Les Corpus en Didactique des langues.

Dominique Macaire, Alex Boulton

RECHERCHES EN DIDACTIQUE DES LANGUES ET DES CULTURES

Recherches en didactique des langues et des cultures
Les cahiers de l'Acedle

11-1 | 2014
Corpus et didactique des langues

Editorial

Notion de corpus et didactique des langues

Dominique Macaire et Alex Boulton

revues.org

Éditeur
ACEDLE

Édition électronique
URL : http://rdlc.revues.org/1665
ISSN : 1958-5772

Référence électronique
Dominique Macaire et Alex Boulton, « Editorial », *Recherches en didactique des langues et des cultures* [En ligne], 11-1 | 2014, mis en ligne le 07 janvier 2014, consulté le 16 mai 2017. URL : http://rdlc.revues.org/1665

Ce document a été généré automatiquement le 16 mai 2017.

Editorial

Notion de corpus et didactique des langues

Dominique Macaire et Alex Boulton

1 La notion de corpus, très présente depuis quelques années dans les recherches en linguistique ainsi qu'en didactique des langues, a retenu notre attention pour ce numéro spécial "Notion en Questions" de la revue *Recherches en didactique des langues et des cultures (RDLC)* : *Les Cahiers de l'Acedle*. La notion de corpus recouvre en effet une large gamme de définitions ne concordant pas toutes les unes avec les autres et que les articles suivants vont s'employer à clarifier en relation avec le champ de la didactique des langues.

2 En linguistique de corpus, le terme de "corpus" renvoie généralement à une grande collection de textes authentiques, présentée sous forme électronique, et censée être représentative d'une langue ou d'une variété de langues (par ex. McEnery et al., 2006 : 5). Cette définition n'est pas sans controverse (cf. Gilquin & Gries, 2009), *a fortiori* pour les enseignants et apprenants de langues (L1 ou L2) qui n'ont pas forcément les mêmes besoins que les chercheurs en sciences du langage. Dès ses débuts, le Crapel parlait de "corpus pédagogique" pour désigner l'ensemble de documents dont disposait l'apprenant – élément crucial à l'époque où l'enseignant avec son centre de ressources était le seul détenteur de ces documents (voir Holec, 1990). De nos jours, existe-t-il désormais des corpus "prototypiques" dans des environnements développant des compétences plurilingues ?

3 Pour les didacticiens des langues, la notion de corpus soulève différentes questions que l'on peut classer sous les rubriques suivantes : Qu'est-ce qu'un corpus en didactique des langues ? Corpus et perspectives pour l'enseignant ; Corpus et perspectives pour l'apprenant ; Corpus et acquisition des langues. À chacune d'elles sont associés un ou deux articles qui se chargent de répondre aux questions que l'on peut se poser. Nous y avons ajouté pour conclure la présentation de l'infrastructure Equipex *Ortolang* qui permet de regarder de l'autre côté du miroir.

1. Qu'est-ce qu'un corpus en didactique des langues ?

4 La question de la nature et de la représentativité des corpus fera l'objet de la première section de ce numéro. Ici, le corpus est interrogé dans sa relation à la langue qu'il éclaire. Y a-t-il une taille optimale d'un corpus en termes de nombre de mots et / ou de nombre de textes ? Doit-on chercher rigoureusement un corpus représentatif, peut-on utiliser le web comme corpus, un corpus *ad hoc*, ou encore une poignée de textes sélectionnés pour démontrer un phénomène précis ? Un étiquetage est-il utile pour les apprenants d'une L2 ? Un corpus écrit peut-il exister sur un support autre qu'électronique, et quels usages faire des corpus oraux ou multimodaux ?

5 Les objectifs pédagogiques et linguistiques ne s'alignant pas parfaitement, quels critères s'imposent pour définir un "corpus" en didactique des langues en termes de forme et de contenu (écrit, oral, parlé, multimodal, parallèle, etc.), de représentativité, de types de textes, de producteurs (natifs, experts, non natifs), de taille, d'annotation, etc. ?

6 L'article de **Kate Beeching** (University of the West of England, Bristol, Grande-Bretagne), par un détour diachronique, propose d'éclairer la définition de la notion de corpus pour la didactique des langues en se plaçant délibérément du côté de l'apprenant et en posant comme essentielle la relation. Pour Kate Beeching, l'engagement garantirait l'authenticité des tâches proposées sur corpus. En particulier, elle décrit différents critères à prendre en compte lorsque l'on choisit ou crée un corpus soi-même : écrit / parlé, sa taille et sa représentativité selon les besoins (spécialisés ou autres) des utilisateurs, l'exploitation des corpus d'apprenants, et ainsi de suite. Si un corpus est une ressource à de multiples affordances, il faut reconnaître qu'il n'y a pas de corpus "parfait" en soi ; l'essentiel est de choisir ou de concevoir un corpus approprié à l'usage que l'on souhaite en faire.

