Linguistic Concerns in Teaching with Language Corpora

Natalie Kübler

University Paris 7 & Paris-Nord University

This paper shows how the Web-based environment developed for language teaching is currently being adapted and extended. It deals with the implications from the linguistics point of view, that if corpora-users hope to extract interesting and useful information, working with corpora requires a firm grasp of linguistics. Corpora must thus be used to teach linguistics, and especially, linguistics related to NLP. The acquisition of a firm grounding in linguistics obtained by intelligent use of corpora, can then lead the users to work more efficiently in their specific fields, such as term extraction, translation, or language teaching with corpora.

0. Introduction

This paper deals with two issues closely related, one leading to the other and vice-versa:

How can a Web-based environment that has been developed for language teaching be extended, augmented, and used to teach linguistics?

Why does the linguistic information we use in our corpora needs some thinking over?

This environment we are referring to was developed at the University of Paris 13 ; it migrated to Paris 7 in Fall 1999 and has been used there since then, in teaching and research at the department of Intercultural Studies and Applied Languages, in LSPs, and the Language Industry and Specialized Translation option. It shall be applied to Technical witing in the coming academic year.

The context in which these issues have been raised will be described, followed by the main aims to be reached. The tools and corpora students and researchers have access to will then be described. This will lead me to detail the method and illustrate it with data and examples. I shall finally conclude with the results obtained and future prospects.

1. Context

This Web-based environment : *WALL*, which stands for " Web-Assisted Language Learning " was originally created to meet needs in teaching authentic and specialized English to French-speaking students in computer science. The general philosophy was to take advantage of the (then) recent developments in all sorts of language ressources, and of the development of the Web, which is now considered as a normal tool for teaching, as well as a source for linguistic resources. This development has proved to be sufficient to offer greater accessibility to varied kinds of "real data", to various corpus linguistic tools, but also to various on-line demonstrations of natural language processing tools.

2. Objectives

Our objectives are based on the concrete needs of the Intercultural Studies and Applied Languages Department at the university of Paris 7. In this department, the students of other departments (e.g. Biology, Litterature, History, Physics, Psychology, etc.) are given English courses for specific purposes. These courses are compulsory.

Future translators in specific subject areas are also trained in this department and are offered two options: a law option and a language industry option. In this option, students must read introductory courses to linguistics and more specialized courses in terminology and translation studies. However, as they will more and more be required to work with corpora, a sound basis in linguistics is essential.

In the LSP section, we aim at introducing compulsory work on computers for specialized English courses and we offer our linguistics teachers language data and tools they can use in their courses. Data manipulation is also automated as much as possible, and lay users are provided with user-friendly tools. Researchers in our department already access corpora for example to study French idioms built with prepositions ; researchers in cultural studies are currently trying to obtain specific corpora to work in their field. But the main research objectives consist in using our already existing corpora to obtain more linguistic information and reinsert it into our tools, and to collect more varied specialized corpora. This means for example describing the syntactic structures and arguments of verbs in LSPs in order to tag them with more information. In the context of our departement, it is most necessary to be able to study the distinction between general language and LSPs on new basis, i.e. data-driven (Habert et al. 1997).

