RECHERCHES EN DIDACTIQUE DES LANGUES ET DES CULTURES

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

11-1 | 2014
Notions en questions - Corpus

Mettre en œuvre la linguistique de corpus à l'université

Vers une compétence utile pour l'enseignement/apprentissage des langues ?

Natalie Kübler

OpenEdition Journals

Édition électronique
URL : http://journals.openedition.org/rdlc/1685
DOI : 10.4000/rdlc.1685
ISSN : 1958-5772

Éditeur
ACEDLE

Référence électronique
Natalie Kübler, « Mettre en œuvre la linguistique de corpus à l'université », *Recherches en didactique des langues et des cultures* [En ligne], 11-1 | 2014, mis en ligne le 07 janvier 2014, consulté le 01 mai 2019.
URL : http://journals.openedition.org/rdlc/1685 ; DOI : 10.4000/rdlc.1685

Mettre en œuvre la linguistique de corpus à l'université

Vers une compétence utile pour l'enseignement/apprentissage des langues ?

Natalie Kübler

Quel type d'enseignement ?

1 Aujourd'hui, tout le monde peut constater la place prépondérante et grandissante de l'anglais dans tous les domaines de la production du savoir. Ceci rend nécessaire et stratégique la mise en place au niveau universitaire d'un enseignement pertinent de l'anglais tirant parti des différentes approches épistémologiques de la linguistique. Dans cette intention, le présent article se propose d'envisager la manière dont les enseignants à l'université peuvent utiliser les corpus pour l'enseignement des langues aux spécialistes d'autres disciplines, en mettant celui-ci en parallèle avec les besoins en traduction spécialisée. En effet, les deux activités ont en commun d'être menées par des acteurs différents des apprenants d'une langue seconde tels que peuvent l'être les étudiants. La linguistique de corpus représente ici une approche théorique et méthodologique qui, appliquée à l'apprentissage des langues, a pour objectif d'amener à se poser les bonnes questions pour obtenir les bonnes réponses. Les deux langues évoquées ici seront l'anglais et le français.

Points communs entre enseignants de Lansad[1] et apprentis traducteurs

2 Tout d'abord, comme les enseignants de L2 aux non spécialistes, les apprentis traducteurs sont des professionnels de la langue, sans avoir forcément reçu une éducation approfondie en linguistique. Leur objectif n'est en effet ni d'étudier, ni de décrire, ni d'expliquer les phénomènes linguistiques. Pour les enseignants, l'objectif est de faire apprendre la langue, mais aussi la culture, qu'il s'agisse de culture générale ou

scientifique, aux apprenants, dans un but communicatif. Ils enseignent donc la langue comme outil de communication. En traduction "pragmatique"[2], l'objectif est de faire passer le message d'un auteur d'une langue source à une langue cible, en respectant l'intention de l'auteur, tout en rendant la traduction la plus idiomatique possible. La langue est donc aussi ici un outil de communication dans laquelle le traducteur est un médiateur. Aujourd'hui, dans les approches didactiques centrées sur l'apprenant, on considère aussi l'enseignant de langue comme un médiateur dans l'apprentissage, plutôt que comme quelqu'un qui dispense le savoir du haut de sa chaire. Les objectifs de l'enseignant et du traducteur sont à l'évidence sensiblement différents, mais l'approche du corpus peut être semblable dans bien des cas.

Attentes et besoins en Lansad

3 Déterminer les objectifs de l'enseignement en Lansad reste une question vaste. En effet, les étudiants ont besoin d'améliorer leur niveau dans les cinq domaines de compétence, mais aussi dans des genres et des registres différents, qui tiennent compte des besoins spécifiques à leurs domaines de spécialité. En outre, on attend d'eux qu'ils soient capables d'obtenir des certifications, comme le *Toeic* (*Test of English for International Communication*) par exemple. La nouvelle loi de l'enseignement supérieur exige qu'en Master, les étudiants reçoivent un enseignement d'au moins une langue (l'anglais le plus souvent), valant trois ECTS non compensables, et prévoit que des cours disciplinaires soient dispensés en anglais. Le besoin d'un travail approfondi sur la langue de spécialité devient une nécessité si l'on veut éviter une discrimination sociale des étudiants par la compétence en anglais. En effet, tous les étudiants n'ont pas forcément joui d'un milieu social leur donnant les moyens d'acquérir une bonne compétence en anglais, avec, entre autres, des séjours linguistiques.

4 On attend des enseignants d'anglais qu'ils soient capables d'apprendre aux étudiants à comprendre et à produire dans leurs domaines de spécialité. Or, la formation des enseignants d'anglais ne les prédispose *a priori* pas à connaître l'anglais de spécialité (sauf pour certaines rares formations en anglais de spécialité). Ceux-ci sont plutôt formés à la langue, la linguistique, la civilisation et la littérature. La question de cette formation est un sujet de réflexion qui mérite encore d'être clarifiée. Par ailleurs, on peut s'interroger sur le niveau de compétence des étudiants à partir duquel on peut enseigner en anglais de spécialité. En effet, certains semblent penser qu'il est nécessaire de connaître d'abord les bases d'une langue avant de s'attaquer à un domaine spécialisé. Pour eux, il faudrait pouvoir d'abord enseigner les bases de l'anglais général et ensuite, passer à la langue de spécialité. Dans la réalité, les étudiants qui entrent en première année d'université ont un niveau en majorité très faible en anglais, entre A1 et B1 pour 80 % d'entre eux[3]. En outre, le nombre d'heures par année qui est dédié au Lansad reste très insuffisant. Concrètement, on ne dispose pas d'assez de temps pour enseigner les bases d'une langue générale et passer ensuite à la langue de spécialité. Une solution possible serait donc d'enseigner les bases en s'appuyant dès le début sur le domaine de spécialité. C'est ici qu'interviennent les corpus, spécialisés ou non. Nous nous limiterons ici à une formation aux corpus à l'écrit, tenant compte des besoins spécialisés, en laissant de côté la question de l'oral et des corpus oraux.

Évolution du public étudiant et enseignant

5 Le public, tant dans l'enseignement des langues que dans la traduction spécialisée, a évolué depuis que l'on a commencé à utiliser des corpus au début des années '90. Les utilisateurs d'aujourd'hui ont connu la généralisation du web, l'utilisation de *Google* comme outil de vérification linguistique, le développement des dictionnaires et des bases de données terminologiques en ligne et le développement de la traduction automatique statistique. Par ailleurs, la simple possibilité d'avoir accès à des ordinateurs pour enseigner dans les universités s'est aussi beaucoup développée. On peut donc espérer avoir affaire à des enseignants qui ont l'habitude de travailler sur ordinateur, qui utilisent partiellement Internet pour préparer leurs cours, et qui, parfois, se servent déjà des ordinateurs dans leurs enseignements, ne serait-ce que pour demander aux étudiants de faire des recherches sur Internet en anglais par exemple. Cependant, le type de recherches que l'on peut faire sur corpus diffère beaucoup des recherches que l'on peut effectuer sur Internet ; en effet, le corpus et les outils d'interrogation de corpus permettent des interrogations beaucoup plus spécifiques et structurées, qui tiennent compte des genres textuels et des domaines de spécialité. Par ailleurs, la recherche sur Internet ne permet pas de maîtriser tous les paramètres des textes interrogés, ce que permet un corpus, car on sait ce qu'il contient. Tout cela demande donc une formation spécifique.

Corpus et enseignement

6 L'utilisation de corpus dans l'enseignement des langues ou de la traduction ne constitue plus une approche innovante. Johns (1990), le premier, a introduit le terme *data-driven learning* il y a plus de vingt ans, ce que Boulton et Tyne (2014) suggèrent de transposer en français par *apprentissage sur corpus* ; en traduction, Aston (1999) est un précurseur dans l' utilisation de corpus pour la formation des traducteurs.

7 On a très vite proposé différentes approches tant dans l'utilisation des corpus pour l' enseignement des langues, aussi bien dans le monde anglophone qu'ailleurs (Minugh, 1997 ; Gavioli, 1997 ; Gavioli, 2005 ; Bernardini, 2004), que dans la manière d'introduire la linguistique de corpus auprès des enseignants (Renouf, 1997) et des traducteurs (Bowker & Pearson, 2002). On a aussi très rapidement développé des outils d'enseignement assisté par ordinateur basés sur des corpus (par exemple Johns, 1997 ; Antoniadis *et al.*, 2010), en France et ailleurs. Foucou et Kübler (2000) proposent un concordancier en ligne avec possibilité de lancer des requêtes dans une syntaxe d'expression régulière, mais aussi de générer automatiquement des exercices à partir de concordances et de contextes. Tono (2011) propose même un enseignement basé sur corpus à la télévision japonaise, avec un personnage nommé "Mr. Corpus".

Utilisation restreinte

8 Cependant, malgré toutes les avancées prometteuses et enthousiastes que l'on rencontre dans la littérature des années 1990 et du début des années 2000 et, malgré le développement de ressources et de corpus, on constate que l'utilisation de ceux-ci n'est pas très répandue chez les enseignants (Römer, 2006) ni chez les traducteurs (Kübler, 2011).

9 Zanettin (2002 : 14) mentionne le fait que les utilisateurs trouvent les outils et les logiciels difficile d'utilisation et qu'il s'agit d'une activité dévoreuse en temps. Dix ans plus tard, Tyne (2012 : 114) fait encore le même constat, cette fois-ci en insistant sur le manque de convivialité des outils. On peut aussi signaler que la linguistique de corpus a été longtemps considérée par de nombreux linguistes formels en France comme une discipline inexistante, suivant en cela l'opinion de Chomsky pour qui la linguistique de corpus "n'existe pas" (Aarts, 2001 : 5). En anglistique, on considérait jusqu'à il y a peu qu' il s'agissait uniquement d'outils. Il est possible que l'origine du retard mis à utiliser les corpus en France puisse partiellement s'expliquer ainsi. Boulton (2008) souligne en effet le retard qu'a pris la France dans ce domaine et l'explique d'ailleurs par des raisons culturelles, mais aussi par le fait qu'il n'existe pas assez de recherche sur l'évaluation de l' efficacité de l'apprentissage sur corpus (Boulton, 2008 : 43). Une autre explication peut tenir au fait qu'en France, on a choisi de développer des dictionnaires comme le *Trésor de la langue française*, sans faire appel aux corpus, comme l'a fait le projet *Cobuild* en Grande-Bretagne. En effet, ce projet a permis de développer la recherche sur corpus de manière extensive dans ce pays et dans les pays scandinaves, ce qui n'a pas été le cas en France. Enfin, tout un pan des réflexions sur la formation des enseignants d'anglais en Lansad penche plutôt pour une formation à "double compétence", avec des enseignements disciplinaires en parallèle à l'enseignement de la langue, de manière à ce que ceux-ci se spécialisent dans un domaine particulier en anglais. C'est aussi le cas de nombreuses formations à la traduction, d'ailleurs.

10 Nous pensons que le développement de l'utilisation des corpus passe par une formation approfondie auprès des étudiants en traduction et des enseignants de langue, avec une approche méthodologique qui leur permette de s'approprier n'importe quel domaine de spécialité, sans qu'il soit nécessaire d'acquérir une formation disciplinaire. Cette approche nécessite cependant une certaine formation théorique et méthodologique à la linguistique de corpus, ainsi qu'une pratique intensive des corpus, avant de pouvoir les utiliser dans leurs professions respectives. Cela est rendu possible d'une part par le nombre croissant de corpus accessibles en ligne, et d'autre part, par les grands progrès qui ont été réalisés dans la conception d'outils d'exploitation de corpus.

11 Dans les sections suivantes, nous reprenons la manière dont les corpus ont été abordés dans l'enseignement, puis, nous développerons l'approche que nous mettons en œuvre. Les exemples illustrant notre propos sont issus à la fois d'expériences en Lansad, mais aussi de formation auprès de traducteurs et d'enseignants Lansad.

Enseignement et corpus

12 Les enseignants et les traducteurs représentent un public soit éduqué dans différents modèles théoriques en linguistique, soit non éduqué en linguistique. Il paraît donc nécessaire de les informer et de les former à certains points théoriques qui sous-tendent l 'utilisation de corpus, quel que soit l'objectif final de cette utilisation. Il convient donc de considérer les enseignants comme des utilisateurs novices des corpus, au même titre que les apprenants, avec des objectifs différents, mais qui se rejoignent sur bien des points. Renouf (1997 : 256) propose déjà les bases d'un cours complet d'introduction théorique, méthodologique et pratique à la linguistique de corpus pour les enseignants non natifs de

l'anglais, en fonction des trois catégories de Fligelstone (1993) et auxquelles elle ajoute une catégorie supplémentaire :

- Teaching about (i.e. the principles and theories of corpus linguistics)
- Teaching to exploit (i.e. the practical aspects of corpus study)
- Exploiting to teach (i.e. deriving language-teaching materials from corpora)
- Teaching to establish resources[4]

13 Pour elle, il faut non seulement enseigner les aspects théoriques et les principes fondamentaux de la linguistique de corpus, la manière d'interroger un corpus et celle de l 'exploiter pour créer du matériau pédagogique, mais aussi, tout la démarche de création de corpus, ressource indispensable à l'exploitation des corpus dans l'enseignement.

