Mélanges CRAPEL n°31

LE WEB COMME CORPUS : DOCUMENTS AUTHENTIQUES ET EXPLOITATION EN FLE

Lucia Drago

Queensland University of Technology (Australie)

Résumé

Dans l'objectif de combiner la linguistique de corpus à la pédagogie de l'interculturel en classe de FLE, un projet d'analyse d'un mini-corpus a été mis en place. Cet article présente un aperçu du travail abordé en partant des extraits d'une quarantaine de journaux en langue française, tels qu'ils sont triés par le concordancier en ligne *GlossaNet* dans une sélection de syntagmes incluant « identité nationale ». Un choix d'expressions circulant actuellement dans les médias pour désigner les personnes d'origine étrangère en France est aussi utilisé avec les étudiants pour stimuler la réflexion sur la description des identités culturelles et ses nuances.

Abstract

Can corpus linguistics serve intercultural education in the foreign language classroom? In order to investigate this issue, a project has been implemented involving the analysis of a mini-corpus. This paper presents work based on data from about forty French-language newspapers, as processed by the *GlossaNet* on-line concordancer when searching for *identité nationale* ("national identity"). In addition, a selection of expressions running through the French media and designating people of foreign origins is used, aiming to engage students in reflection about subtlety in descriptions of cultural identity.

Introduction

L'Internet représente pour les enseignants de français langue étrangère (FLE) une source intarissable pour accéder à toutes sortes de documents authentiques qui reproduisent les expressions les plus courantes de la langue française contemporaine. Pourtant, une sélection faite avec un moteur de recherche standard présente des limites en termes de possibilités d'interrogation à certains niveaux de structure grammaticale. Il serait difficile, par exemple, de trouver quels sont les adjectifs les plus utilisés avec un mot donné sans analyser une quantité de résultats qui comprennent d'autres catégories grammaticales.

Dans cet article, nous présentons ainsi quelques exemples d'applications didactiques utilisant un logiciel en ligne, *GlossaNet*, qui permet de faciliter certaines recherches linguistiques en traitant les versions en ligne des journaux comme des corpus. Les documents authentiques sélectionnés par le logiciel en question sont des articles extraits d'une quarantaine de journaux en langue française. Les résultats sont fournis dans le format classique des corpus informatisés, KWIC (*Key-Words in Context* = mots-clés en contexte). Le langage circulant dans les médias est alors analysé et interprété à plusieurs niveaux avec les étudiants. En particulier, des expressions nouvelles qui sont créées au fur et à mesure que certains événements de l'actualité se produisent et qui sont immédiatement transcrites, diffusées, tracées dans la presse, font l'objet de cette analyse. Les exemples qui sont présentés relèvent d'une collecte de mots et combinaisons de mots utilisés récemment par la presse de langue française dans le débat sur l'immigration et l'intégration en France, à propos de l'identité française ainsi que dans la désignation des Français d'origine étrangère.

L'expérience que nous allons décrire s'insère dans le cadre d'une recherche de doctorat en Communication interculturelle qui prévoit l'exploitation d'un corpus en classe après un travail préparatoire conduit par le chercheur. Les activités proposées envisagent le développement des compétences linguistiques et interculturelles à travers la recherche des unités de signification – « the search for units of meanings » (Sinclair, 1996a) – et la négociation des significations comme pratique communicative.

Lors de cette première expérimentation pédagogique conduite en septembre 2007 à la Queensland University of Technology (Australie), l'attitude des apprenants à l'égard de ce genre d'approche a aussi été explorée, afin de mettre en relief les aspects à prendre en compte dans la suite de la recherche. Cet article inclut donc un aperçu des réactions des participants, ce qui contribue à répondre à la nécessité évoquée par Chambers (2005), Kennedy et Miceli (2001) et Yoon et Hirvela (2004) de montrer comment les investigations sur corpus sont perçues par les apprenants, en particulier pour les langues autres que l'anglais.

1. Les corpus de documents authentiques en classe de FLE : quelle pédagogie pour développer les compétences interculturelles ?

L'introduction des documents authentiques en classe de langue étrangère a encouragé les applications pédagogiques de la linguistique de corpus, qui privilégient les échantillons des discours de la langue orale, en conversation et dans les médias, dans le but d'enseigner « the real language »[1]. Cette approche, par sa nature, implique l'observation de cas multiples d'occurrences des mots sélectionnés, dits collocations. Si d'un côté une telle analyse linguistique des collocations peut servir de formation pour affiner les capacités des étudiants à saisir la palette des significations des mots, d'un autre côté le travail analytique et descriptif effectué en classe finit parfois par négliger les aspects de l'interaction que l'apprentissage des langues étrangères est censé développer, notamment la communication entre pairs et ce qu'on appelle un « information gap », qui motiveraient l'échange. Le risque serait, également, d'encourager les apprenants à une approche trop quantitative et fragmentaire, qui négligerait la globalité des textes et leur contexte de production, ainsi que l'aspect cognitif et culturel de la langue-cible (Prat-Zagrebelsky, 2004 : 29).

L'idée de départ de cette recherche est donc inspirée par la nécessité de développer des activités par lesquelles l'analyse d'un mini-corpus en classe stimulerait l'échange de points de vue et la négociation des significations des mots. En outre, dans le but de lier notre travail sur un corpus linguistique à des activités pédagogiques qui encouragent la réflexion sur une autre culture, nous avons voulu sélectionner des mots liés à la perception de l'étranger et aux représentations de l'identité culturelle. Pour ce faire, la description de l'identité multiculturelle de la France et le sujet de l'intégration des immigrés nous ont semblé offrir un exemple de situation complexe à explorer.

