Bringing corpora to the masses: free and easy tools for interdisciplinary language studies.

Alex Boulton

Alex Boulton. CRAPEL-ATILF/CNRS, Nancy-Université.
In N. Kübler (ed.), *Selected papers from Teaching and Language Corpora 2006*. Frankfurt: Peter Lang, p. 69-96.

Bringing corpora to the masses:
Free and easy tools for interdisciplinary language studies.

Alex Boulton
boulton@univ-nancy2.fr

Keywords
data-driven learning, free tools, practical applications, interdisciplinary, distance learning

Abstract
Despite huge promise, direct applications of corpus linguistics (CL) have yet to make substantial inroads to the language classroom. The main stumbling block is perhaps psychological: even when they are aware of the possibilities, teachers and learners alike may hesitate to change established practices and challenge traditional roles. It is difficult to alter such mindsets, but it should be possible to make corpus linguistics more user-friendly. In particular, we argue here that even using only free on-line tools with minimal training, learners are quick to see a variety of applications in language learning and other areas. This paper describes a CL option in an English MA course where students are allowed considerable freedom to discover the tools and possibilities for themselves, and pursue them to their own ends. An analysis of 30 resulting research papers highlights the diversity of approaches and topics which can be studied through CL, and discusses the wide range of skills learned in the process.

1. Introduction

Few fields of human inquiry have been left untouched by the advent of powerful computer tools for the processing of vast quantities of data. In language studies, this has given rise to complete new fields of study, including corpus linguistics (CL). CL in turn has made its influence felt in various areas connected with language teaching and learning, most especially in informing teaching and reference materials as well as considerations of content. However, and despite a plethora of research articles and projects, when it comes to direct applications in the classroom in the form of "data-driven learning" (Johns 1991a) or DDL, the trickle-down effect has yet to turn into the predicted torrent (Leech 1997:2).

To explain this lack of uptake, we cannot ignore a number of obvious potential disadvantages to DDL, whether real or perceived. For example, the changes to current practices might be thought too big to be absorbed in a single step, and perceived as not "serious". DDL definitely constitutes a challenge to teacher and learner roles: the former might resent a loss of power and control, while the latter might not like having to take greater responsibility for their learning. Both might also lament the "fuzzy" nature of language inherent in corpus work, preferring the certainty and comfort (even if illusory) which accompanies a traditional rule-based approach. Some might argue they have insufficient facilities in terms of computers, software and space, or the money to acquire

them. Even with extensive technological resources, teachers and learners can also feel threatened if they lack sufficient skills to use them. Some training would seem essential for DDL to be of use, but any "learning to learn" can be perceived as a waste of time, serving only to delay the language learning itself.

Many powerful arguments can of course be marshalled to overcome these objections. The widely reported advantages to DDL include promoting autonomy, language awareness and noticing skills in a learner-centred approach using authentic language by applying inductive, hands-on discovery learning. As all of these are central in current research on language teaching and learning, it might be argued that objections to DDL are largely a gut reaction. In a nutshell, learners and teachers simply aren't convinced.

If DDL is to break out of its current research confines and reach a wider audience, it should be seen to offer maximum return on minimum investment. For this to happen, a number of things are necessary:

- **The tools should be free**. Many working teachers are unaware of the considerable resources available free on the internet; all that is needed to exploit them is a single computer with internet access and a printer. A fully-equipped computer laboratory and expensive software certainly allow greater possibilities, but they are by no means essential to DDL.
- **The software should be easy to use**. Specialists aside, few people are likely to exploit the full potential of the sophisticated software available; more important is that it should be user-friendly and ready to use with only a few minutes' exploration, even for those with limited ICT skills. Freely-available software has increasingly intuitive interfaces, and is frequently accompanied by concrete examples, guided tours, tutorials, help facilities, etc.
- **The techniques should be flexible and transferable**. If CL is a methodology rather than a discipline (e.g. Scott & Tribble 2006:4), then it might have applications in a variety of areas above and beyond linguistics and language learning (itself a derivative). The more widely it is used for a variety of purposes, the more likely it is to continue to be used.

This paper presents a course in corpus linguistics which attempts to put these criteria into practice. In particular, we discuss the tools that learners find free on the internet, and look at the topics they choose to explore when allowed to use them to pursue their own interests.

2. An option in corpus linguistics

The course in question is part of an MA in English offered entirely at distance by the Centre de Télé-enseignement Universitaire (CTU) at Nancy Université in France, and the typical student profile is correspondingly quite specific (see also Boulton & Wilhelm 2006). Most of our learners are mature students with French as a mother tongue, but enrolments come from all over the world. They have relatively advanced language abilities as well as a specialist interest in English, and are comparatively autonomous and used to managing their work alone. Motivation is thus a key factor in such an isolated learning environment, as

students try to reconcile their degree with work and family commitments in sometimes very difficult circumstances.

2.1. Course considerations

In accordance with the prevailing paradigm in France, language forms the common core of the MA supported by three major options in literature, cultural studies and linguistics, and a number of subsidiary specialisations (e.g. French as a foreign language). Of the three main options, linguistics is the least well-known, and statistically the least popular. It is important then to find a course which appeals to a wide range of interests, does not require too much prior knowledge, and relates to other studies. CL seems to fit this bill on a number of counts. On a practical level, it is unlikely students will have come across it before, thus placing them all on an equal footing despite their varied prior experience of linguistics in general. Apart from a computer and access to the internet, no extra costs are involved in the form of required reading or expensive tools, nor any specialist knowledge of ICT. The focus is on learning about CL by applying it to language areas students are interested in. This highly practical approach aims to downplay the common perception of linguistics as abstract theorising, and fosters a set of skills which are transferable to and from other areas of study, and indeed to non-academic life.

Evaluation is confined to the final research paper submitted at the end of the semester. This system lends itself well to distance learning at MA level, as students can pursue their own interests and work at their own rhythm, with teacher back-up available as required. The paper is a 20-page report (not counting appendices etc.) on a research project they define and conduct themselves on an area of their choice. We deliberately kept the guidelines for the research paper as flexible and open as possible so that learners could develop their own questions in regard to their own interests, which go well beyond language learning; the only restriction being that it should involve English and CL. Indeed, as Bernardini (2001:228) points out, “it would be inherently contradictory to prescribe a methodology when the aim of the approach is to give learners the instruments to develop their own methodologies and make their own discoveries.” The mark accordingly is less concerned with the product than with the process—defining the question, choosing the corpus, finding the appropriate tools, using their ingenuity to overcome the problems they will inevitably encounter, and reasoning their way through the whole procedure. A number of past research papers are made available on line as well as a list of all past titles and the most popular sites used. Guidance is also given on the appropriate structure, style and format of a research report in English, linguistics generally requiring rather different styles from research in literature or cultural studies.

