Corpus, grammaire et français langue étrangère : une concordance nécessaire

Alain Kamber et Maud Dubois (Neuchâtel)

Research in grammar has probably been influenced
most profoundly by the corpus-based approach.
It has helped to redefine what grammar is.
(Xiao/McEnery 2013)

1 Les corpus, une autre approche du langage

L'avènement de la linguistique de corpus a conduit à une reconsidération radicale des rapports entre langue et discours : rendue accessible à l'analyse à une large échelle par les nouvelles technologies, « la parole » semble maintenant être considérée comme la seule voie d'accès réellement scientifique et fiable pour décrire le système. Le célèbre linguiste de fauteuil (cf. Fillmore 1992: 35) a ainsi été déboulonné de son siège et avec lui la pure introspection du locuteur/scripteur omniscient[1], de préférence natif, qui analyse ses propres énoncés ou collecte plus ou moins aléatoirement des exemples confirmant – rarement infirmant – ses thèses.

Développées à leur origine pour la description de l'anglais, les recherches basées sur corpus sont donc reconnues actuellement comme allant bien au-delà d'une simple nouvelle approche méthodologique. Xiao/McEnery (2013: 1) formulent ce constat dans le domaine de la description grammaticale et évoquent à propos des travaux de Quirk et al. (1985) un véritable changement de paradigme pour l'élaboration des grammaires de l'anglais. Sinclair (2004) reconnait de son côté avoir dans un premier temps sous-estimé l'effet des résultats obtenus par les corpus sur la rédaction du *Collins COBUILD* (premier dictionnaire entièrement basé sur corpus) : des catégories descriptives et des positions théoriques ont dû être modifiées et le dictionnaire reconceptualisé. De manière plus générale, faisant le point sur 40 ans de la discipline, Sinclair (2004: 1) affirme qu'en plus de pousser à revoir certains acquis, les corpus donnent accès à une aire fascinante de nouveautés et d'inattendu. Tyne (2013: 10) souligne « l'écart [très important] qui existe entre la description des données attestées et la manière dont on a pris l'habitude de présenter la langue » et la sortie d'un monde binaire fait de formes 'correctes' ou 'incorrectes' et de cases prédéfinies[2]. Avec le linguiste de fauteuil, « l'éternel grammairien » (Berrendonner 1982) a ainsi lui aussi du plomb dans l'aile.

[1] Douze ans avant Fillmore déjà, Corbin (1980: 121) opposait la « linguistique de bureau » à la « linguistique de terrain ».

[2] Tyne cite ici les positions de Sinclair (2004), Conrad (2010) et Chambers (2013).

L'un des grands changements amenés par la linguistique de corpus est sans conteste la reconsidération de l'oral et la description de ses spécificités et de son fonctionnement. Dépassant la distinction traditionnelle entre code oral et code écrit, les notions de variation et de genre textuel semblent également avoir amené une modification conséquente : la description monolithique de la grammaire d'une langue est remplacée par des descriptions liées à des pratiques langagières et des registres spécifiques (Conrad 2000: 549). L'abolition de la frontière traditionnelle entre lexique et grammaire et l'intégration du premier dans la seconde constituent une autre évolution notable (Xiao/McEnery 2013: 4). Finalement, il est un fait que la prolifération des corpus détrône graduellement le locuteur natif de sa place centrale de modèle et de juge sur l'utilisation d'une langue et permet à des locuteurs non-natifs d'endosser le rôle d'experts (Sinclair 2004, à propos de Mauranen 2004).

2 Un changement de paradigme pour l'enseignement/apprentissage des langues

Dès ses débuts anglo-saxons, la linguistique de corpus a eu partie liée avec l'enseignement/apprentissage des langues (cf. Breyer 2011: 1) dont elle a également modifié profondément les contours : élaboration d'ouvrages de référence basés sur corpus et utilisation directe des corpus en classe (*data-driven learning*) notamment. Conrad (2000: 549) affirme qu'au 21ème siècle, certains résultats de la recherche sur corpus en grammaire ont le potentiel de révolutionner l'enseignement de l'anglais. Si Tyne (2013: 8) préfère l'idée d'évolution à celle de révolution, il n'en qualifie pas moins le *data-driven learning* de nouveau « paradigme de recherche en didactique des langues »[3].

