22
What Corpora Can Offer in Language Teaching and Learning

Tony McEnery and Richard Xiao

Introduction

The corpus-based approach to linguistics and language education has gained prominence over the past four decades, particularly since the mid-1980s. This is because corpus analysis can be illuminating "in virtually all branches of linguistics or language learning" (Leech, 1997, p. 9; cf. also Biber, Conrad & Reppen, 1998, p. 11). One of the strengths of corpus data lies in its empirical nature, which pools together the intuitions of a great number of speakers and makes linguistic analysis more objective (McEnery & Wilson, 2001, p. 103). Unsurprisingly, corpora have been used extensively in nearly all branches of linguistics including, for example, lexicographic and lexical studies, grammatical studies, language variation studies, contrastive and translation studies, diachronic studies, semantics, pragmatics, stylistics, sociolinguistics, discourse analysis, forensic linguistics and language pedagogy. Corpora have passed into general usage in linguistics in spite of the fact that they still occasionally attract hostile criticism (e.g. Widdowson, 1990, 2000).[1]

The early 1990s saw an increasing interest in applying the findings of corpus-based research to language pedagogy. The upsurge of interest is evidenced by the eight well-received biennial international conferences on Teaching and Language Corpora (TaLC) held in Lancaster, Oxford, Graz, Bertinoro, Granada, Paris and Lisbon. This is also apparent when one looks at the published literature. In addition to a large number of journal articles, at least twenty-five authored or edited volumes have recently been produced on the topic of teaching and language corpora: Wichmann, Fligelstone, McEnery and Knowles (1997), Partington (1998), Bernardini (2000), Burnard and McEnery (2000), Kettemann and Marko (2002, 2006), Aston (2001), Ghadessy, Henry and Roseberry (2001), Hunston (2002), Granger, Hung and Petch-Tyson (2002), Connor and Upton (2002), Tan (2002), Sinclair (2003, 2004), Aston, Bernardini and Stewart (2004), Mishan (2005), Nesselhauf (2005), Römer (2005), Braun, Kohn and Mukherjee (2006), Gavioli (2006), Scott and Tribble (2006), Hidalgo, Quereda and Santana (2007), O'Keeffe, McCarthy and Carter (2007), Aijmer (2009) and Campoy, Gea-valor and Belles-Fortuno (2010). These works cover a wide range of issues related to using corpora in language pedagogy, e.g. corpus-based language descriptions, corpus analysis in the classroom and learner corpus research (cf. Keck, 2004).

In the opening chapter of *Teaching and Language Corpora* (Wichmann et al., 1997), Leech (1997) observed that a convergence between teaching and language corpora was apparent. That

convergence has three focuses, as noted by Leech: the indirect use of corpora in teaching (reference publishing, materials development, and language testing), the direct use of corpora in teaching (teaching about, teaching to exploit, and exploiting to teach) and further teaching-oriented corpus development (languages for specific purposes (LSP) corpora, first language (L1) developmental corpora and second language (L2) learner corpora).

In the remainder of this chapter, we will explore the potential uses of corpora in language pedagogy in terms of Leech's three focuses of convergence. The chapter concludes by discussing the debate over the relevance of authenticity and frequency of corpora in language education as well as the future of corpus-based language pedagogy.

Indirect Use of Corpora

The use of corpora in language teaching and learning has been more indirect than direct. This is perhaps because the direct use of corpora in language pedagogy is restricted by a number of factors including, for example, the level and experience of learners, time constraints, curricular requirements, knowledge and skills required of teachers for corpus analysis and pedagogical mediation, and the access to resources, such as computers, and appropriate software tools and corpora, or a combination of these (see the concluding section for further discussion). This section explores how corpora have impacted on language pedagogy indirectly.

Reference Publishing

Corpora can be said to have revolutionized reference publishing (at least for English), be it a dictionary or a reference grammar, in such a way that dictionaries published since the 1990s are typically have used corpus data in one way or another so that "even people who have never heard of a corpus are using the product of corpus-based investigation" (Hunston, 2002, p. 96).

Corpora are useful in several ways for lexicographers. The greatest advantage of using corpora in lexicography lies in their machine-readable nature, which allows dictionary makers to extract all authentic, typical examples of the usage of a lexical item from a large body of text in a few seconds. The second advantage of the corpus-based approach, which is not readily available when using citation slips, is the frequency information and quantification of collocation that a corpus can readily provide (see the section "Syllabus Design and Materials Development" for further discussion of collocation). Some dictionaries, e.g. COBUILD (HarperCollins, 1995) and Longman, 1995, include such frequency information. Frequency data plays an even more important role in the so-called frequency dictionaries, which define core vocabulary to help learners of different modern languages, e.g. Davies (2005) for Spanish, Jones and Tschirner (2005) for German, Davies and de Oliveira Preto-Bay (2007) for Portuguese, Lonsdale and Bras (2009) for French, and Xiao, Rayson and McEnery (2009) for Chinese. Information of this sort is particularly useful for materials writers and language learners alike.

A further benefit of using corpora is related to corpus markup and annotation. Many available corpora (e.g. the British National Corpus, BNC) are encoded with textual (e.g. register, genre and domain) and sociolinguistic (e.g. user gender and age) metadata, which allows lexicographers to give a more accurate description of the usage of a lexical item. Corpus annotations such as part-of-speech tagging and word sense disambiguation also enable a more sensible grouping of words that are polysemous and homographs. Furthermore, a monitor corpus, which is constantly updated, allows lexicographers to track subtle change in the meaning and usage of a lexical item so as to keep their dictionaries up-to-date.

Last but not least, corpus evidence can complement or refute the intuitions of individual lexicographers, which are not always reliable (cf. Sinclair, 1991, p. 112; Atkins & Levin, 1995;

Murison-Bowie, 1996, p. 184) so that dictionary entries are more accurate. The above observations are line with Hunston (2002, p. 96), who summarizes the changes brought about by corpora to dictionaries and other reference books in terms of five "emphases": an emphasis on frequency, an emphasis on collocation and phraseology, an emphasis on variation, an emphasis on lexis in grammar, and an emphasis on authenticity.

It has been noted that non-corpus-based grammars can contain biases while corpora can help to improve grammatical descriptions (McEnery & Xiao, 2005). The *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad & Finegan, 1999) can be considered as a new milestone in reference publishing following Quirk, Greenbaum, Leech and Svartvik's (1985) *A Comprehensive Grammar of the English Language*. Based entirely on the forty-million-word Longman Spoken and Written English Corpus, the book gives "a thorough description of English grammar, which is illustrated throughout with real corpus examples, and which gives equal attention to the ways speakers and writers actually use these linguistic resources" (Biber et al., 1999, p. 45). The new corpus-based grammar is unique in many different ways, for example, by taking account of register variations and exploring the differences between written and spoken grammars.

While lexical information forms, to some extent, an integral part of the grammatical description in Biber et al. (1999), it is the Collins COBUILD series (Sinclair, 1990, 1992; Francis, Hunston & Manning 1997, 1998), that focus on lexis in grammatical descriptions (the so-called "pattern grammar", Hunston & Francis, 2000). In fact, Sinclair and colleagues (1990) flatly reject the distinction between lexis and grammar. While pattern grammars focusing on the connection between pattern and meaning challenge the traditional distinction between lexis and grammar, they are undoubtedly useful in language learning as they provide "a resource for vocabulary building in which the word is treated as part of a phrase rather than in isolation" (Hunston, 2002, p. 106).

