# *Using Language Corpora in Initial Teacher Education: Pedagogic Issues and Practical Applications*

**ANNE O'KEEFFE and FIONA FARR**
*University of Limerick*
*Limerick, Ireland*

The vast increase in the number of corpus-based materials, such as dictionaries and grammars, attests to the importance of corpus linguistics to English language description. Developments are also evident in the use of corpora in the classroom in data-driven learning (Johns, 1991). These rapid developments in the use of language-related technology have not been matched by updated practices in teacher education. This article makes a case for the inclusion of corpus linguistics in initial language teacher education to enhance teachers' research skills and language awareness. The authors offer examples of corpus-based tasks for increasing students' understanding of word classes, register-related grammatical choices, and socioculturally conditioned grammatical choices. Practical considerations for the integration of language corpora in a teacher education program are outlined.

Applied linguists working on technology-related issues have for some time noted the relevance of technological changes in the digital global economy for TESOL (e.g., Chapelle, 2001; Cummins, 2000; Warschauer, 2000). Literacy is no longer just about reading and writing. Society now demands *multiliteracies* (Warschauer, 2000), which include a high proficiency in digital and online competencies (see also Doering & Beach, 2002; Pennington, 2001). Consequently, language teacher educators have a fundamental obligation to educate teachers in a way that empowers them to work in the modern world. The initiation and implementation of many national educational policies and directives targeted at teacher education institutions are testament to such an obligation. Writing about in-service and preservice foreign language teaching, Murray (1998) and Barnes and Murray (1999) argue that information and communication technology (ICT) "can no longer be an added extra but rather [is] an intrinsic part of a teacher's methodological repertoire" (Barnes & Murray, 1999, p. 167). They conclude that "this transition must occur in the initial teacher training period to have the greatest effect" (p. 167) because many novice teachers are too busy with other matters in the first years of teaching to assume the task of developing and integrating ICT into their teaching and learning.

Many researchers concur that promoting critical attitudes and developing conceptual as well as practical frameworks for technology in language learning is the key to meaningful future technology use (see, e.g., Egbert, Paulus, & Nakamichi, 2002; Meskill, Mossop, DiAngelo, & Pasquale, 2002). Doering & Beach (2002) argue that "it is primarily through active participation with technology as opposed to receiving instruction about technology that preservice teachers learn to recognize the value of technology tools" (p. 128). Others (e.g., Egbert et al., 2002; Murray, 1998; Tammelin, 2001) point out that mastery of ICT skills can also foster a positive attitude, increased confidence, and teacher empowerment. However, including technology skills in teacher education adds a complex layer of issues to what is already a full curriculum. This article begins to address some of these issues by describing how students learn to use technology to exploit the linguistic data contained in corpora of English in English language teaching courses at the University of Limerick, in Ireland, where we have been integrating corpora into teacher education courses since 1997.

## CORPUS LINGUISTICS AND LANGUAGE TEACHING

Many applied linguists who conduct corpus linguistic research are convinced of its significant impact on the field. According to McCarthy (2001), corpus linguistics represents cutting-edge change in terms of scientific techniques and methods, and probably foreshadows even more profound technological shifts that will "impinge upon our long-held notions of education, roles of teachers, the cultural context of the delivery of educational services and the mediation of theory and technique" (p. 125). Examination of large quantities of spoken and written texts has revealed language patterns and uses that had hitherto eluded intuition, and has resulted in improved dictionaries (see Fox, 1998) and grammars (see Biber, Johansson, Leech, Conrad, & Finegan, 1999—the *Longman Grammar of Spoken and Written English* [LGSWE], a grammar that draws on a corpus of 40 million words).

