Corpus numériques : Critères pour l'enseignement des langues

Cristelle Cavalla

Cavalla Cristelle, 2024, Corpus numériques : critères pour l'enseignement des langues. Revue HDR, 75-92.

Corpus numériques : Critères pour l'enseignement des langues

Cristelle Cavalla
Université Sorbonne Nouvelle
DILTEC EA 2288

Résumé

Dans cet article, il est question de critères pour le choix de corpus numériques à introduire en classe de langue pour l'enseignement et l'apprentissage. Pour ce faire, il sera question de voir ce qu'est un corpus numérique tel que nous l'entendons ici, puis nous aborderons la question de leur insertion en classe de langue : pourquoi et comment, et enfin, nous proposerons une grille de critères de choix. Une telle réflexion arrive après plusieurs années de réflexions sur l'introduction des corpus numériques en classe de langue. Il est nourri notamment de constats faits en cours avec des étudiants en formation initiale et des enseignants en formation continue pour l'enseignement des langues (essentiellement en master DDL à l'USN) et en formation d'apprentissage de la langue (en Français Langue Étrangère). L'article est donc un récapitulatif de ces constats sous forme de réflexions pour l'aide au choix de ces outils en fonction des objectifs pédagogiques de chacun.

Mots clés

Corpus numériques, classe de langue, enseignement, apprentissage, langue étrangère, formation d'enseignant, formation initiale, formation continue, critères, grille.

Introduction

Dans cet article, nous souhaitons proposer des critères de choix des corpus numériques – écrits et oraux – pour les utiliser en classe de langue. Pour ce faire, notre réflexion est inspirée de Sinclair (Sinclair, 2004a) et nous commencerons par rappeler ce qu'est un corpus numérique, puis pourquoi et comment intégrer les corpus en classe de

langue et enfin nous aborderons les critères de choix des corpus. Nous verrons que les objectifs d'enseignement et d'apprentissage et les choix pédagogiques de l'enseignant sont à prendre très au sérieux pour l'utilisation de ces outils.

Nous allons d'abord expliquer rapidement ce qu'est un corpus numérique sur la base de la linguistique de corpus pour l'enseignement des langues : son contenu, sa fabrication, sa cohérence typologique et linguistique (Rastier, 2002 ; Sinclair, 199§). Nous aborderons ensuite la question du pourquoi, en d'autres termes, ce que les corpus apportent à la classe de langue. Enfin, nous verrons l'intégration pédagogique des corpus en classe de langue (Boulton et Tyne, 2014), c'est-à-dire comment les intégrer dans la classe de langue. L'article sera jalonné d'exemples de corpus numériques oraux et écrits, pour le français uniquement, accessibles et gratuits en ligne pour tous. Cette présentation aboutira à l'extraction de critères de choix des corpus numériques en fonction des objectifs d'enseignement que chacun veut mettre en place dans sa classe. Ces critères seront formulés en prenant en compte les éléments évoqués précédemment afin de toucher au maximum les contextes variés que nous rencontrons tous, sachant qu'il est difficile de tout envisager.

Avant de commencer, j'évoque cette phrase de Michel Serres, énoncée lors de conférences (pour la plupart en ligne) et présente dans son ouvrage « La petite Poucette » (Serres, 2012), qui nous a inspiré pour cette réflexion : « Les nouvelles technologies nous ont condamnés à devenir intelligents ! ». Ce constat optimiste de Michel Serres est à mettre à l'honneur ici, semble-t-il, car nous sommes dans un cadre de nouvelles technologies (Grosbois, 2012) et nous devons être créatifs à leur encontre. Pourquoi ? Parce que ces technologies nous donnent déjà tout, comme le dit Michel Serres, nous avons un grand nombre de savoirs devant nous et il faut désormais penser à comment utiliser, répertorier, ranger, classer cette masse de données. Elles ne sont plus nouvelles en tant que « technologies » – ?même si des avancées techniques se font régulièrement – mais elles le sont en tant qu'éléments à utiliser et à didactiser. Nous devons donc faire preuve d'imagination. Pour un enseignant de langue, l'imagination se déploie autour de l'insertion d'outils nouveaux qui peuvent aider à développer l'enseignement et l'apprentissage dans une classe de langue ; c'est ici

que ce situe la nouveauté. Le constat de Boulton en 2008 – déjà plus de 10 ans – de l'absence de ces corpus dans les classes de langue n'a guère évolué. Notre propos se situe donc bien dans le cadre de l'imagination et de la créativité.

Cet article suivra alors le plan évoqué ci-dessus, selon trois entrées : 1/ Qu'est-ce qu'un corpus pour l'enseignement/apprentissage des langues ? 2/ Pourquoi et comment intégrer les corpus en classe de langue ? 3/ Comment choisir les corpus pour la classe de langue ? L'objectif étant l'élaboration d'une grille de critères de choix de corpus numérique pour la classe de langue.

