CORPUS

Corpus

20 | 2020
Corpus complexes

Les corpus web à travers le prisme de l'ALMT

Web corpora through the lens of Call

Eva Schaeffer-Lacroix

OpenEdition Journals

Electronic version
URL: http://journals.openedition.org/corpus/4579
DOI: 10.4000/corpus.4579
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
Eva Schaeffer-Lacroix, « Les corpus web à travers le prisme de l'ALMT », *Corpus* [Online], 20 | 2020, Online since 22 January 2020, connection on 08 September 2020. URL : http://journals.openedition.org/corpus/4579 ; DOI : https://doi.org/10.4000/corpus.4579

Les corpus web à travers le prisme de l'ALMT

Web corpora through the lens of Call

Eva Schaeffer-Lacroix

Introduction

1 L'intérêt de l'usage de corpus numériques pour enseigner-apprendre une langue étrangère, qu'il s'agisse de l'enseignement postsecondaire ou secondaire, n'est plus à démontrer (Crosthwaite 2019 ; Godwin-Jones 2017 ; Luo et Zhou 2017). Les domaines les plus marqués par ces outils sont l'apprentissage de la production écrite (Tono, Satake, et Miura 2014 ; Dodd 1997), la réflexion sur la langue (Boulton 2011 ; Cheng, Warren, et Xun-feng 2003 ; Schaeffer-Lacroix 2016) et l'apprentissage de la langue de spécialité (Boulton et Landure 2016 ; Charles 2018). Il semble pertinent de se concentrer dorénavant sur à la phase de constitution de corpus dits « pédagogiques » (Braun 2005 ; Polezzi 1994), qui ont les vertus attendues pour une ressource instrumentée pouvant convenir à un public d'apprenants d'une LE : accessibilité, lisibilité et gratuité (Schäfer et Bildhauer 2013). Je propose de rajouter le critère de modificabilité, permettant aux concepteurs de corpus pédagogiques d'adapter les données du point de vue de la qualité et du contenu aux besoins d'apprenants peu avancés et/ou mineurs.

2 Les corpus réflexifs (Mayaffre 2002) répondent bien aux exigences de contrôle de qualité et de contenu. Toutefois, il peut être difficile de trouver des données en quantité suffisante qui sont libres de droit et qui conviennent pour un usage pédagogique.

3 Les corpus web se distinguent par leur taille, qui peut aller jusqu'à plusieurs milliards de tokens. Créés de façon automatique ou semi-automatique, ils permettent de rassembler à moindre coût, à partir d'Internet, des données textuelles dont certaines sont ultra-récentes. On peut toutefois être sceptique par rapport au potentiel pédagogique de tels corpus en raison de la présence potentielle de coquilles et d'erreurs de langue et en raison de leurs contenus disparates d'origine parfois inconnue ou

difficile à détecter. De plus, les apprenants risquent de ressentir comme écrasante la masse de données à laquelle ils sont exposés.

4 Cet article cherche à évaluer dans quelle mesure les corpus web peuvent soutenir l'enseignement-apprentissage d'une langue étrangère. Dans les sections qui suivent, je fournirai des définitions de « corpus pédagogique », « corpus réflexif » et « corpus web ». Ensuite, je contrasterai les apports des corpus web pour l'enseignement-apprentissage des langues étrangères dans le secondaire en France avec les qualités attribuées aux corpus réflexifs (Mayaffre 2002). J'illustrerai enfin le potentiel pédagogique de trois types de corpus, dont un corpus web et un corpus réflexif dérivé d'un corpus web, à l'exemple d'un scénario pédagogique. Ce scénario, visant l'écriture d'une petite annonce pour vendre, donner ou échanger un objet, sera implémenté avec une classe de 22 élèves en quatrième année d'apprentissage de l'allemand inscrits dans un collège parisien.

Définitions

Qu'est-ce qu'un corpus pédagogique ?

5 Selon Polezzi (1994), les corpus pédagogiques contiennent un choix de textes représentant de façon exemplaire des discours et/ou des genres textuels susceptibles de répondre aux besoins d'un public d'apprenants dans une situation d'apprentissage déterminée.

> FL materials can be designed–or, where appropriate, selected from existing sources–on the basis of the academic/professional needs and interests of the learners. To do this we have to locate the target discourse community and its characteristic discourse type(s). The resulting set of texts constitutes a small, pedagogical corpus, which is a subset of the genre(s) in question.
> [Les matériels de LE peuvent être conçus – ou, si indiqué, sélectionnés à partir de sources existantes – sur la base des besoins et des intérêts universitaires / professionnels des apprenants. Pour ce faire, il convient de repérer la communauté discursive cible et le(s) types de discours qui la caractérisent. L'ensemble de textes qui en résulte constitue un petit corpus pédagogique, qui est un sous-ensemble du ou des genres en question.]

6 Un deuxième critère définitoire est celui de la fiabilité des données du corpus pédagogique, qui doivent refléter les normes linguistiques caractérisant le discours et/ou du genre textuel retenu(s). Ce critère est également valable pour les corpus réflexifs.

Qu'est-ce qu'un corpus réflexif ?

7 Un corpus réflexif contient des textes entiers, sélectionnés selon des principes philologiques de proximité générique ou autre. Selon Mayaffre (2002), les constituants d'un tel corpus forment ensemble « un réseau sémantique performant dans un tout (le corpus) cohérent et autosuffisant ». L'analyse de tels ensembles textuels se concentre sur leurs caractéristiques intrinsèques. Rastier (2004) les distingue d'ensembles qu'il appelle « sacs de mots », constitués d'unités plus petites que des textes entiers, et d'archives, contenant l'ensemble des documents accessibles, sans mise en perspective par un projet de recherche identifiable. Le courant de la textométrie s'intéresse, entre autres, aux spécificités des corpus réflexifs. L'outil TXM (Heiden, Magué, et Pincemin

2010) est particulièrement adapté à la constitution et l'exploration de ce type de corpus.

Qu'est-ce qu'un corpus web ?