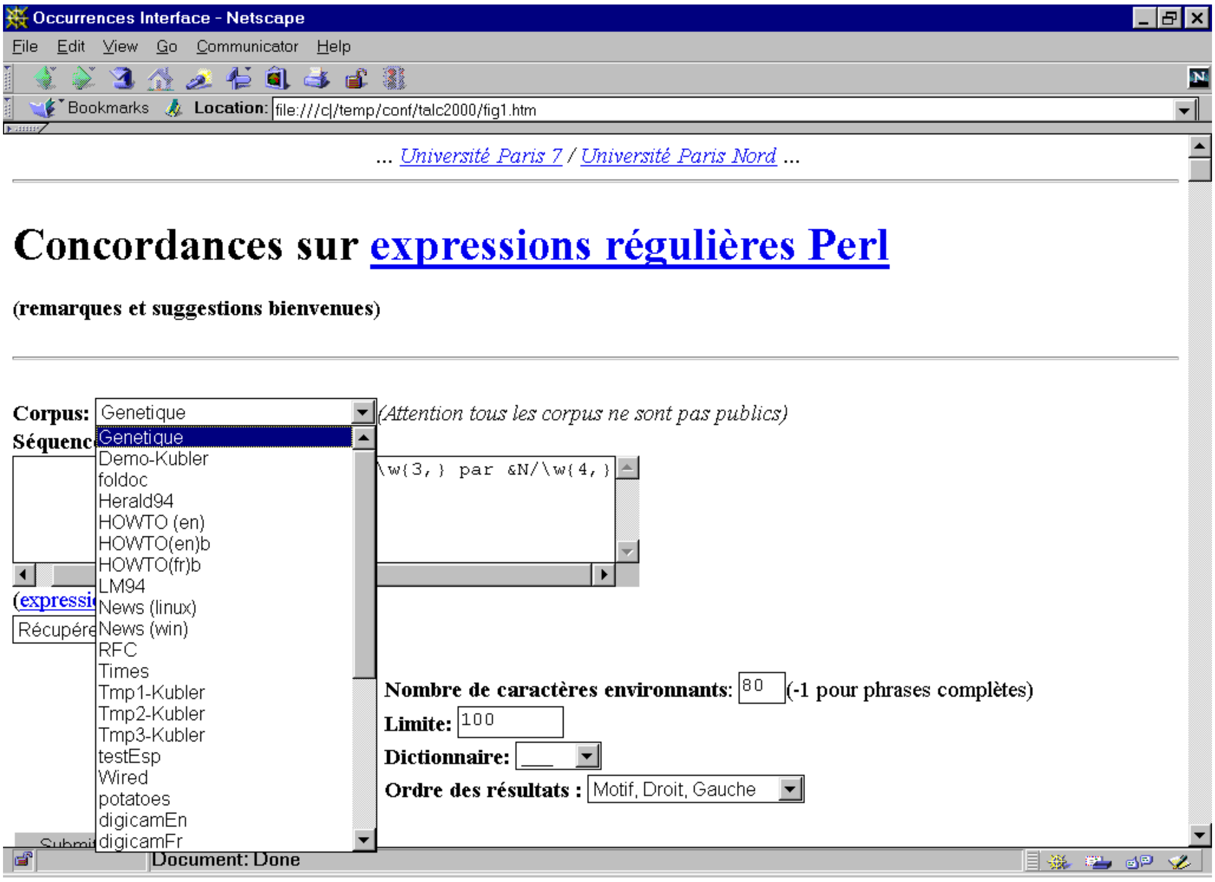
Students in the department work on very different types of subject areas. Our aim is to collect as much different specialized corpora as possible, which is not always easy, because not everything is digitalized. Subjects like wine-making in France for example, are very difficult to deal with, because experts have their secrets they do not want to convey to other people.

Our department's pedagogical objectives are therefore quite varied: teaching LSPs using corpora, teaching how to use corpora in terminology and translation, and using corpora to teach linguistics. This last objetive must be the keystone of the whole program.

3. Tools and Corpora

The tools embedded in the environment are robust enough to deal with various types of texts. Corpora and tools run under Unix, but users access corpora and tools on a Web site, via an Apache server. Current corpora are monolingual and bilingual in French and English. We are currently introducing trilingual corpora in French, English, and German. We use some general language corpora, collected from newspapers and of course to check the degree of specialization of a term, we can use the BNC. But we mainly deal with specialized subject areas, as our department is specialized in this field, especially for the translation and terminology aspects.

Corpus-query tools consist in a concordancer and a tokenizer that can sort words using various criteria ; both are based on perl-like regular expressions using POS, but no disambiguation.



