Postface : Corpus et didactique en France

Alex Boulton

This is a pre-publication version; please refer to the final paper where possible.
Boulton, A. (2024). Postface : Corpus et didactique en France. In V. Privas-Bréauté (Dir.), *Du recueil de données à l'analyse des corpus en didactique des langues* (pp. 173–178). Presses Universitaires de Rennes.

Corpus et didactique en France

Alex Boulton

Tout le monde sait ce que c'est qu'un corpus, comme tout le monde sait ce que c'est que la vie, l'amour, un légume. Toutefois, une définition hermétique et universelle est souvent difficile à établir. Dans notre cas, il y a corpus et il y a corpus. En linguistique de corpus, on trouve souvent des définitions consensuelles comme celle de McEnery *et al*. (2006, p. 5) : « une grande collection de textes authentiques sous forme électronique qui est échantillonnée afin d'être représentative d'une langue ou d'une variété d'une langue. » Une définition très claire mais qui laisse encore planer un certain nombre de questions. Est-ce forcément une collection, et grande comment ? S'agit-il nécessairement d'un corpus langagier ? Ne peut-on pas travailler sur un corpus de textes non « authentiques » ? Tout travail sur du matériel imprimé est-il réellement exclu ? Quels critères pour l'échantillonnage ? Et quel degré de représentativité est adéquat ?

Pour ma part, je m'inspire énormément de ce travail en linguistique de corpus mais à des fins didactiques, en particulier pour ce que Johns avait nommé le « *data-driven learning* » (DDL) dès 1990, glosé par Boulton et Tyne (2014) comme « apprentissage sur corpus » (ASC). Plus précisément, Gilquin et Granger (2010, p. 359) définissent l'ASC comme « l'exploitation des outils et techniques de la linguistique de corpus à des fins pédagogiques », en particulier comme input à l'apprentissage ou à l'utilisation d'une langue étrangère ou seconde par la découverte. En tant que didacticien, ma réponse aux questions ci-dessus est : ça dépend. On peut faire de l'ASC avec : un seul texte (le roman *Swallows and Amazons* dans Johns et al., 2008) ; des corpus très petits (1.200 mots de productions d'apprenant·es dans Seidlhofer, 2002) ; d'écrits produits spécifiquement (des *graded readers* simplifiés dans Allan, 2009) ; sur papier (avec des concordances imprimées dans Hadley & Charles, 2017) ; et quant à l'échantillonnage, quoi de plus confus que l'ensemble du web (avec Google comme substitut de concordancier dans Eu, 2017) ? On peut utiliser des corpus multimodaux, voire même se passer entièrement du texte, toujours à des fins d'apprentissage de langue, comme le démontre l'exemple suivant tiré d'une expérience réelle (Figure 1). En promenade à Taiwan, on me demande si on se retrouve dans des *mountains* ou des *hills* ? Le site Oxford Learner's Dictionaries définit *mountain* comme « a very high hill » et *hill* comme « higher than the land around it, but not as high as a mountain »[1] – une définition circulaire d'une utilité limitée. Avec Google, on peut simplement chercher des exemples de *mountain* et de *hill* et les contraster : les premières sont plus « pointues », rocheuses, dénouées de végétation et souvent couvertes de neige, tandis que les secondes sont plus rondes et vertes. Sans chercher une définition exacte, les tendances sautent littéralement aux yeux.

[1] https://www.oxfordlearnersdictionaries.com


Figure 1 : à Taïwan (à gauche) ; Google images pour des *mountains* (milieu) et *hills* (à droite)

Voilà l'objet de mon intérêt pour les corpus en didactique des langues. En ASC, ce ne sont pas les critères scientifiques qui priment mais l'utilité pédagogique. Pour définir un corpus dans ce contexte, je répète : ça dépend. L'ASC n'est pas idéologique sur la définition d'un corpus, ce sont les outils et les techniques qui lui donnent son sens : les concordances pour l'usage en contexte, les collocations, les listes de fréquence de mots ou de clusters, les mots-clés par rapport à un corpus de référence, la distribution dans des genres et registres, etc. La variété de corpus, d'outils, d'objectifs et d'approches se voit à travers plusieurs synthèses que j'ai produites ces dernières années, dont une méta-analyse statistique de l'ASC (Boulton & Cobb, 2017) et un recensement de l'évolution historique du domaine (Boulton & Vyatkina, 2021).