1.1. La communication interculturelle au quotidien

Nous voudrions insister sur la nécessité d'une pédagogie de la communication interculturelle plus adaptée aux besoins de la société contemporaine, où les défis lancés par la mondialisation aussi bien que les phénomènes migratoires poussent les citoyens à se poser des questions sur la réussite des mélanges culturels et sur l'évolution des identités ethniques. L'échange, le contact, l'interaction entre personnes d'origines culturelles différentes se font de plus en plus au quotidien : à l'école, au travail, dans les espaces publics. Par conséquent, une communication efficace ne consiste plus à savoir utiliser des formules figées pour arriver à bénéficier

1. Nous faisons référence à la langue dans son usage quotidien, ce « real language » qui était l'objet des recherches lexicographiques et pédagogiques de Sinclair, un des pères fondateurs de la linguistique de corpus. Voir, par exemple, l'introduction au dictionnaire COBUILD, réalisé sous sa direction (Sinclair, 1996b).

d'un service essentiel dans un pays étranger (transports, logement ou alimentation). L'apprentissage d'un vocabulaire purement touristique ou commercial ne suffit pas. Les pratiques pédagogiques en langue étrangère portant sur la connaissance des coutumes traditionnelles et des différentes règles de politesse contribuent dans une certaine mesure à se décentrer et à encourager l'acceptation des diversités. Pourtant, lors des projets d'échange scolaires il n'est pas rare de relever des échecs dans la communication, soit face-à-face, soit par Internet (Kramsch & Thorne, 2001). Même les approches basées sur la prédication d'une politique de la tolérance ne parviennent pas souvent à stimuler une vraie curiosité envers les facteurs qui influencent l'enracinement de certaines habitudes plutôt que d'autres, mais, au contraire, peuvent renforcer les stéréotypes négatifs. Dans les sociétés actuelles (multiculturelles, polyculturelles, cosmopolites, ou bien « hybridées » ou « mélangées »), enseigner la « culture invisible » devient de plus en plus essentiel pour la cohésion sociale (Furstenberg *et al*., 2001).

Comme l'affirme Beacco (2000), une approche basée sur l'analyse linguistique des discours des médias représente autrement la clé pour accéder à la « culture-civilisation » d'aujourd'hui et l'interpréter. Cette circulation des traces des événements de l'actualité véhicule en même temps des connotations et des points de vue difficiles à saisir hors contexte (Moirand, 2007). Quoique plutôt complexes, les contenus de l'actualité diffusés par la presse sont intéressants pour les pratiques pédagogiques car ils sont formulés généralement dans un registre de langue assez soigné, qui peut servir de modèle pour les étudiants dans certains contextes de langue écrite[2]. Ils présentent également l'avantage pas négligeable d'englober certains néologismes et locutions que les dictionnaires pourront enregistrer bien plus tard. Si l'on s'interroge donc sur la place que la consultation des corpus en classe peut avoir dans les études de langue étrangère, et notamment par rapport au dictionnaire (Chambers, 2005), nous pouvons affirmer que quand on poursuit le but de l'acquisition d'un vocabulaire très récent, l'intégration de l'analyse de corpus se révèle particulièrement pertinente et même nécessaire si elle est combinée à l'utilisation du web.

1.2. Les actualités triées sur le web

L'utilisation d'Internet facilite le travail de recherche de documents authentiques et rend relativement rapide la collecte des données linguistiques écrites que l'on veut analyser. En effet, désormais l'expression « web as corpus » est utilisée pour faire référence à plusieurs manières d'exploiter la toile à des fins linguistiques (Baroni *et al*., 2006 : 10). Comme le souligne Castagnoli (2006 : 160), en principe, le web offre l'avantage d'une remise à jour constante de la terminologie :

2. Voir aussi Drago (2006).

> since terms are continually being invented and evolving, in relation to both their meaning and usage, it can be argued that a web-based open corpus is more likely to contain up-to-date terms and state-of-the-art concepts than a static corpus.

Pour simplifier les recherches et les comparaisons entre les échantillons de langue, des logiciels gratuits en ligne filtrent les résultats et les organisent en KWICs, la visualisation classique des corpus[3]. Ce format présente l'avantage de permettre l'identification des exemples d'usage par une simple lecture verticale des extraits. En outre, un lien électronique renvoie au texte complet d'où chaque extrait est tiré.

Dans la conception des activités pédagogiques dont nous présentons un aperçu, il a fallu avant tout délimiter le domaine de recherche sur le web, ainsi que la typologie de langage à présenter aux étudiants. Nous avons voulu puiser dans les versions en ligne d'un certain nombre de quotidiens en langue française sur une période récente limitée, qui va du 26 mars jusqu'au 26 juin 2007. Le logiciel utilisé à cette fin, *GlossaNet*, est fourni gratuitement sur Internet par l'Université Catholique de Louvain-la-Neuve, en Belgique[4]. Pour ce qui concerne la langue française, il inclut jusqu'à 37 journaux provenant de 11 pays francophones différents. Lors de ce premier essai, les 37 journaux ont été sélectionnés autant pour tester la performance du logiciel sur tous les sites annoncés que pour relever d'éventuelles défaillances techniques. Le tri fait automatiquement par *GlossaNet* sur le nombre de quotidiens disponibles en français délimite la sélection des extraits linguistiques et définit en même temps la nature du corpus. Par rapport à des pratiques courantes, nous n'avons pas utilisé de système d'annotation des catégories grammaticales sur le corpus. En outre, contrairement aux pratiques habituelles en linguistique de corpus, où des échantillons de langue sont choisis par genre pour en représenter l'usage, nous ne retenons pas l'objectif de représenter le français. D'ailleurs, il faudra remarquer, comme le fait Sinclair (1991 : 17), que :

> in journalism the well-known writers tend to have unusual ways of writing [and] if we are to approach a realistic view of the way in which language is used, we must record the usage of the mass of ordinary writers, and not the stray genius or the astute journalist.

Mais plutôt que proposer le matériel linguistique collecté pour sa valeur de « typicality » (Sinclair, 1991 : 17) ou autrement de référence normative, nous cherchons à reconnaître les traces de l'innovation lexicale reproduite dans les discours des médias. A l'intérieur des discours de la presse en langue française, une sélection a été opérée au niveau des contenus à analyser à travers les mots et leurs significations. En essayant de combiner la linguistique de corpus et une pédagogie de l'interculturel, nous avons donc voulu analyser des expressions qui circulent dans le langage des médias pour désigner les problèmes, les personnes et les lieux impliqués dans les questions d'immigration et d'intégration en France. L'objectif de

3. Pour un exemple du format KWIC, voir l'encadré à la deuxième page de l'activité 1 (annexe 1).
4. Cf. http://ling.fltr.ucl.ac.be/index.php. Nous avons utilisé ce service de mars 2007 à juin 2008. Au moment de la publication de cet article, le nouvelle version *GlossaNet 2* est en cours de finalisation.

cette approche n'est pas tellement d'identifier des attitudes racistes dans les discours de la presse (Van Dijk, 1988), mais plutôt de décrire une variété de formes linguistiques spécifiques qui sont créées dans le langage journalistique et d'en observer l'usage pour mieux le comprendre dans ses significations.

