CORPORA AND THE TEACHING OF LANGUAGES AND LINGUISTICS[1]

JOSÉ RAMÓN VARELA PÉREZ
Universidad de Santiago de Compostela

INTRODUCTION

A corpus is a body or collection of linguistic data used in language research and created to form a representative sample of the kind of language under study. But today the term *corpus* is almost synonymous with the term machine-readable corpus or computer corpus. The ability of computers to find, sort, analyse and quantify linguistic features in corpora has had important consequences for a number of fields in the language sciences such as grammar, lexicography, diachronic studies, language acquisition, natural language processing, language teaching, etc. (McEnery & Wilson 1996: 87-116). Corpus linguistics has thus become a new kind of methodology to be used in descriptive and theoretical linguistics. The relevance of corpora in language study is related to the increasing importance attached to empirical data. Empirical data allow the linguist to make objective findings based on language as it really is rather than based upon the individual's won internalised cognitive perception of the language (McEnery & Wilson 1996: 87). According to Leech (1992: 107-111), the key features on which computer-corpus linguistics focuses are: a) linguistic performance, rather than competence; b) description, rather than linguistic universals; c) quantitative, as well as qualitative models of language; d) a more empiricist, rather than a rationalist view of linguistic inquiry.

Until recently corpus-based methods had nothing to do with teaching and learning. For over the last few years, however, there has been a growing acknowledgement that there must exist a movement from corpus-based research to teaching languages and linguistics. As the number of articles, workshops and conferences dealing with pedagogical activities using corpora has increased year after year, there is now a fruitful exchange of ideas on how the computer corpus can be best exploited to the advantage of teaching (Fligelstone 1993: 97-8; Leech 1997: 5).

The present paper summarises recent research which explores the potential advantages and the effective uses of a corpus-based approach to the teaching of English as a foreign language and the teaching of English linguistics at the university level. The insights gained from corpus-based language descriptions are relevant not only for understanding language structure and use but also for designing data-driven teaching descriptions, materials and classroom activities. First of all, data-driven learning pays special attention to real language by

(1) The research reported in this article was funded by the Galician Ministry of Education (XUGA 20401A97). This grant is hereby gratefully acknowledged.

using authentic examples, which has a bearing not only on what to teach but also in which sequence. Moreover, as computers have become smaller and less costly, learners are now able to explore, investigate, generalise and test hypothesis as to the actual use of the language. Having discovered the rules themselves, students are more likely to remember and use them in real life situations.

CORPORA AND THE TEACHING OF FOREIGN LANGUAGES

Approaches to language teaching are defined among other things by: a) the contents that need to be learned; b) the optimal roles of learners and teachers necessary for language learning; c) the techniques and procedures followed to facilitate the most effective language learning. The following three sections explore how each of these processes is affected by the use of corpora.

a) The contents of learning

The main language teaching methodologies over the last years have been the communicative approach and the task-based approach. For both methodologies the focus must shift from learning vocabulary items and grammatical rules to the use of language to negotiate meaning with other language users and carry out tasks in real-life situations. However, some applied linguists have raised their voices in favour of paying more attention to the linguistic content of teaching (Leech 1997: 1-2; Kennedy 1998: 280-281) and, in general, to systemic accuracy. Grammar and vocabulary, which were neglected in the late 1970s and 1980s, have started to spark off a new interest in the 1990s, and it is here that a computer corpus-based methodology has started to influence mainstream language teaching methodologies.

