PROFILE *7* (2006): *77-86*
Universidad Nacional de Colombia, Facultad de Ciencias Humanas, Departamento de Lenguas Extranjeras

Modals in Brazilian Advanced EFL Learners' Compositions: A Corpus-Based Investigation[1]

Los verbos modales en composiciones de alumnos avanzados de inglés como lengua extranjera de nacionalidad brasileña: Una investigación basada en corpus

Vander Viana*
vander.viana@terra.com.br
Catholic University of Rio de Janeiro (PUC-Rio), Brazil

Corpus Linguistics has become a major trend in Applied Linguistics since the second half of the 20th century due to computing facilities. Nowadays teachers can research and assess their students' production by means of compiling learner corpora. This article describes how this technique was used to investigate the usage of modals in the writing of advanced EFL learners studying at private language schools in Brazil. When the research corpus is compared to the academic prose register studied by Biber et al. (1999), the divergence becomes apparent. The findings, thus, suggest that subjects write in a non-proficient way, which runs counter to previous expectations.

Key words: Corpus Linguistics, learner corpus, modals, proficiency, writing

La Lingüística de *Corpus* se ha convertido en una de las principales tendencias en el área de Lingüística Aplicada desde la segunda mitad del siglo XX debido a la popularización de los ordenadores. Actualmente, los profesores pueden investigar y evaluar la producción escrita de sus alumnos a través de la compilación de *corpora*. Este artículo describe cómo se ha empleado esta técnica para investigar el uso de verbos modales en la producción escrita de alumnos de nivel avanzado de inglés como lengua extranjera de cursos libres en Brasil. Al comparar el *corpus* de investigación con el registro prosa académica estudiado por Biber et al. (1999), se hace evidente una discrepancia. Los resultados sugieren que los sujetos de investigación escriben de una manera no-proficiente, lo que se contradice con las expectativas iniciales.

Palabras claves: Lingüística de *Corpus*, *corpus* de aprendiz, verbos modales, proficiencia, escritura

---

[1] I would like to thank Sonia Zyngier for her illuminating criticism of an earlier version of this article.

***Vander Viana** is currently a MA student in Language Studies at the Catholic University of Rio de Janeiro (PUC-Rio), and has been teaching English as Foreign Language for several years. He has been a member of the APLIERJ Committee (www.aplierj.com.br) since 2004 and of the REDES Project (www.letras.ufrj.br/redes) since 2003.

This article was received on March 31st, 2006 and accepted on May 19th, 2006.

INTRODUCTION

Whenever some kind of language activity is practiced nowadays, corpora tend to come into the picture. It is true, as one may argue, that most English teachers are not fully aware of corpora and corpus-based studies. However, this should not stop researchers from thinking of ways in which corpora may illuminate language teaching. In fact, corpus-based information is part of English teachers' lives. One example is the recently launched six-level course *Top Notch*, a series which includes 'corpus notes' from the beginners' level. In the concise methodology for this course, Saslow and Ascher (2006, p. Txiii) write that

> informed by the Longman Corpus Network – Longman's unique computerized language database of over 328 million words of spoken and written English as well as learner errors – *Top Notch* provides concise and useful information about frequency, collocations and typical native speaker usage.

Materials such as the one illustrated above may help teachers get to know a little bit about the advantages of using the principles of Corpus Linguistics in the classroom. Another possible application of such principles concerns the mapping of students' performance at any stage of the teaching/learning process, as Leech (1998, p. xiv) points out in the preface to *Learner English on Computer*:

> let us suppose that higher education teacher X, in a non-English speaking country, teaches English to her students every week, and every so often sets them essays to write, or other written tasks in English. Now, instead of returning those essays to students with comments and a sigh of relief, she stores the essays (of course with the students' permission) in her computer, and is gradually building up, week by week, a larger and more representative collection of her students' work. Helped by computer tools such as a concordance package, she can extract data and frequency information from this 'corpus', and can analyse her students' progress as a group in some depth.

