Cahiers de praxématique

54-55 | 2010
Corpus, données, modèles

Linguistique de corpus et didactique du F.L.E. Une exploitation du corpus *IntUne*

Corpus Linguistics and French as a Foreign Language : Using the IntUne *Corpus*

Delphine Giuliani et Radia Hannachi

Édition électronique
URL : http://journals.openedition.org/praxematique/1136
DOI : 10.4000/praxematique.1136
ISSN : 2111-5044

Éditeur
Presses universitaires de la Méditerranée

Édition imprimée
Date de publication : 1 janvier 2010
Pagination : 145-160
ISBN : 978-2-36781-012-6
ISSN : 0765-4944

Référence électronique
Delphine Giuliani et Radia Hannachi, « Linguistique de corpus et didactique du F.L.E. Une exploitation du corpus *IntUne* », *Cahiers de praxématique* [En ligne], 54-55 | 2010, document 8, mis en ligne le 01 janvier 2013, consulté le 08 septembre 2020. URL : http://journals.openedition.org/praxematique/1136 ; DOI : https://doi.org/10.4000/praxematique.1136

Cahiers de praxématique 54-55, 2010, 145-160

Delphine Giuliani, Radia Hannachi
Laboratoire H.C.T.I., université de Bretagne Sud

Linguistique de corpus et didactique du F.L.E. Une exploitation du corpus *IntUne*

Introduction

La linguistique de corpus est un paradigme qui connaît depuis les vingt dernières années un succès croissant (Williams 2006). Cependant, sous une même appellation, de nombreuses pratiques coexistent plus ou moins pacifiquement, faisant courir aux corpus le risque d'être galvaudés.

Ces derniers peuvent ainsi être clos ou bien ouverts et dynamiques, échantillonnés ou composés de textes intégraux uniquement, avoir une finalité exhaustive ou non (Mellet 2002), permettant un traitement automatique ou non (Patrascu 2011). Par ailleurs, la notion de corpus dans le domaine des sciences humaines et sociales et plus particulièrement dans le champ de la linguistique, connaît un succès croissant. De nombreux ouvrages de référence ont été produits, différentes revues lui sont dédiées et de nombreux colloques sont organisés autour de cette thématique [1]. Le risque est alors grand de voir la notion de corpus se diluer :

> Le corpus, la notion et l'objet, risque d'être victime aujourd'hui en France de son succès. Plus une discipline, plus un comité scientifique, plus un chercheur qui n'y fasse référence ; plus un linguiste, surtout, qui ne le manipule, le caresse ou le maltraite. (2005 : 1)

1. Concernant les revues, citons *Corpus* : `http://corpus.revues.org/`, *Textes et corpus* : `http://web.univ-ubs.fr/corpus/publi.html` ou bien *International Journal of Corpus Linguistics*, `http://benjamins.com/catalog/ijcl`. Nous pouvons également relever l'organisation au cours du dernier trimestre 2012 de plusieurs rencontres sur cette thématique : « La linguistique de corpus à l'heure de la confrontation entre concepts, techniques et application » organisé par Bordeaux 3, la journée des Jeunes Chercheurs du MoDyCo : outils et méthodes pour le traitement de corpus, le colloque Nomico, nominalisations et corpus organisé à Nancy, etc.

Différents ouvrages de référence ont été rédigés dans le domaine afin de pallier cette difficulté épistémologique. Nous pouvons noter celui de Habert *et al.* (1997) dont le titre : *Les linguistiques de corpus*, met en relief l'hétérogénéité du domaine. D'autres revues ou ouvrages existent, mais ils rassemblent le plus souvent des travaux hétérogènes (Cori *et al.* 2008). Williams (2006) note le besoin que cette discipline a d'être « reconnue », c'est-à-dire qu'il est nécessaire de comprendre les différentes écoles qui travaillent à l'aide de corpus. En effet, la grande diversité des pratiques et des méthodes que l'on peut regrouper derrière l'objet « corpus » interdit l'économie d'une définition précise des positionnements, méthodes, objectifs, et outils employés par le chercheur avant toute démarche de recherche.