In this sense, the first more indirect way in which corpus linguistics has been helpful in applied linguistics has manifested itself in the development of better teaching materials such as grammars and dictionaries. In the field of lexicography, since the 1920s pre-electronic corpus-based research (see Kennedy 1992 for a review of these studies) has shown that teachers and textbooks could best help students to acquire a foreign language by paying attention to the most common lexical items in the language. Currently, frequency lists and concordances derived from corpora are establishing themselves as basic tools in the compilation of English dictionaries such as the 1987 *Collins COBUILD English Language Dictionary* and other advanced EFL dictionaries. Such an approach to lexicography has already offered some advantages (Biber, Conrad & Reppen 1994: 174-9; McEnery & Wilson 1996: 91; Leech 1997: 141). Corpora can be searched quickly and exhaustively so that they can be revised more easily, and they can provide statistical of the frequency of words and word-meanings. Dictionaries such as *Cobuild* show authentic examples for citation, and their definitions can be more complete and precise than those from other dictionaries. Furthermore, these dictionaries can rank the different uses of a word in order of importance, show their most probable collocates and incorporate pragmatic information in the definitions.

As regards grammar, data-driven corpus analysis goes as far back as the early decades of the twentieth century, with the major descriptive grammars of English being written by grammarians such as Kruisinga, Jespersen and Poutsma. It is the work done by these grammarians, who were not native speakers of English and thus based their observations on informal corpora rather than on introspection, that foreshadows the corpus-based grammars of the 1980s and 1990s such as *A Comprehensive Grammar of the English Language*, by Randolph Quirk and his collaborators, or Sinclair's *Collins Cobuild English Grammar.*

On the other hand, by providing data on the likelihood of occurrence and frequency of use, corpora have been taken as a guide for better decisions as to which lexical items, senses and grammatical structures should be included in the teaching syllabus and their sequence of presentation. Such findings often contrast with traditional pedagogical priorities. Thus, a number of scholars working in the field of teaching English as a Foreign Language have used corpus data to look critically at existing language teaching materials (Kennedy 1992, Murison-Bowie 1996). Biber, Conrad & Reppen (1994: 171-3) have shown, for instance, that relatively common linguistic constructions are often overlooked in pedagogic grammars. As an example, they have found that whereas relative constructions are widely discussed in five pedagogic grammars, prepositional phrases, which are much more common in the *Lancaster-Oslo Bergen Corpus*, receive the least attention. As these scholars show, this state of affairs is especially sad if we take into account that prepositional phrases embedded in noun phrases have been found to be rather troublesome for L2 students. Another example is provided by Kennedy (1992: 357), who has shown the importance of discourse items such as hedges, responses and softeners in corpora of spoken English. In fact, as he points out, discourse items have not been part of traditional language teaching, with obvious effect on the naturalness of learners' English.

Despite the widespread importance attached to issues of frequency, as Aston (1995: 259) has remarked, while corpus-based descriptions may reveal new facts about the frequency and distribution of particular forms and meanings, other criteria for the selection and grading items are possible: availability, teachability, class needs, etc.

A related issue is that of authenticity. Data-driven language learning has also stressed the relevance of usage and authentic data, the kind of sentences and vocabulary which students will encounter in real life communication situations but not in textbooks (McEnery & Wilson 1996: 107). The assumption that it is preferable for the student to spend some time working through contextualised examples in authentic texts rather than relying totally or even primarily on isolated paradigms in grammars and course books.

Another reason that explains why corpus linguistics might useful in language pedagogy lies in the fact that corpus data have helped to reconceptualise the units of linguistic description. As Murison-Bowie has pointed out

In using corpora in a teaching context, it is difficult to distinguish what is a lexical investigation and what is a syntactic one. One leads to the other, and this can be used to advantage in a teaching/learning context (Murison-Bowie 1996: 185).

Both corpus linguistics and language pedagogy have started to notice that language exploits multi-word items such as collocations and productive lexical phases, which are co-selected and available for retrieval from memory as ready-made memorised building blocks (Sinclair 1991a; Aston 1995: 261). This has consequences for research on second language acquisition as teachers are forced to reconsider what the units of learning might be.