For language pedagogy the most important developments in lexicography relate to the learner dictionary. Yet corpus-based learner dictionaries have a quite short history. It was only in 1987 that the *Collins COBUILD English Language Dictionary* (Sinclair, 1987) was published as the first "fully corpus-based" dictionary. Yet the impact of this corpus-based dictionary was such that most other publishers in the English language teaching (ELT) market followed Collins' lead. By 1995, the new editions of major learner dictionaries such as the *Longman Dictionary of Contemporary English* (3rd edition) (Longman, 1995), the *Oxford Advanced Learner's Dictionary* (5th edition, Hornby & Crowther, 1999), and a newcomer, the *Cambridge International Dictionary of English* (Procter, 1995) all claimed to be based on corpus evidence in one way or another.

One of the important features of corpus-based learner dictionaries is their inclusion of quantitative data extracted from a corpus. Another important feature, which is also related to frequency information, is that such dictionaries typically select the vocabulary used from a controlled set when defining the entry for a word. Producing definitions in an L2 that language learners can understand is a problem; language learners may not have a very well developed L2 vocabulary. This makes it necessary and desirable for dictionary makers to limit the vocabulary they use when defining words in a dictionary. Nowadays, most learner dictionary makers prepare a list of defining words, usually ranging from 2,000 to 2,500 words, based on the frequency information extracted from corpora as well as on the lexicographers' experience of defining words.

As noted earlier, an important use of corpus data for lexicography is in the area of example selection so that nowadays most dictionaries of English use corpora as the source of their examples. In the case of learner dictionaries, however, there was a tradition of using examples invented by lexicographers, rather than authentic materials, in dictionary production, because they believed that foreign language learners have difficulty understanding authentic materials and therefore have to be presented with simple, rewritten examples in which the use of a given word is highlighted to show its

syntactic and semantic properties. It was corpus-based learner dictionary work that challenged this received wisdom. The COBUILD (Collins Birmingham University International Language Database) project broke with tradition and used authentic data extracted from corpora to produce illustrative examples for a learner dictionary. The use of authentic examples in learner dictionaries is an area where corpus-based learner dictionaries have innovated.

Syllabus Design and Materials Development

While corpora have been used extensively to provide more accurate descriptions of language use, a number of scholars have also used corpus data directly to look critically at existing teaching English as a foreign language (TEFL) syllabuses and teaching materials. Mindt (1996), for example, finds that the use of grammatical structures in textbooks for teaching English differs considerably from the use of these structures in L1 English. He observes that one common failure of English textbooks is that they teach "a kind of school English which does not seem to exist outside the foreign language classroom" (Mindt, 1996, p. 232). As such, learners often find it difficult to communicate successfully with native speakers. A simple yet important role of corpora in language education is to provide more realistic examples of language usage that reflect the complexities and nuances of natural language.

In addition, however, corpora may provide data, especially frequency data, which may further alter what is taught. For example, on the basis of a comparison of the frequencies of modal verbs, future time expressions and conditional clauses in native English corpora and their grading in textbooks used widely in Germany, Mindt (1996) concludes that one problem with non-corpus-based syllabuses is that the order in which those items are taught in syllabuses "very often does not correspond to what one might reasonably expect from corpus data of spoken and written English", arguing that teaching syllabuses should be based on empirical evidence rather than tradition and intuition, with frequency of usage as a guide to priority for teaching (Mindt, 1996, pp. 245–246). While frequency is certainly not the only determinant of what to teach and in what order (see the concluding section for further discussion), it can indeed help to make learning more effective. For example, McCarthy, McCarten and Sandiford's (2005–2006) innovative *Touchstone* book series, which is based on the Cambridge International Corpus, aims to present the vocabulary, grammar and functions students encounter most often in real life.

Hunston (2002, p. 189) echoes Mindt, suggesting that "the experience of using corpora should lead to rather different views of syllabus design". The type of syllabus she discusses extensively is a "lexical syllabus", originally proposed by Sinclair and Renouf (1988) and outlined fully by Willis (1990) and embodied in Willis, Willis and Davids' (1988–1989) three-part *Collins COBUILD English Course*. According to Sinclair and Renouf (1988, p. 148), a lexical syllabus would focus on "(a) the commonest word forms in a language; (b) the central patterns of usage; (c) the combinations which they usually form".

While the term may occasionally be misinterpreted to indicate a syllabus consisting solely of vocabulary items, a lexical syllabus actually covers "all aspects of language, differing from a conventional syllabus only in that the central concept of organization is lexis" (Hunston, 2002, p. 189). Sinclair (2000, p. 191) would say that the grammar covered in a lexical syllabus is "lexical grammar", not "lexico-grammar", which attempts to "build a grammar and lexis on an equal basis". Indeed, as Murison-Bowie (1996, p. 185) observes,

> in using corpora in a teaching context, it is frequently 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.

Sinclair and his colleagues' proposal for a lexical syllabus is echoed by Lewis (1993, 1997a, 1997b, 2000), who provides strong support for the lexical approach to language teaching.

A focus of the lexical approach to language pedagogy is teaching collocations (i.e. habitual co-occurrences of lexical items) and the related concept of prefabricated units. There is a consensus that collocational knowledge is important for developing L1/L2 language skills (e.g. Bahns, 1993; Zhang, 1993; Cowie, 1994; Herbst, 1996; Kita & Ogata, 1997; Partington, 1998; Hoey, 2000, 2004; Shei & Pain, 2000; Sripicharn, 2000; Altenberg & Granger, 2001; McEnery & Wilson, 2001; McAlpine & Myles, 2003; Nesselhauf, 2003). Hoey (2004), for example, posits that "learning a lexical item entails learning what it occurs with and what grammar it tends to have". Cowie (1994, p. 3168) argues that "native-like proficiency of a language depends crucially on knowledge of a stock of prefabricated units". Aston (1995) also notes that the use of prefabs can speed language processing in both comprehension and production, thus creating native-like fluency.

A powerful reason for the employment of collocations, as Partington (1998, p. 20) suggests, "lies in the way it facilitates communication processing on the part of hearer", because "language consisting of a relatively high number of fixed phrases is generally more predictable than that which is not" while "in real time language decoding, hearers need all the help they can get". As such, competence in a language undoubtedly involves collocational knowledge (cf. Herbst, 1996, p. 389).

Collocational knowledge indicates which lexical items co-occur frequently with others and how they combine within a sentence. Such knowledge is evidently more important than individual words themselves (cf. Kita & Ogata, 1997, p. 230) and is needed for effective sentence generation (cf. Smadja & McKeown, 1990). Zhang (1993), for example, finds that more proficient L2 writers use significantly more collocations, more accurately and in more variety than less proficient learners. Collocational error is a common type of error for learners (cf. McAlpine & Myles, 2003, p. 75). Gui and Yang (2002, p. 48) observe, on the basis of the one-million-word Chinese Learner English Corpus, that collocation error is one of the major error types for Chinese learners of English. Altenberg and Granger (2001) and Nesselhauf (2003) find that even advanced learners of English have considerable difficulties with collocation. One possible explanation is that learners are deficient in "automation of collocations" (Kjellmer, 1991). "As a result, learners need detailed information about common collocational patterns and idioms; fixed and semi-fixed lexical expressions and different degrees of variability; relative frequency and currency of particular patterns; and formality level" (McAlpine & Myles, 2003, p. 75).