Ainsi, afin de mieux mieux appréhender et diffuser les bénéfices que la linguistique de corpus peut apporter à l'enseignement des langues, nous situerons dans une première partie parmi les différents courants théoriques présents en linguistique de corpus celui dans lequel la présente expérience se situe. Une seconde partie présentera plus en détail le corpus *IntUne* [1] ici utilisé : les positionnements théoriques qui ont conduit aux choix des textes le composant, à leur format, à leur structuration, ainsi que les différentes contraintes techniques qui ont pu être rencontrées lors de sa constitution. Le logiciel d'exploitation Xaira, sera également présenté. Enfin, l'expérience d'emploi d'un corpus dans le cadre d'un cours de didactique du français langue étrangère (F.L.E.) sera exposée. Les résultats de cette expérience, les questionnements et les pistes de recherche soulevés à cette occasion seront finalement détaillés.

1. La linguistique de corpus ?

Différentes acceptions du corpus

Il existe en sciences du langage différents types de corpus et différentes façons de les aborder. Mayaffre (2005) tente une classification en trois catégories : d'une part, les corpus lexicographiques sont de

1. *IntUne* est un projet européen qui était intégré au sixième programme cadre de recherche de l'Union européenne. Ce projet a notamment eu pour objectif la création et l'exploitation d'un vaste corpus informatisé de textes issus de journaux quotidiens écrits et télévisés dans quatre langues (anglais, italien, français, polonais). Notre recherche a été pratiquée à l'aide de la partie française de ce corpus.

vastes ensembles de mots collectés dans un but de traitement automatique de la langue (TAL), ce type de corpus rejoint l'école de Lancaster décrite par Williams (2006). D'autre part, les corpus phrastiques sont créés par des grammairiens et se composent de phrases construites pour attester ou non des hypothèses linguistiques formées par le chercheur. Ces corpus n'entrent pas dans le domaine de la linguistique de corpus puisqu'ils ne contiennent pas de texte authentique. Enfin, il existe des corpus textuels, qui regroupent de façon ni totalement représentative ni totalement exhaustive des données attestées et reprennent des textes dans leur ensemble et non pas des extraits échantillonnés. Ce dernier type de corpus peut correspondre à ceux réalisés par Sinclair et l'école de Birmingham au sein d'un paradigme contextualiste. Le corpus *IntUne* utilisé dans notre expérience entre dans cette dernière catégorie dans la mesure où il contient des articles complets de presse et des journaux télévisés intégralement retranscrits. Dans les corpus textuels, le texte est l'unité minimale du corpus, l'unité minimale où se construit le sens.

Nous pouvons considérer que coexistent deux pans de la linguistique qui font un usage des corpus, la linguistique « sur » corpus et la linguistique « de » corpus. Nous incluons dans la linguistique « sur » corpus les différentes démarches qui font essentiellement un usage des deux premiers types de corpus décrits, et dans la linguistique de corpus les méthodes de recherche faisant un emploi de corpus textuels et issues du courant contextualiste, notamment anglo-saxon, décrit par Cori *et al.* comme étant une des sources de la linguistique de corpus actuelle :

> La dénomination de « linguistique(s) de corpus » a quant à elle été empruntée au courant britannique *Corpus Linguistics*, l'un des plus anciens et des plus structurés, fondé théoriquement sur la tradition firthienne de la *London School*. (Cori *et al.* 2008)

Dans notre recherche, souhaitant exploiter les apports des corpus dans un cours de didactique du F.L.E., nous nous positionnons en linguistes de corpus et non pas en linguistes sur corpus. Nous allons maintenant préciser ce qu'est le courant contextualiste qui a donné naissance à la linguistique de corpus.

Le contextualisme britannique

Les sources didactiques du contextualisme britannique

Le domaine de l'enseignement des langues a été l'un des plus prolifiques en ce qui concerne la recherche linguistique se basant sur des faits de langue authentiques.

À la fin du dix-neuvième et au début du vingtième siècle, Sweet (1899) a énoncé différents principes se basant sur la nature du langage pour en permettre l'apprentissage et a notamment étudié le lexique et la phraséologie (Williams 2006 : 152). Il proposait dans le domaine de l'apprentissage grammatical d'une langue de procéder de façon inductive, en confrontant tout d'abord les apprenants à des textes authentiques dont ils tireraient les principes grammaticaux de la langue (Véronique 1992 : 180). Le contexte est donc essentiel dans les principes de didactique du langage de Sweet.

Harold Palmer a également travaillé dans ce sens au début du vingtième siècle. Lors de ses travaux, il s'est penché plus particulièrement sur la lexie et la mise en place d'un vocabulaire pour les apprenants. Il a publié en 1937 un travail dans ce domaine avec Hornby, créateur du premier *Advanced Learner's Dictionary*. Il portait dans ses travaux l'idée que la langue est d'abord le véhicule des idées et condamnait une « attention excessive de l'ancienne méthode à la construction grammaticale » (Véronique, *op. cit.* : 176).