1. Qu'est-ce qu'un corpus ?

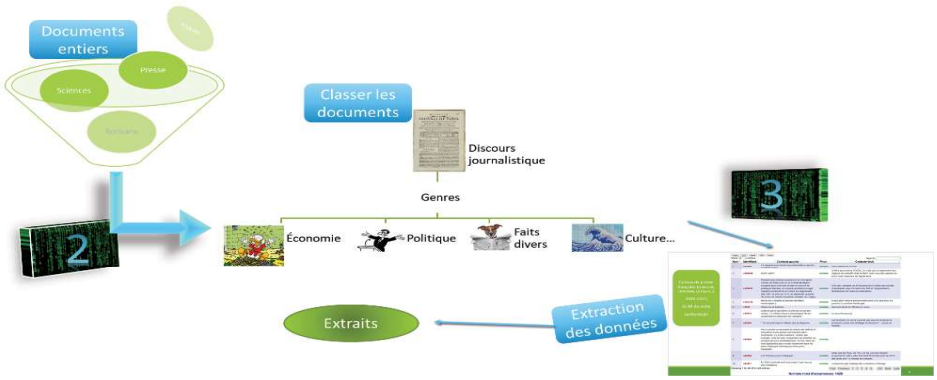
Un corpus numérique tel que nous l'entendons, rassemble des textes auxquels l'utilisateur n'a accès qu'à des extraits. Ces textes sont étiquetés (ce qui permet le repérage par la machine) en fonction de critères définis en amont par les concepteurs selon leurs besoins. Ainsi, les informaticiens et les linguistes (voire les informaticiens-linguistes : les talistes, spécialistes du TAL, le Traitement Automatique des Langues) – souvent concepteurs de ces outils – ne pensent pas toujours à l'utilisation didactique qui peut en être faite. Il est alors de notre ressort d'imaginer comment utiliser ces outils, voire d'en détourner leur utilisation première car nous n'avons pas les mêmes objectifs d'utilisation que les informaticiens et les linguistes.

Les textes rassemblés dans un corpus, ont des éléments en commun, et peuvent présenter une certaine homogénéité discursive quand il s'agit de textes d'un même discours, d'un même domaine de spécialité (par exemple un corpus de presse écrite). En revanche, les corpus davantage généralistes peuvent être très hétérogènes : textes de presse (ex : corpus Lexicoscope, Corpus Français de Leipzig), de littérature (ex : corpus Lexicoscope), de blogs (ex : Corpus Français de Leipzig), d'articles scientifiques (ex. corpus Scientext, corpus Lexicoscope) ou encore sont issus de blogs du web (discours généraliste ; ex. Corpus Français de Leipzig). Enfin, ces textes rassemblés, donc ce corpus, peuvent être rangés/classés de façon à ce que les utilisateurs puissent les interroger en ligne pour diverses études et applications. C'est ici que les didacticiens peuvent intervenir afin de

pouvoir faire des requêtes qui servent à l'enseignement et l'apprentissage des langues. La définition du « corpus » de Sinclair résume bien cela :

Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon au langage [TDA][1]. La dimension « d'échantillon de langage » (« a collection of piece of language ») est importante pour l'enseignement des langues car il faut bien avoir en tête qu'il s'agit d'un regroupement de textes, mais pas de tous les textes possibles écrit dans la langue. Il s'agit d'un échantillon c'est-à-dire d'une partie seulement de la langue et donc le corpus ne peut pas nous donner toutes les acceptions possibles dans l'ensemble des occurrences de la langue. Ainsi, si on interroge un corpus et qu'il ne nous donne pas d'exemple, ceci ne signifie pas que des exemples n'existent pas, ceci signifie que dans ce corpus, des exemples n'existent pas mais qu'on pourrait peut-être en trouver dans d'autres corpus.