Qu'il s'agisse du *data-driven-learning* ou de l'élaboration de matériel didactique (dictionnaires, manuels, etc.), l'intégration de la recherche sur corpus et la reconnaissance de son intérêt semblent maintenant largement acquises pour l'anglais langue étrangère (cf. Breyer 2011: 3). Hunston (2002) détermine l'apport de la linguistique de corpus dans les grammaires de référence et les dictionnaires sur cinq niveaux : ceux de la fréquence, de la collocation et de la phraséologie, de la variation, du lexique-grammaire (*lexis in grammar*), enfin de l'authenticité[4].

Les notions de quantification et de **fréquence** – et par conséquent de représentativité – sont indissociables de la linguistique de corpus, comme le notent McEnery/Wilson (2001: 109–110): « What makes corpora important for syntactic research is, first, their potential for the representative quantification of the grammar of a whole language variety and, second, their role as empirical data also quantifiable and representative, for the testing of hypotheses derived from grammatical theory ». Ceci est vrai à plus forte raison dans une perspective didactique, où il s'agit de permettre à l'apprenant d'une langue de se concentrer sur les phénomènes les plus fréquents ou typiques (Kamber 2014: 4). De ce principe découle également l'idée que l'on n'enseigne plus des règles mais des régularités (cf. Tyne 2013: 10).

[3] A ce propos, voir aussi Boulton et Tyne (2014: 11–15).

[4] Ces éléments sont repris plus récemment par Tyne (2013) lorsqu'il affirme que les corpus renseignent sur le fonctionnement de la langue comme système, les recherches mettant en avant la fréquence d'utilisation, les contextes d'utilisations (collocations), la distribution de la grammaire en fonction du registre ou de la situation d'énonciation.

Les **collocations**, phraséologismes, patterns, etc. sont au cœur même des travaux de linguistique de corpus menés dans la tradition du contextualisme britannique (cf. Firth 1957). Les recherches intégrant le lexique à la grammaire sont nombreuses et ont fait une percée dans l'espace francophone ces dernières années, notamment avec les études basées sur les corpus *Scientext* et *Emolex* (cf. notamment Tutin/Grossmann 2003; Blumenthal/Novakova/Siepmann 2014). Sinclair (2004) rappelle qu'avec les recherches sur corpus, il a dû passer de la notion de mot comme pourvoyeur de sens lexical à celle de *lexical item* (qui peut être constitué de plusieurs mots à la suite). L'intérêt du **lexique-grammaire** pour la didactique des langues est d'intégrer l'enseignement du vocabulaire à celui de la grammaire (Conrad 2000) et de montrer, par exemple, que des mots présentés comme des synonymes ou quasi-synonymes ont des contextes d'utilisation différents (Kamber 2011).

Si, du point de vue de l'enseignant, la **variation** peut être considérée comme une nuisance (Sinclair 2004: 274), elle est un des apports majeurs de la linguistique de corpus à l'enseignement/apprentissage des langues. Selon Conrad (2004), elle ne peut pas être ignorée car elle n'est pas confinée à des variétés spécialisées, mais s'étend à la zone centrale de l'utilisation du langage. Les recherches ayant montré que les choix des locuteurs / scripteurs sont hautement systématiques selon tel ou tel genre du discours, l'un des points importants est de permettre à l'apprenant l'utilisation dans le bon contexte de formes grammaticales alternatives, plus que d'insister sur la correction de ces formes (Conrad 2000: 558). Autrement dit, les formes grammaticalement possibles se distribuent en fonction du registre et de la situation d'utilisation (Tyne 2013).

Mis en avant par le CECR comme source à laquelle les apprenant devraient être exposés directement (CECR: 110), les **énoncés authentiques** sont l'un des arguments forts pour l'utilisation des corpus dans l'enseignement/apprentissage des langues (Breyer 2011: 2). Produits de contextes sociaux d'utilisation réels, ils sont le moyen le plus sûr pour l'apprenant d'accéder à ces contextes : « […] for the most part, the data are largely naturalistic, unmonitored and the product of real social contexts. Thus the corpus provides one of the most reliable sources of naturally occurring data that can be examined » (McEnery/Wilson 2012). En les manipulant et en les observant, dans l'optique du *data-driven learning*, on est soi-même engagé dans une recherche authentique sur de réelles questions liées au langage. Une concordance intervient alors entre l'authenticité de l'énoncé, du but et de l'activité, qui permet *in fine* pour l'apprenant une autonomisation et une conscience linguistique accrues (Johns 1988, commenté par Breyer 2011: 2).