Les débuts de la linguistique de corpus anglo-saxonne moderne

Dans la suite de la première moitié du vingtième siècle, l'importance du contexte a été mise en avant par des linguistes britanniques ayant adapté une partie de la théorie anthropologique de Malinowski. Des années 1930 à 1950, John R. Firth postule une théorie contextuelle du sens basée sur une procédure inductive, et en conséquence, sur l'étude de phénomènes concrets. Il voit le langage comme la répétition d'un procédé social, le travail du linguiste consiste alors à dériver de l'impersonnel à partir du personnel qui est rendu typique. Ce passage de l'observation de faits particuliers à la mise en place de généralités est possible grâce à certaines structures prenant en compte l'observation du contexte, comme les collocations, occurrences répétées et signifiantes de deux mots ensemble ou les colligations, la proximité répétée de mêmes parties du discours (Firth 1957 : 143).

Pour décrire le langage, Firth se base sur une observation de la langue, et non pas sur son introspection. Il donne toute son importance à l'étude du sens, comme étant accessible à travers des usages contextualisés, et en notant que les manifestations de la langue sont observables plutôt que pré-supposables ou ontologiques. Le langage n'est pas dans sa théorie considéré comme un procédé mental ou une entité autonome, mais plutôt comme des événements exprimés par des locuteurs, comme un mode d'action (*ibid.* : 19). Selon lui, des méthodes purement inductives doivent apporter beaucoup à l'étude de la langue, notamment par l'apport de nouvelles catégories permettant de décrire plus scientifiquement, plus précisément la langue (*ibid.* : 28).

Il n'hésite pas à dire que pour être scientifique, la linguistique se doit avant tout de fonctionner de façon inductive, empirique, sans quoi elle risque de s'abîmer dans des hypothèses fallacieuses, ne rendant pas compte de la réalité du langage. Palmer a également travaillé dans ce sens au début du vingtième siècle, il a notamment produit de nombreux travaux sur la mise au point de vocabulaires pour apprenants prenant en compte le contexte et se fondant sur des analyses de corpus (Williams 2006). L'influence de Sweet et de Palmer, qui souhaitaient que les apprenants procèdent par découverte en contexte de la grammaire après la lecture d'un texte, est présente lorsque Firth explique que la linguistique ne doit pas s'éloigner de la parole en la forçant à entrer dans des systèmes artificiels (*ibid.* : 144).

Il a influencé certains linguistes des années suivantes tels que Halliday qui écrira dans les années 1950 une thèse s'intéressant aux liens entre fréquence et sens dans une langue, ou bien Sinclair qui s'est toujours basé sur des observations d'énoncés authentiques en contexte dans ses travaux lexicographiques (Honeybone 2005 : 85).

Ces linguistes ont montré sans se départir de leur intuition qu'ils étaient à même de recevoir les nombreuses informations accessibles à travers des corpus de discours authentique. Ils ont également veillé à rester ouverts aux évolutions technologiques qui ne peuvent qu'influencer toute attitude scientifique vis-à-vis des données observées. En gardant à l'esprit que les possibilités actuellement offertes par le traitement informatique de corpus de textes ne pouvaient seulement être imaginées il y a de cela vingt ou trente ans, on conçoit mieux en quoi cette nouvelle approche a apporté une modification importante dans la façon dont la linguistique était envisagée jusqu'alors.

La linguistique « de » corpus, telle que nous la pratiquons s'inscrit dans ce contextualisme. Il s'agit de procéder de façon inductive, selon une méthodologie *corpus-driven* [1], en se laissant guider par les données du corpus et en utilisant un corpus textuel de grande taille. D'autre part, elle prend sa source dans des réflexions menées par des didacticiens (Palmer, Sweet). Nous allons maintenant voir l'intérêt que la linguistique de corpus peut revêtir en didactique des langues, mais aussi les difficultés qui expliquent son faible emploi dans les salles de classe actuellement.