8 Un corpus web contient des données publiées sur la Toile, récupérées à l'aide de processus automatisés comme le *web-crawling* (Olston et Najork 2010 ; Schäfer et Bildhauer 2013). Afin de respecter les droits d'auteur, les textes de certains corpus web sont réduits à des échantillons : ils sont divisés en phrases, p. ex. les corpus rassemblés dans la *Leipzig Corpora Collection* (Goldhahn, Eckart et Quasthoff 2012), ou en d'autres unités considérées comme tolérables au nom des droits d'auteur du pays dans lequel le corpus a été créé. Dans la littérature, les corpus web sont parfois définis en creux, en les opposant aux corpus dits « traditionnels » (Gatto 2011, 39 ; Schäfer et Bildhauer 2013). Tout comme dans le contexte de la grammaire, le terme de « traditionnel » semble être utilisé pour dire que l'élément qu'il qualifie serait ancien, voire obsolète, ce qui ne correspond pas à une qualification matérielle. Je proposerais plutôt d'opposer les corpus web à des corpus réflexifs (v. section précédente). J'émets également une réserve par rapport au terme d'« authentique », qui peut être compris comme le synonyme de « non inventé ». Ce terme qualifierait, selon Gatto (2011, 4), les données des corpus web, mais il ne serait pas applicable aux données des corpus dits « traditionnels ». Il me paraît plus pertinent de distinguer entre des données non modifiées et des données modifiées. Des modifications peuvent être effectuées lors d'une intervention humaine ayant pour but d'adapter les textes au public cible, par exemple à des apprenants d'une langue étrangère ou à des personnes ayant des besoins particuliers (cf. l'initiative du langage simplifié, décrite dans Bredel et Maaß 2016). Il arrive aussi que l'on intervienne sur des textes pour des raisons pratiques (par exemple, quand un journaliste est prié de raccourcir un texte) ou dans le cadre d'une censure politique. Les corpus web contiennent, en effet – sans que cela ne soit exclusif – une grande proportion de données non modifiées. Les caractéristiques présentées en tableau 1 ne sont pas figées. Je force délibérément le trait afin de montrer les tendances qui opposent les deux entités.

Tableau 1. Caractéristiques des corpus web et des corpus réflexifs

Caractéristiques	**Corpus web**	**Corpus réflexifs**
Critères pour le choix des ressources	- Actualité - Variété - Représentativité	Critères philologiques (textes entiers, formant un ensemble cohérent)
Récolte des données	Données identifiées et récoltées par un robot (*web-crawling*), avec postédition humaine plus ou moins avancée	Données identifiées et récoltées par des personnes, avec l'aide plus ou moins avancée de la machine
Taille	Gigantesque	Moyenne ou petite
État	Dynamique et ouvert	Statique et fermé

Nature du contenu	« Authentique » - Contemporain - Genres textuels émergents - Non modifié	« Non authentique » - Historique - Genres textuels établis - Non modifié

9 La limite entre les corpus réflexifs et les corpus web est parfois floue, comme l'illustrent les caractéristiques du corpus trilingue Bundesblatt / Feuille fédérale / Foglio federale (Elmiger 2015). De taille assez conséquente (en tout presque 500 millions de mots), ses données ont été récoltées de façon automatique, mais elles proviennent d'un seul site, celui du gouvernement suisse. Il s'agit de publications hebdomadaires hétérogènes, mais relevant toutes du discours administratif. Le corpus est stocké sur CQPweb et peut être divisé en sous-corpus, en sélectionnant des années ou d'autres périodes temporelles. Il n'est pas possible de faire des requêtes qui ciblent l'un des différents genres textuels ou types de textes représentés dans ce corpus, à savoir les rapports du Conseil fédéral aux Chambres, les arrêtés et les lois qui ont été votés par les Chambres, les décisions de la Chancellerie fédérale en matière d'initiatives populaires ou de référendums et les notifications des unités de l'administration fédérale et des tribunaux. Le corpus est statique, mais on pourrait l'agrandir vu que les publications continuent à ce jour. Les erreurs de reconnaissance des caractères dans les textes scannés, donc ceux datant d'avant 1999, n'ont pas été rectifiées, ce qui impacte principalement (et de façon modeste) les résultats des requêtes appliquées à la période la plus ancienne (1849-1900). On peut donc considérer la Feuille fédérale comme un corpus hybride, ayant aussi bien des caractéristiques d'un corpus web que d'un corpus réflexif.

Les corpus web du point de vue de la didactique de LE

10 Les corpus web peuvent être appréciés pour la variété, l'étendue et le caractère potentiellement ultra-actuel des données qu'ils contiennent. Ils atteignent une taille allant parfois jusqu'à plusieurs milliards de tokens. Il est également défendable de considérer comme un avantage le fait que ce type de corpus est le miroir de la société actuelle, qui a tendance à s'organiser en réseaux (Gatto 2011).

11 La création de corpus web nécessite de relever des défis variés : comment éviter l'acquisition de données non désirées, par exemple des clauses juridiques ? Comment structurer et comment annoter les données obtenues ? Ces défis vont de pair avec le développement d'outils technologiques de plus en plus sophistiqués. La Toile est organisée de façon extrêmement complexe, pour ne pas dire anarchique. Des contenus pertinents côtoient des contenus que l'on peut trouver non pertinents, comme des sigles ou chiffres ne formant pas de texte, des fragments de textes ou des listes (Fletcher 2004). De plus, l'édition des textes est de qualité variable. On peut trouver que pour un contexte d'enseignement-apprentissage des langues, il y a un risque trop important d'erreurs de reconnaissance de caractères, de coquilles et d'erreurs de langue. Un dernier point est à mentionner qui concerne aussi les corpus réflexifs, mais qui est plus délicat à gérer lors de la constitution de corpus web : les droits d'auteur, variables d'un pays à l'autre, sont particulièrement difficiles à clarifier, par exemple, quand l'origine d'un texte ne peut pas être identifiée avec précision. Certains créateurs

de corpus web optent alors pour la restriction à la taille d'une phrase, ce qui n'est pas dans tous les cas une option acceptable dans le contexte de l'enseignement-apprentissage d'une LE.

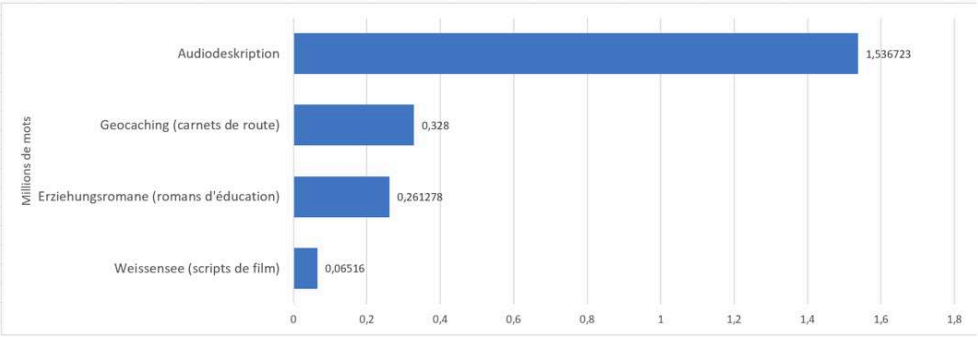
12 Dans les sections suivantes, trois paramètres sont considérés du point de vue de la didactique de LE, à savoir la taille du corpus, son contenu et la qualité des données qu'il contient.