L'ASC se positionne donc comme *input direct* à l'apprentissage, selon l'intention de Johns (1990, p. 8) d' « éliminer le rôle de l'intermédiaire autant que faire se peut afin de donner à l'apprenant un accès direct aux données [langagières]. » Pour l'intermédiaire, on pense tout de suite à l'enseignant·e mais il s'agit en principe de toute médiation entre apprenant·e et la langue cible, y compris les dictionnaires et grammaires, descripteurs et tests, listes de fréquence de mots ou de clusters, manuels et méthodes, textes didactisés voire inventés. Dans tous ces cas, les corpus peuvent contribuer à l'élaboration de telles ressources même si leur apport reste le plus souvent invisible pour l'utilisateur et utilisatrice – un *input indirect*. Dans son chapitre, Clive Hamilton remarque qu'il y a peu d'intérêt pour ces aspects de la linguistique de corpus en France. Pour l'ASC, il a certainement raison ; toutefois, on en trouve de plus en plus de (jeunes) collègues qui s'y mettent, notamment pour les corpus oraux. Pour l'apport indirect des corpus, la France était longtemps pionnier, avec le Français Fondamental (Gougenheim, 1958) qui sort avant la renaissance de la linguistique de corpus à Birmingham sous la direction de Sinclair (ex. 1987) dans les années 1980.

En ce qui concerne la contribution des corpus à l'*output* langagier, visité ici par Margot Kuligowska Esnault, le domaine de la recherche sur les corpus d'apprenants est en plein essor, avec une nouvelle revue qui voit le jour en 2015 chez John Benjamins : *International Journal of Learner Corpus Research*. Ces recherches sont souvent associées avec le travail du CECL, le Centre for English Corpus Linguistics à Louvain-la-Neuve en Belgique. Un apport donc francophone avec une orientation souvent sur l' « erreur », ou du moins les différences entre les productions d'apprenant·es par rapport aux locuteurs et locutrices « natifs/natives », comme dans le chapitre de Pascale Manoïlov et Agnès Leroux-Béal. Ainsi, il s'agit bien toujours de *linguistique* de corpus, un élément présent dans certains chapitres ici, notamment celui de Séverine Behra, Maud Ciekanski, Guillaume Nassau et Dominique Macaire.

Si nous avons déjà cité quelques chapitres du présent ouvrage, le point commun de l'ensemble est une approche toute autre, bien française et fière de l'être. La préoccupation principale n'est pas la linguistique de corpus comme fin en soi mais la collecte et interprétation de données du terrain qui puissent informer les recherches sur les processus d'apprentissage et les pratiques d'enseignement. C'est tout une culture que l'on retrouve moins souvent dans les publications internationales, comme en témoignent les références citées, très majoritairement en français ou produites par des collègues en francophonie (67,9 %). Tout comme en linguistique de corpus, on s'appuie sur des données de terrain réunies selon des protocoles rigoureuse, où la construction est tout aussi importante que l'exploitation. Ces données attestées sont souvent multimodales (comme pour Justine Paris et Pauline Beaupoil-Hourdel), un autre point fort de la recherche française dans ce domaine comme en témoigne l'ampleur du travail sur les corpus oraux émanant de mon laboratoire, l'ATILF (Analyse et Traitement Informatique de la Langue Française), avec TCOF[2] et FLEURON[3] pour ne citer que ces deux. Ces corpus multimodaux sont souvent accessibles directement, permettant de prendre en compte des aspects visuels inaccessibles à travers de simples transcriptions de l'oral (ex. Stéphane Soulaine et Caroline Raymond). Si l'on peut ainsi faire des allers-retours rapides entre occurrence et contexte, la présentation décontextualisée traditionnelle de la linguistique de corpus (des concordances aux listes de fréquence) se retrouve au second plan, permettant de retrouver les textes complets où ces éléments sont attestés. Ainsi, c'est le texte qui est privilégié, texte situé par ses métadonnées, amenant en même temps vers une approche plus qualitative et holistique des situations d'occurrence. En effet, rares sont les pages dans cet ouvrage qui proposent une analyse statistique courante dans les recherches internationales. Cette approche française très riche permet de nourrir la réflexion sur les processus d'apprentissage et ainsi alimenter la réflexion sur la formation des enseignant·es dès la maternelle (comme pour Marie Potapushkina-Delfosse et Anne-Marie Voise).