Par quoi commencer ?

14 La linguistique de corpus, par essence, s'accommode très bien d'une approche empirique de ses fondements théoriques permet d'amener à comprendre. Il s'agira d'emblée de favoriser l'expérimentation sur corpus. Les utilisateurs novices doivent donc être guidés au départ pour apprendre à "lire" un corpus à l'aide des fondements théoriques et méthodologiques qui sous-tendent la recherche en corpus. En cela, nous sommes en accord avec Frankenberg-Garcia (2012 : 479) qui affirme que : *"teachers may need some help to decode the results of their initial corpus searches"*[5]. En effet, la lecture d'une concordance n' est pas forcément limpide dès la première fois ; de même, l'interprétation des statistiques sur les corpus demande une certaine pratique.

15 Gavioli (2005), quant à elle, propose aussi d'enseigner les bases théoriques de la linguistique de corpus afin d'aider les enseignants à les utiliser en classe ou à préparer du matériel pédagogique basé sur l'observation de corpus. Il convient donc d'enseigner, comme l'affirme Frankenberg-Garcia (2012 : 477-479) aux utilisateurs novices à sélectionner le bon corpus, poser les bonnes questions et interpréter correctement les résultats.

Approches corpus

16 La première étape doit donc consister à présenter aux enseignants ce qu'est un corpus et à établir quelques bases théoriques indispensables à l'interprétation des résultats d'une interrogation de corpus. En effet, comme le remarque Tognini-Bonelli (2001), il y a deux approches corpus. On peut mettre en œuvre une approche *corpus-based* dans laquelle on se sert du corpus pour valider une théorie déjà existante. Dans ce cas, le corpus sert de preuve validant la théorie. De même, dans l'enseignement, on peut utiliser le corpus pour exemplifier une structure (Aston, 1997 : 53). L'enseignant reste alors au centre du processus d'apprentissage et délivre aux apprenants une connaissance à l'aide des exemples du corpus. La deuxième approche, *corpus-driven*, place en amont la découverte de régularités dans le corpus qui suscitent ensuite des élaborations théoriques. De même, dans l'enseignement, l'enseignant jouera plutôt le rôle de facilitateur ou médiateur en permettant aux apprenants de tirer leurs propres conclusions lors de l'exploration du corpus (Landure & Boulton, 2010 ; Tyne, 2012). Le choix entre l'une ou l'autre approche dépend de l'activité que l'on veut mettre en place dans la classe. Dans les deux cas de toute manière, il faudra que les enseignants connaissent la méthodologie de base des outils d'interrogation de corpus. Et dans l'approche *corpus-driven*, il est donc

indispensable d'enseigner les bases de l'approche théorique et méthodologique de la linguistique de corpus.

Découvrir le corpus

17 On peut commencer par donner une définition de ce qu'est un corpus, puis passer immédiatement à des exercices pratiques qui permettront à la fois d'exemplifier ce qu'est un corpus, mais aussi d'introduire les principes de base de la linguistique de corpus. Frankenberg-Garcia (2012) montre comment faire découvrir différents corpus en ligne à des enseignants et les aider à comprendre ce que contient un corpus, quelles sont les informations indispensables à connaître sur celui-ci, et quels principes, très différents de ce que l'on apprend dans la grammaire classique, régissent la méthodologie d' interrogation de corpus. Elle propose tout d'abord de rechercher des mots relevant de genres et de registres différents dans le *British National Corpus*, la partie orale du *Bank of English*, le corpus parallèle anglais-portugais *Compara*, le corpus *Europarl* des minutes du parlement européen et le BLC (*Business Letter Corpus*) de Yasumasa Someya.

18 Frankenberg-Garcia pose ensuite une série de questions auxquelles les enseignants doivent répondre et qui font ressortir le fait que l'on trouve des réponses différentes selon le contenu du corpus, et que donc, le type de corpus est important lorsqu'on l' interroge ; cela permet aussi de souligner l'importance de la datation du corpus, par exemple, on ne trouve pas *MP3* dans le BNC, car le corpus s'arrête au début des années 1990, époque à laquelle le format de données audio MP3 n'existait pas encore ou tout juste. Elle propose ensuite des exercices qui font prendre conscience aux utilisateurs que si l'on ne trouve pas forcément toutes les phrases possibles de la langue dans un corpus, on y trouve néanmoins des régularités correspondant à des séquences qui reviennent fréquemment et qui sont parfois très figées. Ce type de travail permet, à notre avis, d' introduire auprès des enseignants, ou des étudiants en traduction, la différence entre le principe de choix ouvert et le principe idiomatique de Sinclair (1991). On explique que les choix langagiers ne sont pas si libres que cela, qu'une grande partie de la langue est régie par le principe idiomatique qui est relativement arbitraire et imprévisible et que tout locuteur natif a à sa disposition un ensemble de préfabriqués du langage que les locuteurs non natifs ne peuvent pas deviner, ni déduire de la grammaire, et doivent donc apprendre, ce qui peut ensuite amener à la question de la collocation qui est à la base de toute étude de linguistique de corpus traditionnelle.

Quels corpus ?

19 A l'heure actuelle, il existe un certain nombre de corpus accessibles gratuitement en ligne, dans différentes langues. D'autres corpus permettent un accès limité dans le temps, avant que d'être proposés par abonnement payant. L'avantage de ces corpus en ligne est évidemment un accès à une grande quantité de données déjà collectées, souvent étiquetées, et d'un accès gratuit. L'inconvénient est que chaque corpus en ligne est doté de ses propres outils d'interrogation ; ceux-ci demandent un certain apprentissage pour pouvoir être utilisés efficacement. Par ailleurs, ils ne proposent pas tous les mêmes fonctionnalités et ne peuvent donc être utilisés de la même manière d'une langue à l' autre, d'un corpus à l'autre.

20 Nous présenterons plus loin les corpus que nous proposons d'utiliser en anglais et en français. Sans vouloir faire une énumération complète des corpus, nous pouvons en citer quelques-uns dans d'autres langues, comme par exemple *Cosmas* pour l'allemand, le corpus de la *Real Academia* ou le *Corpus del Español* (Davies, 2002), le *Corpus do Portugês* (Davies & Ferreira, 2006), le *Lancaster Corpus of Mandarin Chinese* (McEnery & Xiao, 2004), *A Balanced Corpus of Contemporary Written Japanese* (Maekawa *et al*., 2014), ou le *National Croatian Corpus* (Tadić, 2009). Il faut y ajouter les corpus plus spécialisés et multilingues, comme *Europarl* ou bilingues (*Compara*) mentionnés plus haut. Certains corpus utilisent le web comme corpus, *WebCorp* (Renouf, 2003), ou partiellement, comme les corpus de l'université de Leipzig (*Leipzig Corpora Collection*, Biemann *et al*. 2007) dont nous reparlerons plus loin. En outre, on peut vouloir construire son propre corpus manuellement, ou en utilisant des outils comme *WebBootCaT* (Baroni *et al*., 2006) et l'interroger avec des outils d'interrogation de corpus comme *WordSmith Tools* (Scott, 2008) ou *AntConc* (Anthony, 2005) et bien d'autres encore dont nous ne ferons pas la liste ici. L'avantage de construire son propre corpus est que cela permet de choisir les documents qui le constitueront. C'est indispensable lorsque l'on travaille sur un domaine spécialisé, car il existe peu de corpus spécialisés accessibles. Nous considérons donc que les enseignants et les traducteurs doivent apprendre à compiler leur propre corpus et à utiliser au moins un outil d'interrogation de corpus.

21 Pour notre propos, nous utiliserons le *Corpus of Contemporary American English* (*Coca*) et le *British National Corpus* (BNC) dans l'interface d'interrogation http://corpus.byu.edu (Davies, 2008, 2004). L'interface proposée sur le site de BYU permet de rechercher des concordances, des collocats, de comparer deux corpus, ainsi que de nombreuses autres interrogations sophistiquées. La *Leipzig Corpora Collection* comporte, entre autres, le français et l'anglais. Ce corpus ne donne pas de concordances, mais des collocats, ainsi que les collocats immédiatement à gauche et à droite du mot recherché. Ces différents collocats sont calculés de manière à pouvoir donner les collocations fortement liées, mais aussi les plus fréquentes, à savoir :

> *Given two words A, B, each occurring a, b times in sentences, and k times together, we calculate the significance sig (A, B) of their occurrence in a sentence as follows: Two different types of collocations are generated: Collocation based on occurrence within the same sentence as well as immediate left and right neighbors of each word.*[6] *(Biemann et al., 2004 : 219)*

22 Nous tirerons en outre des exemples du corpus français en ligne *Les Voisins de le Monde* basé sur le logiciel *Upery* (Bourigault, 2002). Il s'agit ici d'un corpus constitué de dix années du journal *Le Monde* dont la taille s'élève à environ cent millions de mots. Ce corpus est étiqueté et analysé. Il ne permet pas de rechercher des concordances, mais des collocations et des "voisins", à savoir, l'ensemble de mots qui partagent un certain nombre de collocats avec le lemme recherché ; il s'agit, en quelque sorte, de synonymes. Le corpus d'une année du journal *Le Monde,* installé sur un concordancier en ligne développé à la fin des années '90 (Kübler & Foucou, 1999), ainsi qu'un corpus spécialisé en sciences de la terre, installé sur une version personnalisée de *IMS Corpus Workbench* (Christ *et al*., 1999) seront utilisés.

Sensibilisation préliminaire

23 Cette section se propose de présenter la manière dont on peut approcher un corpus en se basant en priorité sur les collocations, notion fondamentale de la linguistique de corpus, même si elle a été définie avant l'apparition des corpus informatisés. Après avoir présenté quelques corpus monolingues, bilingues et spécialisés et avoir fait réfléchir les utilisateurs novices sur le contenu des corpus, l'importance de la datation des documents, l'existence de séquences récurrentes, etc., il nous paraît important de sortir des concordances et de revenir à la collocation. Nous avons en effet constaté que la lecture de concordances constituait un exercice très difficile pour les utilisateurs novices, car elle fait appel à une appréhension de la langue dont ils n'ont pas l'habitude ; c'est pour cela que nous pensons qu'il faut rappeler la théorie qui sous-tend la linguistique de corpus, tout en confrontant le plus possible les utilisateurs à l'interrogation des corpus.

Sensibilisation au sens contextuel

24 L'une des premières difficultés rencontrées lors de la présentation d'un corpus à des novices, consiste à faire comprendre comment se définit le sens d'un mot par son contexte, son sens contextuel par opposition à son sens conceptuel, au sens firthien du terme :

> Meaning by collocation is an abstraction at the syntagmatic level and is not directly concerned with the conceptual or idea approach to the meaning of words. One of the meanings of night is its collocability with dark, and of dark, of course, collocation with night.[7] (Firth, [1951] 1957 : 196, cité dans Léon, 2007 : 405)

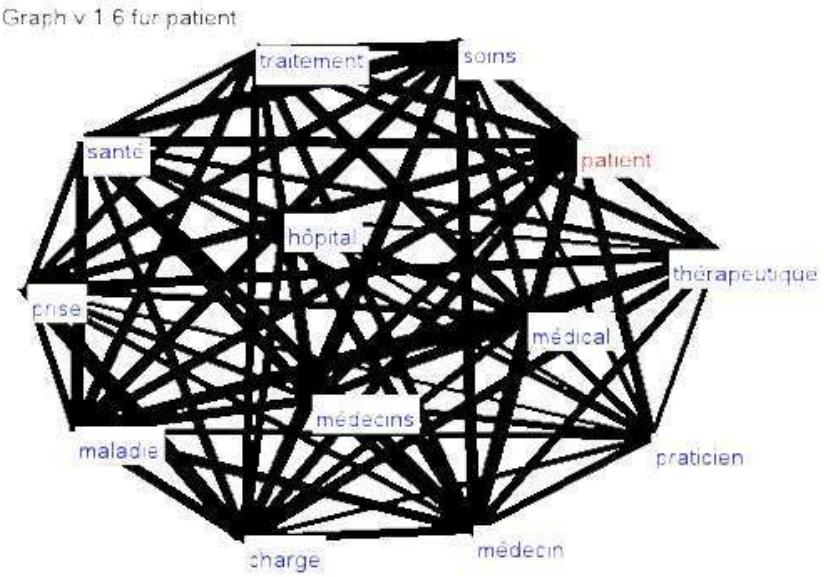
25 Commencer par sensibiliser les utilisateurs novices à la question du contexte du mot, et donc du sens contextuel évoqué par Firth permet de changer l'approche du lexique et des textes. En effet, le sens du mot doit être considéré comme induit par les données, dans une approche empirique allant à l'encontre des réflexes habituels des étudiants, et les enseignants, qui s'appuient sur leur intuition linguistique ou sur les dictionnaires. Dans la définition élargie de la collocation donnée par Sinclair (1991 : 170) : *"collocation is the occurrence of two or more words within a short space of each other in a text"*[8], les novices ont parfois du mal à comprendre le rôle de la collocation qui n'entre pas dans une construction, à savoir une cooccurrence significative de deux mots qui n'entrent pas forcément dans la même construction syntaxique, comme verbe + objet, nom + adjectif, adverbe + adjectif. L'exemple donné par Sinclair est celui de *patient* et *doctor*. Pour arriver à changer les points de vue sur la langue, il vaut mieux commencer par amener le locuteur à faire des recherches dans la langue maternelle, celle qu'il croit très bien connaître. Nous proposons donc aux novices de commencer par rechercher *patient* dans le *Corpus français* de Leipzig et de se demander ce que signifient les différents "cooccurrences significatives" données, ainsi que les "cooccurrences significatives" immédiates à droite et à gauche du mot recherché. Dans le Tableau 1, les chiffres révélant le résultat du calcul statistique autour des cooccurrences ont été supprimés pour la clarté de la lecture ; les cooccurrences significatives, que nous appellerons désormais "collocats", sont classées de la plus haute force de collocation à la plus basse, tout en tenant compte de la fréquence. On peut constater que le mot *médecin* est très fortement lié à *patient*, sans qu'il y ait forcément de lien syntaxique entre les deux.