Corpora are useful in this respect, not only because collocations can only reliably be measured quantitatively, but also because the key word in context (KWIC) view of corpus data exposes learners to a great deal of authentic data in a structured way. Our view is line with Kennedy (2003), who discusses the relationship between corpus data and the nature of language learning, focusing on the teaching of collocations. The author argues that second or foreign language learning is a process of learning "explicit knowledge" with awareness, which requires a great deal of exposure to language data.

In addition to the lexical focus, corpus-based teaching materials try to demonstrate how the target language is actually used in different contexts, as exemplified in Biber, Leech and Conrad's (2002) *Longman Student Grammar of Spoken and Written English*, which pays special attention to how English is used differently in various spoken and written registers.

Language Testing

Another emerging area of language pedagogy that has started to use the corpus-based approach is language testing. Alderson (1996) envisaged the following possible uses of corpora in this area: test construction, compilation and selection, test presentation, response capture, test scoring, and

calculation and delivery of results. He concludes that "[t]he potential advantages of basing our tests on real language data, of making data-based judgments about candidates' abilities, knowledge and performance are clear enough. A crucial question is whether the possible advantages are born out in practice" (Alderson, 1996, pp. 258–259). The concern raised in Alderson's conclusion appears to have been addressed satisfactorily so that nowadays computer-based tests are recognized as being comparable to paper-based tests (e.g. computer-based versus paper-based TOEFL tests).

A number of corpus-based studies of language testing have been reported. For example, Coniam (1997) demonstrated how to use word frequency data extracted from corpora to generate cloze tests automatically. Kaszubski and Wojnowska (2003) presented a corpus-driven computer program, TestBuilder, for building sentence-based ELT exercises. The program can process raw corpora of plain texts or corpora annotated with part-of-speech information, using another linked computer program that assigns the part-of-speech category to each word in the corpus automatically in real time. The annotated data is used in turn as input for test material selection. Indeed, corpora have recently been used by major providers of test services for a number of purposes:

- as an archive of examination scripts;
- to develop test materials;
- to optimize test procedures;
- to improve the quality of test marking;
- to validate tests; and
- to standardize tests.

For example, the University of Cambridge Local Examinations Syndicate (UCLES) is active in both corpus development (e.g. Cambridge Learner Corpus, Cambridge Corpus of Spoken English, Business English Text Corpus, and Corpus of Young Learners English Speaking Tests) and the analysis of native English corpora and learner corpora. At UCLES, native English corpora such as the BNC are used "to investigate collocations, authentic stems and appropriate distractors which enable item writers to base their examination tasks on real texts" (Ball, 2001, p. 7);[2] the corpus-based approach is used to explore "the distinguishing features in the writing performance of EFL/ESL learners or users taking the Cambridge English examinations" and how to incorporate these into "a single scale of bands, that is, a common scale, describing different levels of L2 writing proficiency" (Hawkey, 2001, p. 9); corpora are also used for the purpose of speaking assessment (Ball & Wilson, 2002; Taylor, 2003) and to develop domain-specific (e.g. business English) wordlists for use in test materials (Ball, 2002; Horner & Strutt, 2004).

Teacher Development

For learners to benefit from the use of corpora, language teachers must first of all be equipped with a sound knowledge of the corpus-based approach. It is unsurprising then to discover that corpora have been used in training language teachers (e.g. Allan, 1999, 2002; Conrad, 1999; Seidlhofer, 2000, 2002; O'Keeffe & Farr, 2003). Allan (1999), for example, demonstrates how to use corpus data to raise the language awareness of English teachers in Hong Kong secondary schools. Conrad (1999) presents a corpus-based study of linking adverbials (e.g. *therefore* and *in other words*), on the basis of which she suggests that it is important for a language teacher to do more than using classroom concordancing and lexical or lexico-grammatical analyses if language teaching is to take full advantage of the corpus-based approach. Conrad's concern with teacher education is echoed by O'Keeffe and Farr (2003), who argue that corpus linguistics should be included in initial language teacher education so as to enhance teachers' research skills and language awareness.

Direct Use of Corpora

While indirect uses such as syllabus design and materials development are closely associated with what to teach, corpora have also provided valuable insights into how to teach. Of Leech's (1997) three focuses, direct uses of corpora include "teaching about", "teaching to exploit", and "exploiting to teach", with the latter two relating to how to use. Given a number of restricting factors as noted in the previous section, direct uses have so far been confined largely to learning at more advanced levels, for example, in tertiary education, whereas in general English language teaching (let alone to mention other foreign languages), especially in secondary education (see Braun, 2007 for a rare example of an empirical study of using corpora in secondary education), the direct use of corpora is "still conspicuously absent" (Kaltenböck & Mehlmauer-Larcher, 2005).

"Teaching about" means teaching corpus linguistics as an academic subject like other sub-disciplines of linguistics such as syntax and pragmatics. Corpus linguistics has now found its way into the curricula for linguistics and language related degree programmes at both postgraduate and undergraduate levels in many universities around the world. "Teaching to exploit" means providing students with "hands-on" know-how, as emphasized in McEnery, Xiao and Tono (2006), so that they can exploit corpora for their own purposes. Once the student has acquired the necessary knowledge and techniques of corpus-based language study, the learning activity may become student centred. "Exploiting to teach" means using a corpus-based approach to teaching language and linguistics courses (e.g. sociolinguistics and discourse analysis), which would otherwise be taught using non-corpus-based methods.

If the focuses of "teaching about" and "exploiting to teach" are viewed as being associated typically with students of linguistics and language programmes, "teaching to exploit" relates to students of all subjects which involve language study and learning, who are expected to benefit from the so-called data-driven learning (DDL) or "discovery learning".

The issue of how to use corpora in the language classroom has been discussed extensively in the literature. With the corpus-based approach to language pedagogy, the traditional "three Ps" (Presentation, Practice and Production) approach to teaching may not be entirely suitable. Instead, the more exploratory approach of "three Is" (Illustration, Interaction and Induction) may be more appropriate, where "illustration" means looking at real data, "interaction" means discussing and sharing opinions and observations, and "induction" means making one's own rule for a particular feature, which "will be refined and honed as more and more data is encountered" (see Carter & McCarthy, 1995, p. 155). This progressive induction approach is what Murison-Bowie (1996, p. 191) would call the interlanguage approach: namely, partial and incomplete generalizations are drawn from limited data as a stage on the way towards a fully satisfactory rule. While the "three Is" approach was originally proposed by Carter and McCarthy (1995) to teach spoken grammar, it may also apply to language education as a whole, in our view.

