De l'apport des corpus électroniques pour les analyses lexicales

Organisation : Hélène Vinckel-Roisin & Yvon Keromnes

Ces dernières décennies, la linguistique de corpus a connu un essor exponentiel et a vu naître de grands corpus électroniques annotés, comparables et parallèles, monolingues et multilingues. L'exploitation des corpus existants, selon une approche *corpus-based* ou *corpus-driven*, sous-tend un nombre croissant de recherches menées au sein du laboratoire de l'ATILF : fédérant plusieurs groupes de recherche rattachés notamment à l'Axe disciplinaire « Lexique » et à l'Axe méthodologique transversal « Modélisation, ressources et traitement informatique », les corpus électroniques représentent une ressource incontournable pour mener des analyses linguistiques, quantitatives et qualitatives, focalisées tant sur l'étude de l'usage, d'un phénomène en contexte, que sur l'exploration du lexique (lexèmes simples, néologismes, collocations, moules syntaxiques, constructions etc.).

Cette première demi-journée thématique d'un cycle de manifestations futures est conçue dans l'idée de faire connaître les corpus électroniques contemporains, mono- et multilingues, ainsi que de nouveaux modèles issus de l'intelligence artificielle ; elle donnera ainsi l'occasion de croiser des approches différentes et de faire interagir des chercheuses et chercheurs de l'ATILF sur des problématiques de recherche en lien notamment avec le lexique, la lexicographie, la traductologie et la traduction automatique.

Dans le prolongement de cette première manifestation, une deuxième demi-journée thématique est d'ores et déjà envisagée en 2024 en raison de l'intérêt central accordé aux corpus à l'ATILF ; les corpus électroniques et leurs potentialités seront appréhendés notamment à travers le prisme de l'analyse du discours et de la polysémie.

Bibliographie

Bubenhofer, Noah, 2009. *Sprachgebrauchsmuster. Korpuslinguistik als Methode der Diskurs- und Kulturanalyse*. Berlin/Boston: de Gruyter.

Granger, Sylviane / Meunier, Fanny (eds), 2008. *Phraseology. An interdisciplinary perspective*. Amsterdam / Philadelphia, John Benjamins. Cf. les contributions rassemblées dans le chapitre 2 « Corpus-based analysis of phraseological units », 111-190.

Lemnitzer, Lothar / Zinsmeister, Heike, 2006. *Korpuslinguistik. Eine Einführung*. Tübingen, Narr.

Loock, Rudy, 2016. *La traductologie de corpus*. Villeneuve d'Ascq, Presses Universitaires du Septentrion.

Perkuhn, Rainer / Keibel, Holger / Kupietz, Marc, 2012. *Korpuslinguistik*. Paderborn, Fink.

Steyer, Kathrin, 2004. « Kookkurrenz. Korpusmethodik, linguistisches Modell, lexikografische Perspektiven ». In : Steyer, Kathrin (Hg.), *Wortverbindungen – mehr oder weniger fest*. Berlin / New York, de Gruyter, 87-116.

Tognini-Bonelli, Elena, 2001. *Corpus Linguistics at Work*. Amsterdam, John Benjamins.

Zufferey, Sandrine, 2020. *Introduction à la linguistique de corpus*. Londres, ISTE Editions Ltd.

Programme

9h30 ⬣ 9h45	Introduction / Ouverture
9h45 ⬣ 10h15	*Le traitement automatique des langues au service des analyses lexicales* **Mathieu Constant**, ATILF / UL - CNRS
10h15 ⬣ 10h45	*La place des corpus électroniques en lexicographie ? À la suite de l'introspection lexicographique* **Alain Polguère**, ATILF / UL - CNRS
10h45 ⬣ 11h00	**Pause**
11h00 ⬣ 11h30	*Phraséologie, terminologie et évolution technologique de la traduction AN-FR et AN-AR* **Mehsen Azizi**, ATILF / UL - CNRS
11h30 ⬣ 12h00	Sketch Engine *: le contexte, sinon rien* **Yvon Keromnes**, ATILF / UL - CNRS
12h00 ⬣ 12h30	*Le corpus multilingue de textes parallèles* InterCorp *comme outil d'aide à la traduction* **Hélène Vinckel-Roisin**, ATILF / UL - CNRS