Tableau 1 – Collocats de patient dans le Corpus français de Leipzig.

médecin, traitement, le, soins, du, médical, un, être, charge, thérapeutique, hôpital, maladie, praticien, santé, prise, médecins, doit, peut, hospitalisé, thérapeute, dossier, ou, clinique, état, son, cas, patients, hospitalisation, chez, diagnostic, Argumenter, douleur, traitant, médicale, chaque, consultation, suivi, soignant, anesthésie, risque, malade, symptômes, consentement, médicament, Diagnostiquer, il, médicaments, Le, médicales, l'attitude, intervention, une, chirurgien, infirmière, planifier, prescription, chirurgie, pathologie, et, au, hospitalier, relation, informé, diabétique, faut, cellules, permet, infection, entourage, soignants, dose, informations, chirurgicale, est, décubitus, établissement, pharmacien, examen, par, traitements

26 Ces collocats montrent que les mots ne sont pas isolés dans les productions linguistiques, mais qu'ils se trouvent dans un réseau de collocats fortement liés, ce que d'ailleurs le réseau donné dans l'interface démontre (cf. Tableau 2). On constate en effet que les collocats les plus fortement liés à *patient* (dans le contexte de la phrase), outre le mot *médecin*, relèvent en majorité du domaine médical et hospitalier.

Tableau 2 – Réseau de collocats pour le mot patient.



27 Chez un public novice, l'interprétation des collocats n'est pas immédiate. Il est donc nécessaire de demander aux apprenants d'envisager, pour chaque collocat, comment celui-ci entre en relation avec le mot *patient*, ce qui les amène à constater que la présence de *être* et de *est* sont les deux seuls collocats pouvant être reliés à l'interprétation de *patient* comme adjectif. Cette recherche leur fait aussi prendre conscience qu'une collocation n'est pas forcément inscrite dans une construction, mais reflète la cohésion interne d'un texte et d'un domaine. En effet, le Tableau 1 donne les collocats pris dans une même phrase, leur position n'étant pas déterminée.

28 L'étape suivante consiste à faire comprendre que la position du collocat, à gauche ou à droite du mot recherché, revêt une grande importance. C'est là qu'intervient le travail sur les collocats immédiats du mot, comme dans les Tableaux 3 et 4.

Tableau 3 – Collocats de gauche immédiats de patient.

du, le, un, Le, au, chaque, être, son, Un, montrer, dossier, soyez, d'un, Chaque, Soyez, Aucun, jeune, très, autre, tout, fiche, aucun, même, rester, ce, Tout, seul, leur, forfait, par, votre, premier, travail, sois, ancien, nouveau, quel, Ce, Dossier, suffisamment, qu'un, peu, montré, qu'un, dialogue, assez, Être, identifiant, célèbre, illustre, suis, homme, d'un, qu'un, dressage, Espace, dossiers, Etre, tel, questionnaire, Fiche, soyons, être, capital, montrait, investisseur, Votre, bon, médical, resté, notre, êtes, mystérieux, restant, mon, deuxième

29 On peut ensuite demander à nouveau aux novices d'analyser chaque collocat immédiat de gauche et de droite, et de les classer par grandes catégories grammaticales :

- Det : du, le, un, Le, chaque, son d'un, Aucun, tout, ce, Tout, leur, votre, quel, Ce, tel, Votre, notre
- V : montrer, soyez, Soyez, rester, sois, montré, Etre, suis, soyons, être, montrait, resté, êtes, restant, identifiant
- Adj : jeune, autre, même, seul, premier, ancien, nouveau, célèbre, illustre, capital, bon, médical mystérieux, deuxième
- Adv : très, suffisamment, assez
- N : dossier, fiche, forfait, travail, Dossier, dialogue, identifiant, homme, dressage, Espace, dossiers, questionnaire, Fiche, capital, investisseur

30 Cet exercice a pour objectif d'amener les utilisateurs à prendre conscience pour comprendre le sens du mot, tant de l'importance du contexte, c'est-à-dire de l'ensemble des segments dans lesquels un mot apparaît, que des questions d'ambiguïté syntaxique (cf. *patient* Adj ou N, *capital* N ou Adj, et *identifiant* N ou Adj). Deux formes en apparence identiques peuvent recouvrir des structures différentes, comme *fiche patient* (N N) et *travail patient* (N Adj). Ces observations ne sont pas triviales, car il est indispensable de faire comprendre aux utilisateurs novices qu'un outil d'interrogation de corpus ne recherche que ce qu'on lui demande. Par ailleurs, la linguistique de corpus cherche le sens des mots par les collocations intra- et inter-textuelles. Or, les novices abordent souvent le corpus avec déjà une grammaire en tête et découvrent ici des emplois de *patient* auxquels ils n'auraient pas pensé en s'appuyant uniquement sur leur intuition et sur l'idée que les mots ont un sens "hors-sol", c'est-à-dire, hors contexte, absolu.

Tableau 4 – Collocats de droite immédiats de patient.

hospitalisé, et, doit, est, peut, atteint, diabétique, qui, lui-même, ne, souffrant, âgé, ayant, travail, ou, anglais, porteur, présentant, infecté, immunodéprimé, séropositif, a, reçoit, dyslipidémique, traité, asthmatique, schizophrène, donné, pourra, se, cancéreux, avec, présente, dans, consulte, souffre, devra, ressent, anesthésié, à, afin, vers, douloureux, décédé, puisse, était, déprimé, en, pour, avant, accepte, sera, hypertendu, allongé, amputé, psychotique, Odysseus, respire, bénéficie, greffé, parkinsonien, aphasique, tétraplégique, psychiatrique, coronarien, va, opéré, concerné, discipliné, dépressif, allergique, Décrire, intubé

31 Le même travail avec les voisins de droite peut être présenté aussi (Tableau 4). Ici, on constate que *patient* est suivi de nombreux adjectifs, participes passés adjectivaux ou verbes conjugués. Seule la forme *patient travail* subsiste pour l'adjectif. L'adjectif *patient* a

donc davantage tendance à être en position attribut qu'en position épithète. Cependant, on remarquera aussi que pour le nom *patient*, sauf dans l'occurrence *patient anglais* (qui est le titre d'un film), la majorité des mots à contenu sémantique plein relèvent de la médecine ou de la maladie. Un locuteur non natif du français pourrait donc comprendre le sens de *patient* par ses collocats. C'est ici que l'on peut donner une définition plus précise de la collocation et introduire les phénomènes de colligation, de préférence sémantique ou de prosodie sémantique, qui seront approfondis plus tard.

32 Afin de mieux aider les utilisateurs à intégrer l'idée que le logiciel d'interrogation de corpus ne cherche que ce qu'on lui demande, on peut montrer les résultats que donnent les recherches sur *patiente* et *patients*, dont les collocats les plus fortement liés sont quasiment les mêmes que pour *patient* au masculin singulier, mais dont certains autres diffèrent nettement. En effet, ne rechercher que le lemme d'un mot, sans faire la distinction entre ses différentes formes déclinées ou conjuguées, empêche de mettre en lumière l'importance de chaque forme qui convoque des collocats différents, et donc parfois, des sens différents. Cela permet aussi d'aborder la question de l'étiquetage du corpus et de la différence entre une recherche sur un élément lexical uniquement et une recherche sur un lemme. Pour cela, nous proposons d'utiliser le corpus français *Les Voisins de Le Monde*.

Sensibilisation à la lemmatisation

33 Comme mentionné plus haut, *Les Voisins de Le Monde* est un corpus lemmatisé, étiqueté et analysé. Avant d'expliquer aux utilisateurs novices ce que cela signifie, il suffit de leur demander de rechercher le mot *patiente* dans le corpus, et ensuite le mot *patient.* Rechercher *patiente* ne donnera aucun résultat, alors que *patient* donnera le résultat dans le Tableau 5.

Tableau 5 – Résultats de la recherche de patient dans Les Voisins de Le Monde.

Arguments				
Catégorie	Lemme	Relation	Nb cooccurrents	Nb Voisins
N	patient	_	190	199
A	patient	_	41	9
Prédicats				
Catégorie	Lemme	Relation	Nb cooccurrents	Nb Voisins
N	patient	de	14	5
N	patient	en	2	0
N	patient	mod	32	351
N	patient	à	1	0

34 On peut alors expliquer ce que sont la lemmatisation et l'étiquetage, préciser le type d'analyse auxquels a été soumis le corpus et donc montrer, qu'ici, contrairement à ce qui a été fait dans le corpus de Leipzig, on ne devra pas rechercher des mots mais des lemmes. Ainsi, l'analyse fait immédiatement la différence entre le nom et l'adjectif et donnera tous les résultats trouvés pour toutes les formes fléchies. Par ailleurs, le corpus ne donne que les collocats entrant dans une relation syntaxique avec le lemme, ce qui est donné dans la troisième colonne dans le tableau. Enfin, l'analyse se fait sur le lemme en tant que prédicat ou argument. Il convient alors d'examiner les deux fonctions pour obtenir tous les collocats du nom, l'adjectif n'étant pas prédicat. L'utilisation de ce corpus n'est donc pas tout à fait la même que celle que l'on fait du *Corpus français* de Leipzig. On va ici rechercher directement les constructions les plus probables d'un mot. Le Tableau 6 donne les dix premiers collocats prédicats du nom, triés par information mutuelle. L'information mutuelle est un autre calcul statistique qui a l'avantage de révéler les collocations les plus fortes, mais avec l'inconvénient de donner des collocations rares. On voit ici que lorsque l'on donne uniquement les collocations selon l'information mutuelle, la cohorte de collocats de *patient* est différente de celle que l'on obtient dans le *Corpus français* de Leipzig. Ainsi, on montre que les résultats donnés par les corpus dépendent de la manière dont ceux-ci ont été construits, et de l'outil d'interrogation utilisé.

Tableau 6 – Dix premiers collocats prédicats du nom patient.

Prédicat			Argument			
Catégorie	Lemme	Relation	Catégorie	Lemme	IM	Fréquence
N	séropositivité	de	N	patient	10. 274	6
V	induire	chez	N	patient	10. 274	9
V	prescrire	à	N	patient	8. 746	18
V	transfuser	obj	N	patient	8. 687	9
V	décéder	suj	N	patient	8. 551	5
V	administrer	à	N	patient	8. 273	15
V	survenir	chez	N	patient	8. 61	7
N	abonnement	de	N	patient	7. 907	6
V	observer	chez	N	patient	7. 678	12
V	hospitaliser	obj	N	patient	7. 474	9

35 Observer les collocats prédicats de l'adjectif *patient* montre que le corpus peut être mal étiqueté et mal analysé, comme l'illustre le Tableau 7.

Tableau 7 – Dix premiers collocats prédicats de l'adjectif patient selon l'information mutuelle.