Pour conclure, et en guise de diversion, les textes des chapitres acceptés (mais non finalisés) ont été convertis en format .txt pour en faire un corpus de 60.879 mots compatible avec AntConc (Anthony, 2021). Ce corpus contient tout élément de « texte » du premier mot des introductions au dernier mot des conclusions (sans titres, noms, affiliations, images, références, annexes, remerciements, etc.). Tout d'abord, une simple liste de fréquence atteste que la notion de *corpus* est bien présente en 24^{e} place, le tout premier mot lexical sur la liste, et se retrouve dans chacun des chapitres (de 5 à 6 occurrences chez Soulaine et Raymond, Kuligowska Esnault respectivement, jusqu'à 90 dans Hamilton et 127 dans Behra *et al.*). Ceci représente un contraste marqué avec le mot *linguistique* en 156^{e} place avec 47 occurrences (entre 3 et 18 par chapitre mais entièrement absent dans un cas). De même, des termes intimement liés à la linguistique de corpus sont très peu représentés (Tableau 1), souvent avec un sens non spécialisé (dans le cas de *fréquence*).

f	mots
27	fréquence
15	n-gram(me)(s), bi-grams, tri-gram(me)s, quadri-grams
8	mot(s-)clé(s)
3	collocations, collocats
1	concordancier

[2] Traitement de Corpus Oraux en Français. https://www.atilf.fr/ressources/tcof
[3] Français Langue Étrangère Universitaire : Ressources et Outils Numériques. https://fleuron.atilf.fr

1	clusters

Tableau 1. Exemples de mots en lien avec la linguistique de corpus

Ensuite, en dépassant les mots individuels, les *n*-grams (Tableau 2) attestés dans ce corpos soulignent l'importance de la *didactique* ou de l'*enseignement / apprentissage des langues* ainsi que l'interprétation des données en contexte : *dans le cadre d(e)*, *d'un point de (vue)*, etc.

rang	*f*	distribution	cluster
1	30	6	dans le cadre de
2	23	6	en didactique des langues
3	16	6	dans le cadre d
4	13	3	d'un point de
5	13	6	le cadre de l
6	13	3	un point de vue
7	11	5	c'est-à-dire
8	9	5	en ce qui concerne
9	9	5	enseignement apprentissage des langues
10	9	3	il s'agit d

Tableau 2. Clusters : 4-grams, f ≥2

Enfin, une comparaison avec un corpus d'articles de recherche en français (Chambers & Lebaron, 2007) permet de repérer des mots-clés (Log-Likelihood 4-term >130 ; p<0,05 +Bonferroni). Ainsi, en mettant de côté des mots grammaticaux-fonctionnels pour se concentrer sur les premières familles de mots lexicaux pertinents sur la liste (Tableau 3), on repère bien sûr *corpus* en première position (*f* = 379) avec d'autres termes liés comme *données* et *analyses*, *recherches* et *études* ; le domaine avec *didactique*, *langues* et *anglais* (contre 23 occurrences de *français**) ; un accent fort sur les personnes et processus avec *formation* et *étudiant*, *apprenant* / *apprendre* et *enseignant* / *enseigner*, *élèves* et *chercheurs ;* et les *activités* comme *produire / des productions, interagir / des interactions, transcrire / des transcriptions, écrire / écriture* ; et bien sûr le lieu, avec *classe*.