Schéma 1 : Élaboration d'un corpus

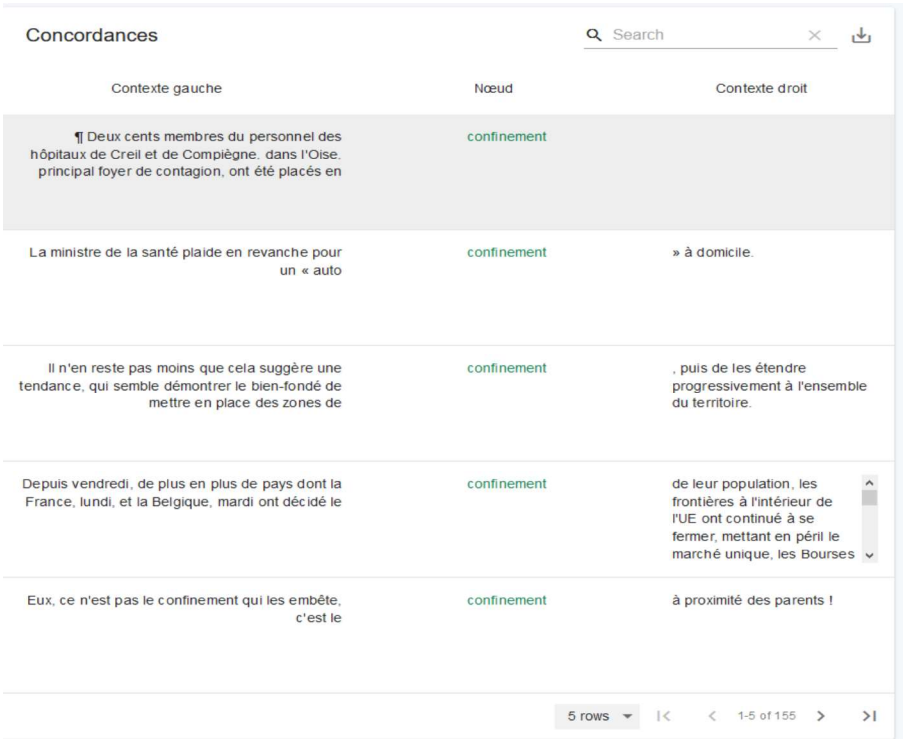


Le schéma 1 permet de visualiser très schématiquement et rapidement, le processus de fabrication d'un corpus. D'abord les concepteurs sélectionnent des textes, puis ils les classent selon des règles linguistiques décidées en amont : par exemple, dans le discours

[1] A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. (Sinclair, 1996).

journalistique se trouve le genre « politique » ou « faits divers ». Enfin, dans le point 3 on voit que tout est rangé et c'est ici que les informaticiens réfléchissent à l'extraction : que veut-on extraire ? à l'aide de quelle requête et que veut-on faire avec ces textes rassemblés et classés ? Par exemple, nous pouvons formuler la requête d'extraction de toutes les phrases contenant le mot « confinement » dans un journal en particulier et durant une période précise ; ceci donne l'exemple suivant extrait du Lexicoscope[2] :

Copie d'écran 1 : Extraction du mot « confinement » dans le Lexicoscope



Cette Copie d'écran 1 donne à voir des phrases contenant le mot « confinement ». Plusieurs informations en amont ont permis cette extraction et surtout plusieurs choix :

[2] URL : http://phraseotext.univ-grenoble-alpes.fr/lexicoscope_2.0/analytics

- Le type de discours : journalistique, le journal Le Monde
- La date des articles : mars-juillet 2020

D'autres informations sont importantes à connaitre pour les futures analyses : le nombre de mots au total (prêt de 60 millions) et le nombre d'occurrences du mot « confinement » (environ 6500). D'autres informations viennent éclairer l'utilisation du mot comme la dispersion ou la spécificité au plan statistique ou le contexte élargit au plan discursif. Nous n'entrerons pas ici dans ces éléments pourtant forts pertinents.

Dans cet exemple, la linguistique de corpus révèle l'un de ses potentiels pour l'enseignement des langues : une grande masse de données. Pour y avoir accès il faut alors que les informations de contenu, de cohérence typologique et linguistique soient connues des utilisateurs. C'est ici que la collaboration entre futurs utilisateurs est essentielle. De ce fait, les linguistes, les talistes et les didacticiens s'associent afin que l'outil final soit approprié aux attentes de chacun. Fabriquer un corpus numérique – sont exclus ici les corpus élaborés par des apprenants – pour l'enseignement des langues, serait peine perdue si les concepteurs ne tenaient pas compte des attentes des didacticiens utilisateurs. Ainsi, que ce soit pour le contenu, pour la fabrication et le classement des données ainsi que pour leur interrogation, dans les trois étapes, chacune des compétences des trois disciplines sont convoquées. Ensuite, les didacticiens peuvent réfléchir à l'intégration de ces outils en classe de langue.

2. Pourquoi et comment intégrer les corpus en classe de langue ?

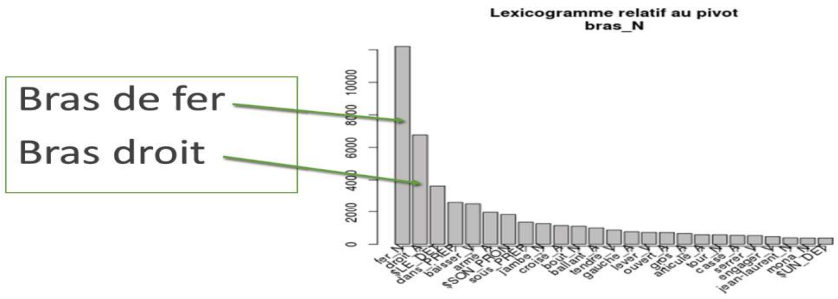
2.1. Pourquoi les corpus en classe de langue ?

