equivalents*. In: Čermák, Fr. – Corness, P. – Klégr, A. (eds). *InterCorp: Exploring a Multilingual Corpus*. Praha: Nakladatelství Lidové noviny/Ústav Českého národního korpusu, 2010, pp. 83–96.

Olohan, Maeve (2004). *Introducing Corpora in Translation Studies*. London – New York: Routledge, 2004.

Rastier, François (2005). Enjeux épistémologiques de la linguistique de corpus. In: Williams, G. (ed.). *La linguistique de corpus*. Rennes: Presses universitaires de Rennes, 2005, pp. 31–47.

Svášek, Martin (2007). *Définition, élaboration et exploitation d'un corpus parallèle bidirectionnel français – tchèque tchèque – français*. Thèse de doctorat, Filozofická fakulta Univerzity Karlovy v Praze (dir. V. Petkevič), INALCO (LALIC – CERTAL), Paris (dir. P. Pognan).