L'utilisation de la linguistique de corpus en didactique des langues

L'emploi de corpus permet d'enseigner tous types de langues, pour peu que l'on soit à même de constituer un corpus informatisé de la langue que l'on souhaite soumettre aux apprenants. Il est alors possible pour un enseignant de réaliser des cours pour tout type de langue générale ou de spécialité. Les spécificités propres à un discours particulier comme par exemple le discours juridique, médical... peuvent être enseignées à des apprenants *via* la constitution de corpus contenant des énoncés issus de ces discours. Landure et Boulton (2010 : 2) exposent notamment le problème des enseignants en anglais LANSAD (LANgues pour Spécialistes d'Autres Disciplines) qui ne sont pas toujours en mesure de maîtriser parfaitement la spécialité de leurs étudiants. Ils préconisent l'emploi de corpus avec les étudiants afin de résoudre une partie de ces difficultés.

Outre une confrontation plus en phase avec la réalité des langues de spécialité, l'usage de corpus par des étudiants représente également un moyen de s'approprier la langue pour des emplois ultérieurs dans diverses situations et il permet le repérage de problèmes spécifiques récurrents liés à l'apprentissage d'une langue donnée.

Landure et Boulton citent Johns lorsqu'il expose les principes du *Data Driven Learning* [2] :

1. Selon l'approche *corpus-driven*, le linguiste ne cherche pas à imposer des modèles déjà existants ou à faire se correspondre son intuition et les données qui sont extraites au cours de l'analyse du corpus, mais il est plutôt question de partir de l'observation des données, et non pas d'une hypothèse fondée sur son intuition. Cette démarche se veut inductive et a été documentée par Tognini-Bonelli (2001).

> Les apprenants peuvent les exploiter à leur tour dans ce que Johns appelle le « data-driven learning » (DDL) qu'il définit comme « la tentative d'éliminer autant que faire se peut l'intermédiaire pour donner à l'apprenant un accès direct aux données [linguistiques] ».
>
> (1991b : 30) (*Ibid.* : 3)

Avec l'emploi des corpus, les étudiants sont amenés à ne plus apprendre les règles selon la façon dont elles sont présentées dans un manuel ou une grammaire, mais plutôt à découvrir des régularités, des tendances plus ou moins typiques au sein de corpus de plusieurs millions de mots, de la même façon que Firth recommandait pour les linguistes de découvrir à partir de données authentiques ce qui était typique dans une langue. L'utilisation de corpus dans l'apprentissage d'une langue constitue une source supplémentaire à utiliser en plus d'une méthode, celle-ci ayant le plus souvent une approche trop générale de la langue. Le corpus donne à l'apprenant la possibilité supplémentaire de créer ses propres savoirs de façon inductive, ce que note également Leech (1997).

> This latter type of task gives the student the realistic expectation of breaking new ground as a « researcher », doing something which is a unique and individual contribution, rather than a reworking and a evaluation of the research of others. (*Ibid.* : 10)

Ce type d'apprentissage, qui fonctionne par découverte, a comme avantage non négligeable d'être un processus autonomisant pour l'apprenant. Celui-ci a alors une motivation plus importante dans son apprentissage, et il s'approprie mieux les notions qu'il découvre. Un outil efficace pour permettre ce type d'apprentissage est le concordancier qui permet de mettre en évidence les différentes structures qui peuvent apparaître de façon typique selon des circonstances précises (Bernardini 2004 : 18).

Tyne (2009 : 93) note encore que l'apprentissage d'une langue aidé par les corpus permet de mieux saisir des phénomènes touchant à la compétence sociolinguistique des apprenants à travers une meilleure sensibilisation à la variation d'autant plus lorsque ceux-ci sont mis en situation d'enquêteurs sur le terrain, de constructeurs du corpus.

Cette énumération non exhaustive de l'intérêt que la linguistique de corpus peut avoir dans le domaine de la didactique des langues doit cependant être relativisée à l'aune de l'emploi effectif des outils et des

méthodologies issus de cette discipline en classe. Landure et Boulton (2010 : 4) notent que ce type d'exploitation pédagogique reste essentiellement cantonné au domaine universitaire et est mené avec des étudiants de haut niveau, dans des contextes où de nombreuses ressources matérielles sont à la disposition des enseignants et des étudiants, ce qui n'est pas toujours le cas de classes ordinaires. Mauranen (2004) propose comme solution à cet usage encore restreint des méthodologies issues de la linguistique de corpus en classe une meilleure formation des enseignants à ces outils :

> Before learners can be introduced to good corpus skills, their teachers need to possess them in the first place. (*Ibid.* : 100)

C'est dans cette optique que notre expérience a été menée avec des étudiants en troisième année de licence F.L.E., se destinant pour la plupart d'entre eux à l'enseignement du Français comme langue étrangère ou seconde. Nous allons maintenant détailler le corpus et les outils qui ont été utilisés lors de cette expérience.