La taille du corpus

13 Une des différences majeures entre les corpus web et les corpus pédagogiques est leur taille potentielle, les uns pouvant être gigantesques, les autres étant parfois minuscules. Dans la littérature, on rencontre les termes de « taille critique » ou « taille minimale » (Chambers 2005 ; Bernardini 2000). Selon Chambers (2005), des corpus dépassant un million de mots ne conviendraient pas à un public d'apprenants peu avancé en langue étrangère. On peut toutefois aussi se demander s'il y a un seuil critique en dessous duquel une collection de textes ne peut pas encore être considérée comme un corpus parce que les recherches que l'on peut y appliquer ne donneront pas satisfaction. Cheng, Warren, et Xun-feng (2003) privilégient des corpus qui dépassent cinq millions de mots parce qu'en dessous de ce seuil, l'étude de la langue via une méthode inductive ne serait pas fructueuse. Dodd (1997, 131) plaide pour des corpus généraux contenant au moins un million de mots, et Aston (2002) recommande l'usage de collections de très grande taille dont on extraira des sous-corpus, selon les besoins du projet pédagogique.

14 La taille de la plupart des corpus pédagogiques que j'ai eu l'occasion de créer entre 2012 et 2017 (v. Graphique 1) ne correspond pas aux seuils mentionnés ci-dessus. Elle est toutefois en augmentation au fil des années : le corpus le plus récent, appelé *Audiodeskription* [audiodescription], contient 1,536 million de mots[1].

Graphique 1. Taille de corpus pédagogiques



15 Pour les chercheurs en linguistique (et, de façon générale, en sciences humaines), la taille du corpus est un critère important, mais elle n'est pas une valeur en soi. La taille est à mettre en relation avec l'objectif poursuivi, la question posée, la tâche à résoudre. Selon Mayaffre (2002), un corpus est utile si l'on y trouve ce dont on a besoin.

> Le corpus est un objet heuristique. C'est une construction arbitraire, une composition relative qui n'a de sens, de valeur et de pertinence qu'au regard des questions qu'on va lui poser, des réponses que l'on cherche, des résultats que l'on va trouver (Mayaffre 2002).

16 Ce positionnement peut, dans certains cas, justifier une taille de corpus très petite. Pour donner un exemple : le corpus pédagogique *Weissensee*, contenant 65 160 mots, a fourni de bons services à de futures documentalistes lors d'un projet d'écriture en allemand LE d'une scène de script de film. La consultation de ce corpus leur a permis de mener des discussions métalinguistiques portant sur le sens d'éléments polysémiques et/ou polycatégoriels (préverbes et prépositions), alimentées par les données observées (Schaeffer-Lacroix 2015). De plus, quand on se place dans la perspective de la formation en langue étrangère, une très grande quantité d'occurrences peut devenir un élément de désorientation et de découragement. Il y a un risque avéré de « noyade » : Robinson (2002 : 7) parle de « *Input flooding* » [submersion par l'input] et Zanettin (2001) de « *Swimming in words* » [prendre un bain de mots]. L'un des participants du projet *Projekt Prospekt*, exécuté avec des apprenants du secondaire ayant 14 ou 15 ans (voir section suivante), dit avoir été "plongé dans le réel" pendant le travail sur corpus (Schaeffer-Lacroix 2009, 194). La rencontre avec « le réel » (ou, autrement dit, la langue en usage), qui est, par définition, complexe, crée chez certains apprenants un sentiment de submersion. Cela ne veut pas dire que les corpus web sont à proscrire pour un public d'apprenants de LE : les évolutions les plus récentes en matière de récolte de données sur la Toile permettent de croiser de façon plus aisée l'univers des corpus web et celui des corpus pédagogiques.

Le contenu du corpus

17 Le *Projekt Prospekt* a invité de jeunes apprenants d'écrire des textes ayant des caractéristiques comparables à ceux publiés sur le site Internet d'un établissement de type culture et loisirs (musée, zoo, etc.). Pour les besoins de ce projet, un corpus pédagogique de très petite taille (12 328 mots) a été utilisé. À l'aide du mot-clé « Eintritt » [entrée], faisant partie du lexique de la rubrique « Informations pratiques », j'ai comparé la performance de cette ressource à celle d'un corpus web et d'un corpus réflexif contenant un sous-corpus du domaine du tourisme. Afin de réduire le nombre très important d'occurrences non pertinentes dans le corpus web, j'ai restreint la requête en ajoutant comme contrainte la co-présence entre « Eintritt » [entrée] et « frei » [libre]. Le tableau 2 compare la pertinence et la quantité des occurrences obtenues pour ces requêtes.

Tableau 2. Pertinence versus quantité des occurrences

Nom du corpus	Praktische Informationen	English-German Translation corpus	Webcorpus 2016c
Créateur du corpus	Schaeffer-Lacroix	Technische Universität Chemnitz	Barbaresi
Type de corpus	Corpus réflexif	Corpus équilibré, sous-corpus "Tourism brochures"	Corpus web
Nombre de mots	12 328	250 000	3,1 milliards

Fréquence absolue de "Eintritt" Occurrences pertinentes	34 34	11 10	86 597 29 057 pour "Eintritt && frei* || frei* && Eintritt"
Fréquence relative des occurrences pertinentes (par million)	2758	40	9,36[2]

18 Les résultats permettent d'illustrer que pour certains types de requêtes, le contenu du corpus prime sur sa taille. Dans le cas présenté, le fait d'avoir sélectionné des données correspondant au genre textuel visé garantit l'obtention d'occurrences fiables et utilisables en vue du projet de production de texte.

19 On peut saluer les initiatives de développement d'outils et de procédures de récolte de données qui prennent en compte le critère de contenu. Selon Volk (2002), on est de plus en plus en mesure de "pêcher dans les eaux de la Toile" (ma traduction) (voir aussi Sharoff 2005). Il est possible de restreindre les requêtes à des domaines web (par exemple, la presse, la culture, etc.) ou à autres éléments distinctifs. L'identification automatique de registres et de genres textuels est également en progrès (Gatto 2011, Sharoff 2018), même si ce geste reste un défi important pour les créateurs de corpus web. Barbaresi (2016) évoque la question des genres web ("the tricky question of web genres") ; on peut, en effet, s'attendre à voir émerger, à côté de genres textuels codifiés, des genres qui sont spécifiques pour la Toile et dont certains correspondent à des évolutions sociétales récentes.

La qualité des données

20 Du fait de leur taille immense, les corpus web sont des objets difficiles à cerner et à contrôler ; la modification des données après leur récolte est une tâche qui peut paraître insurmontable par des non-spécialistes du TAL (traitement automatique des langues). Ces modifications, motivées par un but pédagogique, peuvent consister en une suppression d'éléments non désirés du point de vue de la forme (par exemple, des textes composés de sigles) et/ou du contenu (par exemple, des informations contractuelles ou des textes incitant au racisme) et à la rectification de coquilles et d'erreurs de langue. Toutefois, en matière de postédition de données de corpus web, on peut constater d'importants progrès techniques (Barbaresi 2016 ; Benko 2016 ; Schäfer 2017).