2. Corpus et perspectives pour l'enseignant

7 La seconde section porte sur les usages didactiques des corpus en cours de langue. Ici, le corpus est conçu en tant que ressource. Comment alors exploiter un corpus qui favorise l'apprentissage, comme outil de référence, à travers des concordanciers dédiés ou de logiciels plus généraux, sur papier, etc. ? Que peut-on / ne peut-on pas faire de façon efficace avec un corpus ? Le travail sur corpus se limite-t-il à l'interface lexico-grammaticale ou peut-on s'en servir à d'autres fins (discours, prononciation, etc.) ? Existe-t-il des preuves appuyant les avantages qui sont souvent attribués au travail sur corpus (gains en autonomie, sensibilité linguistique, esprit critique, constructivisme, compétences cognitives et métacognitives, etc.) ?

8 Quelles compétences et quel niveau de formation sont nécessaires pour l'enseignant de langues ? Quels outils conviennent et que peut-on en faire pour ses propres besoins ainsi que pour la préparation d'activités, de tests, etc. de différents types et pour quels objectifs ? Natalie Kübler et Geoffrey Sockett apportent un éclairage sur les relations qu'entretiennent les enseignants avec les corpus en didactique des langues.

9 **Natalie Kübler** (Université Paris Diderot – Paris 7), pour sa part, s'intéresse ici aux corpus d'anglais et de français en abordant la question des compétences nécessaires pour les enseignants de langues souhaitant initier des publics autres que spécialistes en langues à

un travail sur corpus. Elle s'intéresse principalement aux étudiants en Lansad (langues pour spécialistes d'autres disciplines) et aux apprentis traducteurs, qui ont souvent eux aussi besoin d'une langue de spécialité, mais finalement à tout utilisateur professionnel avec des besoins précis. En partant de pratiques qu'elle utilise elle-même, elle préconise une certaine formation théorique et souligne l'importance du travail personnel et de l'apport d'exemples concrets à tout moment. La théorie et la pratique réunies permettent de tirer meilleur profit des atouts d'une approche sur corpus – une sensibilisation à la langue, aux usages probables (plutôt que possibles), des collocations, des formulations plus ou moins figées, la variation entre différents genres, et (pour l'anglais et le français en particulier) la transparence lexicale trompeuse. Les exemples fournis permettent de survoler un grand nombre d'outils et de corpus pour ces deux langues. Pour des besoins spécialisés, même les grands corpus "génériques" se révèlent utiles et contiennent des informations absentes des dictionnaires ; mais Natalie Kübler démontre aussi l'importance des corpus spécialisés – à choisir ou à créer soi-même.

10 **Geoffrey Sockett** (Université Paris Descartes), de son côté, relie l'apprentissage sur corpus et les apprentissages informels, notamment pour l'anglais, chez les apprenants avancés qui regardent des séries télévisées ou films en streaming en version originale, écoutent de la musique en langue étrangère, participent à des échanges en ligne ou des réseaux sociaux, etc. Ces activités permettent l'exposition massive à la langue essentielle pour l'apprentissage selon les modèles dits "basées sur l'usage" et qui est au cœur de l'apprentissage sur corpus. Du fait d'usages de la langue cible qui échappent de plus en plus à la salle de cours et sont peu familiers aux enseignants, il plaide pour des utilisations nouvelles de ces corpus en tant que ressources et activités d'apprentissage. D'un côté, un travail sur corpus encourage une sensibilisation consciente au contact avec la langue qui peut manquer dans les utilisations informelles. En revanche, Geoffrey Sockett constate et pointe un certain nombre d'obstacles ou d'incohérences dans une approche d'apprentissage sur corpus telle qu'elle est souvent perçue par ceux qui la prônent.

3. Corpus et perspectives pour l'apprenant

11 La question de l'adéquation des corpus aux apprenants occupe la troisième section dans laquelle se trouvent les questions sur les apprentissages. Un travail sur corpus convient-il à tous publics d'apprenants (L1, L2, âge, niveau, profil, préférences / styles, cultures, besoins, etc.) ? Faut-il une formation conséquente ou peut-on aborder un corpus même de façon "basique" rapidement ?