2. Un corpus : *IntUne* et un concordancier : Xaira

Le corpus *IntUne*

IntUne est un projet intégré du sixième programme cadre de recherche de l'Union européenne qui a impliqué 32 partenaires dans 15 pays différents. Ce projet a notamment consisté en l'élaboration de deux sous-corpus par un groupe d'universitaires issus de quatre universités : Cardiff en Grande Bretagne, Lodz en Pologne, Lorient en France et Bologne en Italie. Ces chercheurs ont réalisé un corpus, puis l'ont analysé selon les méthodologies issues de la linguistique de corpus, utilisant notamment des méthodes quantitatives de type lexicographiques, ou des méthodes qualitatives se basant sur les cadres théoriques issus des *Computer Assisted Discourse Analysis* (Partington 2007 [1]). Dans le cadre de notre étude, nous nous focaliserons sur ce corpus qui regroupe des articles issus de supports médiatiques papiers qui ont été informatisés.

1. Pour une définition du cadre théorique des CADS, voir : `http://unibo.academia.edu/AlanPartington/Papers/1236069/The_armchair_and_the_machine_Corpus-assisted_discourse_studies_CADS`.

Les articles des journaux choisis pour faire partie du corpus ont été collectés grâce à une base de données sur Internet, sous un format.txt. Le texte était transformé d'un format .txt en .xml pour pouvoir travailler avec une architecture de base TEI [1].

Ainsi, ce corpus, par sa taille (environ 20 millions de mots pour chaque sous-partie en langue française) et son format informatique, suit les recommandations posées par Sinclair en la matière :

> Corpus has to have the material in electronic form [...] The only guidance I would give is that a corpus should be as large as possible.
> (1991 : 14-18)

Un concordancier : Xaira

Une fois le corpus construit, il s'agit de l'exploiter. Différents logiciels sont proposés pour l'étude de corpus. Nous utilisons Xaira [2] étant donné ses capacités à fonctionner sans difficultés majeures sur des corpus de taille importante et son aptitude à gérer des structures complexes de texte. Il s'agit également d'un logiciel gratuit développé par l'université d'Oxford. Xaira permet par ailleurs un fonctionnement de type *corpus-driven* et est utilisable facilement par de nouveaux apprenants, ce qui correspond à notre approche du corpus qui se veut inductive (Tognini Bonelli 2001).

Selon cette approche *corpus-driven*, l'étudiant en position de chercheur ne souhaite pas imposer des patrons déjà existants ou faire se correspondre son intuition et les données extraites au cours de l'analyse du corpus. Il cherche plutôt à extraire des combinaisons typiques de mots et à les étudier pour voir quelles relations existent entre les mots.

1. La TEI (*Text Encoding Initiative*) est une réflexion née d'un besoin de numériser des textes sous un format qui soit le plus accessible possible, et qui donne une représentation riche d'un texte. Le balisage permet de prendre en compte des informations quant à la source du texte, par exemple, ces méta-données permettent un meilleur archivage ainsi qu'un meilleur accès au texte, elles sont généralement stockées dans son en-tête. Ce format a pour avantage, outre une meilleure prise en compte des spécificités du discours traité, d'être largement répandu dans la communauté scientifique (le *British National Corpus* l'utilise) et d'être totalement gratuit.

2. Xaira, *Xml Aware Indexing and Retrieval Architecture*, est un concordancier qui peut se télécharger librement à l'adresse suivante : `www.oucs.ox.ac.uk/rts/xaira/`.

Xaira propose parmi d'autres la fonction de concordancier qui extrait des lignes de contextes au sein desquels les mots choisis apparaissent. L'utilisation de ce logiciel comme concordancier étant assez simple et intuitive et ce mode d'exploitation du corpus permettant aux étudiants de se confronter aux données du corpus sans autre information que le contexte au sein duquel le mot étudié se situe, nous avons privilégié cet outil lors de notre expérience.

Nous allons maintenant exposer le cadre dans lequel cette expérience s'est déroulée.

3. L'utilisation d'un corpus pour résoudre des difficultés de la langue française dans le cadre d'un cours de didactique du F.L.E.

Cette expérience a été menée lors d'un cours de didactique du F.L.E. de trois heures, avec des étudiants en troisième année de licence qui ont fonctionné selon huit binômes. Le cours a eu lieu dans une salle informatique où tous les ordinateurs étaient équipés du logiciel Xaira.