Les corpus numériques représentent de grands échantillons de langue contenant des milliers d'exemples (Sinclair, 1996). Le Corpus français de Leipzig[3] contient environ 1,5 milliard de mots et le Lexicoscope environ 790 millions de mots. De tels outils sont des mannes considérables pour l'enseignement et l'apprentissage d'une langue. Les acteurs de la classe de langue peuvent avoir accès à ces exemples et donc peuvent faire de nombreuses recherches et analyses.

[3] URL: https://corpora.uni-leipzig.de/en?corpusId=fra_mixed_2012

Un tel échantillon est précieux pour l'enseignement et l'apprentissage puisqu'il est admis que la contextualisation multiple aide à a mémorisation des éléments langagiers.
En outre, les corpus révèlent des tournures ou même des utilisations de mots que même le locuteur natif découvre parfois. Ainsi, l'intuition des natifs peut être mise à l'épreuve face à des exemples auxquels il ne s'attend pas. Ainsi, avons-nous découvert que la locution « par exemple » est très fréquente dans le discours scientifique tandis qu'elle n'apparait que très rarement dans le discours journalistique (Tran, 2014). Il apparait aussi que de nombreux mots isolés apparaissent plus souvent dans des expressions plus ou moins figées que de façon isolée avec leur sens premier. Tel le mot « bras » dans un corpus journalistique, apparait essentiellement dans les expressions du graphique 1 :

Graphique 1 : occurrences de « bras », Lexicoscope, Presse, 112 millions de mots



Ici on se rend compte que le contexte, et donc le genre de discours, est primordial. Si nous avions accès à un corpus de médecine, le mot « bras » apparaitrait certainement dans d'autres types de phrases et d'expressions.
Enfin, nombre de ces corpus sont en ligne accessible à tous ; ce qui est un avantage non négligeable pour l'enseignement et l'apprentissage des langues de par le monde. Outre le Corpus Français de Leipzig et le Lexicoscope déjà mentionnés pour l'écrit – il en existe d'autres pour l'écrit tel que Sketch Engine – on trouve des corpus oraux tels que CLAPI FLE[4], IPFC[5], ESLO[6], CFPP2000[7]. Ainsi, tant pour l'écrit que

[4] Corpus de Langue Parlée en Interaction, URL : http://clapi.ish-lyon.cnrs.fr/FLE/accueil.php

pour l'oral, les corpus numériques donnent à découvrir des phénomènes langagiers et leur abord de plus en plus simple permet une autonomie dans la recherche et l'apprentissage.

2.2. Comment intégrer les corpus en classe de langue ?

Nous venons de voir pourquoi intégrer les corpus en classe de langue, voyons désormais comment les intégrer en classe de langue. Une fois le corpus choisi par l'enseignant en fonction du type de discours dont il a besoin (presse, science…), du type de corpus : écrit ou oral, et d'autres critères que nous évoquerons plus tard, il est alors question de la didactisation de l'outil pour les apprenants.

La transposition didactique des corpus en classe de langue (Boulton et Tyne 2014) est une tâche pour le didacticien. Puisque nous connaissons les apports que ces outils permettent d'intégrer à la classe de langue, reste à penser comment les intégrer. Pour répondre à cela, nous présenterons des approches didactiques susceptibles d'aider à intégrer ces outils en classe de langue à la fois directement dans la classe et hors les murs – en autonomie pour les apprenants – et voir quel type d'enseignement et d'apprentissage il est possible de développer (Cavalla, 2018a) selon les contextes et les objectifs d'enseignement et d'apprentissage. Chercher un mot dans un corpus peut se faire de plusieurs façons, ou bien chercher ce que le corpus peut nous montrer est un autre type de requête à envisager (Cavalla, 2019 ; Kübler, 2014).
L'un des critères de choix des corpus est sa taille : grand corpus ou petit corpus. Chambers et al (20011) optaient pour un petit corpus, ce qui est un choix judicieux pour une entrée de type DDL - Data Driven Learning - (Sinclair (ed.) 2004b) ou pour travailler de façon inductive et quantitative. En effet, l'entrée par DDL permet, par exemple, de voir le contenu du corpus et notamment des éléments fréquents quel que soit le type de corpus (les verbes ou les adverbes par exemple). Il s'agit donc de faire travailler les apprenants sur des éléments courants, que le