Location: file:///c|/temp/conf/talc2000/fig1.htm

... Université Paris 7 / Université Paris Nord ...

Concordances sur expressions régulières Perl

(remarques et suggestions bienvenues)

Corpus: Genetique (Attention tous les corpus ne sont pas publics)

Séquenc

```
\w{3,} par &N/\w{4,}
```

(expressi

Récupére

Nombre de caractères environnants: 80 (-1 pour phrases complètes)

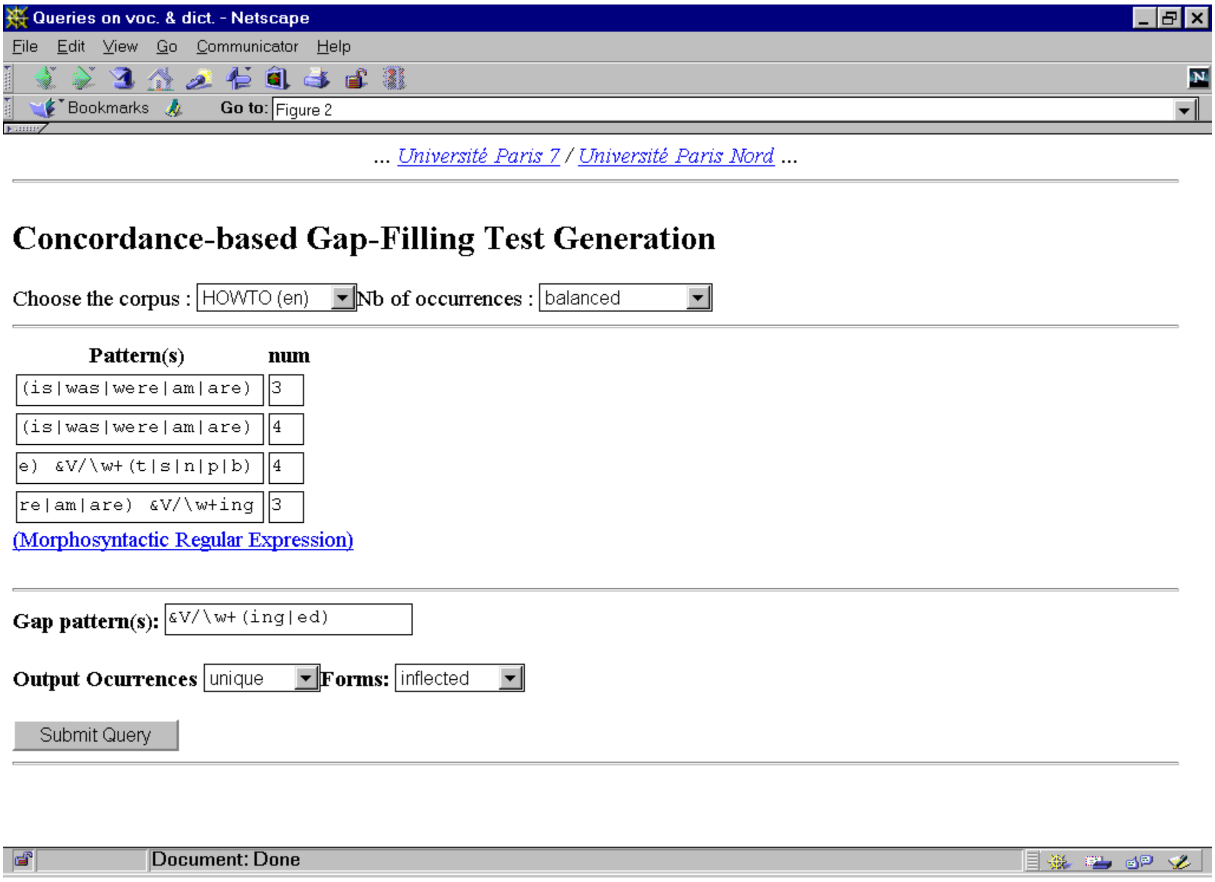
Limite: 100

Dictionnaire: ___

Ordre des résultats : Motif, Droit, Gauche

Figure 1 : Concordancer using perl-like regular expressions

Tools that are normally used by teachers, but also in linguistics courses, allow users to automatically generate varied sorts of exercises, which are then automatically corrected. The most interesting tool for linguistics courses is, apart from the concordancer, the gap-filling exercise generator, which is itself based on concordances. Users must describe the sequences they want to be found, and the words or series of words they want to be deleted in the exercise. Figure two shows the interface of the gap-filling exercise generator.


Queries on voc. & dict. - Netscape
File Edit View Go Communicator Help
Bookmarks
Go to: Figure 2
Document: Done

... Université Paris 7 / Université Paris Nord ...

Concordance-based Gap-Filling Test Generation

Choose the corpus : HOWTO (en) Nb of occurrences : balanced

Pattern(s)	num
(is|was|were|am|are)	3
(is|was|were|am|are)	4
e) &V/\w+(t|s|n|p|b)	4
re|am|are) &V/\w+ing	3

(Morphosyntactic Regular Expression)

Gap pattern(s): &V/\w+(ing|ed)

Output Ocurrences unique Forms: inflected

Submit Query

Figure two: gap-filling exercise generating tool

Using this tool means studying linguistics from a corpus-based point of view. Students have to think of the way language really works in texts to succed in writing exercise-generating queries. In this respect, we aggree with Kettemann (1997 :186), who demonstrates that teaching linguistics with corpoara is “ rewarding and necessary ”.

4. Data and examples

Our students are not linguists, but linguistics is hiding in every corner in their syllabus, as they work with languages, translate many different texts, build term bases etc.; they are not studying Natural Language Processing, but they need to know how it works to become expert users of the NLP tools that they will have to become familiar with. Helping them to become conscious of what kind of linguistic information they, as