Descriptions of a number of CL courses have been published elsewhere. Maia (1997), for example, imposed broad topics (e.g. sport or ecology) on her translation students but required a very specific product for evaluation, namely a bilingual glossary. Others (e.g. Dodd 1997; Jackson 1997; Bernardini 2000; Cheng et al. 2003; Chambers 2005) have used the research paper for evaluation purposes in undergraduate degrees; however, most of these are very explicitly language-oriented courses, often with the language topics for investigation provided by the teachers. Jackson (1997) is the exception in allowing his

students considerable choice in their research, although the projects are confined to a CL analysis of literary style. All of these courses thus impose constraints of one type or another on what students can do, none allowing the freedom in the CTU course. In the light of Johns' (1991a:2) oft-repeated comment that "research is too serous to be left to the researchers," we agree with Leech (1997:10) that individual choice is essential in giving the student "the realistic expectation of breaking new ground as a 'researcher', doing something which is a unique and individual contribution." CL would certainly seem to be a fruitful ground for this, as "the traditional divisions between, on the one hand, teaching and research, and, on the other, literary and linguistic analysis, are to an extent arbitrary" (Burgess & Kohn 1995:63).

2.2. Introducing corpus linguistics

As students are often wary of linguistics, and CL can be terrifyingly technical on first encounter, it is essential to find a way to make the course as unthreatening as possible. Bowker and Pearson, in their "practical guide to using corpora" (2002:108), attempt this by leaving their "introduction to basic corpus processing tools" (frequency lists, concordances, clusters, collocates, etc.) to chapter 7, half-way through the book. Another way is to provide easy, accessible, and familiar examples wherever possible, and leave much of the work to the learner to discover not just the language, but how to go about tackling it through CL. Learners need differing amounts of guidance, and back-up has to be available on demand, but in general students quickly see the point, and discover much of the rest for themselves in a relatively short time (c.f. also Bernardini 2001).

The introductory section to the course thus aims to show how applications of CL can be quite concrete in a familiar environment, which also highlights the variety of applications of corpus methodologies. For example, students are generally familiar with amazon.com, but few have explored the "inside this book" and "search inside" functions, some of which offer insights into CL. Looking at a well-known book such as Dan Brown's *Da Vinci Code*, visitors to the site can read various extracts including the blurb and the first pages, and are introduced to "statistically improbable phrases" with a brief discussion of what that entails (*cilice belt, seeded womb, pope interred, corporal mortification, rosewood box*, etc.); these are hypertext-linked to the actual occurrences in the text. Other features provide an introduction to important concepts including frequency, search strings, concordances, collocations, co-text, and so on. The statistics provided here can be intriguing, and even the more fanciful ones serve to keep the introduction accessible: the book rates 9.1 on the Fog Index, indicating that 80% of books are harder to read; it has 11.1 words per sentence (82% of books have more); the reader gets 14,105 words to the dollar, and 11,425 words per ounce, meaning that each word costs €0.000056 at today's prices and weighs 0.0025 grams. The first 500-odd words of the novel can easily be copied into VocabProfiler[1] to show other features that texts can reveal, highlighting the importance of lexical frequency within the text. The statistics can also be compared to the language as a whole based on the most frequent words in bands of 1000. Other concepts introduced at this stage include types, tokens, lemmatisation and etymology.

[1] URLs to the various sites mentioned can be found in Appendix A.

Standard search engines provide another familiar entry point. Although students know and use google.com, for example, few have ever got around to exploring the advanced search features, which can make all the difference to the relevance of hits. The implications for corpus searches are made explicit, including the importance of domain-specific searches, reliability of websites, etc. The limitations of such tools are also explored, showing how they are information-oriented rather than language-oriented. This allows a lead-in to the web as corpus (see e.g. Kehoe & Renouf 2002), and especially the possibilities of the more linguistic WebCorp. There is an introduction to large public corpora—the British National Corpus (BNC) and the Bank of English (BoE) among others—as well as the common tools associated with them: concordances, collocates and frequency lists, as well as tagging, query language and tips for conducting successful searches. Other "non-linguistic" corpora are mentioned (e.g. Shakespeare, or Old English), as well as the possibilities for students to create and analyse their own corpora. For this, a number of free on-line tools are presented briefly (e.g. Web Frequency Indexer and the demo version of WordSmith Tools), and links are given to numerous other resources on line, including other tutorial pages.

Descriptions of such tools are kept to a minimum as the idea is that students should explore and find out for themselves; in a sense, learners become their own teacher-researchers. This would seem to be particularly appropriate in CL, which, "almost more than any other branch of linguistics, … requires that students have 'hands on' experience of the subject… Only through using corpora can one gain a first-hand sense of their potential" (Leech 1997:7). This works precisely because CL, "by its very nature, has a highly practical and applied orientation" (Hatzidaki 1996:263). We would agree in this with Kirk (2002:156), who also found with experience that his "focus has shifted from *teaching* corpus linguistics to that of *learning* to do corpus linguistics." In a way such a discovery approach is really little more than an extension of the data-driven learning philosophy where learners explore language rather than being told about it; there seems little reason why a similar approach cannot be adopted with the tools and methodologies of CL as well as for language learning.

3. Student research papers

This section provides an overview of the most popular tools, methods and corpora used by our students, followed by an analysis of the research topics chosen. We look at 30 research papers submitted between 2002 and 2005—not an entirely random selection, as it includes only those which were submitted electronically and got a mark of at least 60%. It also excludes papers which adopted a different approach from that described above: for example, some were more theoretical, while others discussed possible applications of CL in language teaching and learning, and so on. A list of the 30 papers can be found in Appendix B along with a brief summary of each.