Prédicat			Argument			
Catégorie	Lemme	Relation	Catégorie	Lemme	IM	Fréquence
N	décédé	mod	A	patient	11. 619	6
N	illustre	mod	A	patient	10. 772	6
N	english	mod	A	patient	10. 284	10
N	labeur	mod	A	patient	9. 432	11
N	élaboration	mod	A	patient	8. 638	7
N	toxicomane	mod	A	patient	8. 252	5
N	zéro	mod	A	patient	8. 218	6
N	reconquête	mod	A	patient	8. 41	5
N	exploration	mod	A	patient	7. 586	7
N	traité	mod	A	patient	7. 428	35

36 On voit dans les trois premières ligne que *patient* a été analysé comme adjectif, alors qu'il s'agit ici d'un nom *patient décédé, illustre patient,* et le film *Le Patient anglais/The English Patient*. Il en va de même pour *patient toxicomane* et sans doute pour *zéro patient.* Cependant, les autres lignes donnent des analyses correctes et montrent aussi une certaine différence dans la cohorte de collocats pour l'adjectif par rapport au *Corpus français* de Leipzig. C'est ici que l'on peut demander aux utilisateurs novices d'aller lire la description de chaque corpus : *Les Voisins de Le Monde* compte 100 millions de mots et représente dix ans du *Monde* uniquement, alors que le *Corpus français* de Leipzig compte 700 millions de mots et est constitué de journaux francophones incluant donc d'autres pays que la France, de pages web en français et de *Wikipédia* en français. L'objectif ici étant de rappeler que ce que l'on trouve dans un corpus dépend de ce que l'on y a mis, de sa taille, de la manière dont il a été construit, analysé, étiqueté, etc., et de la manière dont on peut l'interroger. On pourra constater aussi que le corpus n'a pas été analysé pour les structures attributs et ne montre donc pas la préférence pour la position attribut de *patient* adjectif. Dans la position épithète plus rare, le collocat le plus fréquent est le mot *travail* et ce n'est qu'en neuvième position que l'on trouve *homme*. Le corpus *Les Voisins de Le Monde* donnera des constructions, mais ne donnera pas tous les collocats qui permettent de s'approprier le sens contextuel d'un mot.

Relativité d'un corpus

37 Nous suggérons ici un dernier exercice, comparant deux corpus en français et qui permet à la fois d'insister sur la question du contenu du corpus et d'introduire ce qu'est une

concordance. En effet, montrer des concordances est en général la première activité que l'on fait faire à des utilisateurs novices. Or, comme mentionné plus haut, bien qu'une concordance donne un contexte plus développé qu'une liste de collocats, nous avons pu constater, au fil des ans, que ça n'était pas la meilleure manière d'aborder les collocations, et la question du sens contextuel. En effet, on constate que les débutants cherchent toujours à analyser les lignes de concordance dans une approche "syntagmatique", et ce, bien que dans les concordances, les phrases soient souvent coupées, alors qu'il est plus avantageux d'y projeter une double vision, à la fois syntagmatique et paradigmatique. Pour cette raison, nous n'abordons réellement les concordances qu'à ce stade du travail. Nous prendrons l'exemple du mot *initiative* dans le *Corpus français* de Leipzig et dans le corpus d'une année du journal *Le Monde*, datant de 1994, que nous avons installé sur le concordancier en ligne *Web-Assisted Language Learning* (*Wall*[9]), créé en 1997 par Kübler et Foucou (1999).

38 Avant de faire rechercher le mot dans les deux corpus, il faut commencer par demander quel est le collocat qui vient immédiatement à l'esprit dans ce cas. La majorité des présents répondra par *prendre une* initiative. Afin d'illustrer notre propos, le Tableau 8 montre tout d'abord ce que donne l'interrogation du *Corpus français* de Leipzig.

Tableau 8 – Collocats d'initiative dans le Corpus français de Leipzig.

l, ', cette, populaire, à, Cette, une, d, de, L, UDC, contre-projet, lancée, Conseil, du, a, pris, naturalisations, vise, visant, signatures, ", COSA, propre, AVS, Une, la, parlementaire, qui, des, telle, président, gouvernement, prendre, arabe, soutien, association, créée, référendum, paix, comité, peuple, projet, votation, fédéral, soutenue, par, esprit, cadre, privée, lancé, créé, européenne, minarets, salué, soutenir, prise, pour, faveur, Union, PPTE, louable, cantons, organisée, déposée, le, pays, Association, associations, Commission, lancer, saluer, rejetée, ministre, saluée, nationale, Arle, personnelle

39 On constate que les collocats les plus fréquents, si l'on écarte les mots-outils, sont *populaire, UDC, contre-projet, lancée, Conseil* et que l'on retrouve de nombreux termes liés au système démocratique suisse, comme *référendum, votation, fédéral, cantons*. Le mot *minaret* s'explique par l'initiative assez récente d'un parti suisse contre la construction de minarets en Suisse. Par ailleurs, si l'on examine les concordances du corpus d'une année du *Monde* sur *Wall*, dont nous donnons un extrait en Tableau 9, on ne retrouve que deux lignes dans lesquelles *initiative* a un emploi spécifique aux institutions politiques helvétiques[10].

Tableau 9 – Extrait de 20 lignes de concordances du corpus Le Monde sur Wall.

Emanant de groupes écologistes l'**Initiative** des Alpes " était soutenue par la
du Sundance Festival de Tokyo, de l'**Initiative** latino-américaine et d Equinoxe.
balistiques (BMD), successeur de l'**Initiative** de défense stratégique (IDS) et
notamment veiller au développement de l'**Initiative** lorraine pour l'emploi (ILE) dont le
passer par-dessus, la dague au poing, l'**Initiative**

le Tessin de langue italienne que l'**Initiative** a enregistré ses meilleurs scores, avec
une pratique courante au temps du muet. **Initiative** continuée par Wayne, Lancaster,
la reconstitution de l'empire russe. **Initiative** louable, mais qui se heurte au peu d
de printemps en quelque sorte. **Initiative** intelligente et périlleuse. Bonne idée,
Autre **initiative**, au Fleuve Noir, sous la direction de
Autre **initiative** attendue, celle du mouvement " Forza
France a toujours proclamées. " Autre **initiative**, prise dimanche : la décision de
sont destinés aux chômeurs. Autre **initiative**: la première feria d'économie rurale
sur les déchets nucléaires. Bonne **initiative**. Une trentaine de candidatures se sont
Cette **initiative**, disait-on à l'ONU, doit " compléter "

40 La conclusion à laquelle nous amenons les utilisateurs novices est qu'il est nécessaire de connaître le contenu du corpus et de s'informer sur sa construction. Le *Corpus français* de la *Leipzig Corpora Collection* a été constitué par le groupe de recherche TAL de l'Université de Leipzig, et aménagé avec le concours de Daniel Elmiger et Alain Kamber (Université de Neuchâtel, Suisse). Il contient des journaux francophones et donc des journaux suisses romands, ce qui explique la différence de collocations entre ce corpus et le corpus d'un journal français comme *Le Monde*.

41 Cette petite tâche permettra de montrer en outre qu'on ne peut se contenter d'afficher des concordances. En effet, le corpus d'une année du *Monde* sur *Wall* (10 millions de mots) contient 2.224 occurrences du mot *initiative* au singulier. Il est donc difficile pour un novice de lire ces concordances pour en repérer les collocats les plus fréquents. Ce sera l'occasion de montrer comment trier les concordances en démontrant l'utilité d'une telle fonctionnalité. Cependant, nous avons constaté que, malgré la performance des outils d'interrogation de corpus actuels, il reste nécessaire d'en avoir une longue pratique si l'on veut pouvoir aborder les corpus avec un œil suffisamment critique et informé.

Travail interlinguistique

42 Après avoir tenté de faire comprendre ce qu'était un corpus avec deux exemples différents, nous pouvons proposer de faire tout d'abord un travail similaire sur l'anglais en utilisant, dans la *Leipzig Corpora Collection,* le *Corpus anglais* et toujours avec le même exemple, *patient*. On peut donc commencer par faire étudier les collocats significatifs entre le français et l'anglais, comme dans le Tableau 10.

Tableau 10 – Collocats significatifs de patient dans la Leipzig Corpora Collection / Corpus anglais.

care, hospital, patients, doctor, medical, cancer, treatment, Hospital, doctors, a, hospitals, health, nurse, be, physician, Medical, surgery, safety, clinical, to, nurses, records, Health, physicians, transplant, blood, heart, medication, disease, and, nursing, cells, NHS, emergency, ambulance, therapy, procedure, staff, satisfaction, surgeon, clinic, confidentiality, mental, rooms, diagnosis, condition, room, improve, drug, outcomes, or, information, treat, Center, quality, infection, marrow, medicine, breast, symptoms, cardiac, can, treated, healthcare, privacy, psychiatric, system, surgical, tumor, services, Alzheimer, pain, said, Medicare, A, study, treatments, medications	médecin, traitement, le, soins, du, médical, un, être, charge, thérapeutique, hôpital, maladie, praticien, santé, prise, médecins, doit, peut, hospitalisé, thérapeute, dossier, ou, clinique, état, son, cas, patients, hospitalisation, chez, diagnostic, Argumenter, douleur, traitant, médicale, chaque, consultation, suivi, soignant, anesthésie, risque, malade, symptômes, consentement, médicament, Diagnostiquer, il, médicaments, Le, médicales, l'attitude, intervention, une, chirurgien, infirmière, planifier, prescription, chirurgie, pathologie, et, au, hospitalier, relation, informé, diabétique, faut, cellules, permet, infection, entourage, soignants, dose, informations, chirurgicale, est, décubitus, établissement, pharmacien, examen, par, traitements

43 Cette tâche vise à montrer que *patient* en anglais est surtout utilisé dans son sens médical, comme le montrent ses collocats, et non au sens de 'patience' comme en français. On remarque aussi que le mot entre beaucoup plus dans des constructions nominalisées qu'en français. Dans les deux langues, la majorité des collocats appartiennent au domaine médical et l'on retrouve aussi bien *doctor* en anglais que *médecin* en français.

44 Ensuite, faire réfléchir les utilisateurs novices sur les collocats immédiats de droite et de gauche (cf. Tableaux 11 et 12) leur permet de prendre conscience du fonctionnement de *patient* en anglais et de comparer celui-ci à *patient* en français.

Tableau 11 – Collocats immédiats de gauche dans le Leipzig Corpora Collection / Corpus anglais.

a, be, the, cancer, improve, Be, each, very, mental, more, female, stay, being, The, every, one, per, transplant, remain, leukemia, stayed, psychiatric, One, elderly, another, individual, A, ill, male, improving, dying, dialysis, heart, Each, direct, electronic, former, hospital, citing, improved, Another, hospice, any, enhance, staying, AIDS, ensure, quality, Cancer, trauma, first, Medicare, private, new, diabetic, Alzheimers, protect, dementia, cardiac, stroke, pediatric, and, kidney, fellow, sick, compromise, remained, young, ALS, confidential, average, burn, surgery, affect, were, specific, Alzheimer's, improves, tuberculosis	du, le, un, Le, au, chaque, être, son, Un, montrer, dossier, soyez, d'un, Chaque, Soyez, Aucun, jeune, très, autre, tout, fiche, aucun, même, rester, ce, Tout, seul, leur, forfait, par, votre, premier, travail, sois, ancien, nouveau, quel, Ce, Dossier, suffisamment, qu'un, peu, montré, qu'un, dialogue, assez, Être, identifiant, célèbre, illustre, suis, homme, d'un, qu'un, dressage, Espace, dossiers, Etre, tel, questionnaire, Fiche, soyons, etre, capital, montrait, investisseur, Votre, bon, médical, resté, notre, êtes, mystérieux, restant, mon, deuxième

Tableau 12 – Collocats immédiats de droite de patient dans le Leipzig Corpora Collection / Corpus anglais

care, ', safety, satisfaction, records, who, with, and, confidentiality, rooms,, is, information, outcomes, privacy, deaths, population, visits, at, was, services, data, tower, load, advocacy, died, needs, has, advocates, groups, beds, enough, volume, education, navigator, populations, approach, advocate, files, referrals, can, monitoring, admissions, may, whose, survival, had, access, charts, volumes, dies, experience, navigators, flow, recruitment, simulator, gets, while, transport, compliance, assistance, suffering, registry, arrives, receives, abuse, treatment	hospitalisé, et, doit, est, peut, atteint, diabétique, qui, lui-même, ne, souffrant, âgé, ayant, travail, ou, anglais, porteur, présentant, infecté, immunodéprimé, séropositif, a, reçoit, dyslipidémique, traité, asthmatique, schizophrène, donné, pourra, se, cancéreux, avec, présente, dans, consulte, souffre, devra, ressent, anesthésié, à, afin, vers, douloureux, décédé, puisse, était, déprimé, en, pour, avant, accepte, sera, hypertendu, allongé, amputé, psychotique, Odysseus, respire, bénéficie, greffé, parkinsonien, aphasique, tétraplégique, psychiatrique, coronarien, va, opéré, concerné, discipliné, dépressif, allergique, Décrire, intubé .

45 Les collocats immédiats de droite et de gauche diffèrent du fait autour des modalités de constructions syntaxiques du groupe nominal en anglais et en français. C'est à gauche que l'on trouvera les adjectifs modifieurs. À droite, on trouvera de nombreux noms dus à la construction N N, très répandue en anglais.