Dans le but de saisir les mots utilisés par la presse lors du débat autour de l'identité nationale en France[5], il nous a paru intéressant d'interroger le logiciel *GlossaNet* avec la requête « identité nationale ». Les résultats obtenus sur une période de trois mois (du 26 mars au 26 juin 2007) constituent le mini-corpus à partir duquel cette expérimentation a commencé. Dans la définition de la taille de ce corpus il faut prendre en compte plusieurs facteurs. En premier, nous devrons considérer que *GlossaNet* envoie les concordances triées à l'adresse électronique du destinataire chaque jour, ce qui produit une accumulation des mêmes concordances répétées au cours de plusieurs jours ou mois, selon la permanence de l'article d'origine sur le site-source. La duplication des données, d'ailleurs, est bien connue par les chercheurs qui utilisent le web comme corpus (Baroni *et al*., 2006 : 18). Deuxièmement, des obstacles d'ordre technique ne permettent d'accéder que partiellement aux sites d'un certain nombre de journaux. En conséquence, des statistiques de fréquence basées sur les 37 journaux sélectionnés comme un ensemble homogène ne pourraient pas avoir la portée souhaitée. Pour les raisons décrites ci-dessus, une analyse purement quantitative, soit du nombre des concordances, soit de l'ensemble des articles-source, ne serait pas vraiment significative. Toutefois, cela ne constitue pas un obstacle à notre expérimentation, puisqu'elle s'inscrit dans le cadre d'une recherche de nature qualitative, où la gamme des expressions linguistiques est explorée dans un contexte bien délimité par les contraintes lexico-grammaticales choisies.

L'analyse de ces expressions a été prévue en deux phases : un travail préparatoire, abordé essentiellement par le chercheur/enseignant, et le travail en classe, dont nous présentons la première expérimentation.

1.3. La démarche d'analyse pour le travail préparatoire

La première question que l'enseignant se pose dans l'organisation du matériel pour la classe est sans doute le niveau d'adaptation nécessaire pour que le travail sur un corpus, qui est normalement conduit par un linguiste, devienne accessible aux étudiants, en termes d'équipement technique aussi bien que des capacités d'analyse requises (Chambers, 2005 ; Kennedy & Miceli, 2001 ; Yoon & Hirvela, 2004). La démarche d'analyse choisie a été inspirée par les étapes suggérées par Sinclair

5. Nous faisons référence à la polémique suscitée en 2007 par la juxtaposition du mot « immigration » avec une insolite « identité nationale » au moment de l'annonce de la création d'un Ministère de l'immigration et de l'identité nationale. Voir, par exemple, l'article de Bernard (2007) publié dans *Le Monde*.

(1996a) et adaptées par Tognini-Bonelli (2000 : 215). Lors de cette étude préliminaire, nous avons abordé une définition progressive des structures linguistiques : du lexique à la syntaxe et d'un éventail d'options sémantiques aux interprétations à vérifier et discuter en classe dans une phase suivante. Voici une synthèse des trois passages suivis :

identité nationale
Etape 1 : identification du profil de la collocation
(les réalisations lexicales les plus fréquentes)
→ **de** l'identité nationale

Dans la première étape du processus décrit par Tognini-Bonelli (2000), il s'agit de parcourir les données extraites du logiciel et de retrouver le « collocational profile », ou la « lexical preference », ce qui équivaut ici à repérer les mots qui paraissent plus fréquemment en combinaison avec le groupe « identité nationale » dans le contexte de ce corpus de données. Nous avons donc remarqué que dans notre mini-corpus la préposition « de » était présente à la gauche du noyau « identité nationale » dans la quasi-totalité des cas[6] :

Etape 2 : identification des modèles de colligation
(les structures linguistiques récurrentes et leur articulation dans la syntaxe)
→ Nom + de l'identité nationale

La nouvelle séquence ainsi identifiée, « **de** l'identité nationale », nous a servi de point de départ pour avancer vers la deuxième étape, où l'on cherche les structures syntaxiques typiques plus étendues dans lesquelles cette séquence s'articule. Lors de cette phase, nous avons isolé (encore à gauche) 44 segments différents, dont nous montrons un échantillon dans la figure 1.

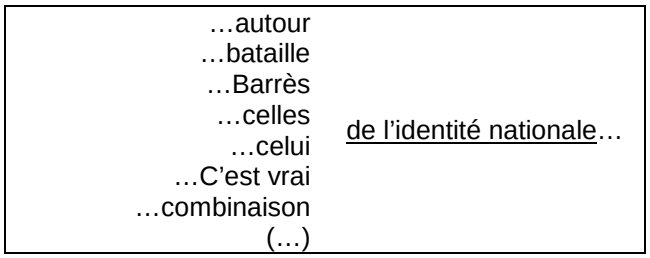


Figure 1. Sélection du modèle de colligation.

Pour identifier ce qui paraissait à la tête du syntagme, ou, autrement dit, ce qui a été dit à propos de l'identité nationale dans son contexte le plus proche, il nous a semblé utile de choisir la structure « Nom [+ Adjectif] + *de l'identité nationale* », ce qui constitue le « modèle de colligation » de notre investigation. Dans ce deuxième passage du processus d'analyse, il a fallu passer à une nouvelle sélection à l'intérieur du corpus pour retenir les séquences de mots qui correspondaient à cette

6. Naturellement, nous n'avons retenu qu'une seule fois les concordances extraites des articles qui se répétaient sur plusieurs jours. Nous faisons référence à la duplication des données sur le web (voir section 1.2).

structure. Sur les 44 différentes séquences précédemment identifiées à la gauche du groupe « de l'identité nationale », nous en avons retenu 33. Un extrait des résultats est reporté dans la figure 2.

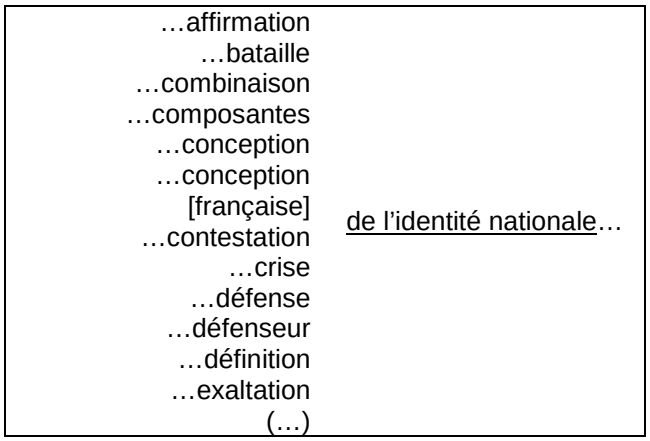


Figure 2. Un modèle de colligation : « Nom [+ Adjectif] + *de l'identité nationale* ».