Leech's words reveal the assumption that teachers can do research in Corpus Linguistics. This may diminish the gap between teaching and researching. Besides, findings are of a twofold nature: they fill a gap in the Applied Linguistics panorama and at the same time help teachers structure their teaching practices.

This study makes use of Corpus Linguistics to investigate a specific area of English grammar, namely, modals. The objective here is to analyze the way Brazilian advanced EFL students from private language schools use modals in their compositions and contrast the results with those obtained by Biber, Johansson, Leech, Conrad, and Finegan (1999) in their mapping of the oral and written production of speakers of English as a first language. Private language schools were chosen because it is in this setting that the author of this paper works.

The main questions which guide the present study are the following:

(a) Do Brazilian EFL learners at an advanced stage make use of modals in their writing in a way which is similar to that of speakers of English as a first language?

(b) If not, what are the differences between these two groups?

These questions will be addressed after the theoretical discussion below.

REVIEW OF LITERATURE

This section covers two main aspects focused on in this paper, namely, Corpus Linguistics and modals. The first sub-section offers a brief explanation of some terms such as corpus and Corpus Linguistics. In addition, it also spells out the purpose of corpus-based research. In the second part, the theory of modals is presented from the perspective of three different descriptions, and their subtypes are explained and exemplified.

Still, in this section, a review of two corpus-based studies is presented, covering, directly or indirectly, the usage of modals by EFL learners.

Corpus Linguistics: An Overview

The notion of corpus as a "collection of written or spoken texts" (Wehmeier, 2000, p. 295) has been around for a very long time, but it was only in the 20th century that it took up new meanings in the area of linguistics. As McEnery and Wilson (1996, p. 21) put it,

> in principle, any collection of more than one text can be called a corpus: the term 'corpus' is simply the Latin for 'body', hence a corpus may be defined as any body of text. […] But the term 'corpus' when used in the context of modern linguistics tends most frequently to have more specific connotations than this simple definition provides for.

A corpus, according to Tognini Bonelli (2001, p. 2), is "a collection of texts assumed to be representative of a given language put together so that it can be used for linguistic analysis".

Corpus Linguistics can be described "as the study of language based on examples of 'real life' language use" (McEnery & Wilson, 1996, p. 1). It is this specific feature of exploiting natural language which distinguishes Corpus Linguistics from Chomskyan tradition. While the former is actually interested in the investigation of real examples of language in use, the latter focuses on artificial, contrived samples. According to Sinclair (2003, p. ix), "before large amounts of data were easily available, most of the generalisation had to be done by intuitive guesswork; pre-Corpus linguists were not able to check their notions". In other words, Corpus Linguistics deals with the probability of language use whereas the rationalist view is concerned with language abstractions. So far, they remain two different ways of looking at language and a common ground has not been arrived at.

The present study is in tune with language use as it analyzes learners' written production. By using such data, teachers can develop their own research, as is the case here, and find out, for instance, which aspects should be focused when teaching a specific language. This objective is in agreement with the following quotation in which Granger (2004, p. 291) states that

> Computer Learner Corpora (CLC) allegedly serve two main purposes: (1) by providing a better description of interlanguage and a better understanding of the factors that influence it, they contribute to Second Language Acquisition theory; and (2) they contribute to the development of pedagogical tools and classroom practices that more accurately target the needs of the learner.

Therefore, corpus-based studies are of great help to language teachers. Once teachers are aware of their students' needs, they will be able to fully achieve their goals.

Modals

For a long time modals have had a relevant place in English grammars. Here is, for instance, Swan's (1998, p. 333) description:

> The verbs *can*, *could*, *may*, *might*, *will*, *would*, *shall* (mainly British English), *should*, *must* and *ought* are called 'modal auxiliary verbs'. They are used before the infinitives of other verbs, and add certain kinds of meaning connected with certainty or with obligation and freedom to act […]. *Need* […] and *dare* […] can sometimes be used like modal auxiliary verbs, and the expression *had better* […] is also used like a modal auxiliary.