Résumés des interventions

⬣ Le traitement automatique des langues au service des analyses lexicales

Mathieu Constant (ATILF / UL - CNRS)

Le traitement automatique des langues a connu des avancées spectaculaires ces dernières années grâce aux progrès de l'intelligence artificielle (IA). Dans cet exposé, nous nous intéressons aux expressions polylexicales qui sont des combinaisons idiosyncratiques de plusieurs lexèmes (ex. *faire face, pomme de terre, en fait*). Nous montrerons tout d'abord des méthodes statistiques/neuronales pour identifier des expressions polylexicales dans des textes. Nous verrons ensuite comment les nouveaux modèles de langues de l'IA appris sur des volumes colossaux de textes peuvent apporter un nouveau regard sur ce type d'expressions.

Phraséologie, terminologie et évolution technologique de la traduction AN-FR et AN-AR

Mehsen Azizi (ATILF / UL – CNRS)

La traduction machine (TM) n'a cessé de s'améliorer surtout depuis l'apparition de la traduction neuronale (TMN), adoptée par Google Translate. Cela ne signifie pas pour autant la fin de la traduction humaine (TH), parce que la TMN est encore loin d'être parfaite, et que pour être efficace, elle demande des ressources et investissements considérables.
L'objectif de cette thèse est donc de mieux comprendre l'impact de cette évolution technologique dans la formation des traducteurs et traductrices : l'activité de traduction étant envisagée comme résolution de problèmes et prises de décision liées à l'accès à des ressources terminologiques et lexicographiques. Alors on se pose, quelle est la spécificité des problèmes posés et des ressources nécessaires dans la post-édition comparée à la traduction humaine ?

Sketch Engine : le contexte, sinon rien

Yvon Keromnes (ATILF / UL - CNRS)

S'il existe aujourd'hui encore des voix pour prétendre après Chomsky que l'introspection est la seule façon scientifique d'étudier le langage, le recours aux corpus s'étend toujours davantage dans les analyses linguistiques, et les études quantitatives se multiplient. Ce recours aux corpus pose cependant de nombreuses questions complexes, tant de nature épistémologique que méthodologique (constitution des corpus, taille, outils d'exploitation et usage). Dans cette présentation, j'explore un corpus « maison » de taille modeste, constitué d'écrits de deux auteurs spécialistes de la théorie de l'évolution et adversaires théoriques, R. Dawkins et S.J. Gould, soit les originaux de ces écrits en anglais et les traductions de ces écrits en allemand et en français. Il s'agit de voir en quoi un outil aussi polyvalent que Sketch Engine permet d'explorer les différences et similitudes entre les deux auteurs dans leur langue, et de voir en quoi celles-ci se reflètent dans la traduction de leurs textes.

La place des corpus électroniques en lexicographie ? À la suite de l'introspection lexicographique

Alain Polguère (ATILF / UL - CNRS)

Mon intervention portera sur la place respective 1) du recours introspectif par les lexicographes à leurs connaissances linguistiques propres et 2) du recours par ces derniers aux données des corpus textuels électroniques. Je défendrai l'idée que l'activité introspective doit impérativement précéder l'exploitation des corpus, en illustrant mon propos à partir de tâches lexicographiques centrales pour la modélisation lexicale. Pour ce faire, je m'appuierai sur la méthodologie de la Lexicographie Explicative et Combinatoire, appliquée à la construction des Systèmes Lexicaux.

Le corpus multilingue de textes parallèles *InterCorp* comme outil d'aide à la traduction

Hélène Vinckel-Roisin (ATILF / UL - CNRS)

Développé par l'Institut du Corpus national tchèque à l'Université Charles de Prague et accessible gratuitement, le corpus parallèle *InterCorp* comprend des textes originaux et leurs traductions dans plus de 40 langues. Ce corpus électronique multilingue offre de multiples possibilités d'exploitation, notamment la mise en évidence de profils collocationnels dans une perspective monolingue, l'étude comparée d'un segment source et d'une ou plusieurs traduction(s) en langue cible, ou encore l'évaluation critique de la qualité de la traduction proposée.
Après avoir i) présenté le corpus *InterCorp* et son originalité par rapport à d'autres corpus électroniques et ii) mis en évidence ses principales fonctionnalités pour des études lexicales, monolingues et multilingues, iii) nous montrerons de quelle manière *InterCorp* peut servir d'outil d'aide à la traduction.