human beings, use to understand or utter sentences can be done using our tools. This leads them to understand the issues raised by NLP tools. After some theoretical introduction about lexicon, morphology, syntax, etc., which they usually more or less understand , they are asked to find concordances on various specific linguistic points that are related to NLP. They have to think of how, with which regular expression, they are going to use to query the corpus and they quickly find out the issues at stake in linguistic analysis. This approach is close to the “ Micro- and Macrolinguistics approach ” Peters (1997 :185) adopts to teach linguistics.

The next subsections show some very simple, but very powerful examples.

4.1. Sentence segmentation

Students are asked to think of possible definitions of sentences from a string point of view. It takes some time for example to find out that a sentence could be defined by punctuation. Students can then test this hypothesis, using a regular expression that defines a sentence as being a certain number of characters between two periods.

(1) \. .{0,100} \. Defines a period, followed by between 0 and 100 characters (all types), followed by a period.

This very simple definition gives the following result in one of the French corpora (LM94):

Ailleurs, Dityvon le restera toujours **. Il est régulièrement oublié des fonds d achat, commandes, bourses, prix, festivals .** Il n appartient pas à une grande

BP 3, 75430 Paris Cedex 09 (tél**. : 42-46-70-38) .**

La succession de M **. Marchais à la tête du parti continue, d autre part, de faire l objet du black-out le plus total .** Pris d une étrange logorrhée, on s

Altman à Paris sur Prêt-à-porter **. Passer à la réalisation ? Elle ne l exclut pas .**" Mais j aime trop le côté " coulisse "

la confédération libérale et centriste **. " C'est un document d union, très proche de la plateforme du PR ", renchérit un négociateur du RPR .**Ainsi se présentent les deux programmes

If the first sentence is complete, the second one is not, as *tél.* is a French abbreviation for “ phone ” and is, of course, followed by a period. With this first simple definition, sentences will be cut after each abbreviation, or words like *.38-colt*, beginning with a period. The third example demonstrates that there are other sentence segmenters, such as questions marks, etc… After this first step, students are ready to confront more complex issues.