In addition, numerous studies have shown that the language presented in textbooks is often based on faulty intuition about how people use language. Holmes (1988, p. 40), for example, looks at epistemic modality in ESL textbooks as compared with corpus data and finds that many textbooks devote an unjustifiably large amount of attention to modal verbs at the expense of alternative linguistic strategies. Boxer and

Pickering (1995) contrast speech acts in textbook dialogues with real, spontaneous encounters found in a corpus. Carter (1998) compares real data from the Cambridge and Nottingham Corpus of Discourse in English (CANCODE) with dialogues from textbooks, finding that they lack core spoken language features such as discourse markers, vague language, ellipsis, and hedges. Kettermann (1995) highlights the mismatch between actual language use and the prescription in pedagogical grammars that reported speech should involve the *backshift rule* for tenses in reported speech constructions (see also Baynham, 1991, 1996; McCarthy, 1998). Hughes and McCarthy (1998) look at the use of past perfect verb forms and find that across a wide range of speakers in CANCODE, the past perfect has a broader and more complex function in spoken discourse than hitherto described. Corpus descriptions have also enhanced the understanding of units of fixed phrasing, collocation, and language patterning (Aston, 1995; Murison-Bowie, 1996; Sinclair, 1991; Svartvik, 1991).

Despite these and many other findings from corpus research, Svartvik (1991) points out that "the attitude to the use of corpora in linguistic research has had its ups and downs" (p. 555). Many practitioners and applied linguists point to the problems of adopting corpus-based material in the language classroom (see, e.g., Cook, 1998; Owen, 1996; Prodromou, 1997a, 1997b; Seidlhofer, 1999; Widdowson, 2000). Sinclair (1991), for example, makes the case for the use of "real" language in the classroom by asserting that "one does not study all of botany by making artificial flowers" (p. 6). However, Widdowson (2000) warns that just because corpus data are "real," one should not assume that using such data in the classroom will bring with it more "reality" (p. 7). The reality that corpus findings represent is, he argues, third- rather than first-person reality, and problems arise when "partial description" of "decontextualised language" (p. 7) is used to determine language prescription for the classroom. We argue, however, that in the end it is teachers who will engage in the process of recontextualising corpora and any useful findings from corpus-based description. It is teachers who will mediate between corpus-based content and the needs of the learners in their individual classroom contexts. To do this, teachers will need to be able to make informed decisions, and, not least of all, they will need to be able to access the validity of the arguments that are made in relation to corpus findings and corpus use.

Carter and McCarthy (1995) and others have argued that language corpora are a "useful resource for teachers and learners" (p. 144). However, Tribble (2000) notes that "despite the best efforts of people like Tim Johns, Guy Aston, John Flowerdew and myself not many teachers seem to be *using* corpora in their classrooms" (p. 31). We argue that if corpus applications and corpus findings are to reach the right audience (i.e., language learners), they must be integrated at the very core of teacher education courses (see also Chapelle, 2001; Conrad, 2000). In the context of teacher education for teachers who are speakers of English as a lingua franca (ELF), Seidlhofer (1999) comments,

> Teachers who have a good idea as to what options are in principle available to them, and have learnt to evaluate these critically, sceptically and confidently, are unlikely to be taken in by the absolute claims and exaggerated promises often made by any one educational philosophy, linguistic theory, teaching method or textbook. (p. 240)

However, many teacher educators themselves have not had extensive experience with corpora. We therefore hope that our experience in integrating corpus linguistics into teacher education courses is informative.