21 Les corpus web contiennent une certaine proportion de données linguistiques qui ne correspondent pas forcément aux normes (p. ex., orthographiques) adoptées dans le milieu de la formation linguistique. Toutefois, doit-on systématiquement exiger une « qualité irréprochable » de la langue, jugée à l'aune d'une norme prescriptive, pour un contexte d'enseignement-apprentissage d'une LE ? Ne serait-il pas défendable de confronter les apprenants au fait que la langue varie selon les genres textuels, les supports de publication et la littératie des auteurs ? Certaines variantes devraient être acceptées parce qu'elles sont constitutives pour un genre textuel donné. Je pense, par exemple, à l'écriture en minuscules des noms en langue allemande dans des livres d'or

sur des sites Internet de musées ou à l'écrit oralisant, qui caractérise les interactions sur un forum en ligne.

Scénario *Kleinanzeigen*

22 Le scénario *Kleinanzeigen* [petites annonces] sert d'objet de démonstration pour mettre en regard le potentiel de trois corpus et les besoins d'apprenants en quatrième année d'apprentissage de l'allemand lors d'un projet de production écrite et de réflexion sur la langue. Son implémentation avec 22 élèves d'une classe de quatrième dans un collège parisien est planifiée pour l'année scolaire en cours. Le projet vise à observer et à documenter la façon dont ce public se saisit de ressources et d'outils numériques pour écrire une petite annonce en ligne destinée à vendre, échanger ou donner un objet. D'une durée totale de cinq heures de cours consécutives, il prendra la forme d'une recherche expérimentale, avec pré- et post-test et entretiens semi-dirigés post-recherche. Il est prévu de séparer la classe en deux. Je prendrai en charge le groupe expérimental, qui se servira d'outils de corpus. L'enseignante habituelle de la classe travaillera avec le groupe témoin, qui aura accès au site d'annonces kleinanzeigen.de. La tâche principale consiste en l'écriture d'une petite annonce pour vendre ou donner un objet dont les apprenants auraient envie de se séparer. Elle nécessite la prise en compte des besoins des apprenants, qui peuvent être d'ordre culturel, pragmatique, linguistique, technique et motivationnel.

Figure 1. Exemple d'une production attendue

Biete ein gut erhaltenes Ravensburger Puzzle, 3000 Teile, mit tollem Bild vom Schloss Neuschwanstein. Gebraucht, aber wie neu. Zu verkaufen für 5 €.
[Offre un puzzle en bon état, 3000 pièces, représentant le beau château de Neuschwanstein. D'occasion, mais comme neuf. En vente au prix de 5 €.]

23 Le troisième cours sera dédié à une phase de réflexion sur la langue, alimentée par des passages repérés dans les brouillons des élèves. Le tableau 3 présente un éventail de ressources et outils possibles, à sélectionner en fonction du profil du groupe d'apprenants.

Tableau 3. Synopsis du scénario *Kleinanzeigen*

Tâche principale	**Créer une petite annonce pour vendre ou donner un objet**
Public	22 élèves d'une classe de quatrième d'un collège parisien (enseignement secondaire), 4e année d'apprentissage de l'allemand
Modalités	Atelier d'une durée de cinq heures, en présence de la chercheuse (groupe expérimental) et de l'enseignante habituelle (groupe témoin)
Objectifs linguistiques	- Observer et comprendre le rôle des terminaisons dans des énoncés comportant des adjectifs - Adopter les conventions discursives du genre textuel "petite annonce" (sections et formules)
Ressources	- Araneum Germanicum Maius - COW-ebay - Schweizer Textkorpus - Site Kleinanzeigen.de
Outils d'exploration de corpus	- Sketch Engine - NoSketch Engine

Questions de travail

24 Le projet cherche à évaluer dans quelle mesure les apprenants tirent profit des ressources et outils proposés pour rédiger et réviser leur texte et pour réfléchir sur la langue. Les recherches et observations se concentreront sur les terminaisons des adjectifs (indiquant le genre, le nombre et le cas), repérées par (Diehl et al. 2000, 198) comme contenus d'apprentissage particulièrement complexes, appris tardivement. De plus, il sera étudié dans quelle mesure l'usage d'outils de corpus ou du site d'annonces favorise le repérage et l'emploi pertinent de formules contenant des adjectifs, p.ex. "in neuwertigem Zustand" [dans un état neuf/comme neuf].

25 Les hypothèses suivantes cadrent le projet.

26 H1 : La manipulation des données d'un corpus numérique aide, mieux que la consultation d'un site Internet tel que kleinanzeigen.de, la conceptualisation de l'adjectif par rapport à ses fonctions (attributives, prédictives, adverbiales : a est b ; b a ; c b a ; c b d a) ;
"Das Spiel ist toll" [Ce jeu est génial]
"tolle Spiele" [des jeux géniaux]
"ein tolles Spiel" [un jeu génial]
"ein toll gemachtes Spiel" [un jeu génialement bien fait]

27 Indices pour H1 : traces d'une activité d'apprentissage invitant les élèves à identifier les fonctions des adjectifs dans les annonces (enregistrements audio des séances entières ainsi que des interactions verbales entre binômes d'élèves) ; traces des actions numériques (écrans d'ordinateur filmés) ; informations recueillies à l'aide des interviews.

28 H2 : La manipulation des données du corpus aide, mieux que la manipulation des données du site Internet, la conceptualisation de l'adjectif par rapport à ses terminaisons (motivées par la fonction, le genre, le nombre et le cas).

29 Indices pour H2 : scores des terminaisons pertinentes obtenus dans des textes contenant des mots inconnus (genre indiqué), proposés lors d'activités d'entrainement ; différence entre le pré-test et le post-test.

30 H3 : Nous nous attendons à ce que le répertoire linguistique du groupe expérimental se distingue de celui du groupe témoin en ce qui concerne le choix de l'adjectif. Il sera plus étendu, plus varié et plus pertinent.

31 Indices pour H3 : nombre et variété des adjectifs pertinents utilisés ; présence d'adjectifs pertinents peu fréquents ; différence entre le pré-test et le post-test.

Méthodologie

32 Le pré-test et le post-test seront le même pour les deux groupes. Il s'agira d'écrire, en situation contrôlée et sans aide lexicale ou autre, quelques lignes destinées à vendre ou donner un objet. Les autres traces récoltées seront les productions des apprenants (pré- et post-test ; toutes les versions de l'annonce produite), les enregistrements audio des cinq séances de cours, des interactions verbales entre binômes d'élèves et les actions effectuées à l'écran, filmées à l'aide de (*Screencast-O-Matic* s. d.). Les entretiens semi-dirigés post-recherche, menés par une formatrice extérieure au collège, seront également filmés. Les données seront analysées de façon quantitative (outils de statistique et de corpus) et qualitative (lecture linéaire des brouillons et productions finales des élèves et catégorisation des éléments à l'aide d'un fichier Excel, par la suite exploité à l'aide de l'outil R (« R: The R Project for Statistical Computing » s. d.)). Les résultats seront comparés à ceux disponibles dans des corpus d'apprenants, p. ex. dans le corpus MERLIN (Wisniewski et al. 2013).