3.1. Corpora and corpus tools

Surprisingly perhaps, only six papers used the internet itself as a corpus with various tools: WebCorp, Edict VLC, and To Google or Not to Google; Google and Yahoo were also used for searching the web as corpus in various languages. This is perhaps disappointing given that

the internet is a "fabulous linguists' playground" (Kilgarriff & Grefenstette 2003:333) and that there is enormous potential for using the web as corpus (see e.g. Kehoe & Renouf 2002; Bergh 2005). However, almost all of the corpora which were used were collected from the internet for obvious reasons of variety, availability and cost (Aston 2002). As table 1 shows, there were two basic types: 18 papers used one or more large, ready-made reference corpora, while 27 involved at least one home-built corpus of their own.

		home-built corpora					
		0	1	2	3	4	total
reference corpora	0		3	6	1	2	12
	1	1	2	4		1	8
	2	2	4	3		1	10
	total	3	9	13	1	4	30

Table 1. Types of corpora used.

The majority of students made use of some kind of large reference corpus, most frequently the BNC (14 papers) or the BoE (12 papers).[2] Two used the Brown corpus, three used different specialist corpora (Old English, Switchboard, W3Corpora). In eleven of these cases, two or three of these large ready-made corpora were used for comparison purposes, or simply to collect more data. Only three students relied entirely on such reference corpora, the others compiling ones of their own. Among these, the majority analysed one corpus (9 papers) or two (13 papers); the remaining five used three or four. In eighteen cases both types of corpus were used, either with the reference corpus as a starting point before moving on to the home-built corpus (12 papers), or the other way round with the reference corpus serving to check the results of the purpose-made corpus (6 papers).

The most popular single source type was inevitably online newspapers for reasons clearly outlined in Rademann (1998). Not only are the texts easily available, but (depending on the newspaper) they have minimum standards of language and content, are up-to-date, cover a wide range of areas, and are arguably more representative than other registers in that they are written by multiple authors and read by millions (c.f. Burnard 2002:58). This does not mean the students are necessarily interested in the news as such, but as long as they are aware of the limitations of newspapers and do not overgeneralise their findings, newspapers would seem to be a fine source in general (Partington 2001a:63). Twelve students based part or all of their corpora on British or American press found on line, although others were considerably more specialised.

It seems then that learners can see the attraction of both large existing corpora and small purpose-built ones (see Aston 1997a for a comparison). In the former case our students used only the dedicated software provided with the corpus (e.g. Sara with the BNC); in creating their own corpora, however, they made use of a variety of different tools available free on the internet.[3] One of the most popular was the Web Frequency Indexer (WFI), used by 15 students. 16 papers made use of WordSmith Tools (WST), although there were occasional

[2] See e.g. Aston (1996) and Bernardini (2000) for exploiting large corpora in language learning.
[3] A list of the most popular websites and software can be found in Appendix A.

alternatives. Tagging was rarely used, perhaps because of the limited availability of free on-line taggers such as CLAWS or the Brill tagger, which are often restricted to very small samples. Word frequency lists from the BNC proved popular, lemmatised or not, as did VocabProfiler and the Compleat Lexical Tutor in more recent papers. Other software included numerous online dictionaries, pgAdmin, Filemaker, the ubiquitous Microsoft Word and Excel, as well as occasional locally-available tools.

With the exception of general-purpose software (notably MS Word and Excel), what all the above tools have in common is that they are free and easily available on the internet. Although the free versions of some are limited, returning a reduced number of hits or allowing only short texts, they were frequently sufficient for most purposes. A number of students showed surprising ingenuity in finding ways around the limitations, especially with carefully formulated search commands, or cutting the text into small sentence or paragraph-length fragments containing the target language; a handful of others were sufficiently hooked to buy full licences for WordSmith Tools. Another point which bears mentioning is that not all of these tools were mentioned in the course: the students located them and found how to use them on their own.

3.2. Research topics

Although CL has permitted many quite revolutionary techniques, in some ways it is "nothing but a methodology" (McEnery & Wilson 2001:1)—at least in the sense that its methods are not limited to linguistics alone. This can be seen in the description of a new journal entitled *Corpora: Corpus-based Language Learning, Language Processing and Linguistics*. One of its "three main features" is:

> Interdisciplinarity: the journal will actively seek to promote a cross fertilization of ideas and techniques across a range of areas… and disciplines… in the belief that these areas have something to offer to each other through their common focus on corpus data.[4]

Hatzidaki (1996:262) argues convincingly for the introduction of CL as a taught course in undergraduate language degrees precisely because of its "strong interdisciplinary character… CL is applicable to a whole gamut of linguistic, and, arguably, non-linguistic disciplines." Partington (2001b) and Adolphs (2006) are among those who point out that the vast majority of social sciences use corpora in one way or another, from psychology to philosophy and beyond. Kettemann and Marko (2004) ask if the L in TaLC can stand for literature, and this can easily be extended to cultural studies, history, sociology, translation, etc., as well as the pursuit of personal interests. All of this suggests that CL may have applications in far wider areas than just "language" (Boulton 2006), a factor which it would seem churlish not to exploit, and even key in ensuring continued use of the methodologies beyond the course itself. Indeed, if corpora are to be effective, it seems extremely important to allow students to use them in domains that are of interest to them (Maia 1997), especially if texts are "only useful insofar as the learner is able to authenticate them, i.e. to create a relationship with the texts" (Braun 2005:53; also Widdowson 1998). And what better way for

[4] http://www.eup.ed.ac.uk/journals/content.aspx?pageId=1&journalId=12801 accessed May 2007.

students to authenticate a corpus than for them to choose the texts based on their own interests?

Given the students' overriding interest in literature and cultural studies, their research projects frequently have a bias towards these and related areas.[5] The tools and procedures needed for the course are quite different from what the students are used to in these other fields, however; even where they have used a "corpus" in the past, the methods involved are likely to be substantially different. The idea is not to oppose the different approaches, but to treat them as complementary (Kettemann & Marko 2004:187). However, the students quickly adapted to the quantitative measures needed, generally supplemented with a sensible qualitative eye. Of the 30 papers looked at here, arguably only 6 started with a purely linguistic question, typically a language point that is difficult to grasp or to translate into the L1. These included phrasal verbs, pairs or groups of close synonyms (e.g. *barely* / *hardly* / *scarcely*; *speak* / *talk*), and spelling differences (*civilise* / *civilize*; *enquire* / *inquire*). A number of others were interested in translations of key items, although only two had a parallel corpus as such. Many others were inspired by their own learning experiences, although again, few made this a major focus of the paper. About two thirds of the students started with a specific question (a bottom-up approach); the other third started with the corpus (a top-down approach).