Swan considers 'ought' to be a type of "modal auxiliary verb". He also holds the notion that 'need', 'dare', and 'had better' can be used as "modal auxiliary verbs" without distinguishing them from

the ones listed in the beginning of the excerpt. This is perhaps due to the fact that Swan's (1998) *Practical English Usage* offers practical presentations of grammar topics. Although it is stated in the introduction that "the book is intended for intermediate and advanced students, and for *teachers of English*" (Swan, 1998, p. xi – my italics), he argues he is not "writing for specialists". Therefore, it is assumed that "where it has been necessary to use grammatical terminology, I [he] have [has] generally preferred to use traditional terms that are well-known and easy-to-understand" (Swan, 1998, p. xi). In spite of being clear, Swan's explanation does not fit the purpose of this study, which requires a more detailed description of modals and their usage.

A different approach to the analysis of the English language is offered by the *Collins Cobuild English Grammar* (Sinclair, 1990), which is corpus-based. It "attempts to make accurate statements about English, as seen in the huge Birmingham Collection of English Texts" (Sinclair, 1990, p. v). In other words, its information is derived from real samples of language in use. In this grammar, modals are described as "a special kind of *auxiliary verb*" (Sinclair, 1990, p. 217) encompassing 'can', 'could', 'may', 'might', 'must', 'ought to', 'shall', 'should', 'will' and 'would'. Other verbs such as 'dare', 'need' and 'used to' are grouped in a subtype labeled 'semi-modals'.

From the perspective of a more recent corpus-based grammar – the *Longman Grammar of Spoken and Written English* (Biber et al., 1999), modals are divided into three groups, namely, 'modals', 'marginal auxiliary verbs' and 'semi-modals'. The first group encompasses 'can', 'could', 'may', 'might', 'shall', 'should', 'will', 'would' and 'must'. These modals (Biber et al., 1999, p. 483) have a number of specific features such as (a) being invariant forms, (b) preceding the subject in yes-no questions and (c) being followed by a verb in the bare infinitive.

Marginal auxiliary verbs correspond to 'need (to)', 'ought to', 'dare (to)' and 'used to'. According to Biber et al. (1999, p. 484), these verbs are rare and almost only present in British English.

Fixed idiomatic phrases as '(had) better', 'have to', '(have) got to', 'be supposed to' and 'be going to' are called semi-modals by Biber et al. (1999, p. 484). They differ from central modals because they can be marked for both tense and person. Besides, they can also occur as non-finite forms.

For the scope of this study, Biber et al.'s (1999) description of modals is taken into account. Their analysis seems to be more accurate since grammatical features are considered within each register analyzed in the grammar (academic prose, newspaper language, conversation and fiction).

The nine modals which are grouped by Biber et al. (1999) in their first category are also referred to as "central modal verbs" in Wilson's (2005) study. The author states these modals have received great attention from scholars due to their high semantic complexity (Wilson, 2005, p. 151).

One example of such type of study is Mindt's (1996) "English Corpus Linguistics and the Foreign Language Teaching Syllabus". In this paper, Mindt argues Corpus Linguistics has had an influence in dictionaries and grammars, but EFL teaching materials remain unchanged. One of the sections of the paper covers the topic of modals. Using a part of the London-Lund Corpus, he argues that 'would', 'can' and 'will' are the most common modals in his research corpus. Considering that "the present forms occur more frequently in main clauses that the past forms" (Mindt, 1996, p. 234) and that 'will' is an extremely frequent modal in conversations in English, he proposes that German EFL textbooks should introduce such modal in the first year of study instead of doing it in the second year. In other words, the presentation of 'will' should not be postponed in favor of the infrequent modals 'must' and 'may'.