Choix de corpus en fonction des besoins des apprenants

33 Dans les lignes qui suivent, je présenterai les avantages et inconvénients de ressources et d'outils d'exploration dont l'utilisation peut être envisagée dans le cadre du projet *Kleinanzeigen*, en les mettant en regard avec les besoins identifiés pour exécuter la tâche choisie. Il s'agit d'un corpus équilibré[3], d'un corpus web et d'un corpus réflexif créé à partir d'un corpus web.

Besoins d'ordre culturel et pragmatique (Schweizer Textkorpus)

34 Le groupe témoin accédera de façon directe à des modèles visuels de petites annonces publiées sur le site kleinanzeigen.de. Afin d'obtenir la visualisation d'une annonce illustrée faisant partie d'un corpus numérique, le groupe expérimental peut parcourir le Schweizer Textkorpus (Bickel et al. 2009), un corpus équilibré contenant

23,5 millions de tokens. Construit sur le modèle du BNC (« British National Corpus (BNC) » s. d.), il a certaines caractéristiques d'un corpus multimodal : sur simple clic, on peut accéder au scan du texte d'origine associé à une requête effectuée dans le corpus. La figure 2 en présente un exemple.

35 Traduction du texte de cette annonce paru en 1950 : « Pelikan – le stylo plume techniquement parfait et le criterium automatique. Stylo plume Pelikan Mod. 100 N 38 Fr. Criterium Auch 200 .. 9 Fr. Stylo plume IBIS, convient particulièrement à des élèves. Mod. 130 25 Fr. ».

Figure 2. Visuel d'une annonce dans le Schweizer Textkorpus

36 Il n'est pas possible de créer de sous-corpus du Schweizer Textkorpus contenant uniquement des annonces. Il faut appliquer des requêtes permettant d'en repérer, par exemple en choisissant le mot-clé « Fr » (abréviation de « Franken » [francs] ou « zu verkaufen » [à vendre]. De plus, à quelques exceptions près, seules les annonces dont les droits d'auteurs ont échu (donc, des annonces datant d'il y au moins 70 ans) sont accompagnées d'un visuel. Cela réduit l'intérêt d'un usage de ce corpus avec un public qui ne s'intéresse pas à la façon dont on illustrait les petites annonces il y a plus de 70 ans.

Besoins d'ordre morphosyntaxique (Araneum Germanicum Maius)

37 Le corpus web Araneum Germanicum Maius (Benko 2014) est accessible via Sketch Engine. Il contient 1,2 milliard de tokens, dont 875 millions sont non étiquetés. Il est possible de restreindre les requêtes à des domaines web ou à d'autres éléments distinctifs et de créer des sous-corpus selon des critères choisis par l'usager. Cela permet de se rapprocher des caractéristiques d'un corpus réflexif et de réduire la masse des données à une quantité susceptible de convenir à un public d'apprenants. La création d'une liste de fréquence des formes de mots, des lemmes ou des étiquettes

permet d'obtenir un aperçu synthétique des occurrences. Il est également possible de sélectionner des lignes de concordance pour créer un échantillon correspondant à des critères déterminés, par exemple, contenant uniquement des noms au pluriel ou au génitif. Cette liste peut ensuite être exportée (voir Tableau 4) dans le but de l'adapter par l'enseignant (si nécessaire) et de la faire observer par les apprenants.

Tableau 4. Liste de fréquence « Déterminant – Adjectif – nom »

# Frequency list # Corpus: preloaded/de_araneum_maius # Query: [tag="ART.*"][tag="ADJ.*"][lemma="Puzzle"] 352 # Frequency limit: 0		
word	**Frequency**	**Traduction**
ein großes Puzzle	19	un grand puzzle
das größte Puzzle	10	le plus grand puzzle
ein kleines Puzzle	9	un petit puzzle
einem großen Puzzle	8	à un grand puzzle
das fertige Puzzle	8	le puzzle terminé
das erste Puzzle	5	le premier puzzle
eines großen Puzzles	4	d'un grand puzzle
ein riesiges Puzzle	4	un puzzle géant
ein neues Puzzle	4	un nouveau puzzle

38 Les apprenants peuvent être invités à observer les différentes terminaisons, qui varient dans cette liste en fonction du déterminant (présence ou absence d'une terminaison ; le type de ses terminaisons) et en fonction du genre, nombre et cas du groupe nominal (GN). L'analyse de certains des items nécessite la consultation d'un contexte plus large, p. ex. afin de déterminer si le GN est au nominatif ou à l'accusatif.

39 Comme activité d'entraînement, l'enseignant pourra proposer au groupe de compléter des amorces de phrases dans lesquelles manque le nom de l'objet à vendre (signalé par « xxx » dans les exemples ci-dessous). Elles peuvent être obtenues à l'aide de la requête [word="Verkaufe"][]{0,5}[tag="ADJA"].
Verkaufe gut erhaltenen xxx [Vends xxx en bon état]
Verkaufe hübsche xxx [Vends jolie / de jolies xxx]
Verkaufe wunderschönes xxx [Vends xxx magnifique]
Verkaufe meinen gebrauchten xxx [Vends mon xxx d'occasion]
Verkaufe diese alte aber noch robuste xxx [Vends cette xxx vieille mais encore solide]

40 Mon analyse des performances du corpus Araneum Germanicum Maius en vue du projet *Kleinanzeigen* mène à la conclusion suivante : les sous-corpus réflexifs que l'on peut créer à partir de cette ressource contiennent trop de bruit. Le corpus entier, de sa part, convient bien à des requêtes ciblées portant sur les caractéristiques morphosyntaxiques. On constate une variété et pertinence satisfaisante de la forme et du sens des adjectifs observés.

Besoins d'ordre discursif (COW-ebay)

41 Les petites annonces se caractérisent par un emploi fréquent de formules (semi-)figées. Afin de les repérer, il paraît utile de se servir d'un corpus réflexif, composé uniquement de petites annonces. C'est le cas pour le corpus COW-ebay contenant 1,272975 million de tokens. Il a été créé à partir des données de DECOW 16B-NANO, d'une taille de plus de 19,8 milliards de tokens (Schäfer et Bildhauer 2013). COW-ebay est accessible, sur demande, via NoSketch Engine (*NoSketchEngine Concordance* s. d.). Cet outil offre un choix restreint des fonctionnalités de Sketch Engine (concordancier, liste de fréquence). Il manque plusieurs fonctionnalités dont la possibilité de créer des n-grams. Toutefois, l'étiquetage de DECOW 16B-NANO est plus fin et plus adapté à l'allemand que celui d'autres corpus étiquetés lors de leur import dans Sketch Engine.