A brief look at the titles of the research papers submitted shows the wide range of other interests pursued (table 2). Top of the list were cultural issues revolving around politics, current affairs, and social issues such as feminism; other cultural topics related to their studies included history, and especially British and American cultural comparisons, confirming or rejecting stereotypes. Five papers were mainly concerned with literature, although many others had this as a minor element, especially in the form of stylistic comparison. Translation and language learning / teaching formed the core of four more, again relating to other courses offered in the degree. Outside interests can be found in music and popular culture, professional life, ICT, religion, and personal contacts with a particular part of the English-speaking world. Many papers managed to combine several of these interests quite explicitly. Such a broad range of topic areas supports Seidlhofer's (2000:208) claim that CL can provide "an overarching theme" relevant to many if not all aspects of studies.

topic area	number of papers
1. politics / current affairs	9
2. music / popular culture	8
3. history	6
4. literature	5
5. GB / US differences	5
6. work	4
7. translation / teaching-learning	4
8. ICT	3
9. personal contacts	3

[5] A list of the titles plus a brief description of each paper can be found in Appendix B

10. religion	3

Table 2: Popular topic areas.

4. What did they learn?

The course outlined here is, first and foremost, a course in *corpus* linguistics for language students. As a result, students should end up with more than a passing knowledge of key concepts involved in CL, and be able to use them in the future. These include, in no particular order, an awareness of concordances, KWICs, frequency lists, keywords, distribution plots, collocates lists and tables, mutual information and *t*-scores, stoplists, picklists, batches, left and right sorting, and so on. Using corpora entails an understanding of the importance of balance and statistical evidence in relation to topic, text type, size, date, language variety and register, number of contributors, text length, representativity, and extrapolation from samples to populations. As with Chambers' (2005:114) course, few if any of our students have an "ambition to become experts in CL, any more than they desire to become lexicographers when using the dictionary;" but it is another potentially useful set of tools for them even without specialist in-depth knowledge (Hatzidaki 1996:256).

In producing their research papers, students knew they would be assessed on the "processes rather than products, on methods rather than outcomes, on resourcefulness, awareness and reflectiveness rather than learnedness" (Bernardini 1997:5). In other words, inconclusive or unexpected results do not automatically entail a low mark. As Maia (1997:5) found in her course evaluation, "although the *product* of each of the assignments varied in quality, the knowledge gained from the *process* was considerable." There is inevitably a degree of serendipity involved (Bernardini 2000), but the students are thus sensitised to the cyclical nature of research, involving continuous feedback and rethinking rather than a strictly linear process. The discovery approach involves defining the question, choosing the corpus, locating appropriate tools and experimenting with their use, using their *nous* to deal with problems they will inevitably encounter, and reasoning their way through the whole process in a clear and rational way. All of this leaves the students better equipped to use CL techniques in the future.

It is also however a course in corpus *linguistics*, and students certainly learned a lot of linguistics in the process—perhaps indeed more effectively as the practical approach can make many quite theoretical linguistic concepts that much more "real". Amongst other things we could mention:

- The concept of "word" (lemma, lexical item, etc.), function vs. lexical words, and the separation of lexis and syntax, and associated ideas (affixation, inflection / derivation, compounds, multi-word units, lexical phrases, chunks, collocations, discourse, cohesion / coherence), including the inherently "fuzzy" nature of language, which allows learners to be more confident with approximations.
- The importance of frequency, lexical density, type/token ratios etc.; semantics, polysemy, homonymy, part of speech, etymology, borrowings, cognates, polysemy, translation, etc.

- The importance of text type (genre / register, style, speech vs. writing, etc.) and variation, especially between different varieties of English (geographic, socio-economic, diachronic, etc.); the concept of "error" and standard vs. non-standard varieties.

Clearly, a course such as this is not primarily concerned with language learning *per se*, but language was by default the mainstay of all projects. Where the specific research question had a language focus, one would hope that the point concerned would be better understood. More importantly as regards language learning, the processes involved are likely to increase noticing and sensitivity to various aspects of language, leading to more efficient learning in the future (Allan 2006). To use a hackneyed analogy, it would seem better in the long term to provide the hungry with a fishing rod rather than with a fish. As Kettemann and Marko (2004:171) point out, "we do not want to make students aware of something, we want them to be able to become aware on their own." This kind of language awareness is what Mair (2002) calls the "hidden surplus value of corpora in continental English departments."

The first port of call for many students was a range of standard reference works, either on paper or on line. These include the dictionary, thesaurus, textbook, usage manual, grammar book, style guide, and so on. Notoriously, such materials are found to be inaccurate descriptions of what actually happens in many languages. Confronting such reference works with corpus evidence, learners might be expected to use them more efficiently or more sensitively in future; even when they are "accurate", they can rarely be complete, and a corpus approach can be seen to add to this (c.f. Dodd 1997). A degree of scepticism and critical thinking is also needed in regard to teacher explanations and folk wisdom—not to mention the learner's own ideas—including an ability to think of alternative explanations for apparently simple phenomena (Kettemann & Marko 2004).

The entire process as outlined here involves a great deal of "authenticity" of text, purpose and activity (Stevens 1995), whether as the main focus or more indirectly, almost as a form of task-based learning. This includes following the course, understanding and using the tools in English, and encountering and manipulating tremendous quantities of English data in their chosen corpora. The lack of an explicit "language learning" aim is therefore not only not a problem, but can be seen as a positive advantage (Kennedy & Miceli 2001); Aston (1997b:210) even claims that learners *should* "approach the corpus with non-linguistic goals." If, as Johns (1991b:30) has claimed, "effective language learning is itself a form of linguistic research," one wonders if the converse might not also be true in many cases, i.e. that linguistic research can be an effective form of language learning (see also Boulton & Wilhelm 2006). Following Maia's 1997 article, "making corpora: a learning process", Mauranen (2004:99) claims that "corpus skills constitute a learning task in themselves… Once acquired, they facilitate learning greatly and need not be constantly refreshed."

In such a course, students inevitably discover an awful lot about using ICT, in particular using the internet more efficiently: finding and evaluating sources of texts and tools, advanced functions of search engines and query language, recovering broken links, downloading and copying. Students often found the tools themselves, which means they should be better able to do so again in the future as resources come and go; they also had to find out how to use them by experimentation, as complete description is not possible—or, as we have argued,

desirable—in the course itself. We could also mention OCR scanning, transcription, screen shots and image editing, among other things.