4.2. Multi-word units

Concordances on multi-word units are an interesting way of explaining the issues raised by word segmentation : human beings are not conscious of the complex processes involved with recognizing a word, categorizing it, understanding it. Students are brought to think of a way of formally defining words. They quite quickly come to the conclusion that a word is a string of characters between two separators. Word segmenters in French are not only spaces, but also hyphens or apostrophes. This segmenting function is relative, as each of those segmenters can also link two units together to build a multi-word unit. The example below that are preceded by “ a. ” are multi-word units, the ones preceded by “ b. ” show several words separated by various segmenters :

(2)	a.	*marge de manoeuvre*	“ room for manoeuvre ”
	b.	*2 millions de francs*	“ 2 million francs”
	a.	*double-cliquez*	“ you double-click ”
	b.	*expliquez-vous*	“ explain yourself ”
	a.	*méga-octet*	“ megabyte ”
	b.	*a-t-il compris ?*	“ did he understand ? ”
	a.	*les modes CRUS de la saisie 12-BIT*	“ the CRUS modes of the 12-bit typing ”
	b.	*Porte de Versailles, 17-20 mars 2000*	“ Porte de Versailles 17-20 March 2000 ”
	a.	*aujourd’hui*	“ today ”
	b.	*j’attendais d’elle*	“ I expected from her”

Students are not always able to make the difference between a multi-word unit and two different words conventionally related by a hyphen, or an apostrophe. Trying to describe and to analyse different word structures on the corpora has proven a good way to open up students’ eyes.

4.3. POS ambiguity

As said before, the corpora that are used are not disambiguated ; it could be possible to use precisely tagged corpora, applying taggers like Cordial for French for example, or using what is done in the department of linguisitics at the university of Paris 7. But disambiguation has been put aside on purpose : from the research point of view, more questions can be asked when the frame is not too rigid. This is one the points Sinclair made clear (Sinclair 1981 for example), and, for certain tasks, I completely agree with him.

For teaching linguistics and NLP issues, it is more fruitful to have no disambiguation. Students have been asked, for example, to look for French multi-word units composed of a noun, followed by the preposition *de*, followeb by another noun, such as *pomme de terre* (“ potato ”) *droit de vote* (“ voting rights ”) etc. This can be obtained with the following query sequence : &N de &N, in which “ &N ” means a noun. Here is a sample of the first results obtained :

(3) publiées à l occasion de l **abrogation de la** loi Falloux ne manquera pas d intriguer
pénales empêchent désormais l **acquisition de la** nationalité. Cette démarche s effectue
L'essentiel de l **action de l** armée algérienne est tourné vers une
et les religieux aussi. Mais, grands **amateurs de certitudes** , ces derniers se sont alliés aux
du marxisme et d un culte de l **argent de plus** en plus effréné.
des fêtes de fin d année, lorsque l **attention de l** opinion publique et de la classe
hors de ce qui leur était permis, à l **attribution de subventions** à l enseignement privé.
de la Banque d Espagne que les **augmentations de capital** en cours seraient insuffisantes.
des citadins bâtissant à la hâte des **barricades de sable** et des secouristes circulant en
la guerre scolaire montre à nouveau le **bout de son** nez, comment rester insensible à la

Because of the absence of disambiguation, words, such as *la*, *l’, son* are tagged both as determiners and as nouns. They therefore appear in the basic definition of compound nouns that was given before. A simple solution consists in asking for words of a minimum of three characters to avoid determiners that can also be considered as nouns. The results would look like the following :