3 Une meilleure description de la langue pour le français

Pour le domaine de l'anglais, Sinclair (2004: 2) note que les corpus sont vus par beaucoup d'enseignants comme des outils utiles et qu'ils font presque partie du paysage pédagogique, la vague de résistance étant largement derrière[5]. Tyne (2013: 8) relève que l'écrasante majorité des études sur l'utilisation des corpus en didactique concerne l'anglais. De fait, le matériel existant – basé sur des corpus – est conséquent. En revanche, du côté du français, on constate

[5] Ce constat optimiste est toutefois nuancé par Breyer (2011: 3–4), qui met en évidence le fait que malgré cet effort considérable, l'impact sur le « mainstream teaching » est resté très limité.

qu'il existe un écart parfois abyssal entre les résultats de la recherche récente et les représentations de la langue dans les dictionnaires, les grammaires et les manuels.

Un rapide survol des dictionnaires destinés aux apprenants de l'anglais basés sur des corpus (*Longman Dictionary of Contemporary English* ou *Collins COBUILD. Advanced Learner's English Dictionary* par exemple) permet de se faire une idée des lacunes dans les dictionnaires du français et des améliorations possibles en termes de présentation, notamment l'introduction de la variation, des collocations, phraséologismes et structures syntaxico-sémantiques. Un constat semblable peut être fait dans le domaine des grammaires, notamment la *Longman Grammar of Spoken and Written English*, dont le seul titre programmatique suffirait à rendre envieux l'utilisateur francophone : ainsi que le rappellent Xiao/McEnery (2013), les grammaires non basées sur des corpus véhiculent des descriptions biaisées qui ne sont pas en accord avec le langage attesté. Ces manques se répercutent fatalement dans les manuels de français langue étrangère et engendrent un certain sentiment de frustration aussi bien chez les enseignants que chez les apprenants : impression d'enseigner ou d'apprendre une langue « artificielle », déconnectée de la réalité, ignorance des écarts entre pratiques de l'oral et de l'écrit, difficulté de communication, connaissance insuffisante des phénomènes de variation, manque de sensibilité face aux registres et aux genres textuels, etc.

Le recours à des corpus oraux et écrits permettra sans conteste d'apporter des correctifs à certaines descriptions encore largement présentes dans les grammaires FLE, voire de modifier en profondeur certaines représentations fondées sur une vision ne tenant pas compte de l'usage. Les articles réunis dans ce numéro constituent un échantillon non-exhaustif, mais néanmoins représentatif des recherches à mener à l'avenir dans le domaine, ainsi que des moyens d'améliorer à la fois les représentations du français dans les dictionnaires, les grammaires et les manuels, et les pratiques d'enseignement.

Les premières contributions mettent en évidence des approximations dans le matériel didactique et les ouvrages de référence, voire des manques dans la terminologie grammaticale, qui demanderaient à être revus à la lumière des corpus. **Christian Surcouf** et **Annick Giroud** analysent les enregistrements audio des trois premières leçons d'un certain nombre de méthodes de français langue étrangère pour apprenants débutants afin d'examiner si des traits marquants de l'oral y sont représentés. Il apparait que la prononciation des manuels ne correspond que très partiellement à celle qui nous est transmise par les corpus oraux. De leur côté, **Mireille Bilger** et **Paul Cappeau** montrent, par l'exemple des prépositions *contre* et *entre*, comment l'analyse d'un corpus échantillonné (oral privé, oral public, presse et littérature) permet de questionner, voire de corriger certaines descriptions des grammaires et des dictionnaires, fondées principalement sur un usage littéraire déjà daté. Partant des résultats d'une recherche sur corpus « pure » sur la question des adjectifs épithètes anté- ou postposés qui montre les lacunes des descriptions traditionnelles, **Marie-Armelle Camussi-Ni**, **Annick Coatéval** et **Juliette Thuilier** rendent manifestes les bénéfices d'une convergence non seulement possible, mais aussi nécessaire, entre une approche théorique et une approche appliquée. Elles proposent une progression didactique pour enseigner aux apprenants les régularités du système sans recourir à des schématisations irréalistes ou à des listes par définition incomplètes. Partant de mots-formes présentés dans un dictionnaire de fréquence, **Daniel Elmiger** et **Alain Kamber** se penchent sur la question des catégories grammaticales du français, une thématique largement simplifiée, voire ignorée dans les