The students learned a lot about more advanced functions of software they thought they were already familiar with, in particular MS Word and Excel. This includes the manipulation of layout and formatting, tables and graphs, images and automatic tables of contents, page numbering, spell-checking in different varieties, keyboard shortcuts, find and replace functions, etc. As the final research paper was in English and in a style similar to professional academic articles, there was the obvious opportunity of providing a style guide similar to those frequently found in professional linguistics journals (c.f. Jackson 1997). If these students do continue to do research, they should be better placed to produce an appropriate paper, with considerations of format, structure, style, the importance of evidence, the respective merits and roles of quantitative and qualitative approaches, interpretation and generalisation, background reading and referencing, and so on. Some students took the initiative to contact researchers, publishers and copyright holders in their chosen area; some explored local university and other resources; some acquired experience in doing fieldwork. Most should also be more sensitive to the dangers of preconceived ideas in the formulation of their research.

Finally, the whole problem-solving, hands-on approach using limited tools encourages all sorts of useful behaviour which may serve in future (see also Kettemann & Marko 2004:185ff). The initial idea has to be formulated and framed in such a way that it becomes a valid topic for research; frequently the original research questions cannot be addressed with a simple command, requiring considerable lateral thinking to overcome the problems—especially given the limits placed on much of the free demonstration software. For example, creative thinking was required to formulate a number of separate but compatible search requests to obtain more than 50 hits with BoE or BNC, how to measure syntactic complexity, alignment of different texts, and so on. Students also needed to be aware of issues of repetition (e.g. with song lyrics), features of spoken vs. written language, non-standard spellings, the presence of other languages or non-Roman scripts. Even if not confronted with these specific problems again, the fact that they have had to solve such problems in the past would seem to put students in better stead to solve new problems as and when they arise in the rest of their studies and beyond.

5. Conclusion: If…

Corpus linguistics is a hugely promising field with a wide variety of applications, but in language courses as a whole, the uptake so far has been extremely limited. If CL is to play any role here beyond materials development and syllabus design, we have to make sure it is not perceived just as "trendy" research that is irrelevant to the working teacher and the "average" learner (c.f. Braun 2005:47, 48). For language learning as such, we need to show that CL can almost immediately produce results which are "illuminating and very helpful" (Sinclair 2004:288).

Clearly, such an approach increases learner autonomy, but more than that "it is now possible for the learner to be autonomous in ways it was previously impossible" (Uzar

2003:153). This applies whether the explicit focus is on language or something else, and the crossover for learners like ours is essential if they are to keep using CL in any form. Several are applying it in their other studies, including the long Master's dissertation and subsequent doctoral thesis; others have provided informal feedback that they also use it outside the academic context. The crossover is by no means guaranteed: Chambers (2005) found her students in Ireland had great difficulty in making the connection. This might be in part because her students were focusing exclusively on language questions, while a wider range of non-linguistic questions and greater room for learner choice to follow their own interests may maximise the versatility of corpus methodologies.

> If corpus consultation… is to become a common activity for learners across the broad spectrum of language studies (general language learning, literary studies, languages for specific purposes, translation, etc.), it would seem necessary for developments to take place in a broader context… It is perhaps outside the classroom that the next important step in research in this area will take place. (Chambers 2007:13)

Like Römer (2006:105), our aim has been to "equip our students with a tool box, containing skills that are transferable from problem to problem across sub-disciplines." Multifunctional tools tend to be kept to hand and used repeatedly, while specialist tools tend to be confined to a dusty drawer and brought out on rare occasions, if they are not forgotten altogether. Although the tools generally used in CL may have been designed with specific purposes in mind, it is a tribute to their power, simplicity and versatility that they can find a range of other uses. Screwdrivers too can be used for opening tins of paint.

Appendix A. Popular websites

The URLs here were provided by the students themselves; all were operational at the time of the TaLC7 conference in 2006.

Large reference corpora:

- BNC (British National Corpus): http://www.sara.natcorp.ox.ac.uk
- BoE (Bank of English): http://www.titania.cobuild.Collins.co.uk
- Brown corpus: http://www.ldc.upenn.edu/cgi-bin/ldc/textcorpus?doc=yes&corpus=BROWN#A
- Old English CorpusSearch: http://www-users.york.ac.uk/ang22/YCOE/doc/corpussearch/CSRefToc.htm
- SwitchBoard: http://www.ldc.upenn.edu/readme_files/switchbrd.readme.html
- W3 Corpora (Project Gutenberg): http://clwww.essex.ac.uk/w3c/

The web as corpus (in addition Google, Yahoo and AltaVista in various languages):

- WebCorp: http://ww.webcorp.org.uk
- To Google or not to Google: http://cli.la.asu.edu/togoogleornot.htm
- Edict Virtual Language Centre: http://www.edict.com.hk

Concordancers:

- WordSmith Tools: http://www.lexically.net/downloads/download.htm

- MonoConc: http://www.athel.com/mono.html
- Concord: http://132.208.224.131/Concord.htm
- Larkin Concordance: http://www.concordancesoftware.co.uk

Taggers and parsers:

- CLAWS demo tagger: http://www.comp.lancs.ac.uk/ucrel/claws/trial.html
- Brill tagger: http://www.ccl.umist.ac.uk/resources/ccl/BrillExec.zip
- Memory-Based Shallow Parser: http://ilk.uvt.nl/cgi-bin/tstchunk/demo.pl
- Tilburg parse tagger: http://pi0657.kub.nl/cgi-bin/tagdemo/tcl.cgi

Other tools:

- Web Frequency Indexer: http://www.georgetown.edu/cball/webtools/web_freqs.html
- BNC frequency lists: http://www.comp.lancs.ac.uk/ucrel/bncfreq
- VocabProfiler: http://www.lextutor.ca/vp/eng
- BNC lemmatised frequency list: http://www.itri.brighton.ac.uk/adam.Kilgarriff/bnceadme.html#lemmatised
- Compleat Lexical Tutor: http://132.208.224.131
- The Internet Archive WayBack Machine: http://www.archive.org/index.php

Appearing in papers not included in this selection:

- VLC Web Concordancer: http://vlc.polyu.edu.hk/concordance/default.htm
- AntConc: http://www.antlab.sci.waseda.ac.jp
- Variation in English Words and Phrases (VIEW): http://view.byu.edu/
- Michigan Corpus of Academic Spoken English (MICASE): http://micase.umdl.umich.edu/m/micase/
- Corpus of British Academic Spoken / Written English (BASE / BAWE): http://www2.warwick.ac.uk/fac/soc/celte/
- EAGLES: http://www.ilc.cnr.it/EAGLES96/home.html
- Corpora4Learning: http://www.corpora4learning.net/
- ICT4LT: http://www.ict4lt.org/en/index.htm
- TropesZoop: http://www.acetic.fr/demo.htm
- KWiCFinder: http://miniappolis.com/KWiCFinder

Appendix B. Research topics

Below are the full titles and a brief description of each of the 30 projects discussed in this paper, most recent first.