Les étudiants ne semblent pas avoir rencontré de difficultés pour utiliser les principales fonctions du concordancier de Xaira, qu'il s'agisse de la recherche de locutions, ou du tri des lignes de concordances en fonction des mots apparaissant à gauche et à droite des nœuds étudiés (voir la figure 2 représentant le concordancier utilisé par les étudiants).

Les consignes consistaient en l'analyse de concordance pour deux paires de mots posant certaines difficultés : très/trop et quand/quant. Les étudiants ont été invités à repérer d'éventuelles régularités pour ces mots, et ce qui pouvait permettre d'opérer un choix entre l'un ou l'autre des mots en question. Ils ont reçu comme conseil d'utiliser le tri en N−1 et en N+1 (un mot avant ou un mot après le mot-clef). À la fin de cet exercice, ils ont rendu les résultats écrits de leurs recherches sur corpus.

Le corpus a été présenté comme une collection d'articles de presse, considérés pour les besoins de cette expérience comme un ensemble de documents écrits sur lesquels il était possible de se baser pour obtenir des informations concernant le comportement de certains items linguistiques du français. Dans le cadre restreint de cette expérience, nous nous situons donc au niveau d'une didactique de l'écrit, mais un corpus tel qu'*Int Une* peut également se révéler intéressant pour étudier

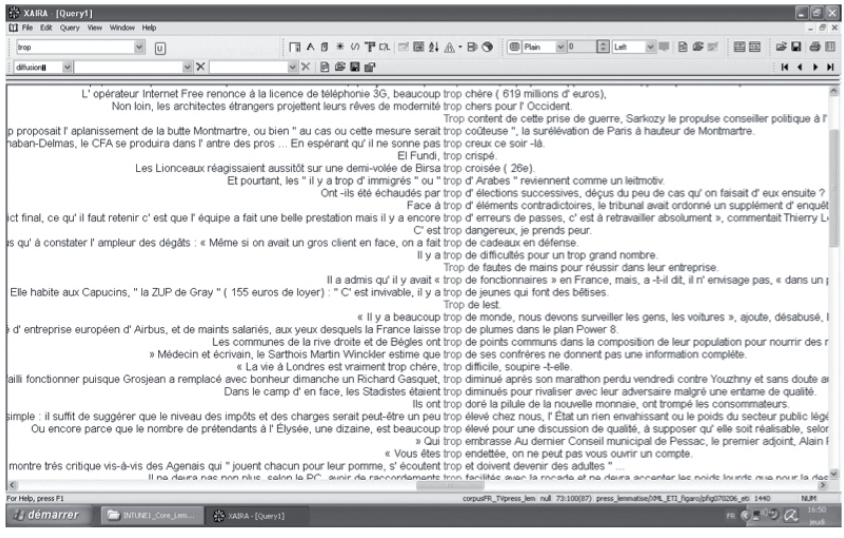

Figure 1. — Vue du logiciel Xaira utilisé dans sa fonction de concordancier.

d'autres aspects didactiques du F.L.E. (interculturels, grammaticaux, oraux, etc.).

Après une première lecture des productions des étudiants, il semble que cinq des binômes sur huit se soient bien adaptés à cette nouvelle méthodologie. Leur trace écrite présentant des mises en forme propres à un travail d'analyse de concordancier : présentation des mots en N−1 ou N+1, présentation des résultats en exprimant des régularités, des tendances. Les trois binômes restant ont donné une trace écrite beaucoup plus rédigée, et qui reprenait assez largement des représentations issues d'années d'enseignement grammatical.

Une analyse plus précise de ces productions nous permet de constater que les observations relatées par les étudiants peuvent se regrouper en quatre catégories :

– Les observations pertinentes par rapport au but de l'étude mentionné.
– Les observations incomplètes.
– Les observations non pertinentes.

– Les observations issues de représentations grammaticales préexistantes à l'étude.

Les résultats obtenus montrent plusieurs choses quant à l'introduction d'étudiants en langue aux méthodologies de la linguistique de corpus.

D'une part, il semble qu'il soit difficile pour certains étudiants de ne pas recourir à leurs connaissances préétablies au niveau grammatical et à les inscrire comme étant le fruit d'une recherche sur corpus. Un des binômes d'étudiants note ainsi comme remarque principale concernant l'observation de « très » dans les lignes de concordances que : « C'est un adverbe qui indique un degré élevé qui sert à intensifier le mot qui le suit. Il marque le superlatif absolu. » Si cette information ne comporte pas d'erreur, elle ne découle cependant pas de l'observation du corpus ce qui était le but de la séance. Cet exercice peut ainsi avoir déconcerté certains étudiants et les avoir poussés à simplement reprendre les connaissances dont ils disposaient déjà sur les mots étudiés.