Another study is Ringbom's (1998) compilation of vocabulary frequencies, which also covered some modals in the writing of learners of English from seven different nationalities (Dutch, Finnish-Swedish, Finnish, French, German, Spanish and

Swedish). The former seven sub-corpora, part of the International Corpus of Learner English (ICLE), were compared to the Louvain Corpus of Native English Essays (LOCNESS) which comprises argumentative essays written by American and British students. Unfortunately, there is not a thorough explanation regarding the 110 most frequent words which are presented in the article. It is possible, however, to notice some differences in terms of usage by the different groups of subjects. It seems that all groups of learners overuse 'can' and underuse 'would' and 'will'. One only exception remains with the French group, which overuses 'will'. In relation to 'should', the French, Finnish and Germans tend to use it more than Americans and the British whereas the Spanish, Finnish-Swedish, Swedish and Dutch generally underuse it. As far as the modal 'could' is concerned, there are three distinct results, namely: (a) Finnish learners use it as much as Americans and British; (b) Spanish EFL students overuse it; and (c) all the other five ethnic groups underuse it. Instead of offering a complete interpretation vis-à-vis the data presented in the article, Ringbom (1998, p. 51) states in the conclusion that the

> chapter has tried to show that a seemingly simple word frequency count may provide a useful starting point for many interesting small-scale projects where the general characteristics of advanced learner language as well as the relative importance of transfer and universal features can be further explored.

Although modals have already been studied by a great number of corpus linguists (cf. Wilson, 2005), there seems to be a lack of research focusing on the written production of Brazilian EFL learners.

METHODOLOGY

As this study is concerned with the usage of modals in compositions by Brazilian advanced EFL learners, it was necessary to compile a corpus representing such production. To this purpose, compositions written in English were collected in three private language schools located in six distinct areas in the city of Rio de Janeiro, Brazil.[2] A decision was made to collect compositions which were parts of exams or tests in order to ensure that the research subjects did not have any help of third parties throughout the writing process or that they did not cheat and/or copy specific parts of their compositions. They could, nonetheless, make use of dictionaries and/or grammar books if they were allowed to do so by the rules of each language school.

As the focus was on advanced students only, participants belonged to the last two terms in each of the three language courses. In other words, only students who were about to graduate were asked to contribute. There was an exception, though. One of the schools offered a specific writing course aimed at teaching students how to write effectively. In this specific language school, students from this special course were also invited to take part in the research.

The topic of the compositions came from the materials chosen by each language school. Therefore, freedom of choice was limited. Writers could choose from a maximum of three topics, but in some occasions they had only one mandatory writing task.

Compositions varied greatly in terms of length. The shortest one had 112 words and the longest, 478 (average 288 words).

After data collection, all compositions were typed so as to probe them by means of a computer program. The digital versions correspond to what was hand-written by students. Mistakes were maintained because they are representative of learners' writing. Only spelling problems were corrected; otherwise, the computer would read,

[2] I am thankful to the people and the institutions who made this research possible by granting their permission and helping me with data collecting. For reasons of privacy, they will be kept anonymous.

for instance, 'should' and 'shuld' as two different words. This would make data analysis more difficult since it would be necessary to go through the list of words in order to identify these problematic cases.

Even though every effort was made to have students produce compositions which would be representative of their own linguistic accomplishments, a few repeated fragments were found in some compositions, e. g. titles. Had these sequences been maintained, they would have constituted a problem in the final counting of lexical items. Therefore, these over-repeated sequences were excluded because they were in fact just a copy of the prompt given by the teacher.

At present the research corpus contains 155 compositions written by Brazilian advanced EFL students from three language courses in six areas of the city of Rio de Janeiro. The corpus totals 30,261 tokens (items) and 2,870 types (different words).

Analysis was performed with WordSmith Tools (Scott, 1999); more specifically, one of its tool, WordList, in order to obtain a list of most frequent words in the corpus. This list allowed the identification of the nine modals to be analyzed ('can', 'could', 'may', 'might', 'must', 'shall', 'should', 'will' and 'would') and their respective frequencies. In the second stage, the tool, Concord, was used to analyze the cotext[3] of these modals.