1. From screen-plays to fan fiction: A corpus-based study.
Interest in reading and writing science-fiction leads to a comparison of home-built corpora from the internet of Star Trek scripts and amateur "fan fiction" (45K words each). WST is used for various functions to check "techno-babble" overall, then focusing on the 10 most frequent items in each.

2. Basic English versus Standard English.
Experience teaching English in Germany provokes interest in the core lexicon of English (in terms of frequency, coverage, Germanic origin). Corpora of literary texts in SE and Ogden's BE (25K words each) are compiled from

the internet using FileMaker; this allows analysis of the definition of "word" claimed for the 850-item vocabulary of BE. Comparative statistical analysis is conducted using Larkin and VocabProfiler.

3. Analysing Sting and the Police lyrics.
Two corpora from different periods are compiled from the internet (10K and 25K words) to analyse maturity of Sting's lyric writing. Focus is on the most frequent nouns using MonoConc (also WST), plus a qualitative interpretation for semantic fields, KWICs and collocates etc.; the results are checked against BNC written / spoken frequency lists.

4. The religious vocabulary in U2's lyrics.
A musical interest in the group instigates background research into their religious inclinations. Lyrics are compiled from the internet in two periods (8K and 12K words), and analysed for religious/spiritual lexical words and collocates using WFI and WST.

5. *Fancy*, *like* and *love*: Expressing liking and feeling.
Words don't always have one-to-one translations, and emotions are a particularly difficult area. After looking in dictionaries and comparing translations, these verbs were checked in BoE and BNC, as well as W3Corpora. To Google or Not to Google was used for the target items plus *someone* or *something*.

6. Building a lexical map of Nick Cave's lyrics: From The Boys Next Door to The Bad Seeds.
The lyrics were found on the web and compiled into two unbalanced corpora from different periods (5K and 30K words), and compared against BNC spoken corpus (keyword frequencies, collocates etc.). WFI and WST were used but a qualitative comparison was necessary too for syntax and meaning, collocation and idiom, concentrating on emotive lexical words.

7. Should we *enquire* or *inquire* about something?
Different varieties of English lead to confusion, and dictionaries and etymology are of little help. Usage is compared in relevant subcorpora of BNC and BoE, then in home-built corpora from GB and US newspapers and literature on line (total 25K words). Analysis of key words in each corpus is conducted with WFI and WST: frequencies, collocates and KWICs are used to compare meanings.

8. The different meanings and uses of the word *concern* and its derivative forms in Tony Blair's and George Bush's speeches.
This paper unites interest in politics and the intuition that *concern* is a "false friend": dictionaries etc. are found unconvincing regarding meaning and use. Two corpora of semi-scripted interviews are compiled from GB and US leaders (120K words each) and tagged with CLAWS. MonoConc and WFI were used to compare different senses and uses in the corpus against frequencies in Brown and BNC, the web as corpus using Virtual Language Centre, KWICs and frequencies etc.

9. The simplified novel, a case study: The Fall of the House of Usher.
The interest here is in literature and reading in a foreign language, as well as vocabulary acquisition. Corpora are compiled from an original text (7K words) and an OCR-scanned simplified reader (2K words). VocabProfiler is used to compare vocabulary (Anglo-Saxon vs. Greek/Roman/French origins), as well as CLAWS and WST for syntax, text structure, omissions and additions; the tagging also helps to check against the simplification guide. The results are compared against 5 other similar novels from the internet.

10. *Civilise* or *civilize*? A comparative study of the verbs ending in *-ise* and those ending in *-ize* and their respective nominalizations.
A common problem where dictionaries, usage manuals, etymology and style guides are not much help. On the claim that it may be a US/GB distinction, examples are checked against Brown and BNC, including BNC frequency lists. Two balanced corpora (12K words each) are then compiled from online newspapers from GB (Guardian, Independent, Telegraph, Mirror) and the US (USA Today, Washington Post, New York Times, LA Times) over two weeks. Other software includes SwitchBoard, VocabProfiler and WST.

11. Stylistic features of verbatim reports of speeches in European Parliament debates: The passage from the tapes to the final transcripts.

Working at the European Parliament with access to in-house and published documents generates an interest in politics and speech/writing differences. The corpora are from unscripted EP speeches by native speakers, comparing transcribed and subsequently edited texts (6K words each). The focus is on spoken features, redundancy, syntactic complexity, connectors, etc. using WFI and WST.

12. Near synonyms for *woman* in Old English: A corpus analysis.
Interests lie in literature, language evolution, the "real" meaning of words, and gender studies. Synonyms from OE dictionaries are checked against online parsed OE corpora (1.5M words of prose, 75K words of verse); relevant segments are analysed with WFI.

13. A corpus-based approach to the speeches made by Gerry Adams since the Good Friday Agreement.
Contacts in Ireland and subsequent interests in politics and history inspire a corpus analysis of Adams' speeches before (45K words) and after (62K words) the GFA, as well as Paisley's speeches (16K words), all from the internet. The corpora are tagged with CLAWS, frequencies checked with WFI, and compared to each other and to BNC subcorpora and frequency lists; the focus is on aspects of scripted speeches with written/spoken styles. All corpora are checked for religious words and loaded language (WFI, WST) and against WebCorp on British newspapers.

14. Some French words and expressions in English.
Motivated by the crackdown on Anglicisms in France, this looks at the causes of borrowing and the shared history of French and English: frequency, prestige, accommodation, semantic shifts, etc. Definitions and etymology of French borrowings and Anglo-Saxon synonyms are compared in monolingual and bilingual dictionaries, then in KWICs and frequencies in different subcorpora of the BoE. There is also a comparison between US and GB English from editorials of newspapers (total 36K words) using WFI.

15. E-words.
From a combination of previous experience in linguistics (including lexicology) and the IT focus of the course, ICT words are noted in general contexts. New coinages are compared against BoE and BNC frequency lists, and subsequently against an international newspaper corpus (35K words) as well as the web as corpus (Yahoo in different languages). Relevant contexts are downloaded and pasted into WFI.