42 En l'absence de la possibilité de repérer les formules à l'aide d'une liste de n-grams, j'ai choisi de chercher d'abord les verbes les plus fréquents et de faire ensuite des requêtes ciblées à partir d'un choix de ces verbes. Parmi les verbes les plus fréquents du corpus, on trouve « bieten » [offrir] et « verkaufen » [vendre]. La requête [lemma="verkaufen"] []{0,4}[tag="NN"] permet de repérer des formules (semi-)figées dont certaines indiquent la raison ou la modalité de la vente.

Verkaufe wegen einem Umzug... [Vends, pour cause de déménagement, ...]
Verkaufe für Selbstabholer... [Vends pour personnes venant chercher la marchandise chez moi...]
Verkaufe im Auftrag von... [Vends pour le compte de...]

43 Une des formules précise l'effet que le projet de vente produit sur le vendeur : *Verkaufe schweren Herzens...* [Vends, le cœur lourd, ...].

Pertinence discursive

44 On peut attendre d'un corpus réflexif que les termes que l'on y trouve recouvrent le sens attendu dans le cadre d'un genre textuel ou discursif donné. L'exemple de la requête [lemma=".*rad"], visant à rassembler les noms composés se terminant par « rad » [vélo] », satisfait cette attente, mis à part quelques occurrences non attendues désignant divers types de roue (« Allrad », « Laufrad ») ou de volants (« Lenkrad », « Lederlenkrad », « Sportlenkrad »), faisant rarement l'objet d'une vente – et encore, dans le secteur « pièces détachées », une vente de tels objets est plausible.

Figure 3. Pertinence discursive des items dans COW-ebay

Pertinence des étiquettes et des formes

45 Quand on a affaire à un corpus web pour l'ALMT, il faut s'assurer du fait que la qualité des formes et/ou de leur étiquetage est suffisante. L'exemple d'étiquettes correspondant aux verbes sans sujet « Verkaufe|Biete » [Vends|Offre] n'est pas concluant : on constate des erreurs pour plus de la moitié des occurrences, qui sont analysées par TreeTagger comme un nom ou comme un verbe à l'impératif au lieu d'être annotées comme verbe fini à la première personne.

Figure 4. Étiquettes pour des verbes sans sujet

46 Au niveau des formes, on peut également remarquer des zones d'erreur, par exemple dans le domaine des noms composés non fusionnés et au niveau des terminaisons : « *Die neuen Adventure Serie* » à la place de « *Die neue Adventureserie* » [la nouvelle série aventure] ; « *eine neuwertige Massage Rückenlehne* » à la place de « *eine neuwertige Massagenrückenlehne* » [un dossier de massage en état neuf] ; « *ein par Motor Teile* » à la place de « ein paar Motorteile » [quelques pièces d'un moteur]. Dans l'ensemble, les imprécisions sont toutefois relativement rares. De plus, on peut se demander si certaines d'entre elles (en particulier l'absence de la fusion de noms) sont constitutives pour le genre textuel tel qu'il a tendance à être pratiqué actuellement.

47 On peut avoir des doutes par rapport à la faisabilité d'une exploration d'un des giga-corpus accessibles via COW dans un contexte d'un projet d'écriture d'une petite annonce avec un public d'apprenants peu avancés. En revanche, le sous-corpus COW-ebay a certaines des qualités attendues : il est possible y détecter des caractéristiques linguistiques et discursives intéressantes, comme l'absence ou la présence d'un déterminant devant les noms. Il contient des formules récurrentes dont certaines

précisent les modalités de la vente : "schweren Herzens" [le cœur lourd], et on y trouve des marqueurs énonciatifs comme dans « Verkaufe hier… » [Vends ici…].

48 Pour résumer, il semble justifié de dire que le corpus COW-ebay convient bien pour créer des matériaux d'apprentissage. Il faudrait tester à l'aide d'une expérimentation s'il est également pertinent de donner aux apprenants un accès direct à ce corpus via l'interface de NoSketch Engine, qui risque d'être éprouvée comme trop sobre par ce type de public.

Conclusion

49 Pour finir, rappelons la question centrale de cet article : dans quelle mesure les corpus web peuvent-ils soutenir l'enseignement-apprentissage d'une langue étrangère ? On peut constater que les trois corpus présentés (dont un corpus web et un corpus réflexif créé à partir d'un corpus web) ont tous des avantages et des inconvénients en vue des besoins associés au projet *Kleinanzeigen*. On peut apprécier la multimodalité du Schweizer Textkorpus, mais regretter la difficulté de faire des requêtes menant à des occurrences pertinentes. À la grande taille de Araneum Germanicum Maius et l'interface conviviale de Sketch Engine, on peut opposer le côté trop généraliste du contenu des données. La grande pertinence des contenus de Cow-ebay est secondée par des caractéristiques moins désirables, à savoir l'interface très technique et le nombre réduit de fonctionnalités de NoSketch Engine.

50 Selon Boulton (2012), d'un point de vue pédagogique, le choix d'outils de corpus et de ressources doit avoir un fondement pragmatique.

> (…) language learners (…) not need to be as scrupulous in their requirements as researchers: the decision should be pedagogically driven rather than based on non-pertinent research criteria (…). If an approach or technique is of benefit to the learners and teachers concerned, it should not be ruled out automatically.
> [(…) les apprenants d'une langue (…) n'ont pas besoin d'être aussi scrupuleux dans leurs exigences que les chercheurs : une décision devrait être prise en fonction de critères pédagogiques plutôt que de critères de recherche non pertinents (…). Si une approche ou une technique profite aux apprenants et aux enseignants concernés, elle ne devrait pas systématiquement être exclue.]

51 Pour des apprenants de LE, il n'est pas forcément nécessaire ni utile d'accéder à une grande quantité de données. Selon le projet d'apprentissage, la pertinence et la fiabilité des formes et formules comptent davantage pour eux que la fiabilité des étiquettes renseignant sur les classes de mots. Certains projets, mettant l'accent sur la conceptualisation, nécessitent toutefois une précision dans le domaine de l'annotation. Les sous-corpus (semi-)réflexifs extraits de corpus web peuvent être de très bonnes ressources pour les apprenants de LE. Certes, on peut leur reprocher un respect trop approximatif des normes linguistiques, qu'il convient de définir en fonction du genre textuel choisi. Néanmoins, il paraît nécessaire de se demander comment et a quoi on veut former les apprenants : doivent-ils tout d'abord de maîtriser le code linguistique ? Ou bien, la priorité est-elle donnée à l'enseignement-apprentissage du patrimoine culturel de la société de la langue étrangère enseignée ? Cherche-t-on principalement à mobiliser leur volonté d'implication dans le travail d'apprentissage en les faisant adhérer à des projets individuels ou collaboratifs ? Ou bien, voudrait-on offrir aux apprenants des contenus et des méthodes près du nerf du temps, du point de vue des contenus et des méthodes employées ? En fonction des choix opérés et des priorités

affichées, la personne qui conçoit le scénario pédagogique optera pour l'un ou l'autre type d'outil et de ressource, qui sont en phase avec les intentions et besoins identifiés.