Another contributing factor in the popularisation of corpus-based methods is the widespread recognition of the sociolinguistic parameters of language acquisition and use, resulting in the need to take account of variation in usage (Kennedy 1998: 281). Thus, corpus-based descriptions of language used in different registers can help create specific-purpose syllabuses, with concordances being used to identify and exemplify the realisation of patterns typical from different text types and situational contexts (Sinclair 1991b; Biber, Conrad & Reppen 1994: 179-83; Aston 1997: 55). Because there are important and systematic diffe-

rences among different text varieties at all linguistic levels, corpus-based research in English for Specific Purposes has shown that "any global characterisation of 'General English' should be regarded with caution." (Biber, Conrad & Reppen 1994: 179). Most language teachers are now conscious of the needs of learners for specific purposes, which are not the same of the learner of a language for general purposes. For this reason, some researchers have started to built corpora of different varieties of English and of other languages with the aim of providing many kinds of domain-specific material for language learning, including quantitative accounts of vocabulary and usage which address the specific needs of students.

b) The roles of the learner and the teacher

Recent cognitive theories of second language acquisition have put forward that the aim of the teacher should not be to communicate a certain amount of knowledge to passive learners, but to help students to interpret and organise the information conveyed to them, fitting it into prior knowledge and revising it in the light of what they have already learned (Van Dijk & Kintsch 1982).

These views on the role of teachers and learners in the learning process tie in nicely with the type of activities promoted by data-driven learning. Thus, corpora can encourage learners to actively explore meaning and recreate the language for themselves. By using corpus search tools the learner becomes a linguistic researcher, testing and revising hypotheses, learning to recognise and interpret clues from the context. As a consequence, the student takes charge of his or her own learning and corpora turn out to be excellent resources for autonomous learning (Aston 1997: 61). However, for this to be possible learners must be given guidelines in using corpora (Gavioli 1997: 83). So, in a way, the role of the teacher becomes that of facilitator, as he or she introduces the learner to the process of interpretation and categorisation of the data. Furthermore, the fact that the computer can customise the learning task to the individual's needs and wishes, rather than simply providing a standard set of examples or data, has been shown to be stimulating for motivated students.

c) The techniques and procedures

Certain techniques and procedures of corpus-based research can have applications inside or outside the language classroom. Thus, working from data not only leads to a radical revision of preconceived ideas about what one should be teaching but also about how one might teach. Corpus linguistics research methodology proceeds by selecting a language point and giving concordanced evidence about it. The evidence is then observed and its salient features are identified and classified. Finally the researcher formulates a rule that accounts for the data (Murison-Bowie 1993: 40). This bottom-up approach can be complemented by a top-down hypothesis testing process of inquiry into language.

The concordancer is a key tool for these activities; it enables one to find all occurrences of a given word, part of a word or a combination of words within a corpus. The learner can then observe how each occurrence functions in context, and note the most frequent senses of each occurrence, and the company the tokens normally keep as part of collocations or grammatical patterns. In short, by observing text patterns and discussing them in class, students gain accuracy about the language and they improve their competence in the language (Murison-Bowie 1993; Murison-Bowie 1996: 191-92). Activities can be created in which students to puzzle things out for themselves. For instance, a concordance on "if" can provi-

de a starting point for a revision unit on conditionals. Students can be given simple research tasks involving the mark-up of the contexts in which a collocation appears with a particular meaning (Sinclair 1997: 101).

Aston (1995: 267-69) claims that corpora can be used in communicative language teaching. Tasks of observation and analysis should be embedded in communicative tasks such as information-gap, reasoning-gaps, opinion-gap, and rapport-gap tasks. Thus, learners derive from data information as to how communicative goals can be achieved. This can be done in three ways:

a) The corpus is used as a reference tool for the solution of problems which emerge in the performance of other tasks.

b) The corpus is used as a source of common tasks. Thus, texts retrieved from corpora provide chances for learners to engage in discourse in different ways. For example, newspaper corpora can be used to find out about people, places, attitudes of the culture of the language that is being learned and then discuss the findings in the classroom. Moreover, corpora can also be used in reading-based activities in which similar texts are compared linguistically. Learners can also find groups of texts dealing with topics of interest to them, which can be discussed in the classroom (Fligelstone 1993).

c) The corpus is used by students in order to browse through texts in an autonomous manner.