16. *Make* or *do*? That is the question.
A long-standing dissatisfaction with traditional explanations of generalisations or idiosyncrasies is reason to check synonyms and translations for common points in BNC and BoE. A corpus is built from the internet Guardian Friday review and analysed with WST, looking also at translations back into French. Semantic, collocational and structural patterns are compared.

17. *Land*, *country*, *nation* and *state*: Scotland in focus.
In interest in politics combines with time spent in Scotland and an awareness of Scottish autonomy. Dictionaries and etymology are found unconvincing for the target words, so two corpora are compiled from the Guardian and SNP websites (27K words), the key words checked for frequency, collocates and KWICs with relation to Scotland using WFI, and compared against BoE.

18. Latin and French derivatives in the Guardian Unlimited and the Sun Online.
French borrowings are frequently claimed to be more formal in English. After historical research, corpora were built using pgAdmin to compare the prestigious Guardian and the popular Sun newspapers (12K words each); the corpora were further subdivided into news (more formal) and sport (more popular). Typically French affixes were the key focus.

19. A corpus comparison of the verbs *speak* and *talk*.
These verbs posed problems in translation which dictionaries did not answer satisfactorily. BoE was used to provide KWICs of the verbs + preposition, as well as collocates, and translations back to French were compared.

20. A study of two phrases: *To go on strike* and *to be on strike*.

Current affairs prompts an interest in phrasal uses of *strike*, which are compared first with French and German equivalents. The key words are analysed in BoE and then a newspaper corpus of Guardian articles covering the event, and the relevant text is checked for various uses, PoS, idioms, argument structure etc.

21. French words in English cuisine.
An enthusiasm for cooking and English recipes prompts interest in borrowings and history. BNC and BoE KWICs allow analysis of accommodation in morphology and structural use as well as semantic coverage against French. A corpus of recipes is compiled from MSN, downloading borrowings for analysis with WFI.

22. *Freedom* and *liberty* in George W. Bush's speeches.
Reactions to 9/11 highlight one Germanic / French doublet in English with a lack of comparable difference in French. This leads to work on etymology, but dictionary differences remain unclear. A corpus is built from the official White House site (27K words) and fed into WST; KWICs, WFI and structure provide insight to differences in meaning and use.

23. Corpus study of non-standard past irregular verbs in Appalachian and Piedmont English.
Living in the US, this student's family is critical of the non-standard local variety (especially *do*). The corpus is largely from a university audiobank supplemented with 3 local informants: interviews, own transcriptions of relevant past forms and focus on the 7 commonest verbs. There is a comparison of (non-)standard usage and variation of form and meaning by register and speaker using tools from the same university.

24. *Clone*, *cloned*, *cloning*: New terms in modern language.
A retired doctor looking at "new" words since Dolly was cloned in 1997, especially as a buzzword outside medicine. WFI and WST were used to compare 4 registers (20K words each) of fiction, newspaper, science and "ethical" texts; collocates by PoS.

25. Technical and non-technical uses of forestry terms.
Working at the Office Nationale des Forêts and reading professional documents, this student compiled two corpora of journal abstracts in fundamental (42K words) and applied (84K words) forestry science over a 4-year period. Lexical word frequencies are compared using WST, as well as technical and sub-technical language (especially compounds), and KWICs for semantic coverage and reference.

26. Is the verb *to maintain* solely reflective of the maintenance department activity or is it part of the group of synonyms *to repair* and *to fix*?
Working at LuxAir with access to trade documents, this paper presents a home-built corpus from professional documents comprising 6 manuals (230K words). The corpus is parsed with the Tilburg parse tagger, and WST is used to check frequencies, KWICs and collocates against BNC and BoE.

27. Investigating use of the adverbs *barely*, *hardly* and *scarcely*.
Meaning and usage (syntax) are confusing for these words, and dictionaries and manuals unconvincing. To find out more, frequencies and collocates are gathered from BNC and BoE, as well as KWICs for word order; also WST for a minicorpus.

28. The initiatory book: A case study of The Little Prince.
An interest in literature and "initiatory" books as well as translation prompts the comparison of two corpora of the novel in French and English (17K words) from the internet. Ten "interesting" words of $f>5$ were compared against BoE (frequencies and KWICs); a quantitative and qualitative analysis of translation is provided, especially pronominal replacement; WFI is also used.

29. British and American spelling—*meter* or *metre*? That is the question.
Different spellings can be intriguing but confusing. After researching historical, cultural and pronunciation differences (homophones etc.) between the two varieties, some of the commonest differences are picked out. These are checked in BoE and BNC subcorpora for frequency and collocates. Two home-built corpora comprise GB and US papers using WFI.

30. *In* in phrasal verbs.

Phrasal verbs often pose problems for French speakers. After checking definitions, the commonest ones are checked in BNC and BoE KWICs and frequencies, and compared against a home-built corpus of fashion magazines (23K words) using WST; other references include Switchboard, ICE and W3Corpora.

References

Adolphs S. (2006) *Introducing electronic text analysis: A practical guide for language and literary studies*. London: Routledge.

Allan R. (2006) Data-driven learning and vocabulary: Investigating the use of concordances with advanced learners of English. Centre for Language and Communication Studies, Occasional Paper 66. Dublin: Trinity College Dublin.

Aston G. (1996) The British National Corpus as a language learner resource. In Botley S., J. Glass, A. McEnery & A. Wilson (eds) *TALC 1996. UCREL Technical Papers* 9, 178-191.

Aston G. (1997a) Small and large corpora in language learning. In Lewandowska-Tomaszczyk B. & J. Melia (eds) *Practical applications in language corpora*. Lodz: Lodz University Press, 51-62.

Aston G. (1997b) Involving learners in developing learning methods: Exploiting text corpora in self-access. In Benson P. & P. Voller (eds) *Autonomy and independence in language learning.* London: Longman, 204-214.

Aston G. (2002) The learner as corpus designer. In Kettemann B. & G. Marko (eds) *Teaching and learning by doing corpus analysis.* Amsterdam: Rodopi, 9-25.

Bergh G. (2005) Min(d)ing English language data on the web: What can Google tell us? *ICAME Journal* 29, 25–46.

Bernardini S. (1997) A 'trainee' translator's perspective on corpora. In Aston G., L. Gavioli & F. Zanetti (eds) *Corpus use and learning to translate*. http://web.archive.org/web/20031231010215/http://www.sslmit.unibo.it/cultpaps/paps.htm accessed May 2007.

Bernardini S. (2000) Systematising serendipity: Proposals for concordancing large corpora with language learners. In Burnard L. & T. McEnery (eds) *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang, 225-234.