(4)	ont signé, jeudi 30 décembre, un	***accord de fusion***	concernant la majeure partie de
	" La simplicité est toujours ce qu'il y	*__a de plus__	difficile à conquérir, explique-t-il.
	pas fermer les yeux sur le risque d un	*__afflux de cent__	à cent cinquante mille réfugiés
	il a su limiter son rôle à celui d'"	***agent de liaison***	" entre les négociateurs catholiques et
	L'	**allocation de parent**	isolé passe à 3 081 francs pour une
	et les religieux aussi. Mais, grands	**amateurs de certitudes**	, ces derniers se sont alliés aux
	du marxisme et d un culte de l	*__argent de plus__	en plus effréné.
	hors de ce qui leur était permis, à l	**attribution de subventions**	à l enseignement privé.
	de la Banque d Espagne que les	***augmentations de capital***	en cours seraient insuffisantes. La
	après Aix-en-Provence, loin des grands	***axes de communication***	, Jouques, 3 000 habitants, pourrait
	des citadins bâtissant à la hâte des	***barricades de sable***	et des secouristes circulant en canots
	la guerre scolaire montre à nouveau le	*__bout de son__	nez, comment rester insensible à la
	évoluent autrement, sur des	**bouts de terrain**	où laisser des traces susceptibles de
	son entreprise et un autre a lancé un	***cabinet de communication***	. Comptable au chômage, une femme
	label bruxellois réputé pour son	**catalogue de pop**	-rock raffiné. De là à parler de
	exclue du lycée Emmanuel-Mounier pour	*__cause de foulard__	islamique ? Que veulent en fait ces
	des examens dans les locaux du Service	**central de protection**	contre les rayonnements ionisants
	apaisement. La représentation française	**change de mains**	. La direction de la coopération va
	chinois font florès, les Indiens sont	***chauffeurs de taxi***	, les Coréens tiennent les kiosques à
	scandales qui avaient entraîné jadis la	*__chute de son__	gouvernement.
	les juges d instance et les	**commandants de brigade**	de gendarmerie. La manifestation de
	. Ces derniers, pas plus que les	***commissariats de police***	ne peuvent accueillir la démarche

Obviously, all problems are not solved: some of the examples above are compounds (in italics), others are parts of compounds (underlined), others are not compound, but collocations bold), and finally, the rest is not composed of “ noun de noun ”, but of other POS (preceded by an asterisk). This kind of issue takes students some time to sort out, but once it is clear, it helps them greatly to understand what linguistic analysis means, and why NLP systems have failures. It can be especially helpful here, because they have to deal with machine translation and analyse translation errors.

As our students work intensively on LSPs, it is very useful to show them that general dictionaries cannot deal with specialised terms, especially terms other than nouns (i.e. specialised verbs, specialised adjectives, adverbs). The following query leads to finding forms that could be verb forms, i.e. ending in *–ing* or *-ed* : \w+(ing|ed). Testing this sequence on a computer science corpus reveals verb forms such as “zipped, unzipped, gzipped ”. The sequence \w*zip\w* finds all occurrences of all derived forms around “ zip ”, which leads to all possible verbs built and derived from the program noun “ zip ” (“ zip ” is a program that allows the user to compress data).

(5) (before the hyphen) as an argument, it **unzips** (keeping the original intact) then
are in /usr/doc/faq/howto/ and are **gzipped** . The file names are XXX-HOWTO.gz, XXX
the CD- ROM) and all the sources as GNU **zipped** tar files. Supporting files such as a

Once these problems are understood, students have to design rules that would disambiguate the words. This means that they must describe the syntactic environment of compound nouns to be able to hit compound nouns only. For verbs, it means analysing the syntactic structures and the various arguments that are allowed with the specific verbs. The verb -- and derived verbs of -- “ to zip ” for example, does not have the same syntactic structure and the same arguments in computer science English (cf. example (5) above) as in general English (cf. example (6) below) :

6) is no small feat for a man who recently **unzipped** his lip rather publicly on such
who roams the globe buying and selling, **unzipped** the silver hood he'd fastened over his
War. By the mid-'30s, photographs could **zip** across continents and oceans by wire or
Three-passenger cars are designed to **zip** along at 30 miles per hour on raised
car, Juice ..."'), heard the theories **zipping** along the communications highway,
$11 an hour - picks up a walkie talkie, **zips** up his blue jacket and leaves the booth

Analysing and describing the syntactic and argumental structure of verbs (or other categories) naturally allow us to link descriptive linguistics with translation studies.