It is clear that the exploratory teaching approach focusing on "three Is" is in line with Johns' (1991) concept of "data-driven learning (DDL)". Johns was perhaps among the first to realize the potential of corpora for language learners (e.g. Higgins & Johns, 1984). In his opinion, "research is too serious to be left to the researchers" (Johns, 1991, p. 2). As such, he argues that the language learner should be encouraged to become "a research worker whose learning needs to be driven by access to linguistic data" (Johns, 1991, p. 2). John's web-based Kibbitzer (www.ling.lancs.ac.uk/corplang/Kibbitzers/Kibbitzers.chw) gives some very good examples of DDL.

DDL can be either teacher-directed or learner-led (i.e. discovery learning) to suit the needs of learners at different levels, but it is basically learner-centred. This autonomous learning process "gives the student the realistic expectation of breaking new ground as a 'researcher', doing something which is a unique and individual contribution" (Leech, 1997, p. 10). It is important to note,

however, that the key to successful DDL, even if it is student-centred, is the appropriate level of teacher guidance or pedagogical mediation depending on the learners' age, experience and proficiency level, because "a corpus is not a simple object, and it is just as easy to derive nonsensical conclusions from the evidence as insightful ones" (Sinclair, 2004, p. 2). In this sense, it is even more important for language teachers to be equipped with the necessary training in corpus analysis.

Johns (1991) identifies three stages of inductive reasoning with corpora in the DDL approach: observation (of concordanced evidence), classification (of salient features) and generalization (of rules). The three stages roughly correspond to Carter and McCarthy's (1995) "three Is". The DDL approach is fundamentally different from the "three Ps" approach in that the former involves bottom-up induction whereas the latter involves top-down deduction. The direct use of corpora and concordancing in the language classroom has been discussed extensively in the literature (e.g. Tribble, 1991, 1997a, 1997b, 2000, 2003; Tribble & Jones, 1990, 1997; Flowerdew, 1993; Karpati, 1995; Kettemann, 1995, 1996; Wichmann, 1995; Woolls, 1998; Aston, 2001; Osborne, 2001, Braun, 2007), covering a wide range of issues including, for example, underlying theories, methods and techniques, and problems and solutions.

Teaching Oriented Corpora

Teaching-oriented corpora are particularly useful in teaching LSP (LSP corpora) and in research on L1 (developmental corpora) and L2 (learner corpora) language acquisition. Such corpora can be used directly or indirectly in language pedagogy as discussed in the previous sections.

LSP and Professional Communication

In addition to teaching English as a second or foreign language in general, a great deal of attention has been paid to domain-specific language use and professional communication (e.g. English for specific purposes and English for academic purpose). For example, Thurstun and Candlin (1997, 1998) explore the use of concordancing in teaching writing and vocabulary in academic English. Hyland (1999) compares the features of the specific genres of metadiscourse in introductory course books and research articles on the basis of a corpus consisting of extracts from twenty-one university textbooks for different disciplines and a similar corpus of research articles.

Likewise, Upton and Connor (2001) undertake a "move analysis" in the business English using a business learner corpus. The authors approach the cultural aspect of professional communication by comparing the "politeness strategies" used by learners from different cultural backgrounds. Thompson and Tribble (2001) examine citation practices in academic text. Koester (2002) argues, on the basis of an analysis of the performance of speech acts in workshop conversations, for a discourse approach to teaching communicative functions in spoken English. Yang and Allison (2003) study the organizational structure in research articles in applied linguistics. Carter and McCarthy (2004) explore, on the basis of the Cambridge and Nottingham Corpus of Discourse in English (CANCODE), a range of social contexts in which creative uses of language are manifested. Hinkel (2004) compares the use of tense, aspect and the passive in L1 and L2 academic texts.

Xiao (2003) reviews a number of case studies using specialized multilingual corpora to teach domain specific translation. Parallel concordancing is not only useful in translation teaching; it can also aid the so-called "reciprocal learning" (Johns, 1997), where two language learners from different L1 backgrounds are paired to help each other learn their language. Studies such as these demonstrate that LSP corpora are particularly useful in teaching LSP and professional communication.

Learner Corpora and Interlanguage Analysis

The creation and use of learner corpora in language pedagogy and interlanguage research has been welcomed as one of the most exciting recent developments in corpus-based language studies. If native speaker corpora of the target language provide a top-down approach to using corpora in language pedagogy, learner corpora provide a bottom-up approach to language teaching (Osborne, 2002).

A learner corpus, as opposed to a "developmental corpus" composed of data produced by children acquiring their mother tongue (L1), comprises written or spoken data produced by language learners who are acquiring a second or foreign language. Data of this type has particularly been useful in language pedagogy and second language acquisition (SLA) research, as demonstrated by the fruitful learner corpus studies published over the past decade (see Pravec, 2002; Keck, 2004; and Myles, 2005 for recent reviews). SLA research is primarily concerned with "the mental representations and developmental processes which shape and constrain second language (L2) productions" (Myles, 2005, p. 374).

Language acquisition occurs in the mind of the learner, which cannot be observed directly and must be studied from a psychological perspective. Nevertheless, if learner performance data is shaped and constrained by such a mental process, it at least provides indirect, observable and empirical evidence for the language acquisition process. Note that using product as evidence for process may not be less reliable; sometimes this is the only practical way of finding about process. Stubbs (2001) draws a parallel between corpora in corpus linguistics and rocks in geology, "which both assume a relation between process and product. By and large, the processes are invisible, and must be inferred from the products". Like geologists who study rocks because they are interested in geological processes to which they do not have direct access, SLA researchers can analyse learner performance data to infer the inaccessible mental process of SLA.

Learner corpora can also be used as an empirical basis that tests hypotheses generated using the psycholinguistic approach, and to enable the findings previously made on the basis of limited data of a small number of informants to be generalized. Additionally, learner corpora have widened the scope of SLA research so that, for example, interlanguage research nowadays treats learner performance data as a category in its own right rather than as decontextualised errors in traditional error analysis (cf. Granger, 1998, p. 6).

At the pre-conference workshop on learner corpora affiliated to the second International Symposium of Corpus Linguistics held at the University of Lancaster, the workshop organizers Tono and Meunier observed that learner corpora are no longer in their infancy but are going through their nominal teenage years—they are full of promise but not yet fully developed.

In language pedagogy, the implications of learner corpora have been explored for curriculum design, materials development and teaching methodology (cf. Keck, 2004, p. 99). The interface between L1 and L2 materials has been explored. Meunier (2002), for example, argues that frequency information obtained from native speaker corpora alone is not sufficient to inform curriculum and materials design. Rather, "it is important to strike a balance between frequency, difficulty and pedagogical relevance. That is exactly where learner corpus research comes into play to help weigh the importance of each of these" (Meunier, 2002, p. 123). Meunier also advocates the use of learner data in the classroom, suggesting that exercises such as comparing learner and native speaker data and analysing errors in learner language will help students to notice gaps between their interlanguage and the language they are learning.

Interlanguage studies based on learner corpora which have been undertaken so far focus on what Granger (2002) calls "Contrastive Interlanguage Analysis (CIA)", which compares learner data and the data produced by native speakers of the target language, or the learner's L1. The first type of comparison typically aims to evaluate the level of under- or overuse of particular linguistic

features in learner language while the second type aims to uncover L1 interference or transfer. Corpus data produced by learners from different L1 backgrounds can also be compared against one another with the aim of uncovering common features of SLA process by discarding L1-specific peculiarities. In addition to CIA, learner corpora have also been used to investigate the order of acquisition of particular morphemes. Readers can refer to Granger et al. (2002) for recent work in the use of learner corpora, and read Granger (2003) for a more general discussion of the applications of learner corpora such as the International Corpus of Learner English (ICLE).