BIBLIOGRAPHY

Aston G. (2002). « The Learner as Corpus Designer », in *Teaching and Learning by Doing Corpus Analysis*, édité par Bernhard Kettemann et Georg Marko, 9-26. Amsterdam, New York : Rodopi B.V.

Barbaresi A. (2016). « Efficient construction of metadata-enhanced web corpora », in *Proceedings of the 10th Web as Corpus Workshop*, 7-16. Berlin : Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-2602.

Benko V. (2014). « Aranea: Yet Another Family of (Comparable) Web Corpora », in *Text, Speech and Dialogue*, édité par Petr Sojka, Aleš Horák, Ivan Kopeček, et Karel Pala, 8655 : 247-256. Cham : Springer International Publishing. https://doi.org/10.1007/978-3-319-10816-2_31.

———. (2016). « Two Years of Aranea: Increasing Counts and Tuning the Pipeline ». *LREC*, 4245-48.

Bernardini S. (2000). « Systematising serendipity: Proposals for concordancing large corpora with language learners », in *Rethinking language pedagogy from a corpus perspective*, édité par Lou Burnard et Tony McEnery, 225-234. Frankfurt am Main, Berlin, Bern, Bruxelles, New York, Oxford, Wien : Peter Lang.

Bickel H., Gasser M., Häcki Buhofer A., Hofer L. et Schön C. (2009). « Schweizer Text Korpus – Theoretische Grundlagen, Korpusdesign und Abfragemöglichkeiten ». *Linguistik Online* 39 (3). https://bop.unibe.ch/linguistik-online/article/view/474.

Boulton A. (2011). « Language awareness and medium-term benefits of corpus consultation », in *New trends in computer-assisted language learning: Working together*, édité par A. Gimeno Sanz, 39-46. Madrid.

———. (2012). *What Data for Data-Driven Learning?* European Association for Computer-Assisted Language Learning (EUROCALL). https://eric.ed.gov/?id=ED544438.

Boulton A. et Landure C. (2016). « Using Corpora in Language Teaching, Learning and Use ». *Recherche et pratiques pédagogiques en langues de spécialité - Cahiers de l APLIUT*, 35 (2). https://doi.org/10.4000/apliut.5433.

Braun S. (2005). « *From Pedagogically Relevant Corpora to Authentic Language Learning Contents* ». *ReCALL* 17 (1) : 47-64. https://doi.org/10.1017/S0958344005000510.

Bredel U. et Maaß C. (2016). *Leichte Sprache: theoretische Grundlagen, Orientierung für die Praxis*. Sprache im Blick. Berlin : Dudenverlag.

« British National Corpus (BNC) ». s. d. Consulté le 20 octobre 2019. https://www.english-corpora.org/bnc/.

Chambers A. (2005). « Integrating Corpus Consultation in Language Studies ». *Language Learning & Technology* 9 (2) : 111-125.

Charles M. (2018). « Using do-it-yourself corpora in EAP: A tailor-made resource for teachers and students ». *Journal of Teaching English for Specific and Academic Purposes*, octobre, 217. https://doi.org/10.22190/JTESAP1802217C.

Cheng W., Warren M. et Xun-feng X. (2003). « The Language Learner as Language Researcher: Putting Corpus Linguistics on the Timetable ». *System* 31 (2) : 173-186. https://doi.org/10.1016/S0346-251X(03)00019-8.

Crosthwaite P., éd. (2019). *Data-driven learning for the next generation: corpora and DDL for pre-tertiary learners*. London, New York, NY : Routledge.

Diehl E., Christen H., Leuenberger S., Pelvat, et Studer T. (2000). *Grammatikunterricht, alles für der Katz? Untersuchungen zum Zweitsprachenerwerb Deutsch*. Reihe Germanistische Linguistik 220. Tübingen : Niemeyer.

Dodd B. (1997). « Exploiting a Corpus of Written German for Advanced Language Learning ». In *Teaching and Language Corpora*, édité par Anne Wichmann, Steven Fligelstone, Tony McEnery, et Gerry Knowles, 131-145. London : Longman.

Elmiger D. (2015). « Les corpus Feuille fédérale / Bundesblatt / Foglio fédérale. V. 1.2 ». Université de Genève. https://archive-ouverte.unige.ch/unige:80593.

Fletcher William H. (2004). « Facilitating the Compilation and Dissemination of Ad-Hoc Web Corpora ». In *Studies in Corpus Linguistics*, édité par Guy Aston, Silvia Bernardini, et Dominic Stewart, 17 : 273-300. Amsterdam : John Benjamins Publishing Company. https://doi.org/10.1075/scl.17.21fle.

Gatto M. (2011). « The 'Body' and the 'Web': The Web as Corpus Ten Years On ». *ICAME Journal*, 35 : 35-58.

Godwin-Jones R. (2017). « Data-Informed Language Learning ». *Language Learning & Technology* 21 (3) : 9-27.

Goldhahn D.k, Eckart T. et Quasthoff U. (2012). « Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages », in *Proceedings of the 8th International Conference on Language Resources and Evaluation*, édité par Nicoletta Calzolari, 759-765. Istanbul : European Language Resources Association (ELRA).

Heiden S., Magué J.P. et Pincemin B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement », in *Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, 2(3) :1021-1032. Edizioni Universitarie di Lettere Economia Diritto. https://halshs.archives-ouvertes.fr/halshs-00549779/document.

Kilgarriff A., Rychlý P. et Pomikálek J. s. d. *Sketch Engine | language corpus management and query system*. Lexical Computing. Consulté le 31 juillet 2018. https://www.sketchengine.eu/.

Luo Q. et Zhou J. (2017). « Data-driven Learning in Second Language Writing Class: A Survey of Empirical Studies ». *International Journal of Emerging Technologies in Learning (iJET)* 12 (03) : 182. https://doi.org/10.3991/ijet.v12i03.6523.

Mayaffre D. (2002). « Les corpus réflexifs : entre architextualité et hypertextualité ». *Corpus* 1 (novembre). http://journals.openedition.org/corpus/11.

McEnery T., McEnery A., Xiao R. et Tono Y. (2006). *Corpus-Based Language Studies: An Advanced Resource Book*. London ; New York : Taylor & Francis.