46 Un exercice dérivé de ces observations peut consister à trier les voisins de *patient* en français et en anglais pour détecter les emplois adjectivaux au sens de 'qui fait preuve de patience' et les emplois nominaux au sens de 'malade', tout en comparant ce qui revient le plus fréquemment d'une langue à l'autre. Cette activité permet d'illustrer l'approche des différences de sens contextuel pour un mot qui pourtant semble équivalent d'une langue à l'autre.

Affiner le sens contextuel à l'aide des concordances

47 Dans les sections précédentes, nous avons montré comment utiliser les collocats donnés par deux types de corpus pour illustrer la recherche de sens contextuel dans les corpus et dans des langues différentes. La notion de concordance a aussi été abordée pratiquement. Pour des enseignants travaillant en Lansad, il est ensuite indispensable de les faire travailler sur l'interface de Mark Davies, sur le BNC et le *Coca*. En effet, cette interface permet d'afficher, outre des collocats, des lignes de concordance contenant ces collocats. Un exemple de tâche consiste à rechercher les collocats de *patient* dans le *Coca*, afin de démontrer que, malgré une recherche dans deux corpus différents, le sens contextuel reste le même par suite du caractère plus ou moins généraliste de ces corpus. Le Tableau 13 ci-dessous illustre l'intérêt de cette tâche.

TABLEAU 13 – Collocats du mot *patient* dans le *COCA*.

PATIENT, DOCTOR, CANCER, MEDICAL, HOSPITAL, TREATMENT, SAFETY, SATISFACTION, PHYSICIAN, HANDLING, SURGERY, AIDS, RECORDS, OUTCOMES, SYMPTOMS, CONDITION, CLINICAL, NURSE, UNDERWENT, TREATED, THERAPY, ILL, TREAT, CHARACTERISTICS, DIAGNOSIS, ALZHEIMER, CONSENT, DISCHARGE, PROCEDURE, FOLLOWUP, SURGICAL, EXAMINATION, REFERRED, ADMITTED, NURSES, DISCHARGED, POSTOPERATIVE, TUMOR, THERAPIST, VISITS, NURSING, COMPLIANCE, AUTONOMY, TREATING, ELDERLY, DIAGNOSED, ADVOCATE, MEDICATION, ADVOCATES, CLINIC, POSTOPERATIVELY, TRANSPLANT, ADVOCACY, WISHES, PHYSICIANS, INFECTION, PROVIDER, CONFIDENTIALITY, CARING, DIABETES, CHRONIC, COMPLICATIONS, PSYCHIATRIC, CAREGIVER, PERSISTENT, RECOVERED, SURGEON, DIALYSIS, MANUAL, TRANSFERS, ACUTE, SELFDETERMINATION, TERMINALLY, UNDERGONE, ARTHRITIS, QUESTIONNAIRE, HEALTHCARE, SYNDROME, EXHIBITED, ANESTHESIA, HOSPICE, CT, INFECTED, CHEMOTHERAPY, CARDIAC, CONSULTATION, TRANSFERRED, BEDSIDE, DEMENTIA, NASAL, PRESCRIBED, MEDICATIONS, UNDERGO, LESION, COMPETENT, LIVER, ASYMPTOMATIC, NAVIGATION, DIES

48 L'observation des collocats de *patient* dans le *Coca*, permet de démontrer la relativité d'un corpus et de rappeler que tout corpus, même le plus grand, ne donne que ce qu'il contient. Aucun corpus n'est exhaustif, ni complètement représentatif. Cependant, de très grands corpus, contenant des textes correspondant à un grand éventail de situations de communications, sont suffisamment représentatifs pour obtenir une représentation utilisable de la langue générale. Dans le cas présent, le *Corpus anglais* de Leipzig et le *Coca* rendent un sens contextuel à peu près semblable du mot recherché. Cependant, les deux interfaces présentent des fonctionnalités différentes. L'intérêt du *Coca*, et de tous les corpus sur cette interface en ligne de BYU, est de pouvoir mener des recherches allant du très simple au très sophistiqué.

49 L'exemple des collocations du *Coca* démontre l'intérêt d'avoir accès à une concordance. En effet, dans la liste des collocats qui accompagnent *patient*, on trouve par exemple *navigation*. Si l'on comprend bien les liens entre *hospital, treatment* ou *chronic*, il est difficile, si l'on ne le connaît pas, de cerner le lien avec *navigation*. On peut supposer qu'il s'agit d'une erreur, mais toute hypothèse demande à être vérifiée dans le corpus. L'interface du *Coca* permet de cliquer sur le collocat et d'obtenir toutes les lignes de concordances dans lesquelles se trouvent *patient* et *navigation*. On découvre alors que l'on parle de *patient navigation* et plusieurs des lignes de concordance en donnent une définition (Tableau 14).

Tableau 14 – Extrait de concordances pour patient navigation dans le Coca.

member said, " **Patient navigation** is a type of case management and works congruently with the social worker
the concept of **patient navigation** is presumed to be a promising strategy to reduce racial and ethnic disparities
The effect of **patient navigation** on time to diagnosis, anxiety, and satisfaction in urban minority women with

50 On ajoute ici un argument supplémentaire révélant l'utilité d'une concordance et comment celle-ci peut être utilisée comme un complément aux dictionnaires. En effet, *patient navigation* n'est pas dans le *Merriam-Webster Online Dictionary*, ni dans les *Cambridge Dictionaries Online*. Ce type de recherche est indispensable lorsque l'on souhaite enseigner l'anglais de spécialité, comme nous le verrons plus loin.

La langue, un phénomène social observé dans les corpus

51 Chaque tâche à effectuer doit mener à une nouvelle découverte et à un pas supplémentaire dans la compréhension de ce qu'est un corpus et des questions que l'on peut lui poser. Les résultats obtenus jusqu'ici montrent comment un mot fait appel dans un corpus à toute une cohorte de mots qui lui sont associés, syntaxiquement ou non. Les collocats directs, donc, la plupart du temps, dans une structure syntaxique, montrent à quel point le sens du mot dépend du contexte. Cela permet aussi de montrer que la langue est un phénomène social et que l'on ne recherche pas dans le corpus la représentation mentale qu'a le locuteur du sens, ni celle qu'en a le destinataire, comme le souligne Teubert (1999), mais le sens par les mots qui se retrouvent avec le mot recherché. Prenons l'exemple de *fumer* dans le *Corpus français d*e Leipzig (Tableau 15).

Tableau 15 – Collocats de fumer dans le Corpus français de Leipzig.

interdiction, lieux, publics, arrêter (), fumeurs, interdit, restaurants, tabac, cigarette, dans, boire, bars, de, cigarettes, les, cafés, vigueur, manger, cesser, tabagisme, interdisant, interdire, loi, totale, discothèques, fumeur, cigare, janvier, établissements, pipe, non-fumeurs, buralistes, narguilé, tabacs, décret, santé, nicotine, fermés, cannabis, crack, Evin, clope, train, interdictions, entrée, Santé, cigares, bars-tabac, arrêté, marijuana, joint, fumoirs, anti-tabac, Interdiction, calumet, bars-tabacs, l, abstenir, Fumer, jeunes, fumée, Arrêter, casinos, alcool, ne, arrête, travail, dehors, espaces, joints, tous, déclarent, à, fument, arrêtent, où, envie

52 On peut tout d'abord faire le même exercice que précédemment et ensuite souligner les interprétations sociolinguistiques que l'on peut en faire. Cet exemple souligne tout d'abord que fumer est mal considéré aujourd'hui et que c'est un trait dominant dans notre société. On constate en effet que l'on veut l'interdire et qu'il faut cesser de fumer, et que fumer est associé à des drogues. Le *Corpus français* de Leipzig est constitué de journaux, de sites et de *Wikipédia* en français, mais ne contient pas de blogs qui expriment des opinions individuelles. Eensoo-Ramdani *et al.* (2011), dans une recherche menée sur la fouille de texte pour détecter l'opinion, ont analysé de nombreux blogs sur la question de la cigarette/fumer pour déterminer les caractéristiques de ce genre textuel dans un domaine hautement polémique. Ils ont démontré que, dans les blogs qui, comme nous l'avons mentionné, expriment une opinion personnelle, se dessine très clairement un axe militant/non-militant, plutôt qu'un axe pro-/anti-tabac. Nous insisterons sur le fait que le résultat de la recherche reflète ce qu'il y a dans les types de textes sélectionnés. Les réponses du corpus reflètent également son contenu en termes de genres textuels, qui se définissent, de manière extralinguistique, par l'identité de l'auteur et celle du public visé. On peut ensuite vérifier que le mot *smoking* a aujourd'hui le même sens en anglais qu'en français, comme c'est le cas dans le Tableau 16.

Tableau 16 – Collocats de smoking dans le Leipzig Corpora Collection / Corpus anglais de Leipzig.

ban, cigarette, quit, tobacco, bars, cigarettes, smokers, restaurants, bans, cessation, marijuana, places, drinking, public, smoke, health, casinos, statewide, workplaces, banned, banning, lung, indoor, crack, effect, quitting, obesity, alcohol, habit, cancer, cigar, risk, smoke-free, stop, ordinance, factors, bar, Health, nicotine, law, pot, non-smoking, areas, smoker, disease, casino, enclosed, Tobacco, prohibit, cocaine, study, anti-smoking, Smoking, pubs, pipe, people, Winehouse, habits, cigars, establishments, gun, dangers, weight, prohibits, designated, and, nonsmoking, researchers, outdoor, cannabis, in, hookah, taverns, restrictions, diabetes, smoked, emphysema, eating, diet, buildings

53 L'objectif ici est d'introduire la notion de préférence sémantique sous forme d'une tâche dans laquelle les utilisateurs doivent regrouper les collocats selon des classes sémantiques intuitives. On montre ainsi que les mots entrent en co-occurrence de manière significative avec des classes sémantiques spécifiques qui contribuent à son sens, ce qui est le cas avec *smoking* et *fumer*.

54 Cependant, *smoking* ne présente pas tout à fait les mêmes classes sémantiques que *fumer*. On y trouve la classe sémantique des maladies et celle du risque, comme le montre le Tableau 17.

Tableau 17 – Collocats de smoking triés par classes sémantique intuitives.

Anglais	Français
• ban, bans, banned, banning, prohibits, restrictions, prohibit	• interdiction, interdit, interdisant, interdire, Interdiction, interdictions
• stop, quit, quitting, cessation	
• smoke-free, non-smoking, anti-smoking	• cesser, Arrêter, arrête, arrêtent arrêter, arrêté, abstenir
• obesity, cancer, disease, diabetes, amphysema	• cannabis, crack, marijuana, joint, joints
• risk, dangers	• tabagisme, anti-tabac
• cannabis, crack, marijuana	

55 *Smoking* n'est pas toujours l'équivalent syntaxique exact de *fumer* puisque l'infinitif *to smoke* est aussi possible en raison des différences de constructions entre le français et l'anglais. De plus, *smoking* peut aussi être le résultat d'une nominalisation représentant 'le fait de fumer'. Cela nous amène à attirer l'attention des utilisateurs novices sur la prudence linguistique avec laquelle il faut aborder les résultats du corpus et la nécessité de développer une conscience linguistique fine pour observer les phénomènes en corpus et pouvoir les utiliser correctement avec les étudiants. La tâche suivante illustre ceci.

Développer la conscience linguistique

56 L'exercice *consiste* à faire rechercher les collocats de *manger* et *eating* en français et en anglais (Tableau 18) et de trier les collocats par catégories syntaxiques et classes sémantiques intuitives.