In addition to SLA research, learner corpora can also be used directly in classroom teaching. For example, Seidlhofer (2002) and Mukherjee and Rohrbach (2006) demonstrate how a "local learner corpus" containing students' own writings can be used directly for learning by coping with students' questions about their own or classmates' writings, or analysing and correcting errors in such familiar writings.

Conclusions

Before we close the discussion of using corpora in language pedagogy, it is appropriate to address some objections to the use of corpora in language learning and teaching. While frequency and authenticity are often considered two of the most important advantages of using corpora, they are also the locus of criticism from language pedagogy researchers. For example, Cook (1998, p. 61) argues that corpus data impoverishes language learning by giving undue prominence to what is simply frequent at the expense of rarer but more effective or salient expressions. Widdowson (1990, 2000) argues that corpus data is authentic only in a very limited sense in that it is de-contextualized (i.e. traces of texts rather than discourse) and must be re-contextualized in language teaching. On the other hand, it can be argued that

> using corpus data not only increases the chances of learners being confronted with relatively infrequent instances of language use, but also of their being able to see in what way such uses are atypical, in what contexts they do appear, and how they fit in with the pattern of more prototypical uses.
>
> (Osborne, 2001, p. 486)

This view is echoed by Goethals (2003, p. 424), who argues that "frequency ranking will be a parameter for sequencing and grading learning materials" because "frequency is a measure of *probability* of usefulness" and "high-frequency words constitute a core vocabulary that is useful above the incidental choice of text of one teacher or textbook author". Hunston (2002, pp. 194–195) observes that "items which are important though infrequent seem to be those that echo texts which have a high cultural value", though in many cases "cultural salience is not clearly at odds with frequency".

While frequency information is readily available from corpora, no corpus linguist has ever argued that the most frequent is most important. On the contrary, Kennedy (1998, p. 290) argues that frequency "should be only one of the criteria used to influence instruction" and that "the facts about language and language use which emerge from corpus analyses should never be allowed to become a burden for pedagogy". As such, raw frequency data is often adjusted for use in a syllabus, as reported in Renouf (1987, p. 168).

It would be inappropriate, therefore, for language teachers, syllabus designers and materials writers to ignore "compelling frequency evidence already available", as pointed out by Leech (1997, p. 16), who argues that: "Whatever the imperfections of the simple equation 'most frequent' = 'most important to learn', it is difficult to deny that frequency information becoming available from corpora has an important empirical input to language learning materials".

Kaltenböck and Mehlmauer-Larcher (2005, p. 78) downplay the role of frequency in language learning, arguing that "what is frequent in language will be picked up by learners automatically, precisely because it is frequent, and therefore does not have to be consciously learned". This is not true, however. Determiners such as *a* and *the* are certainly very frequent in English, yet they are difficult for Chinese learners of English because their mother tongue does not have such grammatical morphemes and does not maintain a count-mass noun distinction.

Clearly, frequency is not "automatically pedagogically useful" (Kaltenböck & Mehlmauer-Larcher, 2005, p. 78); decisions relating to teaching must also take account of overall teaching objectives, learners' concrete situations, cognitive salience, learnability, generative value and, of course, teachers' intuitions (cf. Kaltenböck & Mehlmauer-Larcher, 2005, p. 78). However, frequency can at least help syllabus designers, materials writers and teachers alike to make better-informed and more carefully motivated decisions (cf. Gavioli & Aston, 2001, p. 239).

If we leave objections to frequency data to one side, Widdowson (1990, 2000) also questions the use of authentic texts in language teaching. In his opinion, authenticity of language in the classroom is "an illusion" (1990, p. 44) because even though corpus data may be authentic in one sense, its authenticity of purpose is destroyed by its use with an unintended audience of language learners (see Murison-Bowie, 1996, p. 189). Widdowson (2003, p. 93) makes a distinction between "genuineness" and "authenticity", which are claimed to be the features of text as a product and discourse as a process respectively: corpora are genuine in that they comprise attested language use, but they are not authentic for language teaching because their contexts (as opposed to co-texts) have been deprived.

We will not be engaged in the debate here, but would like to draw readers' attention to Stubbs' (2001) metaphor of product versus process as cited in the section "Learner Corpora and Interlanguage Analysis". The implication of Widdowson's argument is that only language produced for imaginary situations in the classroom is "authentic". Even if we do follow Widdowson's genuineness-authenticity distinction, it is not clear why such imaginary situations are authentic because authenticity, as opposed to genuineness, would mean real communicative context. Situations conjured up for classroom teaching obviously do not take place in really communicative contexts; then how can they be authentic, if we choose to keep this distinction? When students learn and practise a shopping "discourse", they are actually by no means doing shopping! Furthermore, as argued by Fox (1987), invented examples often do not reflect nuances of usage. That is perhaps why, as Mindt (1996, p. 232) observes, students who have been taught "school English" cannot readily cope with English used by native speakers in real life. As such, Wichmann (1997, p. xvi) argues that in language teaching, "the preference for 'authentic' texts requires both learners and teachers to cope with language which the textbooks do not predict".

The discussions in the previous sections suggest that corpora appear to have played a more important role in helping to decide what to teach (indirect uses) than how to teach (direct uses). While indirect uses of corpora seem to be well established, direct uses of corpora in teaching are largely confined to advanced levels such as higher education. Corpus-based learning activities are nearly absent general teaching English as a foreign language (TEFL) classes at lower levels such as secondary education. Of the various causes for this absence mentioned earlier, perhaps the most important are the access to appropriate corpus resources and the necessary training of teachers, which we view as priorities for future tasks of corpus linguists if corpora are to be popularized to more general language teaching context.

While there are a wide range of existing corpora that are publicly available (see Xiao, 2008 for a recent survey), the majority of those resources have been developed "as tools for linguistic research and not with pedagogical goals in mind" (Braun, 2007). As Cook (1998, p. 57) suggests, "the leap from linguistics to pedagogy is [...] far from straightforward". To bridge the gap between corpora and language pedagogy, the first step would involve creating corpora that are pedagogically

motivated, in both design and content, to meet pedagogical needs and curricular requirements so that corpus-based learning activities become an integral part, rather than an additional option, of the overall language curriculum. Such pedagogically motivated corpora "should not only be more coherent than traditional corpora; they should, as far as possible, also be complementary to school curricula, to facilitate both the contextualisation process and the practical problems of integration" (Braun, 2007, p. 310). The design of such corpus-based learning activities must also take account of learners' age, experience and level as well as their integration into the overall curriculum.