The reference corpus is the Longman Spoken and Written English (henceforth LSWE) Corpus on which the Longman Grammar of Spoken and Written English (Biber et al., 1999) is based. This grammar "describes the actual use of grammatical features in different varieties of English: mainly conversation, fiction, newspaper language, and academic prose" (Biber et al., 1999, p. 4). In the present research, it was decided to compare the results being reported here to those obtained by Biber et al. (1999) in their mapping of the academic prose register since both represent the written medium.

DATA ANALYSIS AND DISCUSSION

Eight out of the nine modals analyzed here can be grouped into two categories: those which refer to non-past time and those which can refer to past time (cf. Biber et al., 1999, p. 484-485). In the first group, there are 'may', 'can', 'will' and 'shall'; and in the second, there are, respectively, 'might', 'could', 'would' and 'should'. The difference in usage between modals which refer to non-past time and the ones which can refer to past time is noteworthy. Table 1 summarizes this contrast.

'May', 'can' and 'will' are at least three times more common than their counterparts, namely, 'might, 'could' and 'would'. The only exception is the pair 'shall' and 'should', the latter being much more common than the former. As a matter of fact, there are no instances of 'shall' in the research corpus. These results are similar to the ones found by Biber et al. (1999, p. 486) who state that "considering the pairs of central modals, the tentative/past time member is less frequent than

Non-past time		Past time		
Modal	Percentage	Modal	Percentage	Total
May	86.96%	Might	13.04%	100%
Can	84.33%	Could	15.67%	100%
Will	75.94%	Would	24.06%	100%
Shall	0.00%	*Should*	100%	100%

Table 1. Distribution of non-past/past modals in the research corpus.

[3] Sinclair (2003, p. 174) defines cotext as "the group of words that occur on either side of it in a text".

its partner in all cases except shall/should". The difference between the results of their study and the one being reported here is that in the learner corpus there are no instances of 'shall' whereas in the reference corpus this modal is present albeit rarely.

Modals may also be grouped into three categories according to the ideas they convey (cf. Biber et al., 1999, p. 489). They can express (a) permission, possibility or ability ('can', 'could', 'may' and 'might'), (b) volition or prediction ('will', 'would' and 'shall'), and (c) necessity or obligation ('should' and 'must'). Table 2 shows the number of times each modal occurs in the learner corpus.

Ideas	Modals	Occurrences
Permission Possibility Ability	Can	253
	Could	47
	May	20
	Might	3
Volition Prediction	Will	202
	Would	64
	Shall	0
Necessity Obligation	Should	58
	Must	31

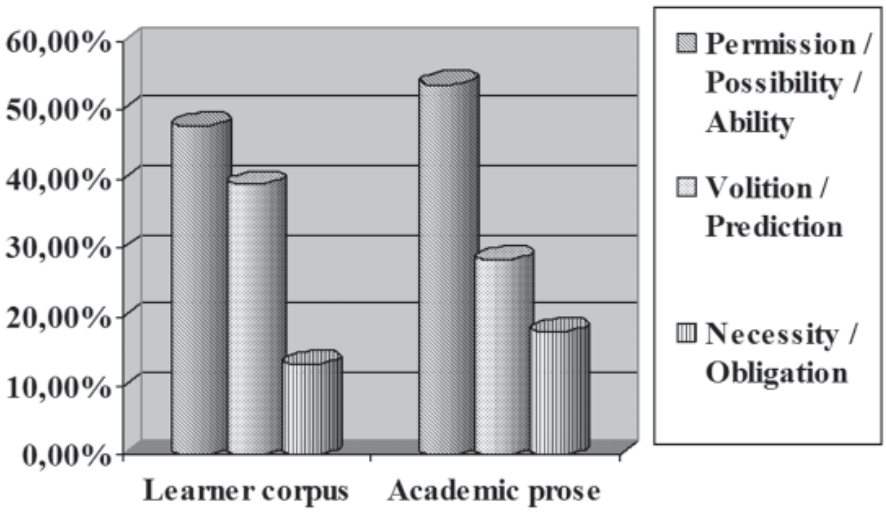
Table 2. Distribution of modals in the learner corpus.