Tableau 18 – Collocats de manger et eating dans la Leipzig Corpora Collection/ Corpus français et anglais.

boire, salle, à, dormir, cuisine, viande, repas, faim, on, nourriture, pain, salon, légumes, restaurant, fumer, fruits, table, ne, et, quoi, aliments, je, chambres, coucher, pas, ils, il, cheminée, aller, midi, pour, ou, sainement, bien, laver, lui, rien, poulet, sans, faire, gens, se, plats, plaisir, terrasse, enfants, donner, faut, poisson, qu, peut, soupe, abstenir, me, chocolat, nous, salles, vivre, mange, déjeuner, ni, chambre, vous, aime, envie, tout, gras, On, Salle, animaux, train, cantine, acheter, ..., chez, maison, jour, vais, restaurants	habits, food, healthy, disorders, drinking, foods, meat, lunch, disorder, diet, and, vegetables, meals, fish, breakfast, meal, restaurant, exercising, eat, weight, healthier, dinner, sleeping, exercise, contaminated, people, fruits, or, restaurants, health, ill, fat, you, your, calories, are, tomatoes, obesity, raw, fruit, contest, hot, cooking, pizza, out, chocolate, dumplings, cream, less, lifestyle, bulimia, dog, unhealthy, sick, anorexia, smoking, children, pie, while, chicken, they, day, sandwich, after, watching, consumers, Eating, nutrition, Takeru, re, like, healthful, dining, beef, she, their, bread, shopping, sitting

57 Si les deux mots ont un grand nombre de collocats en commun (nous ne les détaillerons pas ici, mais la tâche doit être exécutée de manière complète), il en est cependant un certain nombre qui diffèrent et frappent en anglais : *disorders, anorexia, bulimia, obesity, ill, unhealthy, sick, nutrition*. Une petite remarque humoristique sur le fait qu'en anglais, on parle immédiatement de maladie lorsque l'on parle de manger permet de revenir sur les différences syntaxiques entre l'anglais et le français en mettant au jour le composé *eating disorders* qui est le terme superordonné (l'hyperonyme) de *bulimia, anorexia*. Trouver la classe sémantique de la maladie avec *eating* provient aussi de l'ambiguïté syntaxique de ce mot qui peut être à la fois verbe et nom, ce qui n'est pas le cas en français (à part *le manger* qui est un nom obsolète ou maintenu dans des locutions comme *le boire et le manger*). Cela se confirme par la recherche des collocats de *to eat* qui se superposent à ceux de *manger*. Les enseignants novices dans l'emploi des corpus sont donc rendus attentifs aux interprétations culturelles rapides qui ne tiennent pas compte des phénomènes linguistiques.

Concordances et domaines de spécialité

58 L'exemple de *patient navigation* en Section 4.1 soulevait de manière implicite la question des domaines de spécialité. En effet, il s'agit ici d'un terme lié au domaine médical et donc spécialisé, dont on ne trouve pas de définition dans les dictionnaires généraux. La présente section introduit auprès des enseignants la notion de domaine spécialisé et de termes du domaine, à savoir, le vocabulaire spécialisé. Avant de travailler sur des corpus spécialisés, nous proposons de commencer par une recherche de collocation sur un échantillon de concordance qui permet d'expliciter la notion de domaine spécialisé par

les différences de collocations d'un même mot. Ce travail peut se faire avec des concordances sur papier, triées en fonction de l'objectif pédagogique. Utiliser des concordances éditées et imprimées permet tout à fait de faire découvrir des collocations potentielles et d'analyser les contextes d'apparition des mots (Boulton, 2010), surtout quand le travail sur collocations a déjà été longuement réalisé avec un corpus comme celui de Leipzig.

59 Prenons l'exemple édité de *dépression* en Tableau 19, extrait du corpus composé d'un an du journal *Le Monde* et qui est accessible sur le concordancier en ligne, *Wall.*

Tableau 19 – Échantillon de concordances pour dépression.

	Dépression	centrée sur l Europe
dans les orages de la "Grande	**Dépression**	", l existence itinérante qui suivit
par la réalité sauvage de la Grande	**Dépression**	
plus souvent que les banlieusards de	**Dépression**	et de troubles de l anxiété. 6 %
Pearl est tombée malade (un début de	**Dépression**	nerveuse), il a bien fallu placer
S'étant réfugié dans une	**Dépression**	naturelle du terrain, l'animal
puissant du village et gagne. En cas de	**Dépression**	économique ou politique, cette
peut espérer qu'entre-temps le cycle de	**Dépression**	mondiale sera terminé, que la
semble sombrer dans une sorte de	**Dépression**	collective, avec des conséquences
après quelques jours la	**Dépression**	tropicale s'est éloignée
juridique adjoint souffrait de	**Dépression**	depuis plusieurs semaines sinon
Saïd Magri, souffrirait de	**Dépression**	et aurait entamé une grève de la faim.
1957, il traverse une période de grave	**Dépression**	: épuisement, répercussions de la
Au bord d'une grave	**Dépression**	, frère Jacob éclate de rire en
explique-t-elle, souffre d'une grave	**Dépression**	. N'obtenant ni le passeport requis ni
de la Résistance, victime d'une grave	**Dépression**	après la publication de son
peine, après avoir traversé une grave	**Dépression**	qui lui fit remettre en question son
à l'année 1992, au plus profond de la	**Dépression**	boursière, pour retrouver des cours
et ont contribué à déclencher "une	**Dépression**	économique sans précédent en temps de
mais ne nous plongera pas dans une	**Dépression**	économique, affirme Nancy Bolton,

à 900 mètres en retrait, dans une	**Dépression**	naturelle du terrain, il a fallu leur
front des Flandres avant de céder à une	**Dépression**	nerveuse, le confirme dans ses plus
rigidité par Michèle Moretti) fait une	**Dépression**	nerveuse après que le soldat qu'elle
plaquée par son amant, fait une	**Dépression**	nerveuse, tente de se suicider, puis se
avaient été fortement perturbé par une	**Dépression**	tropicale. Les caprices de la météo
forte	**Dépression**	centrée sur l'Europe

60 Cette concordance permet de faire prendre conscience aux utilisateurs novices des points suivants :

- dépression se retrouve souvent avec les mêmes collocats qui le suivent ou le précèdent immédiatement ;
- un locuteur natif du français peut reconnaître des collocations, même si celles-ci ne sont pas répétées dans la concordance ;
- selon le collocat, le mot n'a pas le même sens et appartient à des domaines différents : médical, économique, géographique, météorologique – par exemple grave dépression n'appartient pas au même domaine que forte dépression ;
- certains éléments lexicaux, tels que des verbes, ont souvent dépression pour objet – ces verbes ne sont pas accolés directement à dépression ;
- certains éléments qui sont accolés à dépression peuvent être considérés comme venant construire un nom composé lexicalisé.

61 Ce type de travail peut être fait plusieurs fois, sur des concordances éditées en français et en anglais, dans des corpus encore non spécialisés, comme le *BNC* et le *Coca* pour l'anglais.

62 Effectuer ce type de tâche sur des concordances conduit ensuite à pointer les différentes fonctionnalités et les différents langages des concordanciers en ligne ou téléchargeables. Nous travaillons pour notre part sur l'interface du *Coca*, de *WebCorp* et *d'IMS Corpus WorkBench*, ainsi que sur *AntConc* qui est gratuit, s'installe facilement, permet d'introduire les expressions régulières de manière simple et propose un tutorat en ligne.

63 Outre ces question d'outils, il ne faut pas s'arrêter à la question de la collocation, mais bien sûr, aborder les autres phénomènes collocationnels qui ne se situent plus au niveau du lexique comme pour la collocation, mais au niveau syntaxique pour la colligation, sémantique pour la préférence sémantique et pragmatique pour la prosodie sémantique que nous avons déjà partiellement évoquées.

Google ou corpus ?

64 Outre que les corpus demandent un apprentissage, les arguments que l'on oppose souvent à l'utilisation de corpus par les enseignants de Lansad relèvent du fait qu'il est déjà possible de trouver des exemples et des réponses sur *Google*. La différence entre les corpus et *Google* réside dans le fait que les corpus ont été collectés avec un objectif précis et que l'on connaît le contenu des documents qu'ils réunissent. Sur *Google*, et en anglais particulièrement, on trouve de nombreux documents rédigés par des locuteurs non natifs. Par ailleurs, de nombreux sites, comme les blogs justement, sont rédigés dans un

registre de langue qui ne correspond pas toujours aux objectifs de l'enseignement. Enfin, notamment en ce qui concerne les langues de spécialité, les informations trouvées sur *Google* ne sont pas toujours fiables. On peut lancer des recherches sur *Google*, mais en vérifiant la validité des sources et en abordant les résultats linguistiques avec précaution. Nous donnons ici un exemple issu d'une interrogation d'un étudiant sur la validité de l'expression *it's noted* en anglais, comme équivalent de *c'est noté* en français. Une recherche de l'expression anglaise sur *Internet* donne 17 000 résultats. Cependant, ces résultats sont présentés de telle sorte qu'il faut aller sur chaque site pour vérifier si le contexte d'emploi correspond bien à celui dans lequel on trouve *c'est noté* en français. Certains contextes correspondent, mais on ignore si les auteurs sont ou non des locuteurs natifs. Une simple recherche sur le *Coca* ou le *BNC* donnera des résultats indiquant que l'expression n'est pas utilisée en anglais comme en français, comme le montre le Tableau 20.

TABLEAU 20 – *it's noted* dans *Google*, *Coca* et le *BNC*.

Google : 170 000 occurrences
It's noted, it's noted down, it's noted that
COCA : 5 occurrences:
It 's noted that Jack is on oxygen
It 's noted for its long, fleshy snout
it 's noted for helping parents bridge the gap
BNC : 1 occurrence: and already **it 's noted** for its cuisine and cellar

65 On constate tout d'abord qu'il y a très peu d'occurrences dans les deux corpus, contrairement à *Internet*. En outre, celles-ci entrent dans des structures qui ne reflètent pas l'emploi du français et ont donc un autre sens.

S'approprier un domaine de spécialité

66 Une fois cette introduction aux corpus généraux terminée, on peut aborder les domaines de spécialité. Le travail sur corpus en langue de spécialité ne diffère que peu du travail sur la langue générale au niveau des questions de collocations et d'autres phénomènes. La différence réside dans le fait que les enseignants d'anglais sont en général peu familiers des domaines disciplinaires et doivent donc se les approprier du point de vue de la compréhension et de la différence entre la phraséologie de la langue générale et celle d'une langue de spécialité donnée. En anglais de spécialité, les corpus accessibles en ligne et gratuits ne sont pas légion. Il est donc fort probable que l'enseignant doive constituer lui-même son corpus spécialisé. On peut donc former à des outils comme *WebBootCaT* (Baroni *et al.*, 2006), mais aussi aux logiciels permettant de transformer des fichiers pdf en txt, afin de constituer un corpus "manuellement".

67 L'utilisation du logiciel *AntConc* (Anthony, 2011) est adaptée à cette problématique, car celui-ci permet d'effectuer un grand nombre de manipulations complexes, tout en restant très facile d'accès pour des débutants. Son seul défaut est de limiter la taille des corpus à environ six millions de mots ; au-delà de ce seuil, le traitement des corpus s'arrête. Ce

seuil reste raisonnable pour arriver à s'approprier un domaine de spécialité. La maîtrise d'*AntConc* donne l'accès à d'autres ressources en ligne dans lesquelles l'interrogation peut se faire en utilisant des expressions régulières, comme le corpus *Europarl* sur *IMS Corpus WorkBench* et le corpus de sciences de la terre que nous avons mis en ligne sur une version personnalisée de IMS. Les exemples donnés ci-dessous sont tirés de ce corpus que nous avons créé au fil des ans, avec l'aide des étudiants de Master 1 en traduction spécialisée, mais aussi d'autres petits corpus spécialisés créés par des étudiants de Master 2.

68 Les enseignants d'anglais de spécialité, comme les traducteurs pragmatiques, peuvent donc être confrontés à des domaines qu'ils ne connaissent pas encore et qu'ils doivent s'approprier. La création et l'interrogation de corpus spécialisés les aide à découvrir un domaine. En outre, cette phase de découverte peut leur permettre de préparer du matériau pédagogique pour des exercices de compréhension de lexique spécialisé par exemple.

Termes et définitions

69 Nous proposons d'abord un travail sur la définition de termes du domaine. En effet, les termes d'un domaine ne sont généralement pas définis dans les dictionnaires généraux, et les lexiques spécialisés ou banques de données terminologiques ne sont jamais complets, en raison de l'évolution constante des différents domaines qui génèrent des néologismes.

70 La recherche de définitions de termes en corpus se fait à l'aide de marqueurs linguistiques de définition comme l'ont déjà démontré Pearson (1998) et Péraldi (2010). Nous en donnons quelques exemples dans le Tableau 21.

Tableau 21 – Recherche de définitions à l'aide du marqueur is called.

Recently, the most typical activity of Nakadake Volcano has been continuous fallout of black sandy ash from dark eruption plume. This **is called** an ash eruption
For gas liquid flow in vertical columns with low gas flow rates and small bubbles relative to the conduit diameter, bubbles are more or less randomly dispersed and move upward through the liquid phase without much dynamical interaction. This **is called** the bubble flow regime

71 Dans ce tableau, on découvre deux termes et leurs définitions : *ash eruption,* qui est relativement facile à appréhender, et *bubble flow regime*, plus complexe. Cela amène à se poser la question de la définition de *regime* et à découvrir tous les autres types de *regime* du domaine, comme par exemple, *differential flow regime, melting regime, dislocation creep regime*. À partir de là, on recherche les définitions de *creep, flow, melting*, etc. en avançant petit à petit dans la compréhension du domaine.

72 La recherche de contextes définitoires à l'aide de marqueurs permet aussi de comprendre l'organisation du vocabulaire spécialisé. Le Tableau 22 par exemple montre une organisation hiérarchique entre un terme superordonné (un hyperonyme) et ses termes subordonnés (les co-hyponymes).

Tableau 22 – Contextes hiérarchisés à parti du marqueur main types.