Given the situation of learners (e.g. their age, level of language competence, level of expert knowledge and attitude towards learning autonomy) in general language education in relation to advanced learners in tertiary education, even such pedagogically motivated corpus-based learning activities must be mediated by teachers. This in turn raises the issue of the current state of teachers' knowledge and skills of corpus analysis and pedagogical mediation, which is another practical problem that has prevented direct use of corpora in language pedagogy. As Kaltenböck and Mehlmauer-Larcher (2005, p. 81) argue, "mediation by the teacher is a necessary prerequisite for successful application of computer corpora in language teaching and should therefore be given sufficient attention in teacher education courses" (cf. also O'Keeffe & Farr, 2003). However, as the integration of corpus studies in language teacher training is only a quite recent phenomenon (cf. Chambers, 2007), "it will therefore at least take more time, and perhaps a new generation of teachers, for corpora to find their way into the language classroom" (Braun, 2007, p. 308).

In conclusion, if these two tasks are accomplished, it is our view that corpora will not only revolutionize the teaching of subjects such as grammar in the twenty-first century as Conrad (2000) has predicated, they will also fundamentally change the ways we approach language education, including both what is taught and how it is taught. As Gavioli and Aston (2001) argue, corpora should not only be viewed as resources that help teachers to decide what to teach, they should also be viewed as resources from which learners may learn directly.

Notes

1. In this chapter, we will not be concerned with the debate over the use of corpus data in linguistic analysis and language education. Readers interested in the pros and cons of using corpus data should refer to Sinclair (1991), Widdowson (1991, 2000), de Beaugrande (2001) and Stubbs (2001). While Widdowson, Sinclair and de Beaugrande characterize two extreme attitudes towards corpora, there are many milder (positive or negative) reactions to corpus data between the two extremes. Readers can refer to Nelson (2000) for a good review.
2. "Stem" is a technical term in language testing that refers to "the top part of a multiple-choice item, usually a statement or question" (Fulcher & Davidson, 2007, p. 53). As a collection of attested language data, a corpus is a good resource for test writers as it can provide abundant authentic stems.

References

Aijmer, K. (2009). *Corpora and language teaching.* Amsterdam: John Benjamins.

Alderson, C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds.), *Using corpora for language research* (pp. 248–259). London: Longman.

Allan, Q. (1999). Enhancing the language awareness of Hong Kong teachers through corpus data. *Journal of Technology and Teacher Education*, *7*(1), 57–74.

Allan, Q. (2002). The TELEC secondary learner corpus: a resource for teacher development. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 195–212). Amsterdam: John Benjamins.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*(2), 173–195.

Aston, G. (1995). Corpora in language pedagogy: matching theory and practice. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 257–270). Oxford: Oxford University Press.

Aston, G. (Ed.) (2001). *Learning with corpora*. Houston, TX: Athelstan.
Aston , G., Bernardini, S. & Stewart, D. (Eds.) (2004). *Corpora and language learners*. Amsterdam: John Benjamins.
Atkins, B., & Levin, B. (1995). Building on a corpus: A linguistic and lexicographical look at some near-synonyms. *International Journal of Lexicography, 8*, 85–114.
Bahns, J. (1993). Lexical collocations: A contrastive view. *ELT Journal, 47*(1), 56–63.
Ball, F. (2001). Using corpora in language testing. *Research Notes, 6*, 6–8.
Ball, F. (2002). Developing wordlists for BEC. *Research Notes, 8*, 10–13.
Ball, F., & Wilson, J. (2002). Research projects relating to YLE speaking tests. *Research Notes*, 7, 8–10.
Bernardini, S. (2000). *Competence, capacity, corpora: A study in corpus-aided language learning*. Bologna: CLUEB.
Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
Biber, D., Johansson S., Leech G., Conrad S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.
Biber, D., Leech, G. & Conrad, S. (2002). *Longman student grammar of spoken and written English*. Harlow: Longman.
Braun, S. (2007). Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL, 19*(3), 307–328.
Braun, S., Kohn, K. & Mukherjee, J. (Eds.) (2006). *Corpus technology and language pedagogy*. Frankfurt: Peter Lang.
Burnard, L., & McEnery, A. (Eds.) (2000). *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang.
Campoy, M., Gea-valor, M. & Belles-Fortuno, B. (2010). *Corpus-based approaches to English language teaching*. London: Continuum.
Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics, 16*(2), 141–158.
Carter, R., & McCarthy, M. (2004). Talking, creating: interactional language, creativity, and context. *Applied Linguistics, 25*(1), 62–88.
Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda & J. Santana (Eds.), *Corpora in the foreign language classroom: Selected papers from the sixth International Conference on Teaching and Language Corpora (TaLC 6)* (pp. 3–16). Amsterdam: Rodopi.
Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *CALICO Journal, 16*(2–4), 15–33.
Connor, U., & Upton, T. (Eds.) (2002). *Applied corpus linguistics: A multidimensional perspective*. Amsterdam: Rodopi.
Conrad, S. (1999). The importance of corpus-based research for language teachers. *System, 27*, 1–18.
Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34*, 548–560.
Cook, G. (1998). The uses of reality: A reply to Ronald Cater. *ELT Journal, 52*(1), 57–64.
Cowie, A. (1994). Phraseology. in R. Asher (Ed.), *The Encyclopaedia of language and linguistics* (Vol. 6) (pp. 3168–3171). Oxford: Pergamon Press Ltd.
Davies, M. (2005). *A frequency dictionary of Spanish*. London: Routledge.
Davies, M., & de Oliveira Preto-Bay, A. (2007). *A frequency dictionary of Portuguese*. London: Routledge.
de Beaugrande, R. (2001). Interpreting the discourse of H. G. Widdowson: A corpus-based critical discourse analysis. *Applied Linguistics, 22*(1), 104–121.
Flowerdew, J. (1993). Concordancing as a tool in course design. *System, 21*(3), 231–243.
Fox, G. (1987). The case for examples. In J. Sinclair (Ed.), *Looking up: An account of the COBUILD project* (pp. 137–149). London: HarperCollins.
Francis, G., Hunston, S. & Manning, E. (1997). *Collins COBUILD grammar patterns 1: Verbs*. London: HarperCollins.
Francis, G., Hunston, S. & Manning, E. (1998). *Collins COBUILD grammar patterns 2: Nouns and adjectives*. London: HarperCollins.
Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.
Gavioli, L. (2006). *Exploring corpora for ESP learning*. Amsterdam: John Benjamins.
Gavioli, L., & Aston, G. (2001). Enriching reality: language corpora in language pedagogy. *ELT Journal, 55*(3), 238–246.
Ghadessy, M., Henry, A. & Roseberry, R. (Eds.) (2001). *Small corpus studies and ELT: Theory and practice*. Amsterdam: John Benjamins.
Goethals, M. (2003). E.E.T.: the European English Teaching vocabulary-list. In B. Lewandowska-Tomaszczyk (Ed.), *Practical Applications in language and computers* (pp. 417–427). Frankfurt: Peter Lang.
Granger, S. (1998). The computer learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3–18). London: Longman.
Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3–33). Amsterdam: John Benjamins.

Granger, S. (2003). Practical applications of learner corpora. In B. Lewandowska-Tomaszczyk (Ed.), *Practical applications in language and computers* (pp. 291–302). Frankfurt: Peter Lang.

Granger, S., Hung, J. & Petch-Tyson, S. (Eds.) (2002). *Computer learner corpora, second language acquisition, and foreign language teaching.* Amsterdam: John Benjamins.