[5] Inter-phonologie du français contemporain, URL : http://cblle.tufs.ac.jp/ipfc/index.php?id=83
[6] Enquêtes SocioLinguistiques à Orléans, URL : http://eslo.huma-num.fr/
[7] Corpus de Français Parlé Parisien des années 2000, URL : http://cfpp2000.univ-paris3.fr/Corpus.html

corpus donne à voir facilement car ces éléments seront présents dans tous les types de discours. En revanche, si l'enseignant veut amener les apprenants à découvrir des éléments spécifiques (un mot en particulier ou un élément figé spécifique), parfois peu présents dans certains corpus mais très fréquents dans certains types de discours, alors il faudra un grand corpus voire, un corpus spécialisé. Quand on envisage ce dernier type de recherche, il s'agit d'une entrée dite « corpus-based » (Léon, 2008), c'est-à-dire davantage déductive et qualitative. Afin de trouver des éléments singuliers, il convient que la taille du corpus soit importante. Cela se vérifie avec les formes figées (Tutin et Kraif, 2017) par exemple. Certaines vont être présentes dans des corpus dits « généraux » tels ceux qui rassemblent des textes de presse (Lexicoscope : « le grand chassé-croisé (des vacances) », « un monstre sacré du cinéma ») tandis que d'autres seront spécialisées et apparaitront principalement dans des discours scientifiques (Scienquest : « émettre l'hypothèse », « par exemple », « des résultats pertinents ») ou professionnels (Cuisitex : « écraser l'ail », « râper le fromage », « écaler l'œuf »). La taille du corpus est importante à prendre en compte car ce sont des éléments linguistiques présents dans de nombreux discours spécialisés mais présents une seule fois dans chaque texte. Par exemple, si on prend « émettre une hypothèse », cette collocation (sorte de forme figée) apparait quasiment dans tous les écrits scientifiques, mais seulement une à deux fois par texte donc au final la fréquence d'une telle forme est peu élevée. Pourtant il s'agit d'un élément primordial dans l'écrit scientifique et tout apprenant qui rédige un article scientifique en français sera amené à utiliser cette collocation.

Une fois la taille et le type de corpus choisi, la question pédagogique suivante apparait : les apprenants interrogent-ils directement le corpus en ligne ou pas ? Pour cette question, deux entrées sont encore possibles : soit l'enseignant édite des extractions du corpus et fait travailler les apprenants à partir de ce choix limité. Soit l'enseignant place les apprenants devant le corpus et ils doivent faire les requêtes d'interrogation. La taille du corpus va donc être un choix crucial ici car sur un petit corpus choisi et imprimé par l'enseignant, les apprenants auront peu d'exemples tandis que sur un grand corpus ils risquent d'avoir à trier des centaines d'exemples. Tout dépend de ce qu'on veut leur montrer.

Si on place les apprenants devant un corpus, il est important de savoir que ces outils sont comme des applications sur les téléphones par exemple, ils ne sont pas compliqués à interroger. Désormais on trouve des interfaces simples avec une case dans laquelle on inscrit le mot ou l'expression recherchée (Lexicoscope ou Corpus français de Leipzig). Ce qu'il est important de faire comprendre aux apprenants est l'apport de ces outils : ils permettent d'extraire des exemples précis et pas des sites dans lesquels se trouvent le mot comme dans des moteurs de recherche classiques tels que Google ou Qwant. Toutefois, on peut demander aux apprenants de chercher un exemple sur un moteur de recherche afin de partir de ce qu'ils connaissent (Boulton et Tyne, 2014). Une grande quantité de sites va leur être proposée et il sera difficile (voire impossible) d'extraire des exemples dans chacun d'entre eux. En revanche, une telle recherche fait émerger des questions pour des requêtes plus précises afin d'extraire précisément ce qu'on cherche (Kübler, 2014). C'est donc à ce moment-là que nous pouvons montrer des corpus aux apprenants afin qu'ils cherchent et trouvent le mot ou l'expression qui les intéresse.

C'est alors que le choix pédagogique de l'enseignant va se mettre en place et prendre en compte les questions des apprenants : soit une entrée par mot et on rejoint l'approche corpus based (Leech, 1992) qui s'inscrit dans une entrée dite sémasiologique ; soit une entrée par thème ou sémantique qui s'apparente à l'approche du corpus driven (Firth, 1957) et qu'on nomme aussi onomasiologique. Deux questions didactiques sont alors associées à ces entres : pour l'approche sémasiologique on se demande dans quel contexte le mot est utilisé ; donc on cherche le mot et on étudie les contextes possibles. Pour l'approche onomasiologique, la question est la suivante : de quels mots a-t-on besoin pour dire telle notion ? Voici un exemple de question que les étudiants posent : comment introduire mon sujet ou mes hypothèses ? Les apprenants commencent souvent par des questions de type sémantique (onomasiologique) puis de type lexical (sémasiologique) afin de trouver des exemples du mot qu'ils ont choisi.