The teacher can also incorporate data from corpora and adapt it according to his or her requirements in order to create written exercises, which can be retrieved from corpora automatically (Wilson 1997: 116). These exercises can refer to lexis (word use, idioms, collocations...), syntax (prepositions, verb forms...) or discourse (cohesion above the sentence and paragraph level).

CORPORA AND THE TEACHING OF LINGUISTICS

Corpora have been used not only in foreign language pedagogy, but also in the teaching of linguistics. The use of corpora in teaching disciplines such as grammar, historical linguistics, sociolinguistics, discourse analysis, etc., ties in nicely with the recent reconsideration of the issues of what should be taught and how it should be taught in undergraduate and graduate linguistic courses.

The corpus can be a tool for investigating linguistic phenomena and testing competing linguistic theories. The description of particular linguistic features and their functions in discourse by computational techniques enables a variationist and empirical perspective on any field of language study. Again, the boundary between teaching and research is productively blurred. Typically, students undertake assignments in which they select their own topic or replicate the results obtained by the research of professional linguists. Then they are provided with contextualised corpus examples of sufficient variety and scope for the study of that topic. This task gives the student the realistic expectation of breaking new ground as a 'researcher', doing something which is a unique and individual contribution, rather than a reworking and evaluation of the research of others (Leech 1997: 10).

An interesting example of this kind of approach to teaching is provided by Kirk (1994), who reports on courses on varieties of English he has taught at Queen's University of Belfast. He uses projects rather than traditional essays and exams to test their students, who

are required to base their projects on corpus data which they must analyse in the light of a theoretical model –Grices's co-operation principle, Biber's multidimensional approach to linguistic variation, Brown and Levinson's politeness theory, etc.

Kettermann (1997) takes a similar approach in his teaching different theories of child language acquisition. He argues that the use of a specific corpus in his classroom, the *Polytechnic of Wales Corpus of Child Language*, has taught students to determine the importance of qualitative as well as quantitative factors in the acquisition sequence of grammatical morphemes in English, thus using corpus-driven research to decide on the respective explanatory power of universal grammar (UG) vs. cognitive constructivism and self-organisation (CC).

Knowles (1997) provides an example of how to use easily available diachronic corpora of English (the *Helsinki Corpus*, different versions of the King James Bible, etc.) in classroom activities such as selecting a word or group of words and examine a number of examples in context. Among some of the problems tackled by students were the following: changes in the meaning of individual words; patterns of word formation over the centuries; the distribution of morphological variants (*-eth* vs. *-s* verb endings, e.g. *saith* vs. *says*); and the connection between degrees of formality and the origin of words.

A further application of computer-corpus linguistics comes from the availability of multilingual parallel corpora, an important tool in the teaching of translation: a multilingual corpus can provide side-by-side examples of style and idiom in more than one language and generate exercises in which students can compare their own translations with an existing professional translation or original.

CONCLUSION

The present paper has discussed recent research which explores the potential advantages and the effective uses of a corpus-based approach to the teaching of English as a foreign language and the teaching of English linguistics. These advantages can be summarised as follows:

a) The evidence from corpora is then a valuable resource for improving descriptions of language.

b) Data-driven approaches to research have repeatedly shown that greater attention should be paid to the frequency of linguistic items. This statistical information should enable better decisions as to which lexical items, grammatical structures, etc. should be taught and in which sequence these items should be presented in the syllabus.

c) This research has also pointed out the importance of presenting authentic materials in the classroom.

d) Corpus linguistics has helped to reconceptualise the units of linguistic description: a better understanding of fixed phrases, collocations and other lexical phases.

e) A new role is created for both the learner and the teacher. This involves teaching students to search texts effectively and ask the appropriate questions, that is, teaching learners to become researchers themselves. Using corpora also creates opportunities for autonomous learning, which can be tailored to the learner's individual needs (Knowles 1990; Fligelstone1993; Biber, Conrad & Reppen 1994).

f) It has been suggested that certain methods of this research may be applied to the language and linguistics classroom, with concordances listing data from which learners can either infer or test the rules which govern the use of data. Thus, concordancing allows for "data-driven learning" where students take the role of researchers constructing their own lexical and grammatical descriptions instead of relying entirely on textbooks.