D'autre part, certains phénomènes peuvent parfois être mis en relief par les étudiants alors qu'aucune information dans les fréquences relevées ne semble pouvoir justifier qu'ils soient relevés. C'est le cas de la plupart des observations classées comme « incomplètes », un binôme a ainsi relevé que « quand est toujours suivi par des noms » alors que cela n'apparaissait pas dans les lignes de concordances observées : quand y était parfois suivi de noms, mais pas dans la majorité des cas. Un autre binôme a noté que « très est placé devant un nom » alors que de la même façon, cela n'apparaissait pas dans le corpus *IntUne*.

Il semblerait enfin qu'une analyse fine du corpus et le rapport de cette analyse, telle que nous l'avons proposée aux étudiants dans cette expérience, pose comme pré-requis de bonnes aptitudes et connaissances linguistiques au niveau grammatical. Ces connaissances sont en effet nécessaires à la formulation et à la conceptualisation d'une règle (Besse et Porquier 1984 : 197), ne serait-ce que par l'usage du métalangage.

Cependant, il est à noter qu'il s'agissait d'une simple introduction aux outils et méthodes propres à la linguistique de corpus, que les étudiants ne connaissaient pas avant cette séance de leur cours de didactique du F.L.E. Les résultats obtenus, si l'on tient compte de ce contexte, sont plutôt bons puisque la plupart des observations

menées sont correctes et peuvent être utiles pour mieux appréhender l'usage des paires de mots étudiées. Ainsi, tous les étudiants ont vu que pour expliquer à un apprenant comment utiliser le mot « quand » ou « quant », il suffisait de voir s'il était suivi de à, au ou aux. Cette expérience n'a pu être menée qu'avec des étudiants natifs, et il serait intéressant de la poursuivre avec des non natifs pour confronter les résultats obtenus.

Ces résultats sont également plus complets qu'une règle qui était donnée dans un manuel de F.L.E. [1] où seuls les adjectifs sont mentionnés comme suivant « trop ». Ici, le fait que l'on puisse trouver des participes passés devant les adjectifs par dérivation impropre a été noté par les étudiants. Le fait que « trop » puisse être suivi d'un adverbe a également été bien relevé. À la fin de leur étude, nous avons présenté aux étudiants cette règle issue d'un manuel et ceux-ci ont pu prendre conscience de l'utilité des corpus dans des situations où des méthodes de F.L.E. peuvent ne pas rendre compte de la multiplicité des usages réels d'une langue. L'utilisation de corpus permet alors un accès très rapide à une ressource linguistique importante (plusieurs millions de mots).

Conclusion

Cette expérience, restreinte dans le temps, nous a permis de voir quelles difficultés et quels intérêts l'utilisation d'un corpus pouvait rencontrer avec un public de futurs enseignants en langue.

Il s'agissait de mener une expérience se situant dans le cadre de la linguistique de corpus et non pas de faire de la linguistique sur corpus. Les étudiants ont ainsi pratiqué une démarche inductive, partant de l'observation de données sur un corpus textuel, pour ensuite mettre en relief les régularités du langage étudié. Cet exercice est très proche du *« Data Driven Learning »* exposé par Johns (1991) dans la mesure où l'enseignant a très peu joué le rôle de médiateur entre les données linguistiques et les apprenants.

Certaines difficultés ont été rencontrées lors de cette séance avec des étudiants en didactique du F.L.E. L'importance des représentations linguistiques que ceux-ci ont mis en place tout au long de leur cursus

1. Sylvie Poisson-Quinton, Michèle Maheo-Le Coadic, (2005), *Festival*, Paris, CLE International.

scolaire a ainsi parfois pu agir comme une sorte de bruit parasite lors de l'observation du corpus. Cela est plus particulièrement apparu chez des binômes pour qui l'exercice semblait déroutant et pour qui le recours à des connaissances déjà bien en place a pu apparaître comme la réponse idéale à fournir au travail demandé. Cette difficulté pourrait se résoudre par une sensibilisation accrue des étudiants à l'utilisation de concordanciers, ce type d'analyse nécessitant souvent un apprentissage (Sinclair 2003). À partir de là, les connaissances préalables des étudiants pourraient être pleinement exploitées par ceux-ci, et moins interagir de façon malheureuse avec les observations menées sur le corpus. Néanmoins, certaines des difficultés couramment admises dans le cadre d'une pratique didactique basée sur des méthodologies issues de la linguistique de corpus n'ont pas été observées. Les étudiants n'ont ainsi pas rencontré de problèmes techniques lors de cet exercice, une fois le corpus installé et démarré, l'utilisation de Xaira dans ses fonctions de concordancier a eu lieu avec une parfaite autonomie de la part des étudiants (un didacticiel sommaire avait été rédigé au tableau dans la salle).