Bernardini S. (2001) 'Spoilt for choice': A learner explores general language corpora. In Aston G. (ed) *Learning with corpora*. Houston: Athelstan, 220-249.

Boulton A. (2006) When linguistics isn't linguistics: Interdisciplinary approaches to corpus linguistics. *Interdisciplinarité dans les études anglophones*. Nancy Université.

Boulton A. & S. Wilhelm (2006) Habeant Corpus—They *should* have the body. Tools learners have the right to use. *Asp* 49, 155-170.

Bowker L. & J. Pearson (2002) *Working with specialized language: A practical guide to using corpora.* London: Routledge.

Braun S. (2005) From pedagogically relevant corpora to authentic language learning contents. *ReCALL* 17(1), 47-64.

Burgess J. & J. Kohn (1995) The use of parallel concordancing for literary and linguistic text analysis. In Gimeno A. (ed) *EUROCALL '95. Technology enhanced language learning: Focus on integration*, 61-72.

Burnard L. (2002) Where did we go wrong? A retrospective look at the British National Corpus. In Kettemann B. & G. Marko (eds) *Teaching and learning by doing corpus analysis.* Amsterdam: Rodopi, 51-70.

Chambers A. (2005) Integrating corpus consultation in language studies. *Language Learning & Technology* 9(2), 111-125.

Chambers A. (2007) Popularising corpus consultation by language learners and teachers. In Hidalgo E., L. Quereda & J. Santana (eds) *Corpora in the foreign language classroom.* Amsterdam: Rodopi, 3-16.

Cheng W., M. Warren & X. Xun-Feng (2003) The language learner as language researcher: Putting corpus linguistics on the timetable. *System* 31(2), 173-186.

Dodd B. (1997) Exploiting a corpus of written German for advanced language learning. In Wichmann A., S. Fligelstone, T. McEnery & G. Knowles (eds) *Teaching and language corpora*. Harlow: Addison Wesley Longman, 131-145.

Hatzidaki O. (1996) Corpus linguistics as an academic subject. In Botley S., J. Glass, A. McEnery & A. Wilson (eds) *TALC 1996. UCREL Technical Papers* 9, 254-265.

Jackson H. (1997) Corpus and concordance: Finding out about style. In Wichmann A., S. Fligelstone, T. McEnery & G. Knowles (eds) *Teaching and language corpora*. Harlow: Addison Wesley Longman, 224-239.

Johns T. (1991a) Should you be persuaded: Two examples of data-driven learning. In Johns T. & P. King (eds) *Classroom concordancing. English Language Research Journal* 4, 1-16.

Johns T. (1991b) From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. Revised version in Johns T. & P. King (eds) *Classroom concordancing. English Language Research Journal* 4, 27-45.

Kehoe A. & A. Renouf (2002) WebCorp: Applying the web to linguistics and linguistics to the web. *WWW2002 Conference*, Hawaii. http://www2002.org/CDROM/poster/67/ accessed May 2007.

Kennedy C. & T. Miceli (2001) An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology* 5(3), 77-90.

Kettemann B. & G. Marko (2004) Can the L in TALC stand for literature? In Aston G., S. Bernardini & D. Stewart (eds) *Corpora and language learners*. Amsterdam: John Benjamins, 169–193.

Kilgarriff A. & G. Grefenstette (2003) Introduction to the special issue on web as corpus. *Computational Linguistics* 29(3), 333-347.

Kirk J. (2002) Teaching critical skills in corpus linguistics using the BNC. In Kettemann B. & G. Marko (eds) *Teaching and learning by doing corpus analysis.* Amsterdam: Rodopi, 155-164.

Leech G. (1997) Teaching and language corpora: A convergence. In Wichmann A., S. Fligelstone, T. McEnery & G. Knowles (eds) *Teaching and language corpora*. Harlow: Addison Wesley Longman, 1-23.

Maia B. (1997) Making corpora: A learning process. In Aston G., L. Gavioli & F. Zanetti (eds) *Corpus use and learning to translate.* http://web.archive.org/web/20031231010215/http://www.sslmit.unibo.it/cultpaps/paps.htm accessed May 2007.

Mair C. (2002) Empowering non-native speakers: The hidden surplus value of corpora in continental English departments. In Kettemann B. & G. Marko (eds) *Teaching and learning by doing corpus analysis*. Amsterdam: Rodopi, 119-130.

Mauranen A. (2004) Spoken corpus for an ordinary learner. In Sinclair J. (ed). *How to use corpora in language teaching*. Amsterdam: John Benjamins, 89-105.

McEnery T. & A. Wilson (2001) *Corpus linguistics,* 2nd edition. Edinburgh: Edinburgh University Press.

Partington A. (2001a) Corpus-based description in teaching and learning. In Aston G. (ed) *Learning with corpora*. Houston: Athelstan, 63-84.

Partington A. (2001b) Corpora and discourse strategies in action: From 'footing' to 'fooling'. In Lewandowska-Tomaszczyk B. (ed) *PALC 2001: Practical applications in language corpora.* Frankfurt: Peter Lang, 263-279.

Rademann T. (1998) Using online electronic newspapers in modern English-language press corpora: Benefits and pitfalls. *ICAME Journal* 22, 49-71.

Römer U. (2006) Where the computer meets language, literature, and pedagogy: Corpus analysis in English studies. In Gerbig A. & A. Müller-Wood (eds) *How globalization affects the teaching of English: Studying culture through texts*. Lampeter: E. Mellen Press, 81-109.

Scott M. & C. Tribble (2006) *Textual patterns: Key words and corpus analysis in language education.* Amsterdam: John Benjamins.

Seidlhofer B. (2000) Operationalizing intertextuality: Using learner corpora for learning. In Burnard L. & T. McEnery (eds) *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang, 207-223.

Sinclair J. (2004) New evidence, new priorities, new attitudes. In Sinclair J. (ed) *How to use corpora in language teaching*. Amsterdam: John Benjamins, 271-299.

Stevens V. (1995) Concordancing with language learners: Why? When? What? *CAELL Journal* 6(2), 2-10.

Uzar R. (2003) A toolbox for translation quality assessment. In Lewandowska-Tomaszczyk B. (ed) *Practical applications in language and computers: PALC 2003*. Frankfurt: Peter Lang, 153-162.

Widdowson H. (1998) Context, community, and authentic language. *TESOL Quarterly* 32(4), 705-716.