It is then possible to compare the results of the present study to the mapping of modals in academic prose carried out by Biber et al. (1999, p. 489).

Graph 1 indicates that the Brazilian learners of English investigated in this study tend to underuse modals which mark both permission, possibility or ability as well as necessity or obligation. On the other hand, they show a tendency to overuse modals signaling either volition or prediction, especially with the use of 'will', as illustrated in Figure 1.

According to Biber et al. (1999, p. 489), 'will' and 'would' are least frequent in academic prose. The register which contains the highest frequencies of such modals is conversation. Therefore, the overuse of such modals in the research corpus may suggest that the research participants write in a way which is similar to the way speakers of English as a first language talk.

Another feature Biber et al. (1999) argue to be characteristic of academic prose is the use of verb phrases incorporating modals in the passive voice. As they hold, "passive voice with modals is rare in conversation and fiction, but relatively common for some modals in academic prose" (Biber et al., 1999, p. 499). They also conclude that "with the



Graph 1. Ideas conveyed by modals.

Volition	Prediction
But, even so, I *will* try to reach happiness and success, no matter what this may cost. [4]	"Christmas trees" *will* also be decorated with lights and Christmas ornaments.
After finishing my studies, I *will* start to work with my sister, who is a doctor, and has a little emergency hospital.	I'm sure that this *will* be a great experience and we will never forget it!
I *would* try to help them anyway.	because the most powerful country *would* impose its culture and, consequently, its language.

Figure 1. Examples of modals expressing volition and prediction.

passive, *can* and *should* are particularly common, *could* and *must* are also fairly common".

In the learner corpus, however, the picture is quite different. Most of the verb phrases which incorporate modals are actually in the active voice as can be seen in Graph 2.

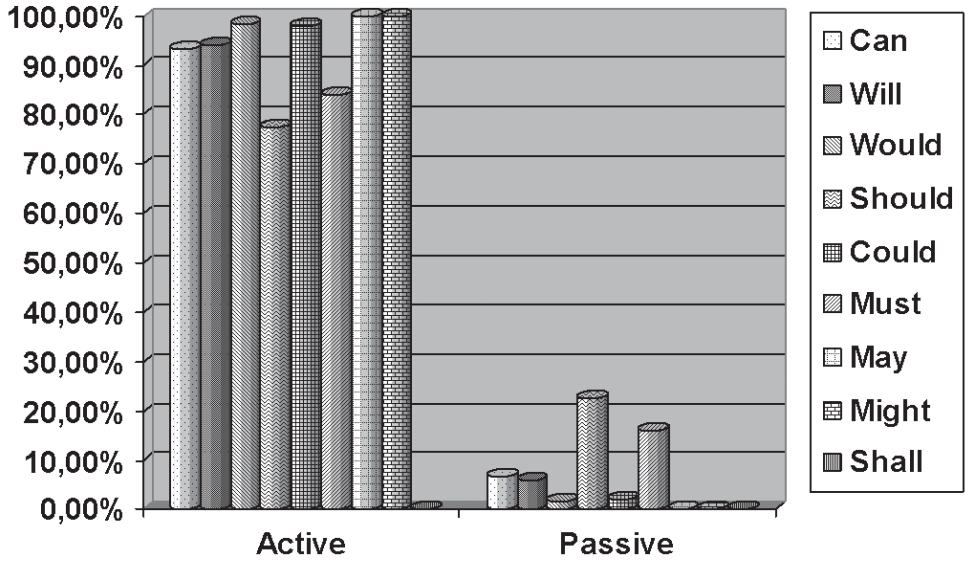
There are few instances of verb phrases in the passive voice as illustrated by the following examples:

That is why death punishment it is not a good idea, this *should not be accepted* in any country, in any constitution.

When a baby becomes a child, his or her growing *must be accompanied* attentively by his parents.

There are a lot of things that *can be done* to better your health.

If happiness brings health, friends and positive views of life, nothing *could* really *be considered* more important than it.