The three **main types** of terrestrial lava are basaltic, andesitic, and rhyolitic . Terrestrial basaltic flows erupt at temperatures between 1000 and 1400 °C .
The viscosity of basaltic lava amounts to View the MathML source at the liquidus temperature. Basaltic lavas are high in Ca, Mg, and Fe, and low in Si, Na, and K. The high temperature and low silica content allow basaltic lavas to flow readily .
Rhyolitic lavas display viscosities of View the MathML source at their liquidus temperature.

73 On comprend ici qu'il existe des *terrestrial lava* de différents types dont les principaux sont *basaltic lava, andesitic lava* et *rhyolitic lava*.

74 Enfin, les deux exemples suivants (Tableaux 23 et 24) permettent de montrer aux enseignants que le vocabulaire spécialisé, la terminologie d'un domaine, n'est pas, contrairement à ce que l'on pense, complètement figé et que certains termes ne sont pas totalement acceptés par la communauté de discours. La conséquence en est qu'il faut rendre les étudiants Lansad attentifs à la terminologie qu'ils utilisent dans leur domaine qu'ils croient souvent bien connaître. Les marqueurs permettant de trouver des contextes déterminant cette incertitude terminologique sont, par exemple, *a type of*, *a sort of*.

TABLEAU 23 – Exemples d'incertitude terminologique ; les termes sont en italiques.

For example, [Ibanez et al., 2000] initially used the term *hybrids* for **a type of** LP events with a strong high-frequency initial phase, due to their spectral properties.
These results suggest that the active volcanoes in northeast China are not hotspots but **a sort of** *back-arc volcanoes* which are closely related to the subduction process of the Pacific slab.

75 Le marqueur *the term* est utilisé aussi bien pour évoquer un terme bien installé dans la communauté de discours, que pour préciser les conditions dans lesquelles un terme doit s'utiliser. On précise donc le statut des termes, c'est-à-dire que ceux-ci et non d'autres doivent être utilisés. Cela reflète un problème d'imprécision dans l'utilisation des termes par la communauté de discours. Dans ce cas, l'utilisation de guillemets n'est pas significative, alors qu'on l'évoque souvent comme marque d'un terme non encore installé.

TABLEAU 24 – Exemples d'un terme en cours d'installation et d'un terme installé.

It is therefore suggested here to use **the term** 'wrinkle structure' only if microbial participation is likely but a clear classification not possible.
Otherwise, use of **the** well introduced **terms** 'Kinneyia' and 'elephant skin' is recommended here if these structures can be clearly identified

76 Ces exemples montrent d'une part comment mieux aborder un domaine en recherchant des contextes définitoires et explicatifs, mais aussi comment l'utilisation de ces marqueurs linguistiques amène à mieux comprendre la structuration terminologique du

domaine et les différentes relations existant entre les termes : hyperonymie/hyponymie, cause-effet, partie-tout, synonymie/antonymie, terme recommandé / terme à éviter, etc.

Phénomènes collocationnels en langue de spécialité

77 Une fois le domaine découvert, les enseignants doivent être capables d'aborder le corpus de la même manière que les corpus généraux. Il s'agit notamment de relever les phénomènes collocationnels, comme dans un exemple issu d'un corpus de prévisions météorologiques dans lequel on voit que *rain* présente un phénomène de colligation (Tableau 25), car il se trouve fréquemment en position de sujet de verbes et qu'il présente aussi un phénomène de préférence sémantique puisque la majorité des verbes appartiennent à la classe sémantique du mouvement. S'y ajoute un phénomène de prosodie sémantique. La prosodie sémantique est le fait, pour un mot d'entrer en cooccurrence avec un ensemble de mots partageant la même connotation. Ici, le terme *rain* a une prosodie sémantique négative, car il est entouré de mots à connotation négative, comme par exemple *fog, mist, threaten, cloud, cloudy, outbreaks, heavy*.

Tableau 25 – Extrait de concordances du mot rain.

parts of England towards dawn. Outlook: On Sunday, rain will push southeast across Scotland, Northern Irela
/ night with clear spells but some mist or fog will form. Rain will threaten the far north-west of Scotland later o
dry and bright, with some pleasant sunshine in the W. Rain moving into the far SW later in the day. Mild. Out
uch milder everywhere, although it will also be windy. Rain will sweep eastwards over much of the country, bu
me drier and brighter through the day as the cloud and rain clears south-eastwards. The rest of England and W
Today A band of cloud and rain will move east over northwestern Scotland and No
n on Wednesday. Turning cloudy with patchy late-day rain, however, in Northern Ireland, westernmost Wales
ID: Mostly cloudy and chilly Sunday with isolated light rain and sleet showers or patchy drizzle, mostly in the w
along the east coast of England with very isolated light rain patches.
vercast on Sunday with a few patches of drizzle or light rain, but dry and bright in some parts. Gentle to moder
to the NE of the UK will give outbreaks of patchy light rain and drizzle across northern parts of Scotland tonigl
Bright spells early Monday in Leinster with patchy light rain elsewhere, spreading eastward later in the day. Son
sheltered glens. Mild and wet tonight with outbreaks of rain, heavy at times across England and Wales. Lows ot
tish Isles through Thursday. There will be outbreaks of rain, most of it light. The highest chance of rain is over
Tonight UK: Outbreaks of rain, a few maybe heavy, on Thursday night in norther
Tuesday there will be showers and longer outbreaks of rain across England, Wales and Northern Ireland, movi
s may be merging into some more pe rsistent periods of rain, turning increasingly sleety. Thunder is likely at an
neavy, on Thursday night in northern Scotland. Patchy rain in southern Scotland, Northern Ireland and in far n

78 L'étude des concordances est nécessaire pour relever ces phénomènes et faire prendre conscience aux enseignants de ce à quoi ils doivent rendre leurs étudiants attentifs.

Phénomènes généraux en langue de spécialité

79 Nous avons argumenté plus haut que l'acquisition de structures générales pouvait se faire à partir de la langue de spécialité. L'exemple suivant montre que le corpus spécialisé permet aussi de traiter des questions de grammaire et qu'il est possible de le faire sur des corpus spécialisés, même avec des apprenants de faible niveau.

80 Une connaissance de base des expressions régulières est alors nécessaire aux enseignants, car elle permet de lancer des requêtes plus complexes afin de préparer du matériel pédagogique portant sur des questions simples ou complexes. Les apprenants de faible niveau ont par exemple tendance à confondre les formes en *-ed* et en *-ing*. Une recherche à l'aide d'expressions régulières sur des formes passives et des formes continues dans un corpus spécialisé (en Tableau 26, le corpus de sciences de la terre) permet à la fois de

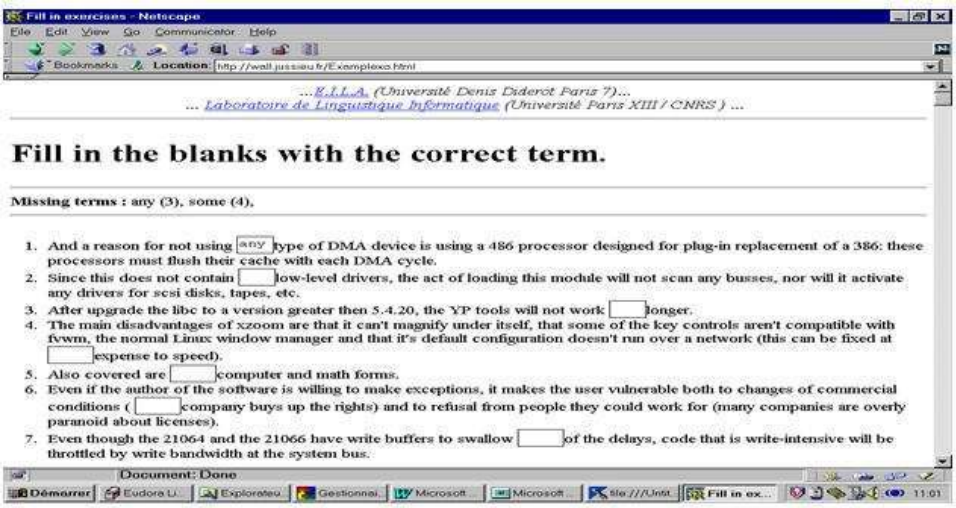
présenter des exemples de leur utilisation, mais aussi de préparer des exercices lacunaires par exemple, dans lesquels l'apprenant doit rétablir le verbe à la forme correcte.

Tableau 26 – Exemples de recherche de structures en –ed ou –ing.

Expression régulière simple : ha(s|ve|d) been \w+(ed|ing)
2008 Wenchuan earthquake **had been accumulating** for a much longer time.
a deeper silicic magma chamber that **had been cooling** slowly.
the lava dome,which **had been growing** since 1995, had not resumed growth
lava dome **had been growing** since the large
a pre-event **had been observed** about 2.8 s before the main rupture
Such branching **had been studied** earlier on the basis of static crac
This regional amplification **had been suggested** by the analysis of the very few records

81 Nous ne donnons ici qu'un exemple du type de travail ou d'exercices pouvant être réalisés à l'aide de corpus. Cet exemple est issu de l'outil *Wall*, qui permet non seulement de générer des concordances à partir d'expressions régulières, mais aussi de générer automatiquement des exercices lacunaires à partir des concordances, comme l'illustre le Tableau 27.

Tableau 27 – Exemples de génération de texte lacunaire sur some et any.



82 Il ressort clairement de ces exemples que les enseignants doivent apprendre à maîtriser la syntaxe des expressions régulières.

Conclusion

83 Nous avons montré comment amener les enseignants à traiter par le corpus des questions lexicales, dans une approche lexico-grammaticale, des questions de grammaire et des questions de contenu. Cependant, pour arriver à un tel traitement, il est nécessaire de former les enseignants à des compétences relativement avancées en linguistique de

corpus, tant sur les plans théorique et méthodologique, que sur le plan technique. Nous argumentons en faveur d'une approche corpus qui commence par un travail sur les collocats, avant de montrer des lignes de concordances. La décision de former les enseignants à cette approche ne relève pas seulement de la didactique, mais avant tout de choix politiques. Les étudiants qui se destinent à l'enseignement des langues pourraient être formés à la linguistique de corpus dans leur cursus universitaire. Cela peut paraître un point de vue partisan, car pourquoi cette approche plutôt qu'une autre ? La réponse réside dans les fondements de la linguistique de corpus, à savoir que l'on découvre et comprend la manière de communiquer dans une langue à partir du corpus, le corpus étant un réservoir inépuisable de découvertes, d'exemples et d'activités. L'exploitation de corpus représente l'une des approches possibles, parmi les nombreuses approches à mettre en œuvre dans la formation en langues. C'est en fait une question beaucoup plus vaste sur la formation des enseignants qui se pose ici.

BIBLIOGRAPHIE

Aarts, B. 2001. "Corpus linguistics, Chomsky and fuzzy tree fragments". In Mair, C. & Hundt, M. (dir.). *Corpus linguistics and linguistic theory*. Amsterdam: Rodopi. pp. 5-13.

Anthony, L. (2005). "AntConc : design and development of a freeware corpus analysis toolkit for the technical writing classroom". In *Professional Communication Conference Proceedings*. pp. 729-737. Disponible en ligne. http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=1494244&url=http%3A%2F%2Fieeexplore.ieee.org%2Fstamp%2Fstamp.jsp%3Ftp%3D%26arnumber%3D1494244

Anthony, L. (2011). *AntConc*, version 3.2.4. Tokyo: Waseda University. Disponible en ligne. http://www.antlab.sci.waseda.ac.jp/

Antoniadis, G., Ponton, C. & Zampa, V. (2010). "Exxelant et Mirto : deux exemples d'environnement d'ALAO intégrant des outils TAL". In Biskri, I. & Jebali, A. (dir.). *Multilinguisme et traitement des langues naturelles*. Montréal : Presses de l'Université du Québec. pp. 165-179.

Aston, G. 1997. "Enriching the learning environment : corpora in ELT". In Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (dir.). *Teaching and language corpora*. Harlow: Addison Wesley Longman. pp. 255-266.

Aston, G. 1999. "Corpus use and learning to translate". *Textus*, vol. 12. pp. 289-313.

Baroni, M., Kilgarriff, A., Pomikálek, J. & Rychlý, P. (2006). "WebBootCaT : instant domain-specific corpora to support human translators". In *Proceedings of EAMT 2006: 11th Annual Conference of the European Association for Machine Translation*. Oslo. pp. 247-252.

Bernardini, S. (2004). "Corpora in the classroom: an overview and some reflections on future developments". In Sinclair, J. M. (dir.). *How to use corpora in language teaching.* Amsterdam: John Benjamins. pp. 15-36.

Biemann, C., Bordag, S., Heyer, G., Quasthoff, U. & Wolff, C. (2004). "Language-independent methods for compiling monolingual lexical data". In *Computational linguistics and intelligent text processing*. Berlin/Heidelberg: Springer. pp. 217-228.

Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). "The Leipzig corpora collection: monolingual corpora of standard size". In *Proceedings of Corpus Linguistics 2007*. Disponible en ligne. http://ucrel.lancs.ac.uk/publications/CL2007

Boulton, A. (2008). "Esprit de corpus : promouvoir l'exploitation de corpus en apprentissage des langues. *Texte et Corpus*, vol. 3. pp. 37-46.

Boulton, A. (2010). "Data-driven learning : taking the computer out of the equation". *Language Learning*, vol. 60, n° 3. pp. 534-572.

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

Bourigault, D. (2002). "Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus". In *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2002)*. pp. 75-84. Disponible en ligne. http://www.loria.fr/projets/JEP-TALN/TALN

Bowker, L. & Pearson, J. (2002). *Working with specialised language: a guide to using corpora*. Londres : Routledge.

Christ, O., Schulze, B. M., Hofmann, A. & König, E. (1999). *The IMS Corpus Workbench Corpus Query Processor (CQP): user's manual*. Stuttgart: University of Stuttgart.

Davies, M. (2002). *Corpus del Español : 100 million words, 1200s-1900s*. Disponible en ligne. http://www.corpusdelespanol.org

Davies, M. (2004). *BYU-BNC*. Disponible en ligne. http://corpus.byu.edu/bnc/

Davies, M. (2008-). *The Corpus of Contemporary American English: 450 million words, 1990-present*. Disponible en ligne. http://corpus.byu.edu/coca/

Davies, M. & Ferreira, M. (2006). *Corpus do Português : 45 million words, 1300s-1900s*. Disponible en ligne. http://www.corpusdoportugues.org

Eensoo-Ramdani, E., Bourion, E., Slodzian, M. & Valette, M. (2011). "De la fouille de données à la fabrique de l'opinion : enjeux épistémologiques et propositions". *Les Cahiers du Numérique*, vol. 7, n°2. pp. 15-39.

Fligelstone, S. (1993). "Some reflections on the question of teaching, from a corpus linguistics perspective". *ICAME Journal*, vol. 17. pp. 97-109.

Foucou, P.-Y. & Kübler, N. (2000). "A web-based environment for teaching technical English". In Burnard, L. & McEnery, T. (dir.). Rethinking language pedagogy from a corpus perspective. Francfort : Peter Lang. pp 65-74.

Frankenberg-Garcia, A. (2012). "Raising teachers' awareness of corpora". *Language Teaching*, vol. 45, n° 4. pp. 475-489.

Gavioli, L. (1997) "Exploring texts through the concordancer : guiding the learner". In Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (dir.). *Teaching and language corpora*. Harlow: Addison Wesley Longman. pp. 83-99.

Gavioli, L. (2005). *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.

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

Johns, T. (1997). "Contexts: the background, development and trialling of a concordance-based CALL program". In Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (dir.). *Teaching and language corpora*. Harlow: Addison Wesley Longman. pp. 100-115.

Kübler, N. (2011). "Working with different corpora in translation teaching". In Frankenberg-Garcia, A., Flowerdew, L. & Aston, G. (dir.). New trends in corpora and language learning. Londres : Continuum. pp. 62-80.

Kübler, N. & Foucou, P.-Y. (1999). "A web-based language learning environment: general architecture". In Schulze, M., Hamel, M.-J. & Thompson, J. (dir.). Language processing in CALL. *ReCALL*. pp 31-39.

Landure, C. & Boulton, A. (2010). "Corpus et autocorrection pour l'apprentissage des langues". *ASp*, vol. 57. pp. 11-30.

Léon, J. (2007). "Meaning by collocation: the Firthian filiation of corpus linguistics". In Kibbee, D. (dir.). *Proceedings of ICHoLS X, 10th International Conference on the History of Language Sciences*. Amsterdam: John Benjamins. pp. 404-415.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M. & Den, Y. (2014). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, vol. 48, n° 2. pp 345-371.

McEnery, A. & Xiao, Z. (2004). "The Lancaster Corpus of Mandarin Chinese: a corpus for monolingual and contrastive language study". Communication présentée à LREC 2004. Lisbonne.

Minugh, D. (1997). "All the language that's fit to print: using British and American newspaper CD-ROMs as corpora". In Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (dir.). *Teaching and language corpora*. Harlow: Addison Wesley Longman. pp. 255-266.

Pearson, J. (1998) *Terms in context.* Amsterdam : John Benjamins.

Péraldi, S. (2011). *Indétermination terminologique et multidimensionnalité dans le domaine de la chimie organique.* Thèse de doctorat, Université Paris Diderot.

Renouf, A. (1997). "Teaching corpus linguistics to teachers of English". In Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (dir.). *Teaching and language corpora*. Harlow: Addison Wesley Longman. pp. 255-266.

Renouf, A. (2003). "WebCorp : providing a renewable data source for corpus linguists". *Language and Computers*, vol. 48, n° 1. pp. 39-58.

Römer, U. (2006). "Pedagogical applications of corpora: some reflections on the current scope and a wish list for future developments". *Zeitschrift für Anglistik und Amerikanistik*, vol. 54, n° 2. pp. 121-134.

Scott, M. (2008). *WordSmith Tools*, version 5. Liverpool: Lexical Analysis Software. Disponible en ligne. http://www.lexically.net/publications/citing_wordsmith.htm

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Tadić, M. (2009). New version of the Croatian National Corpus. In Hlaváčková, D., Horák, A., Osolsobě, K. & Rychlý, P. (dir.). *After half a century of Slavonic natural language processing*. Brno: Masaryk University Press. pp. 199-205.

Teubert, W. (1999). "Corpus linguistics : a partisan view". *TELRI Newsletter*, vol. 8. Disponible en ligne. http://telri.nytud.hu/telri2/newsletter/newsl8.html

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: John Benjamins.

Tono, Y. (2011). "TaLC in action: recent innovations in corpus-based English language instruction in Japan". In Frankenberg-Garcia, A., Flowerdew, L. & Aston, G. (dir.). New trends in corpora and language learning. Londres: Continuum. pp. 3-25.

Tyne, H. (2012). "Corpus work with ordinary teachers: data-driven learning activities". In Thomas, J. & Boulton, A. (dir.). *Input, process and product: developments in teaching and language corpora*. Brno: Masaryk University Press. pp. 114-129.

Zanettin, F. (2002). "Corpora for translation practice". In Yuste-Rodrigo, E. (ed.). *Language resources for translation work and research, LREC 2002 Workshop Proceedings*. Las Palmas de Gran Canaria. pp. 10-14.

Corpus et outils

AntConc : http://www.antlab.sci.waseda.ac.jp/software.html

Balanced Corpus of Contemporary Written Japanese : http://www.ninjal.ac.jp/english/products/bccwj

BNC, *Coca, Corpus del Español, Corpus do Português* : http://corpus.byu.edu/

Business Letter Corpus : http://www.someya-net.com/concordancer

COMPARA : http://193.136.2.104/COMPARA/psimples.php?language=en

Croatian National Corpus : http://www.hnk.ffzg.hr

Europarl : http://opus.lingfil.uu.se/cwb/Europarl7/frames-cqp.html

IMS Corpus Workbench P7 : https://cwb-test.eila.univ-paris-diderot.fr/ims/index.php

Lancaster Corpus of Mandarin Chinese : http://www.lancaster.ac.uk/fass/projects/corpus/LCMC

Leipzig Corpora Collection : http://corpora.uni-leipzig.de

Les Voisins de Le Monde : http://redac.univ-tlse2.fr/voisinsdelemonde

SketchEngine : http://www.sketchengine.co.uk

Wall : http://wall.eila.univ-paris-diderot.fr/dyn/Context2

Web as Corpus : http://webascorpus.org

WebBootCaT : http://www.sketchengine.co.uk/documentation/wiki/SkE/Help/WebBootCat

WebCorp : http://www.webcorp.org.uk/live

NOTES

1. Langues pour spécialistes d'autres disciplines.
2. La traduction pragmatique, évoquée par Newmark (1988 : 133) se concentre sur l'effet perlocutoire qu'elle a sur le lecteur. La manière dont un texte est traduit dépend du type de lecteur à qui elle s'adresse, ainsi que des objectifs du texte source (Kübler & Aston, 2012).
3. Ce pourcentage est tiré des tests de niveau obligatoire que tous les étudiants de L1 (sauf les étudiants en anglais et LEA) doivent passer en entrant à l'Université Paris Diderot.
4. "Enseigner sur", c'est-à-dire, enseigner la théorie et les principes fondamentaux de la linguistique de corpus.
"Enseigner l'utilisation", à savoir enseigner les aspects concrets, les outils de la linguistique de corpus.
"Utiliser pour enseigner", à savoir créer du matériau pédagogique à partir des corpus.

"Enseigner pour créer des ressources", c'est-à-dire, apprendre à créer des corpus. (notre traduction)

5. Les enseignants ont sans doute besoin d'aide sur la manière d'interpréter les résultats de l'interrogation de corpus.

6. Prenons deux mots, A, B apparaissant chacun un nombre de fois a, b, dans les phrases du corpus et k fois ensemble. Nous calculons la signifiance sig (A, B) de leurs occurrences dans une phrase de la manière suivante. Deux types de collocations sont générés : les collocations basées sur le fait d'apparaître dans la même phrase et celles qui comporte le voisin droit immédiat et le voisin gauche immédiat du mot recherché.

7. Comprendre le sens à partir de la collocation représente une abstraction au niveau syntagmatique. Ça n'est pas directement lié à l'approche conceptuelle ou "idéale" du sens du mot. L'un des sens de *night* (nuit) est le fait que ce mot a pour collocation *idarki* (noire); de même pour *dark*, le fait qu'il peut avoir pour collocat *night*.

8. La collocation concerne l'occurrence de deux ou plusieurs mots à proximité dans un texte.

9. http://wall.eila.univ-paris-diderot.fr

10. Par ailleurs, ces deux occurrences se retrouvent dans le même article, comme le montre le contexte élargi suivant : *(...) écologistes l'" Initiative des Alpes " était soutenue (...) C'est (...) dans le Tessin de langue italienne que l'Initiative a enregistré ses meilleurs scores, avec respectivement 87,6 % et 63,8 % de " oui ".*

RÉSUMÉS

Natalie Kübler est Professeur des Universités à l'Université Paris Diderot Paris 7. Ses recherches portent sur la linguistique de corpus, les langues de spécialité, l'apprentissage sur corpus et la traduction spécialisée ; elle s'intéresse particulièrement à la phraséologie en anglais et en français, ainsi qu'à la prosodie sémantique en langues de spécialité. Il y a 20 ans, elle a introduit la linguistique de corpus dans l'enseignement de l'anglais de spécialité à l'Université Paris 13. Par la suite, elle a été la première à introduire la linguistique de corpus dans une formation en traduction spécialisée, il y a presque 15 ans. Aujourd'hui, elle enseigne la linguistique de corpus appliquée à la terminologie et à la traduction spécialisée. Elle dirige par ailleurs le laboratoire CLILLAC-ARP et le Centre de Ressources Informatiques en Langue de l'Université Paris Diderot.
Courriel : nkubler@eila.univ-paris-diderot.fr.
Toile : http://www.eila.univ-paris-diderot.fr/user/natalie_kuebler.
Adresse : Laboratoire CLILLAC-ARP, case 7002, Université Paris Diderot, 8 Place Paul Ricoeur, 75205 Paris cedex 13, France.

Dans cet article, nous posons la question de la formation des enseignants de langues aux spécialistes d'autres disciplines à l'université, à la linguistique de corpus. L'objectif est d'amener les enseignants à adopter une approche lexico-grammaticale pour aborder les questions de lexique, de grammaire, de contenu et de phraséologie dans la langue générale et les langues de spécialité. Nous cherchons à démontrer la nécessité de former les enseignants à l'approche théorique et méthodologique de la linguistique de corpus, afin qu'ils acquièrent des compétences relativement avancées, tant sur les plans théorique, que méthodologique et technique. Cela leur permet ensuite d'utiliser les corpus dans leurs enseignements, soit dans une approche inductive directe avec les étudiants, soit dans une approche indirecte pour préparer du matériau pédagogique.

This paper deals with the issue of training language teachers at university level into adopting a corpus linguistics approach in their teaching. The aim is to help them use a lexcio-grammatical approach in dealing with lexicon, grammar, phraseology, content in general language and in languages for specific purposes. This paper tackles the issue of training language teachers to relatively advanced levels of skills and competences in theoretical, methodological and technical aspects of corpus linguistics. It will be shown that this is needed in order to allow them to use corpora in the classroom, either in a direct inductive approach, or in an indirect approach for preparing teaching materials.

INDEX

Mots-clés : linguistique de corpus, formation des enseignants, acquisition des langues, langues de spécialité
Keywords : corpus linguistics, teacher training, language acquisition, languages for specific purposes