REFERENCES

ASTON, G. 1995. "Corpora in language pedagogy: matching theory and practice". In G. Cook & Seidlhofer, B. (eds). 1995. *Principle and Practice in Applied Linguistics: Studies in Honour of H. G. Widdowson*. Oxford: Oxford University Press.

ASTON, G. 1997. "Enriching the learning environment". In Wichmann, A., Fligelstone, S., McEnery, T. & Knowles, G. (eds). *Teaching and Language Corpora*. London: Longman.

BIBER, D., Conrad, S. & R. Reppen. 1994. "Corpus-based approaches to issues in applied linguistics". *Applied Linguistics* 15: 169-189.

BIBER, D., Conrad, S. & R. Reppen. 1996. "Corpus-based investigations of language use". *Annual Review of Applied Linguistics* 16: 115-136.

FLIGELSTONE, S. 1993. "Some reflections on the question of teaching, from a corpus linguistics perspective". *ICAME* 17: 97-109.

GAVIOLI, Laura. 1997. "Exploring texts through the concordancer". In Wichmann, A., Fligelstone, S., McEnery, T & Knowles, G. (eds). *Teaching and Language Corpora*. London: Longman.

KENNEDY, G. 1992. "Preferred ways of putting things with implications for language teaching". In Svartvik, J. (ed). *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter.

KENNEDY, G. 1998. *An Introduction to Corpus Linguistics*. London & New York: Longman.

KETTERMAN, B. 1997. "Using a corpus to evaluate theories of child language acquisition". In Wichmann, A., Fligelstone, S., McEnery, T & Knowles, G. (eds). *Teaching and Language Corpora*. London: Longman.

KIRK, J. M. 1994. "Teaching and language corpora: the Queen's Approach". In Wilson, A. & McEnery, T. (eds). *Corpora in Language Education and Research: A Selection of Papers from Talc 94*. Lancaster: UCREL.

KNOWLES, G. 1990. "The use of spoken and written corpora in the teaching of language and linguistics". *Literary and Linguistic Computing* 5: 45-48.

KNOWLES, G. 1997. "Using corpora for the diachronic study of English". In Wichmann, A., Fligelstone, S., McEnery, T & Knowles, G. (eds). *Teaching and Language Corpora*. London: Longman.

LEECH, G. 1992. "Corpora and theories of linguistic performance". In Svartvik, J. (ed). *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter.

LEECH, G. 1997. "Teaching and language corpora: a convergence". In Wichmann, A., Fligelstone, S., McEnery, T & Knowles, G. (eds). *Teaching and Language Corpora*. London: Longman.

MCENERY, T. & Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburh University Press.

MURISON-BOWIE, S. 1993. *MicroConcord: Manual*. Oxford: Oxford University Press.

MURISON-BOWIE, S. 1996. "Linguistic corpora and language teaching". *Annual Review of Applied Linguistics* 16: 182-199.

SINCLAIR, J. M. 1991a. *Corpus, Concordance and Collocation*. Oxford: Oxford University Press.

SINCLAIR, J. M. 1991b. "Shared knowledge". In Alatis, J. E. (ed). *Linguistics and Language Pedagogy: The State of the Art*. Washington, D.C: Georgetown University Press.

SINCLAIR, J. M. 1997. "Corpus evidence in language description". In Wichmann, A., Fligelstone, S., McEnery, T & Knowles, G. (eds). *Teaching and Language Corpora*. London: Longman.

VAN DIJK, T. A. & Kintsch, W. 1982. *Strategies of Discourse Comprehension*. Orlando: Academic Press.

WILSON, Eve. 1997. "The automatic generation of CALL exercises from general corpora". In Wichmann, A., Fligelstone, S., McEnery, T & Knowles, G. (eds). *Teaching and Language Corpora*. London: Longman.