Reprenons l'exemple de l'introduction des hypothèses : il existe un corpus et son interface, ScienQuest (Firth, 1957), dans lequel les apprenants peuvent d'abord trouver des réponses à la question « comment introduire ses hypothèses » puis, une fois qu'ils ont vu les

différentes formes possibles, ils peuvent aller chercher des exemples plus précis de la forme qu'ils auront choisie. L'enseignant sert ici de guide dans l'utilisation de l'outil et donne à l'apprenant les clés pour ensuite chercher seul ce dont il a besoin (Cavalla, 2018b).

L'objectif d'une telle démarche est l'autonomie de l'apprenant face à l'outil afin que chacun puisse mener son écriture comme il l'entend. Pour ce faire il parait indispensable de prendre le temps à la fois de la prise de conscience de l'existence d'outils de ce genre et de la maitrise de leur manipulation. Ce temps est finalement assez court depuis quelques années. Ensuite laisser le temps aux apprenants de poser les bonnes questions pour aborder leur propre écriture sans penser qu'il faille impérativement copier les exemples qu'ils extraient des corpus mais qu'ils peuvent s'approprier l'écrit scientifique (dans l'exemple de l'hypothèse) petit à petit et que l'abondance d'exemples les aidera à cela. Enfin, leur donner le temps nécessaire à cette appropriation à la fois du discours visé et de la manipulation de l'outil entraine parfois une sorte d'instrumentalisation (Rabardel, 1999) de l'outil c'est-à-dire qu'ils pourront manipuler l'outil comme ils l'entendent voire en détourner l'utilisation.

De ces constats sur l'intégration des corpus numériques en classe de langue, nous pouvons désormais extraire des critères de choix.

3. Les critères de choix des corpus pour la classe de langue

Comme indiqué en introduction, cette présentation aboutit à des critères de choix des corpus numériques en fonction des objectifs d'enseignement que chacun veut mettre en place dans sa classe. En prenant en compte les éléments évoqués précédemment nous pouvons établir le tableau suivant (Tableau 1) qui peut se décliner sous forme de grille servant de diagnostique pour aider l'enseignant à choisir son corpus.

Tableau 1 : Critères de choix des corpus numériques pour la classe de langue

CRITÈRES CORPUS ÉCRITS	CRITÈRES CORPUS ORAUX
Types de documents : presse, littérature, articles scientifiques…	Types d'enregistrements : audio, vidéo / transcription ou pas Types d'interactions : dialogues / monologues / joués ou pas
Taille : fréquence, représentativité	
Date : synchronie, diachronie	
Variation : région, authenticité	
Scripteurs / Locuteurs : natif ou allophone / âge / sexe / plurilingue ou pas	
Accessibilité : gratuit / payant / inscription	
Type de requête : simple ou pas	
Type d'extractions : concordanciers / statistiques / graphiques	

Nous retrouvons dans le Tableau 1 des critères déjà évoqués comme l'écrit et l'oral ou le type de document (presse écrite ou vidéo) ou encore la taille. L'ordre d'apparition des critères ne parait pas important pour le choix final ; le classement – et donc le choix 1er en oral / écrit – peut même être revu plus tard si finalement c'est un autre critère qui prime. On se rend compte de cela dès l'instant où les critères sont identiques sauf bien sûr pour le type de document puisqu'il serait dommage de choisir un texte écrit pour faire de la compréhension orale (choisir « audio » dans ce cas) même si la lecture à haute voix est intéressante à travailler.

La date et le caractère synchronique ou diachronique du corpus, est un critère important selon l'objectif pédagogique de l'enseignant et les besoins des apprenants. En effet, un document du XIXe siècle – dans le corpus Frantext[8] par exemple – contiendra des éléments linguistiques différents d'un document du XXIe siècle. De façon moins prototypique, un article de journal de 1998 (dans Lextutor[9] par exemple) ne traitera pas les mêmes thèmes, ou si c'est le cas, pas de la même façon, qu'un journal de 2020 (dans Lexicoscope par exemple). Dans tous les cas, la date est importante et peut d'ailleurs servir à comparer des évolutions (lexicales, sémantiques, syntaxiques) comme le font des chercheurs

[8] URL : Frantext
[9] URL : WEB CONCORDANCE FRENCH (lextutor.ca)

autour du corpus ESLO[10] qui comparent des oraux des années 1970 à ceux des années 2010.