f	famille	*f*	famille
379	corpus	146	élève*
290	lang*	128	chercheu*
265	form*	125	didacti*
253	appren*	117	transcri*
226	analys*	108	étude*
225	recherch*	107	étudi*
218	enseign*	103	classe*
217	données	102	activité*
201	produ*	99	anglais*
148	intera*	96	écri*

Tableau 3. Mots-clés par famille

Pour conclure, l'intérêt des corpus dépasse largement la seule linguistique de corpus, comme en témoignent ces chapitres rassemblés ici par Virginie Privas-Bréauté. Dès que l'on a affaire à des textes (électroniques), les outils et techniques propres à la linguistique de corpus peuvent s'avérer utiles, que l'on soit étudiant·e, enseignant·e ou chercheur·e, en langues ou dans d'autres disciplines. En didactique des langues, une approche sur corpus peut servir directement ou indirectement à fournir un input, mais aussi à analyser l'output

afin de mieux comprendre les pratiques d'apprenant·es et processus d'apprentissage afin d'informer les recherches et les pratiques enseignantes, une spécificité de ce livre et, dans une certaine mesure, de la didactique française. Une synergie de différentes approches peut s'avérer utile : en France, nous pouvons être plus conscient·es de ce qui se fait dans des domaines très proches en ASC et plus largement en linguistique de corpus ; mais les chercheurs et chercheuses ailleurs pourraient bénéficier tout autant d'une meilleure connaissance d'autres possibilités en corpus et didactique explorées ici où l'approche sur corpus permet de dépasser la langue afin d'accéder à des contenus, pertinents à d'autres fins didactiques.

Références

Allan, R. (2009). Can a graded reader corpus provide 'authentic' input? *ELT Journal, 63*, 23–32. https://doi.org/10.1093/elt/ccn011

André, V. (2021). Corpus : exploitation, médiation et autonomisation. L'utilisation du concordancier de la plateforme FLEURON en cours de FLE. *Bulletin Suisse de Linguistique Appliquée*, 114, 129–148.

Anthony, L. (2020). AntConc (Version 3.5.8m). Tokyo : Waseda University. https://www.laurenceanthony.net/software

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*(2), 348–393. https://doi.org/10.1111/lang.12224

Boulton, A., & Tyne, H. (2014). Des documents authentiques aux corpus : Démarches pour l'apprentissage des langues. Paris : Didier.

Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions. *Language Learning & Technology, 25*(3). Sortie en octobre 2021.

Chambers, A., & Le Baron, F. (2007). *Le Corpus Chambers-Le Baron d'articles de recherche en français*. Oxford Text Archive. http://ota.ox.ac.uk/desc/2527

Eu, J. (2017). Patterns of Google use in language reference and learning: A user survey. *Journal of Computers in Education, 4*(4), 419–439. https://doi.org/10.1007/s40692-017-0094-5

Gilquin, G., & S. Granger. (2010). How can data-driven learning be used in language teaching? In A. O'Keeffe & M. McCarthy (Dir.), *The Routledge Handbook of Corpus Linguistics* (pp. 359–370). Londres : Routledge. https://doi.org/10.4324/9780203856949.ch26

Gougenheim, G. (1958) *Dictionnaire fondamental de la langue française*. Paris : Didier.

Hadley, G., & Charles, M. (2017). Enhancing extensive reading with data-driven learning. *Language Learning & Technology, 21*(3), 131–152. https://doi.org/10125/44624

Johns, T. (1990). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria, 10*, 14–34.

Johns, T., Lee, H. C., & Wang, L. (2008). Integrating corpus-based CALL programs and teaching English through children's literature. *Computer Assisted Language Learning, 21*(5), 483–506. https://doi.org/10.1080/09588220802448006

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Londres : Routledge.

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. In S. Granger, J. Hung & S. Petch-Tyson (Dir.), *Computer learner corpora, second*

language acquisition and foreign language teaching (pp. 213–234). Amsterdam : John Benjamins. https://doi.org/10.1075/lllt.6.14sei

Sinclair, J. (Dir.). (1987). *Looking up: An account of the COBUILD project in lexical computing*. Londres : Collins.