manuels d'enseignement du FLE. Leur démarche empirique, exemplifiée par plusieurs mots sujets à une certaine variation catégorielle adjectif – nom, permet d'apporter des réponses au problème sur le plan théorique et d'esquisser une réflexion didactique pour une représentation plus précise du phénomène dans des dictionnaires à l'attention d'apprenants avancés du français.

Les trois articles suivants s'inscrivent dans les recherches en lexique-grammaire actuelles. Les corpus – dédiés à une thématique ou reliés à un genre particulier – y sont utilisés pour identifier les phénomènes fréquents ou saillants qui seraient donc susceptibles d'être enseignés avec profit. Dans cette perspective, **Monika Bak Sienkiewicz** étudie les patrons « verbes causatifs + noms d'émotion » dans le corpus *Emolex*, composé d'écrits journalistiques et littéraires. Elle procède à l'analyse statistique de l'attraction entre les verbes et les substantifs, mais met aussi en évidence leur profil combinatoire au sens plus large, comprenant les adjectifs, les adverbes, etc. Les deux autres contributions s'appuient sur le corpus de textes scientifiques *Scientext*. Dans le cadre de leur collaboration à *Dicorpus*, outil d'aide à la rédaction scientifique pour les apprenants du français, **Rui Yan** et **Sylvain Hatier** analysent les patrons lexico-syntaxiques du verbe *montrer* et de ses deux quasi-synonymes *démontrer* et *indiquer*, extrayant et mettant en lumière les structures spécifiques de chacun de ces trois verbes. **Thi Thu Hoai Tran**, **Agnès Tutin** et **Cristelle Cavalla** quant à elles s'intéressent au rôle du corpus en classe de langue pour identifier les propriétés syntaxiques et sémantiques des marqueurs discursifs, propriétés qui ne sont pas présentées dans les grammaires traditionnelles. L'utilisation des corpus pour aider les apprenants à développer leurs compétences transdisciplinaires et à améliorer leurs performances à l'écrit entraine des questions méthodologiques pour l'enseignement/apprentissage du FLE.

Enfin, élargissant la thématique, **Maï Leray** et **Henry Tyne** testent une démarche d'apprentissage sur corpus (ASC) – habituellement réservée aux L2 – dans le domaine du français L1, chez des écoliers de 9-10 ans. L'expérience porte plus particulièrement sur l'enseignement des homophones, une difficulté bien connue de l'orthographe française. L'étude montre que, s'il ne semble pas y avoir de différence dans un premier temps entre les élèves ayant bénéficié d'un enseignement traditionnel et ceux qui ont été exposés aux données issues de corpus, les résultats du post-test semblent plaider pour le recours à l'ASC.

Les articles réunis dans ce numéro démontrent, à travers des objets de recherche variés, que l'utilisation de corpus (oraux ou écrits) permet de répondre à une gamme importante d'interrogations des apprenants et des enseignants des langues vivantes. En favorisant une approche inductive qui procède par observation, découverte et vérification, la linguistique de corpus contribue à une représentation du français plus proche des faits de langue. Elle permet donc d'affiner, corriger ou parfois même modifier radicalement les descriptions des dictionnaires, grammaires ou manuels. Cette exigence scientifique et didactique concerne tous les objets grammaticaux d'une recherche empirique visant à la description de la langue ou à une élaboration de théories selon une approche inductive, du lexique aux relations syntaxiques complexes.

Nous tenons ici à remercier chaleureusement les collègues qui ont accepté d'expertiser les articles soumis pour cette publication : Peter Blumenthal (Cologne), Alex Boulton (Nancy), Juliette Delahaie (Lille), Marie-Paule Jacques (Grenoble), Laure Anne Johnsen (Neuchâtel),

Christopher Laenzlinger (Genève), Anton Näf (Neuchâtel), Chantal Parpette (Lyon), Jean-Christophe Pellat (Strasbourg) et Corinne Rossari (Neuchâtel). Nos sincères remerciements vont aussi à la rédactrice en chef de *Linguistik online*, Elke Hentschel, qui a accepté de consacrer un numéro purement francophone à la thématique des corpus et de la grammaire dans l'enseignement/apprentissage du français.