La variation est un élément à prendre en compte car selon le contexte, la langue peut être différente du standard parfois attendu. En effet, la variation régionale (diatopique) apparait dans certains corpus comme Varitext[11] pour le français écrit d'Afrique francophone ou IPFC[12] et la francophonie à l'oral. Au plan diastratique (variation selon le groupe social), nous rencontrons de nombreux corpus oraux avec des interactions entre jeunes et moins jeunes par exemple. Ces informations peuvent être croisées avec celles concernant le type de scripteur ou de locuteur qui sont souvent très pertinentes pour l'enseignement. En effet, savoir que le locuteur vient de telle région (de France ou du monde ; niveau diatopique) pourra aider à reconnaitre les accents francophones, et à l'écrit, l'utilisation de telle expression permettra de savoir qu'elle est utilisée dans telle région du monde et pas dans d'autres. En outre, son âge ou sa profession seront des informations pertinentes si on travaille sur des discours spécialisés (écrits ou oraux) ou sur la nature des interactions entre enfants et adultes ou entre adolescents.
Reconnaissons qu'à l'écrit, les informations sur l'âge ou la région du scripteur ne sont pas toujours simples à trouver voire inexistantes quand il s'agit de corpus du web et des blogs ; les auteurs eux-mêmes n'indiquent que rarement de telles informations pourtant parfois précieuses pour la recherche. De petits corpus sont constitués pour des recherches spécifiques mais ils ne sont que rarement accessibles au grand public. Enfin, au plan diaphasique, la situation d'énonciation, et donc la prise en compte du locuteur, est intéressante par exemple pour des discours spécialisés écrits ou oraux. En effet, un scientifique qui écrit un article pour ses pairs le rédigera différemment pour un journal de vulgarisation, ou bien un menuisier ne s'adressera pas de la même façon à un confrère et à ses clients (à l'écrit et à l'oral). Enfin, les critères de genre et de plurilinguisme sont très intéressants pour différentes entrées par exemple littéraires pour le plurilinguisme : savoir

[10] URL : http://eslo.huma-num.fr/
[11] URL : http://syrah.uni-koeln.de/varitext/
[12] URL : https://www.projet-pfc.net/le-projet-pfc-ef/ressources-linguistiques/corpus-thematique/

si tel auteur écrit dans sa L1 ou L2 ou L*n* et étudier en quoi cela peut influencer son écriture.

Reste l'accessibilité de ces corpus. Certains sont payants (Frantext) et d'autres gratuits avec ou non une inscription pour l'accès à toutes les fonctionnalités. L'inscription sert le plus souvent aux chercheurs – créateurs du corpus – à avoir une idée du nombre de connexions et donc de l'utilisation quantitative de l'outil. L'inscription n'est pas liée à des publicités – du moins pour les corpus français que je connais.

Enfin, la question du type de requêtes attendues et possibles en ligne pour l'extraction des données est parfois difficile à résoudre. En effet, les requêtes dépendent de l'utilisation attribuée à l'outil ; un linguiste ne l'utilisera pas de la même façon ni pour les mêmes raisons qu'un apprenant. Ainsi cette question est à gérer en amont soit entre les concepteurs et les utilisateurs si possible, soit les utilisateurs peuvent réfléchir à un détournement de la fonction initiale de l'outils (Rabardel, 1995).De ce fait, le contexte et les objectifs de l'enseignement-apprentissage sont essentiels à prendre en compte à ce niveau de la réflexion afin de faire les choix pertinents pour le type de requêtes à développer.

Le diagnostic possible à faire pour le choix du corpus servirait de vérification des éléments importants à retenir pour l'objectif linguistique d'une séance par exemple et permettrait le choix raisonné d'un corpus parmi une liste. Si quelqu'un pouvait proposer une telle possibilité en ligne avec un menu déroulant permettant le choix de corpus trié en amont, alors je veux bien tester l'outil.

Conclusion

Les corpus numériques sont des ressources à la fois intéressantes et importantes pour l'enseignement des langues. Rappelons que, au plan qualitatif, les enseignants et les apprenants peuvent avoir accès à des documents écrits ou oraux authentiques et au plan quantitatif, le nombre d'exemples peut souvent dépasser l'imagination. Le souci qui pourrait guetter, c'est l'utilisation exclusive des corpus et dans ce cas, nous serions limités non pas par le nombre d'exemples, mais par la taille des extractions et donc ne travailler que sur des extraits courts (de la taille

d'une phrase pour un concordancier de base, ou d'1 à 2 paragraphes dans certains corpus) ; ce qui entrainerait d'autres questions pour l'enseignement d'une langue, questions à traiter ultérieurement. Notons qu'actuellement cette question ne se pose pas puisque très peu d'enseignants connaissent ces outils.

Pour conclure, nous avons répondu aux questions de l'introduction, grâce aux collègues enseignants et chercheurs qui ont développé des outils, qui ont réfléchi à leur utilisation, qui ont voulu leur développement pour un accès libre, gratuit, simple voire pédagogique. La grille proposée contient des éléments primordiaux de notre point de vue et qu'il faudra surement revoir d'ici quelques temps (au plan sémantique peut-être… pourquoi pas). En effet, la vitesse du développement des outils informatiques, et donc des nouvelles possibilités qui nous seront offertes dans un avenir proche, augure l'entrée en lice d'autres critères de sélection dans les années (voire les mois) à venir.