4.4. Translation problems

In specific subject areas, terminology is a big problem when translating. Students use aligned and comparable corpora to extract the terminology of a subject, and to find possible translations. They have worked in the field of computer science and digital cameras, and will probably have to deal with other subjects in the coming years.

Below are two translation examples from English into French. In both cases, it is not possible to separately translate each element of the multi-word unit French. Literal translation will not be correct :

7) plastic card /= *carte en plastique => carte de crédit, carte bleue (France only)

smart card /= *carte futée => carte intelligente

The “ smart card ” example is a good example of collocation : it is not a completely set phrase, at least in French. However, the translation must definitely be *carte intellingente* ; *futée* represents a level of language that cannot be used in the context. More complex problems appear with translating (Kübler et al. forthcoming).

5. Conclusion

To conclude, it can be stated that using corpora to teach linguistics prompts very interested reactions among students : since they can test hypotheses, make mistakes, and finally obtain results, they grasp possible problems much more quickly and are able to take some distance from their mothertongue. When they are later shown how taggers or more complex machine translation software works, they can understand what is going on, without having a complete theoretical background in natural language processing. Corpora seem then to be a “ natural ” tool to approach linguistics and NLP from an applied point of view.

For the future, as our students are constantly collecting corpora for terminology projects, we are going to augment the available corpora. Since the Web-based environment we use is quite easy to modify, it is possible to quickly add more corpora. The system can also be augmented with all linguistic information that has been collected while working with the already existing corpora.

Another path lead to studying and elaborating the methodology with which corpu use must be approached and taught.

References

Foucou, Pierre-Yves and Natalie Kübler. 1999. “ A web-based language learning environment: General architecture. ” In Schulze, Mathias, Marie-Jose Hamel, and June Thompson, (eds.) 1999. *Language Processing in CALL. ReCALL Special Publication,* Hull, pp 31-39.

Kettemann Bernhard. 1997. “ Using a Corpus tau Evaluate Theories of Child Language Acquisition. ”. In Wichmann, Anne, Steven Fligelstone, Tony Mc Enery, Gerry Knowles (eds). *Teaching and Language Corpora*. 1997, Longman, London, pp.186-194.

Kübler, Natalie and Pierre-Yves Foucou, forthcoming. “ Teaching English Verbs With Bilingual Corpora : Examples in the Computer Science Area ”. in S. Granger (ed) : *Contrastive Linguistics and Translation Studies*, Benjamins Publishing : Dordrecht.

Peters Pam, 1997. “ Micro- and Macrolinguistics for NLP ”. In Wichmann, Anne, Steven Fligelstone, Tony Mc Enery, Gerry Knowles (eds). *Teaching and Language Corpora*. 1997, Longman, London, pp.175-185.

Renouf, A. 1997. “ Teaching Corpus Linguistics tau Teachers of English. ” In Wichmann, Anne, Steven Fligelstone, Tony Mc Enery, Gerry Knowles (eds). *Teaching and Language Corpora*. 1997, Longman, London, pp.255-266.

Johns T. and P. King. (eds), 1991. “ Classroom Concordancing ”. 4. *English Language Research Journal*, Birmingham University.

Sinclair J. 1991. *Corpus, Concordance, Collocations*. Oxford University Press, Oxford.

Habert Benoît, Adeline Nazarenko and André Salem, 1997. *Les linguistiques de corpus*. Armand Colin, Paris.