12 Quels usages d'un corpus peut en faire l'apprenant(e) directement ou indirectement (des activités préparées sur papier jusqu'à l'utilisation autonome d'un concordancier) en tant qu'aide à l'apprentissage ou bien comme outil de référence ? Dans quelles conditions ces approches conviennent-elles – à qui, pourquoi, etc. ? Ana Frankenberg-Garcia et Maud Ciekanski se centrent sur les corpus du point de vue de l'apprenant et des utilisations qu'il en fait dans son apprentissage des langues.

13 **Ana Frankenberg-Garcia** (University of Surrey, Guildford, Grande Bretagne) identifie trois approches possibles pour les usages de corpus pour l'apprentissage des langues. La première concerne l'utilisation de ressources conçues en amont par des professionnels et informées par des corpus (dictionnaires, grammaires, manuels, etc.). Si l'impact des

corpus peut être invisible à l'utilisateur dans ces cas, il en va tout autrement pour les deux approches suivantes où l'apprenant travaille directement sur des données de corpus. Ce contact peut être médié par l'enseignant, notamment sous forme d'activités sur des lignes de concordances imprimées pour répondre à une question précise ; mais certains apprenants peuvent aller plus loin et interroger directement un corpus, lorsqu'ils ont des questions plus individualisées. Pour Ana Frankenberg-Garcia, cette dernière approche est plus pertinente et plus autonomisante, mais ne conviendra pas à tout public – en effet, il est important pour l'enseignant de ne pas se laisser emporter par les multiples affordances des corpus, outils et méthodologies et de respecter à tout moment les besoins pédagogiques et humains des apprenants.

14 **Maud Ciekanski** (Université de Lorraine) s'intéresse, pour sa part, à l'utilisation directe des corpus par les apprenants, souvent qualifiée d' "utilisation en autonomie". Elle y retrouve bon nombre de concepts-clés issus de la didactique des langues d'aujourd'hui, tels l'exposition à la langue, l'authenticité, la découverte, une sensibilité à la variation et aux régularités, un traitement cognitif profond, l'apprentissage de séquences plus ou moins figées, etc. Elle s'arrête sur la question de l' "authenticité" en se demandant si les corpus répondent aux caractéristiques d'une ressource adaptée à l'apprentissage en autonomie et retient un certain nombre de freins dans ce sens, notamment "l'accessibilité cognitive des corpus", y compris son interface et la présentation de masses importantes de données. L'auteure décrit deux dispositifs qui se donnent pour objectif de surmonter ce type de barrière avec l'accès à des corpus vidéo sur des sujets bien spécifiques : Sacodeyl et Fleuron. Mais pour Maud Ciekanski, un usage autonome des corpus devra être étayée par l'intervention de l'enseignant pour un meilleur guidage, et une médiation plus efficace ; or, l'enseignant s'avère, selon l'auteure, être lui-même l'un des principaux freins à l'utilisation des corpus.

4. Corpus et acquisition des langues

15 La question du niveau d'exploitation de corpus en didactique est abordée dans la quatrième section. On peut en effet identifier des corpus en amont pour informer le contenu des syllabus, l'évaluation, les matériels, les outils, des corpus constitués par les enseignants en dehors de la classe (en langue de spécialité, comme source d'exemples, etc.), voire même des corpus construits par les apprenants directement. Quel type d'informant (le natif, l'expert, le novice, etc.) convient à différentes utilisations ? Les corpus d'apprenants servent-ils uniquement à déceler des problèmes linguistiques à éviter, ou sont-ils utiles comme référentiels à différents niveaux ? Les apprenants peuvent-ils se servir de corpus d'apprenants, voire de corpus de leurs propres productions, ou créer des corpus eux-mêmes. Peut-on justifier l'exploitation en classe d'un corpus de textes non authentiques, quelle que soit sa définition ?

16 Que nous montrent les corpus d'apprenants en termes de développement langagier ? Dans quelle mesure l'analyse des erreurs (et des formes non erronées) peut-elle informer l'élaboration de programmes, voire inspirer directement des activités pédagogiques ?