Le but de cette succession d'étapes ainsi abordées n'est pas seulement de repérer la structure dans laquelle un mot se présenterait de préférence (« collocational profile » et « colligational pattern »), pour savoir ce qui est dit sur le sujet choisi, mais aussi de rendre compte d'un choix d'expressions qui suggèrent des liens avec des chaînes différentes de signification :

Etape 3 : identification des préférences sémantiques
(groupage par champs sémantiques)

→ **par ex**.	**bataille**	de l'identité nationale
	troubles	de l'identité nationale
	contestation	de l'identité nationale
		= **Allusion à des conflits**

L'étape suivante de Sinclair (1996a) consistait à observer les données et à les regrouper par significations communes. Au cours de ce travail préparatoire, effectué avant l'intervention en classe, les 33 segments représentant dans ce corpus le modèle « Nom [+ Adjectif] + *de l'identité nationale* » ont été distribués dans six champs sémantiques distincts. Les deux groupes dans la figure 3, par exemple, nous ont semblé faire référence respectivement aux questions relatives à l'identification du sujet de discussion et aux conflits suscités dans la polémique.

Ensuite, dans l'exploitation en classe qui est décrite au paragraphe suivant, les mêmes segments ont été également interprétés et puis groupés différemment par les étudiants. Une phase préalable d'introduction au sujet de discussion (multiculturalisme, immigration et identité nationale) et une activité d'approche au travail sur corpus ont assuré le lien entre les concordances et un contexte plus large culturel et cognitif.

modèle de colligation	**« Nom [+ Adjectif] + *de l'identité nationale* »**	préférences sémantiques **champs sémantiques identifiés**
question **question [controversée]** **question [« importante »]** **conception** **conception [française]** **définition** **symboles** **vision** **vision [française]** **thème** **thématique** **versions**	de l'identité nationale	**identification du sujet de discussion**
crise **bataille** **contestation** **défense** **défenseur** **troubles**	de l'identité nationale	**allusion à des conflits**
(…)		**(…)**

Figure 3. Exemples de préférences sémantiques.

2. L'exploitation en classe des unités lexicales étendues : repérer, grouper, négocier

L'expérimentation pédagogique qui suit le travail préparatoire décrit ci-dessus a été menée auprès d'une classe de 22 étudiants de FLE inscrits à la Queensland University of Technology en Australie. Les étudiants provenaient de plusieurs filières non linguistiques et, selon leur niveau de français à l'entrée à l'université, avaient déjà suivi quatre semestres de cours (si débutants), ou deux semestres (non-débutants). Deux lycéens avec une bonne maîtrise de la langue française faisaient aussi partie de la classe. Le niveau de connaissance de la langue était intermédiaire (une moyenne de B1-B2, selon le Conseil de l'Europe, 2001).

Bien que la Queensland University of Technology soit dotée de laboratoires suffisamment équipés pour permettre aux participants une analyse directe du corpus, la médiation de l'enseignant/chercheur a été préférée essentiellement pour éviter un long entraînement technique et linguistique (« corpus training »). L'objectif du cours de français dans lequel l'expérimentation s'inscrit (sur une courte période, d'ailleurs) est plutôt de former les apprenants à une lecture avertie de la presse, notamment au sujet du multiculturalisme. Le développement des compétences linguistiques pour exprimer son opinion et pour argumenter est aussi encouragé. Néanmoins, le « corpus training » a été remplacé par l'activité 1, où l'on fait activer les étudiants à la découverte des unités lexicales étendues, des rapports que les mots-clés entretiennent avec leur co-texte (dans les collocations) et leur contexte (dans la

transcription d'un document oral, puis dans des extraits du corpus de la presse française et à nouveau dans la même transcription pour en vérifier les similarités). La circularité du processus engendré lors des activités 1 et 2 peut être retrouvée dans la synthèse reproduite dans la figure 4.