NoSketchEngine Concordance. s. d. Lexical Computing. Consulté le 24 avril 2014. http://nl.ijs.si/noske/wacs.cgi/first_form.

Olston C. et Najork M. (2010). « Web Crawling ». *Foundations and Trends® in Information Retrieval* 4 (3) : 175-246. https://doi.org/10.1561/1500000017.

Polezzi L. (1994). « Concordancers in the Design and Implementation of Foreign Language Courses ». *Computers & Education* 23 (1-2) : 89-96. https://doi.org/10.1016/0360-1315(94)90036-1.

« R: The R Project for Statistical Computing ». s. d. Consulté le 31 juillet 2018. https://www.r-project.org/.

Rastier F. (2004). « Enjeux épistémologiques de la linguistique de corpus ». *Texto !*, 2004, sect. Dits et inédits. http://www.revue-texto.net/1996-2007/Inedits/Rastier/Rastier_Enjeux.html.

Schaeffer-Lacroix E. (2009). « Corpus numériques et production écrite en langue étrangère. Une recherche avec des apprenants d'allemand ». Thèse de doctorat, Université de la Sorbonne nouvelle - Paris III. https://tel.archives-ouvertes.fr/tel-00439095/document.

———. (2015). « Impact de discussions métalinguistiques sur l'apprentissage de la production écrite en allemand, langue étrangère ». *Linx. Revue des linguistes de l'université Paris X Nanterre* 72 (septembre) : 233-237. https://doi.org/10.4000/linx.1676.

———. (2016). « Talking about German Verb Particles Identified in Concordance Lines – from Spontaneous to Expert-like Metatalk ». *Language Awareness* 25 (1-2) : 127-143. https://doi.org/10.1080/09658416.2015.1122023.

Schäfer R. (2017). « Accurate and Efficient General-Purpose Boilerplate Detection for Crawled Web Corpora ». *Language Resources and Evaluation* 51 (3) : 873-889. https://doi.org/10.1007/s10579-016-9359-2.

Schäfer R. et Bildhauer F. (2013). « Web Corpus Construction ». *Synthesis Lectures on Human Language Technologies* 6 (4) : 1-145. https://doi.org/10.2200/S00508ED1V01Y201305HLT022.

Screencast-O-Matic. s. d. Big Nerd Software. Consulté le 31 juillet 2018. https://screencast-o-matic.com.

Sharoff S. (2005). « Open-Source Corpora: Using the Net to Fish for Linguistic Data ». *International Journal of Corpus Linguistics* 11 : 435-462.

———. (2018). « Functional Text Dimensions for the Annotation of Web Corpora ». *Corpora* 13 (1) : 65-95. https://doi.org/10.3366/cor.2018.0136.

Tono Y., Satake Y. et Miura A. (2014). « The Effects of Using Corpora on Revision Tasks in L2 Writing with Coded Error Feedback ». *ReCALL* 26 (2) : 147-162. https://doi.org/10.1017/S095834401400007X.

Volk M. (2002). « Using the Web as Corpus for Linguistic Research », in *Ähendusepüüdja. Catcher of the Meaning. A Festschrift for Professor Haldur Õim*, édité par Renate Pajusalu et Tiit Hennoste, 1-10. Tartu : University of Tartu. https://doi.org/info:doi/10.5167/uzh-20339.

Wisniewski K., Schöne K., Nicolas L., Vettori C., Boyd A., Meurers D., Abel A. et Hana J. (2013). « MERLIN: An Online Trilingual Learner Corpus Empirically Grounding the European Reference Levels in Authentic Learner Data ». Limena : Libreriauniversitaria.it, 1-5. https://bia.unibz.it/handle/10863/8846.

NOTES

1. La terminologie se réfère aux usages de Sketch Engine (Kilgarriff, Rychlý, et Pomikálek s. d.) : un *token* est soit un mot soit un signe ; le terme de *word* correspond à une forme lexicale. Je propose de le traduire ici par le terme de « mot ».
2. Résultat correspondant à la requête restreinte.
3. Un corpus équilibré vise à contenir une quantité comparable pour plusieurs types de discours et/ou de genres textuels jugés représentatifs pour une langue donnée (McEnery et al. 2006, 16).

ABSTRACTS

This article focuses on the potential of web corpora for foreign language teaching and learning. Web corpora are big datasets retrieved from the Internet. They are created in a largely automated way, which goes along with characteristics that may confuse researchers specialising in computer-assisted language learning. However, the following claims can convince even applied linguists: web corpora contain very large amounts of data that support, with the help of text statistics tools, the observation of language features which may be hidden in other forms of text representation. The production method of such resources is both efficient and inexpensive. Moreover, they can be considered as the mirror of a perpetually changing society, organized as networks. In this paper, the terms "pedagogical corpus", "reflexive corpus" and "web corpus" will be defined, and the contribution of the latter two for foreign language teaching and learning will be evaluated. I will conclude by presenting elements of a learning scenario dedicated to writing classified ads for selling, giving or exchanging an item. Designed for a secondary school audience in the fourth year of German language learning, this scenario provides the framework for assessing the potential of web corpora in a given pedagogical context.

Cet article s'intéresse au potentiel des corpus web pour l'enseignement-apprentissage des langues étrangères. Les corpus web sont de très grands ensembles textuels provenant d'Internet. Leur constitution est largement automatisée, ce qui entraîne certaines caractéristiques qui peuvent laisser perplexes les spécialistes de l'apprentissage des langues médiées par les technologies (ALMT). Toutefois, les arguments suivants en leur faveur peuvent convaincre non seulement les linguistes, mais aussi les didacticiens : les corpus web contiennent des quantités de données très importantes permettant d'observer, à l'aide d'outils de statistique textuelle, certaines caractéristiques de la langue en usage, qui sont moins visibles dans d'autres formes de représentation textuelle. Le mode de production de telles ressources est rationnel et peu coûteux. De plus, il semble légitime de les considérer comme le reflet d'une société en perpétuel changement, ayant tendance à s'organiser en réseaux. Dans cet article, je définirai les termes de « corpus pédagogique », « corpus réflexif » et « corpus web » et je comparerai les apports des deux derniers pour l'enseignement-apprentissage des langues étrangères. Je terminerai par la présentation d'éléments d'un scénario dédié à l'écriture d'une petite annonce destinée à vendre, donner ou échanger un objet. Conçu pour un public de l'enseignement secondaire en quatrième année d'apprentissage de l'allemand, ce scénario fournit le cadre pour l'évaluation du potentiel des corpus web dans un contexte pédagogique déterminé.

INDEX

Mots-clés: ALMT, corpus web, production écrite, allemand
Keywords: CALL, web corpora, writing, German

AUTHOR

EVA SCHAEFFER-LACROIX

Inspé de l'académie de Paris (école interne de Sorbonne Université)