Michel Serres nous « condamnait à devenir intelligents » et donc à innover pour nous servir de toute la technologie à notre portée et j'espère que la réflexion pour l'utilisation des corpus numériques va se poursuivre dans ce sens. De grands progrès peuvent être mentionnés à ce propos :

- Il y a 20 ans on se posait la question de l'insertion ou pas de ces corpus en classe de langue et les enseignants de FLE découvraient les concordanciers ;
- Aujourd'hui on met les apprenants devant les corpus et on les aide à s'approprier ces outils ;
- Laissons pour demain notre imagination nous entrainer vers des détournements de ces outils pour des applications encore plus performantes et pertinentes pour l'enseignement et l'apprentissage des langues.

Cette dernière étape – optimiste – invite à la collaboration entre chercheurs et enseignants de disciplines différentes pour l'élaboration de tels outils. Il faudrait donc penser à davantage rassembler des talistes, des didacticiens (enseignant et formateurs de formateurs) et des linguistes autour de tels projets.

Références bibliographiques

Boulton A. 2008. « Esprit de corpus : Promouvoir l'exploitation de corpus en apprentissage des langues». *Texte et Corpus*, 3, 37-46.

Boulton A. et **Tyne H**. 2014. *Des documents authentiques aux corpus. Démarches pour l'apprentissage des langue*s. Didier.

Cavalla C. 2018a. « Exemple d'enseignement de la phraséologie transdisciplinaire à l'aide de corpus numériques en FLE». La lettre de l'AIRDF, Association internationale de recherche en didactique du français, 43-47.

Cavalla C. 2018b. « Lexique transdisciplinaire et enseignement aux étudiants allophones. In Lexique transversal et formules discursives des sciences humaines» (Agnès Tutin, Marie-Paule Jacques, p. 191-214). ISTE Editions.

Cavalla C. 2019. « Comment former les étudiants de Master FLE à l'utilisation pédagogique des corpus numériques ? ». In Apports et limites des corpus numériques en analyse de discours et didactique des langues de spécialité (J. Goes et al. p. 79-92). Editura Universitaria. http://editurauniversitaria.ucv.ro/apports-et-limites-des-corpus-num

Chambers A., **Farr F**. et **O'Riordan S**. 2011. « Language teachers with corpora in mind : From starting steps to walking tall ». *The Language Learning Journal,* 39:1, 85-104.

Firth J. R. 1957. « Modes of Meaning. In Papers in Linguistics 1934-1951 (p. 190-215). Oxford University Press.

Grosbois M. 2012. « Formation en langues des enseignants et usage des TIC : quelle réalité ? » P*rocedia-Social and Behavioral Sciences*, 34, 79-83.

Kübler N. 2014. « Mettre en œuvre la linguistique de corpus à l'université— Vers une compétence utile pour l'enseignement/apprentissage des langues ? » *Les Cahiers de l'Acedle*, 11(1), 37-77.

Leech G. 1992. « Corpora and Theories of Linguistic Performance». In Corpus Linguistics. Proceedings of Nobel Symposium (J. Svartik, p. 105-122). Mouton de Gruyter.

Léon J. 2008. « Aux sources de la *Corpus Linguistics* » : Firth et la London School». *Langages*, 171(3), 12-33. Cairn.info. https://doi.org/10.3917/lang.171.0012

Rabardel P. 1995. *Les hommes et les technologies, approche cognitive des instruments contemporains*. Armand Colin.

Rabardel P. 1999. « Le langage comme instrument ? Éléments pour une théorie instrumentale élargie». In Y. Clot (Éd.), Avec Vygotski (p. 241-266). Éditions La Dispute.

Rastier F. 2002. « Enjeux épistémologiques de la linguistique de corpus. Texto ! » http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html

Serres M. 2012. *La petite poucette*. Le Pommier.

Sinclair J. M. 2004a. « Corpus and Text—Basic Principles». In M. Wynne (ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, (pp.1–21). Oxford Text Archive. https://web.archive.org/web/20160825230039/

http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter1.htm

Sinclair J. M. (ed.). 2004b. *How to use corpora in language teaching.* Benjamins.

Sinclair J. M. 1996. P*reliminary Recommendations on Corpus Typology* [Scientific]. EAGLES - Expert Advisatory Group on Language Engineering Standards. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Tran T. T. H. 2014. Description de la phraséologie transdisciplinaire des écrits scientifiques et réfle*xions didactiques pour l'enseignement à des étudiants non-natifs. Application aux marqueurs discursifs* [Université Grenoble Alpes]. https://tel.archives-ouvertes.fr/tel-01330952/document

Tutin A. et **Kraif O**. 2017. « Comparing recurring lexico-syntactic trees (RLTs) and ngram techniques for extended phraseology extraction: A corpus-based study on French scientific articles ». Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017) at the European Chapter of the Association for Computational Linguistics Conference (EACL 2017), Stoudsburg (PA), Association for Computational Linguistics, 176-180.