Bibliographie

Berrendonner, Alain (1982): *L'Eternel grammairien. Etude du discours normatif*. Frankfurt a. M.: Lang.

Biber, Douglas et al. (1999): *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Blumenthal, Peter/Novakova, Iva/Siepmann, Dirk (eds.) (2014): *Les émotions dans le discours / Emotions in Discourse*. Frankfurt a. M.: Lang.

Breyer, Yvonne Alexandra (2011): *Corpora in Language Teaching and Learning: Potential, Evaluations, Challenges*. Frankfurt a. M.: Lang. (= *English Corpus Linguistics* 13)

Boulton, Alex/Tyne, Henry (2014): *Des documents authentiques aux corpus : démarches pour l'apprentissage des langues*. Paris: Didier.

Chambers, Angela (2013): "Learning and Teaching the Subjunctive in French: The Contribution of Corpus Data". *Bulletin VALS-ASLA* 97: 41–58.

Collins Cobuild. Advanced Learner's English Dictionary (2006): Glasgow: Harper Collins.

Conrad, Susan (2000): "Will Corpus Linguistics Revolutionize Grammar Teaching in the 21st Century?". *TESOL Quaterly* 34/3: 548–560. doi: 10.2307/3587743

Conseil de l'Europe (2001): *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Paris: Didier. www.coe.int/t/dg4/linguistic/Source /Framework_FR.pdf [22.02.2016].

Corbin, Pierre (1980): « De la production des données en linguistique introspective ». In: Dessaux-Berthonneau, Anne-Marie (ed.): *Théories linguistiques et traditions grammaticales*. Lille, Presses universitaires de Lille: 121–179.

Firth, John Rupert (1957): *Papers in Linguistics* 1934-1951. Oxford: Oxford University Press.

Hunston, Susan (2002): *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Johns, Tim (1988): "Whence and Whither Classroom Concordancing?" In: Bongaerts, Theo/Haan, Pieter de/Lobbe, Sylvia/Wekker, Herman (eds.): *Computer Applications in Language Learning*. Dordrecht, Foris: 9–27.

Kamber, Alain (2011): « Contexte et sens: utilisation d'un corpus écrit dans l'enseignement / apprentissage du FLE ». *Travaux neuchâtelois de linguistique* 55: 199–218.

Kamber, Alain (2014): « *Prendre*, un verbe support dans l'enseignement du FLE: une analyse sur corpus ». *Revue Mosaïques, Au cœur du verbe. Discours, syntaxe et didactique* (hors-série n°2): 3–16.

Longman Dictionary of Contemporary English (2009). London: Longman.

Mauranen, Anna (2004): "Spoken Corpus for an Ordinary Learner". In: Sinclair, John McH. (ed.): *How to Use Corpora in Language Teaching*. Amsterdam, Benjamins: 89–105. (= *Studies in Corpus Linguistics* 12)

McEnery, Tony/Wilson, Andrew (2001): *Corpus Linguistics: An Introduction*. Edinburgh: Edinburgh University Press. (= *Edinburgh Textbooks in Empirical Linguistics*)

McEnery Tony/Wilson Andrew (2012): “Corpus Linguistics. Module 3.4”. In: Davies, Graham (ed.): *Information and Communications Technology for Language Teachers* (ICT4LT). Slough, Thames Valley University. www.ict4lt.org/en/en_mod3-4.htm [22.02.2016].

Sinclair, John McH. (ed.) (2004): *How to Use Corpora in Language Teaching*. Amsterdam: Benjamins.

Tutin, Agnès/Grossmann, Francis (2003): *Les collocations : analyse et traitement*. In: Marque-Pucheu, Christiane (ed.) (2007): *L’Information Grammaticale* 112. Amsterdam, de Werelt: 58–60.

Tyne, Henry (2013): « Corpus et apprentissage-enseignement des langues ». *Bulletin VALS-ASLA* 97: 7–15.

Xiao, Richard/McEnery, Tony (2013): “Grammar and Corpora”. In: *The Encyclopedia of Applied Linguistics*. Wiley/Sons. DOI: 10.1002/9781405198431.wbeal1411.