17 L'article de **Sylvie De Cock** (Université catholique de Louvain et Université Saint-Louis à Bruxelles, Belgique) et de **Henry Tyne** (Université de Perpignan Via Domitia) s'intéresse aux "corpus d'apprenants", recueils électroniques de productions orales ou écrites d'apprenants de langue seconde ou étrangère (L2). Ces corpus nous permettent une

approche beaucoup plus systématique et transparente pour aborder l'acquisition de différentes L2 par des locuteurs de différentes langues maternelles dans différents contextes, à l'oral comme à l'écrit. De meilleures descriptions nous permettent de meilleures théories ; en particulier, des corpus (quasi)longitudinaux nous informent sur le développement langagier, la construction dynamique et complexe des savoirs langagiers. Ce travail peut également informer des ressources (dictionnaires, grammaires, manuels, etc.) et activités, des programmes et des descripteurs en lien, par exemple, avec *le Cadre européen commun de référence pour les langues* (Conseil de l'Europe, 2000). Par ailleurs, les apprenants peuvent, dans une perspective inductive, travailler sur des corpus locaux de leurs propres productions ou de celles de leurs pairs, se rapprochant ainsi du *data-driven learning*. Sylvie De Cock et Henry Tyne ne négligent toutefois pas les obstacles logistiques (écologie de collecte de données écrites mais surtout orales, corpus et outils chers ou indisponibles, systèmes d'étiquetage multiples et ad hoc), ni théoriques, voire philosophiques (la nature floue des données qui résistent à une classification hermétique, distinction natif" / "non natif", statut de l' "erreur", etc.). Sur ce dernier point, grâce aux corpus d'apprenants, nous pouvons dépasser la position simpliste selon laquelle toute déviation par rapport aux normes natives constitue une erreur ; on peut distinguer des erreurs qui entravent la communication, d'autres qui n'ont pas d'effets, et d'autres encore qui sont même éventuellement bénéfiques si elles compensent d'autres problèmes. Ainsi, l'erreur n'est pas forcément à éliminer à tout prix, et nous pouvons aussi regarder ce que les apprenants savent faire de façon efficace en complément des défaillances.

5. L'infrastructure Equipex Ortolang

18 Internet permet un partage de ressources inconcevable autrefois, mais toute nouvelle technologie qui se généralise est en danger de fragmentation et de dispersion, et l'utilisateur peut se trouver devant un trop grand nombre de corpus et d'outils incompatibles. Dans l'article qui occupe la dernière section de ce numéro spécial sur la notion de corpus, Jean-Marie Pierrel (Université de Lorraine) présente la genèse et la mise en place d'une infrastructure d'importance, *Ortolang (Open Resources and Tools for LANGuage / Outils et Ressources pour un Traitement Optimisé de la LANGue* : www.ortolang.fr). L'objectif d'Ortolang est d'assurer la gestion, la mutualisation, la diffusion et la pérennisation de ressources linguistiques de type corpus, dictionnaires, lexiques et outils de traitement de la langue, avec une focalisation particulière sur le français et les langues de France. La contribution de Jean-Marie Pierrel permet aux didacticiens de passer pour ainsi dire de l'autre côté de la scène, en regardant comment de grands corpus se constituent grâce aux efforts concertés de plusieurs structures de recherche.

19 Ce numéro spécial de *Recherches en didactique des langues et des cultures : Les Cahiers de l'Acedle* est une contribution en France à un questionnement largement porté par le monde anglo-saxon et cependant peu présent en France (voir aussi Boulton & Tyne, 2014). Comme tous les numéros inspirés par les journées NeQ (*Notions en questions*) de l'Acedle, il contribue à l'état des lieux à un moment donné et engage à aller plus loin encore dans la réflexion, en tissant des relations entre chercheurs du domaine et avec la didactique des langues.

BIBLIOGRAPHIE

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

Conseil de l'Europe (2000). *Un cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Disponible en ligne. http://www.coe.int/t/dg4/linguistic/Source/Framework_FR.pdf

Gilquin, G. & Gries, S. (2009). "Corpora and experimental methods: a state of the art review". *Corpus Linguistics and Linguistic Theory*, vol. 5, n° 1. pp. 1-26.

Holec, H. (1990). "Des documents authentiques, pour quoi faire ?". *Mélanges Pédagogiques*, vol. 1990. pp. 65-74.

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

AUTEURS

DOMINIQUE MACAIRE

Laboratoire ATILF, équipe Didactique des langues et sociolinguistique (Crapel) ; UMR 7118, Université de Lorraine & CNRS

ALEX BOULTON

Laboratoire ATILF, équipe Didactique des langues et sociolinguistique (Crapel) ; UMR 7118, Université de Lorraine & CNRS