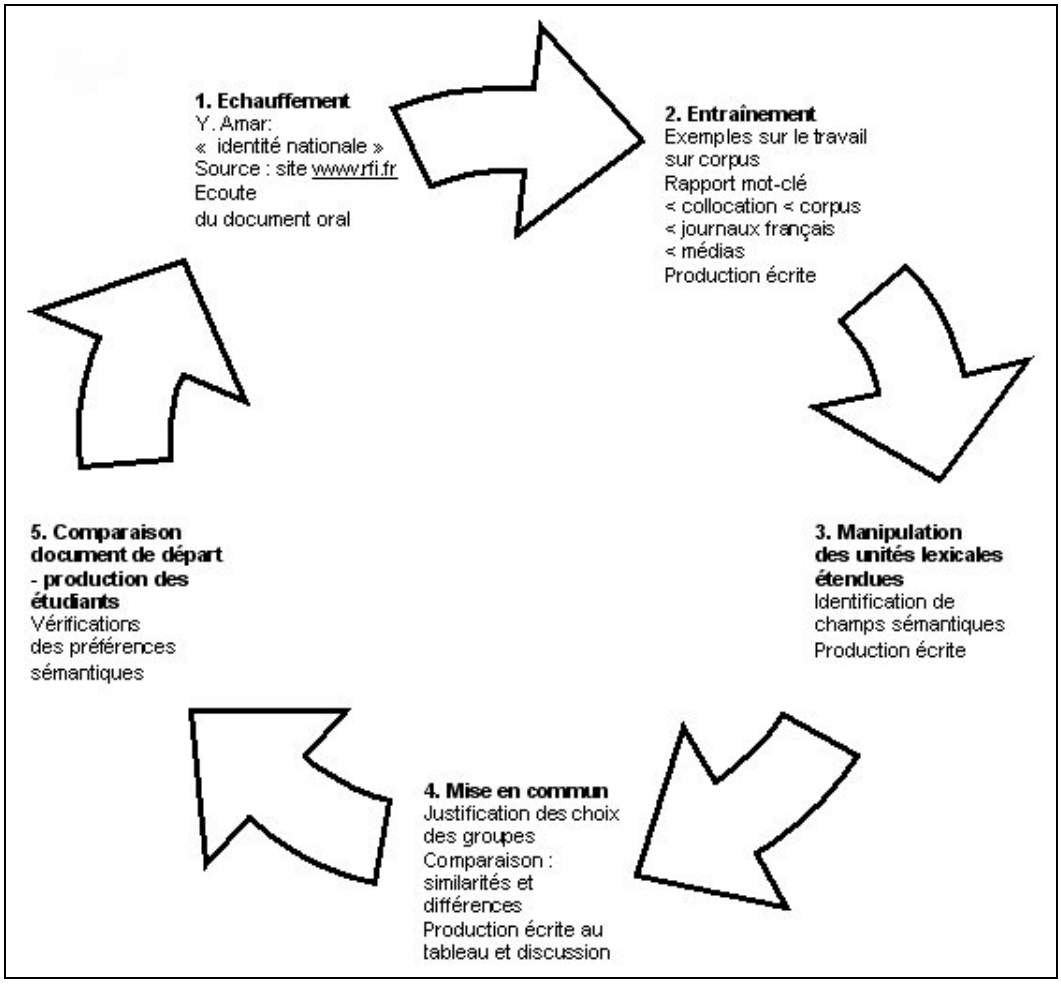


Figure 4. Exploitation en classe.

2.1. Analyse linguistique et négociation des significations

Le passage de l'interprétation individuelle des données du corpus à la négociation des significations est un aspect crucial de l'expérimentation. Il a demandé un choix préalable de notre part pour ce qui concerne les modalités d'observation et d'évaluation des étudiants en classe. Pour ce faire, nous avons exclu l'enregistrement vidéo des échanges oraux, puisqu'une telle démarche aurait empêché la spontanéité des discussions. Il nous a semblé plutôt préférable de demander aux étudiants de s'accorder pour écrire leurs interprétations dans un dossier de groupe[7], ce qui signifie que la négociation des significations devait se

7. Les dossiers sont reproduits en annexe.

vérifier d'abord au sein de chaque groupe pour définir les mots à reporter sur papier. Ensuite les deux groupes de la classe devaient trouver les similitudes et les différences entre leurs résultats respectifs. La justification de chaque choix d'un groupe par rapport à l'autre représentait donc une mise en commun des résultats et en même temps visait une deuxième négociation des significations. La comparaison finale avec la transcription du document oral utilisé comme déclencheur en début de cours a permis une vérification ultérieure des hypothèses de signification concernant les syntagmes extraits du corpus. Finalement, l'observation des interprétations des étudiants de la part du chercheur a été conçue comme analyse de la production écrite sur les points sollicités par l'enseignant et dans le dossier. Le choix de travailler avec des matériels sur papier a permis de garder des traces aisément analysables dans la recherche.

L'efficacité de la négociation des significations pour le développement des capacités communicatives et interculturelles ne nous a pas semblé mesurable de façon précise lors d'une si courte période d'expérimentation, qui ne prévoyait que trois interventions de deux heures chacune sur trois semaines. Néanmoins, la production écrite de plusieurs documents a été demandée aux étudiants pour qu'ils réfléchissent expressément sur la portée du travail abordé :

- un questionnaire initial sur les connaissances préalables, respectivement au sujet de l'immigration (*remue-méninges : association de mots*), sur le mot « banlieue » (*explication*) et sur la représentation (*stéréotypée ou personnelle*) des Français (*compléter la phrase*…) ;
- un questionnaire final pour la prise de conscience d'un certain nombre de stratégies d'apprentissage lexicale ;
- une entrée du blog personnel sur le réseau de classe expressément consacrée à la description de l'identité australienne, stimulant la réflexion « en miroir » sur les représentations des identités nationales ;
- une entrée ultérieure du blog sur les stratégies de lecture habituelles et sur l'expérience des activités en classe.

Ces dernières consignes concernant la production écrite sur un blog dédié[8] ont amené à une variété d'observations dont nous présentons ici un aperçu.

2.2. Aperçu des réactions des étudiants

Nous avons sélectionné de courts extraits des réflexions écrites par les participants à l'expérimentation. D'abord on essaiera de passer rapidement en revue quelques aspects problématiques relevés par les apprenants, puis on reportera certains des aspects positifs. Pour protéger la confidentialité des étudiants, selon le règlement de l'université, toutes les citations sont anonymes. A l'intérieur des

8. Voir les consignes dans la partie finale des dossiers des activités, en annexe.

citations, nous avons indiqué en gras ce que nous identifions comme mots-clés pour les observations à suivre.

- L'aménagement des documents authentiques : faut-il simplifier ?

Je trouve qu'une approche plus **simple** est mieux pour moi.

Cette remarque nous a semblé renvoyer aussi bien au caractère multiforme et complexe des langues qu'à la question de l'organisation des phases d'apprentissage ou bien à l'aménagement adapté des documents. L'analyse de corpus, avec les multiples exemples fournis et leur variété de significations, reproduit la nature combinatoire créative de la langue (attestée par l'usage en structures semi-idiomatiques), ce qui peut contraster avec une version traditionnelle simplifiée à l'usage pédagogique (normée par une logique rationnelle et cohérente, souvent inefficace dans le passage à la traduction).

- Le travail de déduction : faut-il prendre du temps pour arriver à des certitudes ?

J'ai pris beaucoup de **temps** pour faire des activités **certainer** [sic].

Le temps pour la familiarisation avec la « pré-méthodologie » (Tognini-Bonelli, 2000) du travail sur corpus semble constituer le problème majeur, autant pour les apprenants (Kennedy & Miceli, 2001 ; Yoon & Hirvela, 2004) que pour les enseignants qui doivent intégrer ces interventions dans un curriculum de langue étrangère. Nous observons, en outre, que les étudiants préfèrent parfois recevoir des certitudes directement du professeur plutôt que de se lancer à la découverte des significations en autonomie. Mais cette attitude peut aussi bien changer en cours de route, comme c'est le cas d'ailleurs pour cet étudiant, qui déclare, plus tard, avoir fini par se prendre au jeu :

Mais après le premier choc, je l'aimais parce que je pense que c'était intéressant.

- Compréhension globale vs. analyse lexicale : pourquoi et comment chercher les nuances de signification ?

Je ne comprends pas vraiment le **but** des activités de corpus.

Les finalités des analyses linguistiques peuvent contraster parfois avec les stratégies de compréhension globale, et, en conséquence, dérouter les étudiants dans leur apprentissage. Apparemment élémentaire, la compréhension du but de l'utilisation des corpus en classe représente en effet un autre point critique de ce genre d'investigations (Yoon & Hirvela, 2004 : 271).