Graph 2. Distribution of modals in the active and passive voices.

[4] This example and all the following ones were taken from the research corpus and have not been corrected in any way as stated in the methodology section. The only exception concerns spelling.

Altogether there are only 49 instances of modal verbs being used in the passive voice, which represents 7.22% of the verb phrases containing modals in the research corpus. Once more it is possible to observe the gap between the learner corpus and the academic prose register investigated by Biber et al. (1999).

CONCLUSION

The results of the present study reveal that participants use modals in ways which diverge from those of speakers of English as a first language. When writing compositions, Brazilian learners of English tend to use structures which characterize the oral production of speakers of English as a first language. One indication of such result is the frequent usage of modals which signal either volition or prediction, especially the modal 'will' in the writing of Brazilian learners. Another indication which was reported in this article is the rare frequency of marked voice in verb phrases containing modals, a feature of the academic prose studied by Biber et al. (1999).

Besides the grammatical description, this study also has some pedagogical implications. The findings reported here may cast some light in the way modals should be taught to Brazilian learners of English. Teachers should raise their students' awareness of the topics discussed in this article, namely, the overuse of 'will' and the underuse of marked voice. By doing so, these learners will be able to write more proficiently and communicate their ideas more fluently.

Conducting corpus-based studies is of great importance to language teachers. Such a type of investigation makes it possible for the teacher/researcher to spot the most troublesome areas of English language as regards particular groups of students. As Tribble and Jones (1990, p. 23) put it, "even with very small classroom-based studies it is possible to come to some very interesting conclusions about the way students are dealing with English". These studies highlight learners' production, that is, they are based on what students actually write or say instead of considering abstract models of language. This can only be accomplished by means of Corpus Linguistics.

As a final comment, it is worth citing Granger's (2004, p. 299) words about the potential of learner corpora:

> Learner corpora may not yet have given rise to a large number of teaching and learning applications, but the buzzing activity in the field and the CLC-informed reference and teaching tools that have already been produced are concrete evidence of an ongoing trend which should result in highly innovative pedagogical application in the years to come.

REFERENCES

Berber Sardinha, A. P. (2004). *Lingüística de corpus*. São Paulo: Manole.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Granger, S. (1998). *Learner English on computer*. London / New York: Longman.

Granger, S. (2004). Practical applications of learner corpora. In B. Lewandowska-Tomaszczyk (Ed.), *Practical applications in language and computers: PALC 2003* (pp. 291-301). Frankfurt am Main: Peter Lang.

Leech, G. (1998). Preface. In S. Granger (Ed.), *Learner English on computer* (pp. xiv-xx). London / New York: Longman.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. Edinburgh: Edinburgh University Press.

Mindt, D. (1996). English corpus linguistics and the foreign language teaching syllabus. In J. Thomas & M. Short (Ed.), *Using corpora for language research* (pp. 232-247). London / New York: Longman.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 41- 52). London / New York: Longman.

Saslow, J., & Ascher, A. (2006). *Top notch fundamentals: Teacher's edition and lesson planner*. New York: Pearson Longman.

Scott, M. (1999). *WordSmith tools 3.0*. Oxford: Oxford University Press.

Sinclair, J. (1990). *Collins cobuild English grammar*. London / Glasgow: Collins.

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

Sinclair, J. (2003). *Reading concordances: An introduction*. Oxford: Oxford University Press.

Swan, M. (1998). *Practical English usage*. Oxford: Oxford University Press.

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

Tribble, C., & Jones, G. (1990). *Concordances in the classroom: A resource book for teachers*. Harlow: Longman.

Wehmeier, S. (2000). *Oxford advanced learner's dictionary of current English*. Oxford: Oxford University Press.

Wilson, A. (2005). Modal verbs in written Indian English: A quantitative and comparative analysis of the Kolhapur corpus using correspondence analysis. *ICAME Journal 29 (April 2005)*. Retrieved March 19, 2006, from http://gandalf.aksis.uib.no/icame/ij29/ij29-page151-170.pdf