Gui, S., & Yang, H. (2002). *Zhonguo Xuexizhe Yingyu Yuliaoku* [*Chinese learner English corpus*]. Shanghai: Shanghai Foreign Language Education Press.

HarperCollins (1995). *Collins COBUILD English dictionary* (2nd ed.). London: Collins COBUILD.

Hawkey, R. (2001). *IIS student questionnaire.* Cambridge: UCLES.

Herbst, T. (1996). What are collocations: sandy beaches or false teeth? *English Studies*, *4*, 379–393.

Hidalgo, E., Quereda, L. & Santana, J. (2007). *Corpora in the foreign language classroom: Selected papers from the sixth International Conference on Teaching and Language Corpora (TaLC 6).* Amsterdam: Rodopi.

Higgins, J., & Johns, T. (1984). *Computers in language learning.* Oxford: Oxford University Press.

Hinkel, E. (2004). Tense, aspect and the passive voice in L1 and L2 academic texts. *Language Teaching Research*, *8*(1), 5–29.

Hoey, M. (2000). A world beyond collocation: new perspectives on vocabulary teaching. In M. Lewis (Ed.), *Teaching collocations* (pp. 224–245). Hove: Language Teaching Publications.

Hoey, M. (2004). Lexical priming and the properties of text. In A. Partington, J. Morley & L. Haarman (Eds.), *Corpora and discourse* (pp. 385–412). Bern: Peter Lang.

Hornby, A., & Crowther, J. (1999). *Oxford advanced learner's dictionary* (5th ed.). Oxford: Oxford University Press.

Horner, D., & Strutt, P. (2004). Analyzing domain-specific lexical categories: evidence from the BEC written corpus. *Research Notes*, *15*, 6–8.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English.* Amsterdam: John Benjamins.

Hyland, K. (1999). Talking to students: Metadiscourse in introductory coursebooks. *English for Specific Purposes*, *18*(1), 3–26.

Johns, T. (1991). "Should you be persuaded": Two samples of data-driven learning materials. In T. Johns and P. King (Eds.), *Classroom Concordancing ELR Journal*, *4* (pp. 1–16). University of Birmingham.

Johns, T. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). Harlow: Longman.

Jones, R., & Tschirner, E. (2005). *A frequency dictionary of German.* London: Routledge.

Kaltenböck, G., & Mehlmauer-Larcher, B. (2005). Computer corpora and the language classroom: On the potential and limitations of computer corpora in language teaching. *ReCALL*, *17*, 65–84.

Karpati, I. (1995). *Concordance in language learning and teaching.* Pecs: University of Pecs.

Kaszubski, P., & Wojnowska, A. (2003). Corpus-informed exercises for learners of English: The TestBuilder program. In E. Oleksy & B. Lewandowska-Tomaszczyk (Eds.), *Research and scholarship in integration processes: Poland–USA–EU* (pp. 337–354). Łódź: Łódź University Press.

Keck, C. (2004). Corpus linguistics and language teaching research: Bridging the gap. *Language Teaching Research*, *8*(1), 83–109.

Kennedy, G. (1998). *An introduction to corpus linguistics.* Harlow: Longman.

Kennedy, G. (2003). Amplifier collocations in the British National Corpus: Implications for English language teaching. *TESOL Quarterly*, *37*(3), 467–487.

Kettemann, B. (1995). On the use of concordancing in ELT. *TELL&CALL*, *4*, 4–15.

Kettemann, B. (1996). Concordancing in English Language Teaching. In S. Botley, J. Glass, T. McEnery & A. Wilson (Eds.), *Proceedings of teaching and language corpora* (pp. 4–16). Lancaster: Lancaster University.

Kettemann, B., & Marko, G. (2002). *Teaching and learning by doing corpus Analysis.* Amsterdam: Rodopi.

Kettemann, B., & Marko, G. (Eds.) (2006). *Planning, gluing and painting corpora: Inside the applied corpus linguist's workshop.* Frankfurt: Peter Lang.

Kita, K., & Ogata, H. (1997). Collocations in language learning: Corpus-based automatic compilation of collocations and bilingual collocation concordancer. *Computer Assisted Language Learning*, *10*(3), 229–238.

Kjellmer, G. (1991). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), *English corpus linguistics: Studies in honour of Jan Svartvik* (pp. 111–127). Harlow: Longman.

Koester, A. (2002). The performance of speech acts in workplace conversations and the teaching of communicative functions. *System*, *30*, 167–184.

Leech, G. (1997). Teaching and language corpora: A convergence. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 1–23). London: Longman.

Lewis, M. (1993). *The lexical approach: The state of ELT and the way forward.* Hove: Language Teaching Publications.

Lewis, M. (1997a). *Implementing the lexical approach: Putting theory into practice*. Hove: Language Teaching Publications.

Lewis, M. (1997b). Pedagogical implications of the lexical approach. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition: A rationale for pedagogy* (pp. 255–270). Cambridge: Cambridge University Press.

Lewis, M. (Ed.) (2000). *Teaching collocation: Further developments in the lexical approach*. Hove: Language Teaching Publications.

Longman (1995). *Longman Dictionary of Contemporary English* (3rd ed.). Harlow: Longman.

Lonsdale, D., & Bras, Y. (2009). *A frequency dictionary of French*. London: Routledge.

McAlpine, J., & Myles, J. (2003). Capturing phraseology in an online dictionary for advanced users of English as a second language: A response to user needs. *System*, *31*, 71–84.

McCarthy, M., McCarten, J. & Sandiford, H. (2005–2006). *Touchstone* (Books 1–4). Cambridge. Cambridge University Press.

McEnery, A., & Xiao, R. (2005). Help or help to: What do corpora have to say? *English Studies*, *86*(2), 161–187.

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

McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Routledge.

Meunier, F. (2002). The pedagogical value of native and learner corpora in EFL grammar teaching. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 119–142). Philadelphia: John Benjamins.

Mindt, D. (1996). English corpus linguistics and the foreign language teaching syllabus. In J. Thomas & M. Short (Eds.), *Using Corpora for language research* (pp. 232–247). Harlow: Longman.

Mishan, F. (2005). *Designing authenticity into language learning materials*. Chicago: Chicago University Press.

Mukherjee, J., & Rohrbach, J. (2006). Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In B. Kettemann & G. Marko (Eds.), *Planning, gluing and painting corpora: Inside the applied corpus linguist's workshop* (pp. 205–232). Frankfurt: Peter Lang.

Murison-Bowie, S. (1996). Linguistic corpora and language teaching. *Annual Review of Applied Linguistics*, *16*, 182–199.

Myles, F. (2005). Interlanguage corpora and second language acquisition research. *Second Language Research*, *21*(4), 373–391.

Nelson, M. (2000). A corpus-based study of business English and business English teaching materials. PhD thesis, the University of Manchester, Manchester. Accessed on 15 July 2010 at the URL http://users.utu.fi/micnel/thesis.html.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, *24*(2), 223–242.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

O'Keeffe, A., & Farr, F. (2003). Using language corpora in initial teacher education: pedagogic issues and practical applications. *TESOL Quarterly*, *37*(3), 389–418.

O'Keeffe, A., McCarthy, M. & Carter, R (2007). *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press.