J'ai eu un problème avec cela, parce que ce n'était pas quelque chose que je n'ai **jamais pensé** de [sic].

Comme dans le témoignage de Sammi dans l'étude de Yoon et Hirvela (2004 : 274), une approche jamais adoptée à l'avance provoque sans doute des difficultés

d'adaptation, mais après l'effort initial, généralement, la satisfaction d'avoir augmenté ses compétences garantit la validité de la démarche.

- Repérer les liens entre les mots : sémantique, syntaxe et modèles.

> Maintenant j'ai plusieurs manières d'identifier de nouveaux mots et d'observer les **liens** entre les mots.

Les liens entre les mots en termes de structures combinatoires semi-figées sont évidemment l'objet de l'étude des corpus : « most everyday words do not have an independent meaning, or meanings, but are components of a rich repertoire of multi-word patterns that make up text » (Sinclair, 1991 : 108).

> [Cette expérience] m'a fait réfléchir encore plus sur l'importance… de chercher à **contextualiser** chaque mot en considérant le message que l'auteur veut transmettre, et de ne pas aller seulement chercher la première traduction qu'on trouve sur le **dictionnaire**.

Malgré le fait que l'investigation ne comprenait aucune sollicitation spécifique pour définir le rôle de la consultation d'un corpus par rapport à la consultation d'un dictionnaire, l'apprenant ouvre une discussion qui est en fait souhaitée dans les suggestions de Chambers (2005 : 121) pour la poursuite des recherches en ce sens et qui est aussi prise en compte dans les interviews de Yoon et Hirvela (2004 : 273). Ensuite, l'importance donnée au contexte le plus proche (co-texte) et en même temps l'extraction des données d'analyse de leur propre contexte plus large constituent ce qui peut apparaître autrement comme le paradoxe de la linguistique de corpus (cf. Chomsky, 1965). Néanmoins, les travaux sur corpus ont apporté un progrès remarquable à la lexicographie, notamment pour aider les apprenants à bien choisir un mot parmi les différentes options de traduction, comme signalé aussi dans l'extrait qui suit :

> Les **stratégies**… sont utiles si on veut découvrir le **vrai sens** d'un mot pour bien le comprendre.

Comme l'affirment Corda et Marello (2004 : 66), dans les pratiques pédagogiques pour l'apprentissage lexical, isoler les exercices de vocabulaire est contre-productif, tandis que manipuler le vocabulaire dans son contexte d'usage permet une meilleure acquisition car l'apprenant est impliqué dans l'élaboration des nouvelles informations.

- Les stratégies d'apprentissage : négocier pour interpréter ?

Les deux derniers extraits des réflexions des étudiants font référence à la mise en commun des interprétations et aux effets des échanges de points de vue. Nous voudrions y voir les bienfaits d'une pratique qui conduirait au développement des compétences de communication en interaction. Dans le premier l'accent est posé sur les différences :

> C'était intéressant de voir qu'une personne pourrait regrouper certains mots ensemble et **une autre personne** les grouperait avec d'autres choses.

Le dernier témoigne de l'appréciation du partage :

> La situation où nous avons travaillé ensemble dans les petits **groupes** était très bien.
> J'aime les **discussions**.

Nous souhaitons que l'intérêt suscité par la découverte de nouvelles interprétations, basées sur d'autres valeurs de jugement, ouvre le chemin vers une meilleure compréhension interculturelle.

Conclusion

Nous avons abordé ici quelques observations préliminaires sur les données obtenues lors d'une expérimentation pédagogique en didactique du français. Cette expérimentation prévoyait des applications de la linguistique de corpus conçues de manière à favoriser la négociation des significations, dans le but d'encourager le développement des compétences communicatives et interculturelles. Des activités basées sur la compréhension et la production écrites ont servi de déclencheurs pour les comparaisons entre pairs des interprétations linguistiques, tout en analysant des mots et des expressions qui relevaient des représentations des étrangers. En plus des supports pédagogiques sur papier, fournis dans les annexes, cette étude inclut quelques-unes des réactions que les participants ont reportées dans leur blog scolaire à propos du travail fait en classe.

ANNEXES

ANNEXE 1. Dossier activités 1 et 2

Activité 1

Objectif : Préparation à l'utilisation de concordances. Approche des notions de « noyau », « fréquence », « concordances » et « collocations » comme extraits partiels d'un corpus de mots.

1. Vous allez **écouter** un extrait tiré de la rubrique « Les mots de l'actualité », daté du 17 mars 2007. Le journaliste, Yvan Amar, fait une courte description de la séquence de mots « identité nationale ».
2. **Relisez** le texte et **repérez les lignes** où Yvan Amar cite les mots ou séquences de mots qui suivent :

Liste par ordre alphabétique :

- étrangers ……..ligne/s _____
- identité ……..ligne/s _____
- identité française ……..ligne/s _____
- identité nationale ……..ligne/s _____
- jeunes ……..ligne/s _____
- immigration ……..ligne/s _____
- ministère ……..ligne/s _____
- ministère de l'immigration et de l'identité nationale ……..ligne/s _____

Ne comptez pas le titre.

Nous avons ainsi relevé et localisé les « noyaux », c'est-à-dire les mots-clés de notre recherche.

---

3. **Comptez** combien de fois les « noyaux » paraissent dans le texte et classez-les par ordre de **fréquence**.

_____	_____
_____	_____
_____	_____
_____	_____
_____	_____

4. **Comparez** vos résultats avec l'autre groupe.

Nous avons donc une idée de la *fréquence* de ces mots dans ce texte.

5. Retrouvons maintenant le **contexte** le plus proche du « noyau » **<u>identité nationale</u>**, 3 mots à gauche et 3 mots à droite :

l'identité nationale
identité nationale,
L'identité nationale,
l'identité nationale
l'identité nationale

Nous avons trouvé les « concordances » du « noyau » **<u>identité nationale</u>**.
Chacune de ces 6 lignes représente une « collocation » du « noyau » que nous avons choisi.

6. Ensuite, nous observons les « concordances » suivantes du groupe **identité nationale**. Elles sont tirées des articles de plusieurs quotidiens en langue française par le biais d'un logiciel en ligne, *GlossaNet*. L'échantillon ici reproduit concerne les résultats obtenus de la presse du 19 mai 2007.
   Repérez les mots qui se répètent.