En outre, cet exercice constituait la première confrontation à des outils issus de la linguistique de corpus pour les étudiants. Malgré un temps d'introduction très court, ils ont réussi à mettre en place en trois heures une démarche d'observation linguistique nouvelle et à obtenir des résultats pertinents. La plupart d'entre eux se sont montrés très intéressés et ont posé des questions quant à l'existence de corpus consultables dans différentes langues.

Cette expérience bénéficie donc d'un bilan positif et nous souhaiterions pouvoir l'élargir en procédant à de futures séquences permettant aux étudiants de mieux maîtriser ces outils. Différentes pistes pourraient alors être exploitées : la création par les apprenants eux-mêmes de corpus écrits ou oraux, la création de séquences pédagogiques où les futurs formateurs se projetteraient dans l'emploi de corpus en classe, etc.

On pourrait également penser à poursuivre ce type d'expérience en classe, et à les développer, notamment à destination d'un public de futurs formateurs en langue ou d'apprenants de L2 et de L3.

Références bibliographiques

BERNARDINI S., 2004, « Corpora in the classroom. An overview and some reflections on future developments », *How to use corpora in language teaching*, éd. SINCLAIR, coll. « Studies in corpus Linguistics », Amsterdam, John Benjamins.

BESSE H. & PORQUIER R., 1984, *Grammaires et didactique des langues*, Paris, Hatier-Didier.

CORI M., DAVID S., LÉON J., 2008, « Présentation : éléments de réflexion sur la place des corpus en linguistique », *Langages* 2008/3, nº 171, 5-11.

FIRTH J. R., 1957, *Papers in Linguistics 1934-1951*, Oxford, Oxford University Press.

HABERT B., NAZARENKO A., SALEM A., 1997, *Les linguistiques de corpus*, Paris, Armand Colin.

HONEYBONE P., 2005, *Key Thinkers in Linguistics and the Philosophy of Language*, Edinburgh, Edinburgh University Press, 80-86.

JOHNS T., 1991, « From Printout to Handout : Grammar and Vocabulary Teaching in the Context of Data-Driven Learning », *English Language Research Journal* 4, 27-45.

LANDURE C. BOULTON A., 2010, « Corpus et autocorrection pour l'apprentissage des langues », *ASP* nº 57, Nancy, 11-30.

LEECH G., 2007, « New Resources, or Just Better Old Ones ? The Holy Grail of Representativeness », *Corpus Linguistics and the Web*, éd. M. HUNDT, N. NESSELHAUF et C. BIEWER, Amsterdam, Rodopi, 133-149.

MAURANEN A., 2004, « Spoken Corpus for an Ordinary Learner », How to Use Corpora in Language Teaching, éd. SINCLAIR, coll. « Studies in corpus Linguistics », Amsterdam, John Benjamins.

MAYAFFRE D., 2005, « Rôle et place des corpus en linguistique : réflexions introductives », *Actes des journées d'études toulousaines JETOU 2005*, Toulouse, 5-17.

SINCLAIR J., 1991, *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

SINCLAIR J., 2003, *Reading Concordances*, Londres, Pearson.

SWEET H., 1899, *The Practical Study of Languages*, Londres, J. M. Dent & Co.

Tognini Bonelli E., 2001, *Corpus Linguistics at Work, Studies in Corpus Linguistics*, Amsterdam, John Benjamins.

Tyne H., 2009, « Corpus oraux par et pour l'apprenant », *Mélanges Crapel*, n° 31, 91-111.

Véronique D., 1992, « Sweet et Palmer, précurseurs de la didactique des langues ? », *Cahiers Ferdinand de Saussure*, n° 46/1992, 173-192.

Williams G., 2006, « La linguistique de corpus, une affaire prépositionnelle », revue *Texto* consulté en juin 2011 sur :
`www.revue-texto.net/Parutions/Livres-E/Albi-2006/Williams.pdf`.