Osborne, J. (2001). Integrating corpora into a language-learning syllabus. In B. Lewandowska-Tomaszczyk (Ed.), *PALC 2001: Practical applications in language corpora* (pp. 479–492). Frankfurt: Peter Lang.

Osborne, J. (2002). Top-down and bottom-up approaches to corpora in language teaching. In U. Connor and T. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 251–265). Amsterdam: Rodopi.

Partington, A. (1998). *Patterns and meanings: Using corpora for English language research and teaching*. Amsterdam: John Benjamins.

Pravec, N. (2002). Survey of learner corpora. *ICAME Journal*, *26*, 81–114.

Procter, P. (1995) *Cambridge international dictionary of English*. Cambridge: Cambridge University Press.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Harlow: Longman.

Renouf, A. (1987). Moving on. In J. Sinclair (Ed.), *Looking up: An account of the COBUILD project* (pp. 167–178). London: HarperCollins.

Römer, U. (2005). *Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics*. Amsterdam: John Benjamins.

Scott, M., & Tribble, C. (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 L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 207–224). New York: Peter Lang.

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning driven data. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213–234). Philadelphia: John Benjamins.

Shei, C., & Pain, H. (2000). An ESL writer's collocational aid. *Computer Assisted Language Learning*, *13*(2), 167–182.

Sinclair, J. (1987). *Collins COBUILD English language dictionary*. London: HarperCollins.
Sinclair, J. (1990). *Collins COBUILD English grammar*. London: HarperCollins.
Sinclair, J. (1991). *Corpus, concordance, collocation: Describing English language*. Oxford: Oxford University Press.
Sinclair, J. (1992). *Collins COBUILD English usage*. London: HarperCollins.
Sinclair, J. (2000). Lexical grammar. *Naujoji Metodologija*, *24*, 191–203.
Sinclair, J. (2003). *Reading concordances*. Harlow: Longman.
Sinclair, J. (Ed.) (2004). *How to use corpora in language teaching*. Amsterdam: John Benjamins.
Sinclair, J., & Renouf, A. (1988). A lexical syllabus for language learning. In R. Carter & M. McCarthy (Eds.), *Vocabulary and language teaching* (pp. 140–158). Harlow: Longman.
Smadja, F., & McKeown, K. (1990). Automatically extracting and representing collocations for language generation. In B. Berwick (Ed.), *Proceedings of the 28th annual meeting of Association for Computational Linguistics* (pp. 252–259). Pittsburgh: University of Pittsburgh.
Sripicharn, P. (2000). Data-driven learning materials as a way to teach lexis in context. In C. Heffer, H. Sauntson & G. Fox (Eds.), *Words in context: A tribute to John Sinclair on his retirement* (pp. 169–178). Birmingham: University of Birmingham.
Stubbs, M. (2001). Texts, corpora, and problems of interpretation: A response to Widdowson. *Applied Linguistics*, *22*(2), 149–172.
Tan, M. (2002). *Corpus studies in language education*. Bangkok: IELE Press.
Taylor, L. (2003). The Cambridge approach to speaking assessment. *Research Notes*, *13*, 2–4.
Thompson, P., & Tribble, C. (2001). Looking at citations: Using corpora in English for academic purposes. *Language Learning & Technology*, *5*(3), 91–105.
Thurstun, J., & Candlin, C. (1997). *Exploring academic English: A workbook for student essay writing*. Sydney: NCELTR.
Thurstun, J., & Candlin, C. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes*, *17*, 267–280.
Tribble, C. (1991). Concordancing and an EAP writing program. *CAELL Journal*, *1*(2), 10–15.
Tribble, C. (1997a). Corpora, concordances and ELT. In T. Boswood (Ed.), *New ways of using computers in language teaching*. Alexandria, VA: TESOL.
Tribble, C. (1997b). Improving corpora for ELT: Quick and dirty ways of developing corpora for language teaching. In B. Lewandowska-Tomaszczyk & P. Melia (Eds.), *Practical applications in language corpora—Proceedings of PALC '97* (pp. 107–117). Łódź: Łódź University Press.
Tribble, C. (2000). Practical uses for language corpora in ELT. In P. Brett & G. Motteram (Eds.), *A special interest in computers: Learning and teaching with information and communications technologies* (pp. 31–41). Kent: IATEFL.
Tribble, C. (2003). The text, the whole text … or why large published corpora aren't much use to language learners and teachers. In B. Lewandowska-Tomaszczyk (Ed.), *Practical applications in language and computers* (pp. 303–318). Frankfurt: Peter Lang.
Tribble, C., & Jones, G. (1990). *Concordances in the classroom: A resource book for teachers*. Harlow: Longman.
Tribble, C., & Jones, G. (1997). *Concordances in the classroom: Using corpora in language education*. Houston TX: Athelstan.
Upton, T., & Connor, U. (2001). Using computerized corpus analysis to investigate the text-linguistic discourse move of a genre. *English for Specific Purposes*, *20*, 313–329.
Wichmann, A. (1995). Using concordances for the teaching of modern languages in higher education. *Language Learning Journal*, *11*, 61–63.
Wichmann, A. (1997). General introduction. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. xvi–xvii). London: Longman.
Wichmann, A., Fligelstone, S., McEnery A. & Knowles, G. (Eds.) (1997). *Teaching and language corpora*. London: Longman.
Widdowson, H. (1990). *Aspects of language teaching*. Oxford: Oxford University Press.
Widdowson, H. (1991). The description and prescription of language. In J. Alatis (Ed.), *Georgetown University Round Table on Languages and Linguistics 1991*, (pp. 11–24). Washington, DC: Georgetown University Press.
Widdowson, H. (2000). The limitations of linguistics applied. *Applied Linguistics*, *21*(1), 3–25.
Widdowson, H. (2003). *Defining issues in English language teaching*. Oxford: Oxford University Press.
Willis, D. (1990). *The lexical syllabus: A new approach to language teaching*. London: HarperCollins.
Willis, J., Willis, D. & Davids, J. (1988–1989). *Collins COBUILD English course* (Parts 1–3). London: HarperCollins.
Woolls, D. (1998). Multilingual parallel concordancing for pedagogical use. In *Teaching and language corpora* (pp. 222–227). Keble College, Oxford, 24–27 July.
Xiao, R. (2003). Use of parallel and comparable corpora in language study. *English Education in China*, *1*. Accessed on 8 September 2010 at the URL http://pub.sinoss.net/portal/webgate/_CmdArticleList?QUERY=1126&JOURNALNO=1126&JournalID=1125.

Xiao, R. (2008). Well-known and influential corpora. In A. Lüdeling & M. Kyto (Eds.), *Corpus linguistics: An international handbook* (pp. 383–457). Berlin: Mouton de Gruyter.

Xiao, R., Rayson, P. & McEnery, T. (2009). *A frequency dictionary of Mandarin Chinese.* London: Routledge.

Yang, Y., & Allison, D. (2003). Research articles in applied linguistics: Moving from results to conclusions. *English for Specific Purposes, 22*, 365–385.

Zhang, X. (1993). *English collocations and their effect on the writing of native and non-native college freshmen.* PhD thesis, Indiana University of Pennsylvania.