Corpus: L'Union de Reims - Date: 2007/05/19
e a changé : à « l'immigration » et « l'identité nationale » ont été ajoutés « l
un « ministère de l'immigration et de l'identité nationale », déclenchant la col
rsé combinant Immigration, Intégration, Identité nationale et codéveloppement. A
e l'Immigration, de l'Intégration, de l'Identité nationale et du Codéveloppement
Corpus: L'Express - Date: 2007/05/19
e l'immigration, de l'intégration, de l'identité nationale, et du codéveloppemen
Corpus: Le Progrès de Lyon - Date: 2007/05/19
coopération, immigration, intégration, identité nationale > Xavier Darcos : Edu
Corpus: Le Figaro - Date: 2007/05/19
e l'Immigration, de l'Intégration, de l'Identité nationale et du Codéveloppement
Corpus: Le Journal du Jura - Date: 2007/05/19
d'un Ministère de l'immigration et de l'identité nationale, un concept forgé par
Corpus: Le Temps - Date: 2007/05/19
questions économiques, sur celles de l'identité nationale et de la politique ét
e l'Immigration, de l'intégration, de l'identité nationale et du codéveloppement
obtient l'Immigration, l'intégration, l'identité nationale et le codéveloppement
autres: le travail, avant tout, puis "l'identité nationale" et enfin le rejet de
ndividuelles, comme le lien fait entre "identité nationale" et immigration, ont
eux Ministère de "l'immigration et de l'identité nationale" reviendrait au fidèl
efeux, qui regroupe l'immigration et "l'identité nationale", mais il y a adjoint
e la ministre de "l'immigration et de l'identité nationale"? Il est trop tôt pou
capitalisme et au marché. La crise de l'identité nationale, dont les signes sont

_____
_____
_____
_____
_____

L'édition en ligne de chaque quotidien est ici considérée comme un *corpus* : un ensemble de textes électroniques disponibles pour analyser l'usage de la langue des médias. Cette approche fait partie de la « Corpus Linguistics ».

Activité 2

Objectif : Aborder les mots utilisés par la presse et faire des hypothèses sur le débat autour de l'identité nationale en France.

1. Pour retrouver le sujet de la discussion autour de l'identité nationale, nous allons examiner maintenant les parties des « concordances » qui présentent la structure :

   nom + **de l'identité nationale**

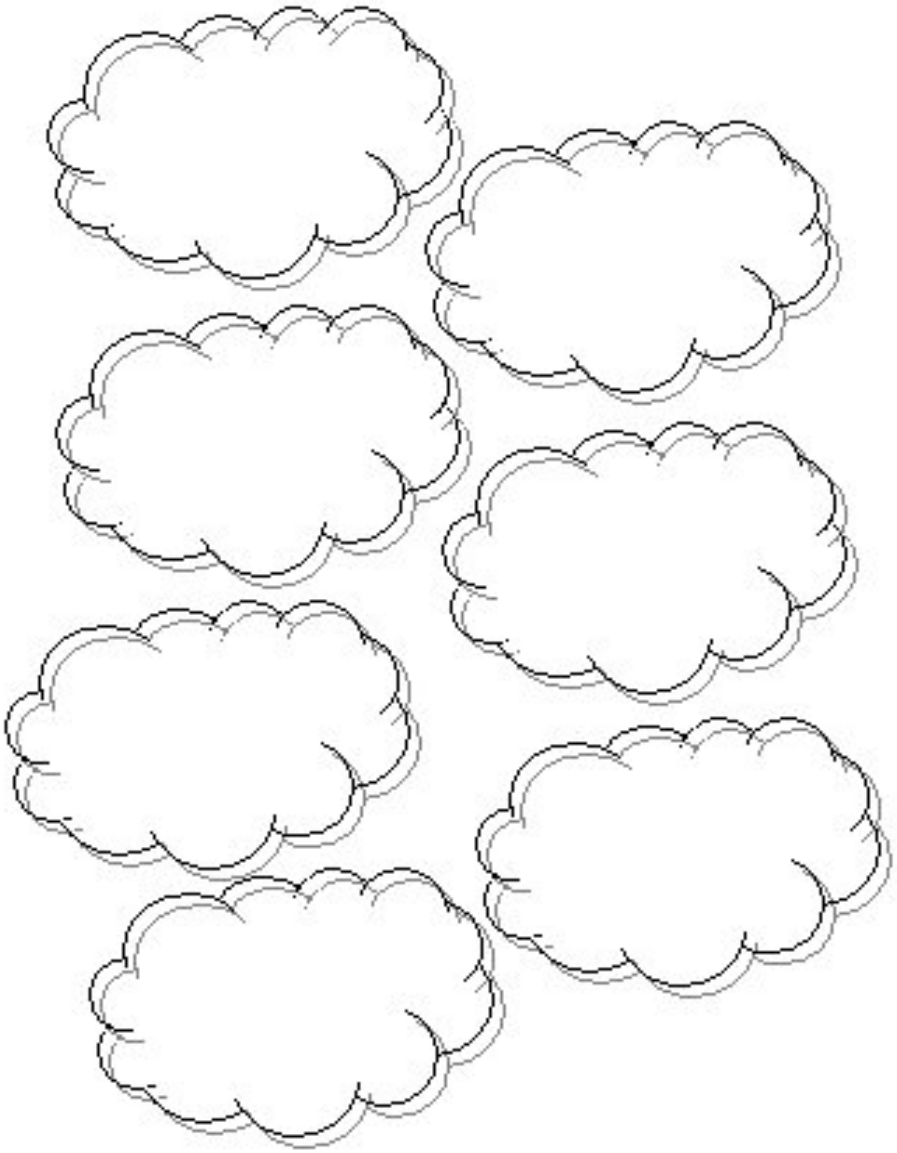
   comme, par exemple *la crise de l'identité nationale.*

   A cette fin, vous trouverez ici une sélection par type des mots parus dans la presse entre le 26 mars et le 26 juin 2007.

affirmation bataille combinaison composantes conception conception [française] contestation crise défense défenseur définition exaltation immigration [et] Immigration [et] intégration, Intégration,	de l'identité nationale

ministère ministères obsession question question [controversée] question [« importante »] raffermissement refrain remise [en cause] symboles thème thématique troubles versions vision vision [française]	de l'identité nationale

2. En utilisant cette liste, relevez les mots qui ont une signification commune et transcrivez-les dans les bulles, par groupes.



3. Décrivez la signification commune à chaque groupe.

_____ _____
_____ _____
_____ _____

_____ _____
_____ _____
_____ _____

_____ _____
_____ _____
_____ _____

_____ _____
_____ _____
_____ _____

4. Comparez vos choix avec l'autre groupe.

5. Reprenons le texte d'Yvan Amar et comparons-le avec les sujets de la discussion sélectionnés dans la presse.

6. Discussion.

Dans votre blog sur Blackboard, vous réfléchissez à ce qui d'après vous caractérise l'identité australienne. Par exemple, vous pourriez parler de la population, du territoire, des symboles, des stéréotypes.

Merci beaucoup pour votre collaboration
et bon courage !

Lucia Drago
PhD Student
in Intercultural Communication

ANNEXE 2. Dossier activité 3

Objectif : Explorer des expressions utilisées dans la presse pour représenter certains aspects de la diversité culturelle en France.

1. Vous avez ici une sélection des termes choisis par la presse pour nommer la population qui réside actuellement en France. Essayez de préciser quelles caractéristiques sont effectivement rappelées exclusivement par les mots utilisés. Associez donc aux expressions suivantes les numéros correspondants aux caractéristiques que vous relevez.

Arabes _____
Beurs _____
beurettes _____
clandestins _____
enfants de l'immigration _____
enfants d'immigrés _____
étrangers _____
Français d'origine algérienne _____
Français de souche _____
Français musulmans _____
France 'Black, Blanc, Beur' _____
immigrés de deuxième génération _____
issus de la diversité culturelle _____
issus de l'immigration _____
issus de l'immigration maghrébine et d'Afrique noire _____
jeunes des banlieues _____
jeunes des cités _____
jeunes franco-algériens _____
jeunesse des banlieues françaises _____
jeunesse issue de l'immigration _____
migrants _____
musulmans _____
noirs de France _____
nord-africains _____
primo-arrivants _____
ressortissants étrangers _____
sans-papiers _____

2. Essayez maintenant d'expliquer les adjectifs qui sont utilisés pour désigner les quartiers populaires où l'on observe de graves problèmes sociaux.

cités sensibles
zones urbaines sensibles
quartiers sensibles / difficiles / défavorisés
banlieues sensibles / difficiles / défavorisées

sensible : _____
_____
_____

difficile : _____

_____

_____

défavorisé : _____

_____

_____

Dans votre blog sur Blackboard, vous réfléchissez à vos stratégies de lecture habituelles. Pouvez-vous les décrire ? Pensez-vous que l'expérience de cette approche linguistique pourrait améliorer vos stratégies ? Pourquoi ?

Merci encore une fois pour votre collaboration!

Lucia Drago
PhD Student
in Intercultural Communication

BIBLIOGRAPHIE

BARONI, Marco, Silvia BERNARDINI & Stefan EVERT, 2006. « A WaCky Introduction. » In M. Baroni & S. Bernardini (éds.), *Wacky! Working Papers on the Web as Corpus*. Bologna : Gedit, p. 9-40.

BEACCO, Jean-Claude, 2000. *Les dimensions culturelles des enseignements de langue*. Paris : Hachette.

BERNARD, Philippe, 2007. « Nicolas Sarkozy et l'identité nationale. » *Le Monde.* http://www.prochoix.org/cgi/blog/index.php/2007/03/19/1281-nicolas-sarkozy-et-l-identite-nationale-philippe-bernard-le-monde, page consultée le 29/06/08.

CASTAGNOLI, Sara, 2006. « Using the web as a source of LSP corpora in the terminology classroom. » In M. Baroni & S. Bernardini (éds.), *Wacky! Working Papers on the Web as Corpus*. Bologna : Gedit, p. 159-172.

CHAMBERS, Angela, 2005. « Integrating corpus consultation in language studies. » *Language Learning & Technology,* 9/2, p.111-125.

CHOMSKY, Noam, 1965. *Aspects of the Theory of Syntax*. Cambridge, MA : MIT Press.

CONSEIL DE L'EUROPE, 2001. *Cadre européen commun de référence pour les langues : apprendre, enseigner, évaluer*. Paris : Didier.

CORDA, Alessandra & Carla MARELLO, 2004. *Lessico : Insegnarlo e impararlo*. Perugia : Guerra.

DRAGO, Lucia, 2006. « Corpus en ligne : mode d'emploi. Des idées pour enseigner le français avec des corpus. » *Carnet Austral*, 25, p. 12-19.

FURSTENBERG, Geneviève, Sabine LEVET, Kathryn ENGLISH & Katherine MAILLET, 2001. « Giving a virtual voice to the silent language of culture: the Cultura project. » *Language Learning & Technology*, 5/1, p. 55-102.

GlossaNet. http://ling.fltr.ucl.ac.be/index.php, page consultée le 18/06/08.

KENNEDY, Claire & Tiziana MICELI, 2001. « An evaluation of intermediate students' approaches to corpus investigation. » *Language Learning & Technology*, 5/3, p. 77-90.

KRAMSCH, Claire & Steven THORNE, 2001. « Foreign language learning as global communicative practice. » http://language.la.psu.edu/~thorne/KramschThorne.html, page consultée le 29/06/08.

MOIRAND, Sophie, 2007. *Les discours de la presse quotidienne : observer, analyser, comprendre*. Paris : Presses Universitaires de France.

PRAT-ZAGREBELSKY, Maria Teresa, 2004. « I corpora nella descrizione e nella didattica delle lingue : una nuova risorsa per gli insegnanti. » *Lingua e nuova didattica,* 33/1, p. 22-30.

SINCLAIR, John, 1991. *Corpus, Concordance, Collocation*. Oxford : Oxford University Press.

SINCLAIR, John, 1996a. « The Search for Units of Meaning. » *Textus*, 9/1, p. 75-106.

SINCLAIR, John, 1996b. *COBUILD English Dictionary: Helping learners with real English*. Londres : HarperCollins Publishers.

TOGNINI-BONELLI, Elena, 2000. « Corpus classroom currency. » *Darbai ir Dienos*, 24, p. 205-244.

VAN DIJK, Teun Adrianus, 1988. « How 'they' hit the headlines: ethnic minorities in the press. » In G. Smitherman-Donaldson & T. van Dijk (éds.), *Discourse and Discrimination.* Detroit, MI : Wayne State University Press, p. 221-262.

YOON, Hyunsook & Alan HIRVELA, 2004. « ESL student attitudes toward corpus use in L2 writing. » *Journal of Second Language Writing,* 13/4, p. 257–283.