## CORPUS APPLICATIONS FOR TEACHER EDUCATION

We discuss our understanding of corpora in teacher education through the three characteristics Sternberg and Horvath (1995) ascribe to an expert teacher. To be an expert teacher, one must be more *knowledgeable,* be more *efficient,* and have better *insight* than nonexperts (either experienced or inexperienced). Whether or not one accepts this characterization, the responsibility of initial teacher education courses is ultimately to aim to produce teachers who have at least started their journey along the road to expertise. To do so, we at the University of Limerick have attempted to increase students' pedagogic, linguistic, and sociocultural awareness by examining how linguistic choices are realized in the ESL/EFL classroom. We came to use corpora for this purpose because the materials that we had been using for methodological skills acquisition (i.e., commercially available classroom transcripts and video recordings) have two major shortcomings: (a) They have traditionally lent themselves almost exclusively to qualitative scrutiny, the conclusions of which may sometimes be elusive to and oversubjectified by inexperienced students; and (b) they fail to allow the practices of teaching to be interpreted within their contexts of realisation because many local contextualization cues are lost in their reproduction and extraction for third-party analysis operating in far-removed realities. In other words, nonpresent third parties in different educational or cultural surrounds cannot easily capture in their entirety the sociocultural and environmental factors that create and cast the lesson. This is particularly true in our Irish context, as many of the teacher education materials available commercially are either British or U.S. produced and often mismatch the conditions experienced by our students.

To rectify the contextual mismatch, we have been engaged in the process of building our own English language teaching classroom corpus to use in teacher education. For example, Farr (2002) reports on a study in which teachers' classroom interactions were recorded and then transcribed to form a minicorpus, which in turn was used as the basis for analysis of the correlation between question forms and productivity (i.e., the length of student response in numbers of words) in the language classroom. Our classroom corpus will ultimately include four types of transcriptions: experienced teachers operating in different sociocultural settings from our students, experienced teachers operating in the same sociocultural settings, other students operating in different sociocultural settings, and our students during their on-site teaching practice sessions.

Another area of application for corpora in language teacher education that we look at is raising linguistic awareness (relating to the knowledge category as detailed by Sternberg & Horvath, 1995). However, in addition to pedagogical and linguistic awareness, and fundamental to the evolution of corpus use in the context of English language classrooms around the world, teachers need to develop a critical awareness of what corpus findings represent. As we illustrate, corpus investigations can engender enquiry in prospective teachers so that they do not readily accept corpus findings as absolute truths.

## TECHNOLOGICAL EXPERTISE FOR CORPUS EXPLORATION

To work with corpora, students need some basic technological expertise. At first, corpus linguistics can seem very daunting, and teacher educators should be careful not to frighten students off with seemingly complex statistics and computations. It is crucial, we have found, to start with a basic distinction between a *corpus,* which is essentially a collection of texts (see Biber, Conrad, & Reppen, 1998), and the *software* that one can use to analyse it. Teachers who choose to use corpora in their language classrooms will need to be discerning about software and corpora, and, at the most basic level, they will need to know the common functions and applications of the available software.

### Concordancing

We always begin with concordancing as it is a core tool for analysis in corpus linguistics. Concordancing is the process of using software to search for all the occurrences of one word (or phrase) in a corpus. All of the occurrences are presented with the *node word/phrase* (the one searched for) in the centre of the line, with seven or eight words presented at either side of the node word. Depending on the software, the number of words at either side of the node word or phrase can be adjusted to allow for more context. The sample of concordance lines for the word *made* in Figure 1 was produced with the Collins Cobuild (n.d.) *Corpus Concordance Sampler* (freely available online; see the Appendix for Web sites and software relevant to corpus linguistics). It provides 40 examples based on any or all of the following corpora: British books, ephemera, radio, newspapers, magazines (26 million words); U.S. books, ephemera, and radio (9 million words); and British transcribed speech (10 million words). Apart from free Internet concordancing sites, many commercially available software packages allow the user to go back to the original source text of any one of lines or at least provide a much larger amount of context if required.

A key manipulation of a concordance involves sorting alphabetically to the left and to the right of the node word or phrase. Figures 2 and 3 show an example produced with WordSmith Tools (Scott, 1996) analysing the *Corpus of Spoken Professional American English* (CSPAE, 2000, a 2-million-word corpus on CD-ROM made up of academic discussions, committee meetings, and White House press conferences). The node word is still *made,* but this time we present the line samples in two different sorting formats: sorted to the left (Figure 2) and sorted to the right (Figure 3) of the node word.

By looking to the left and to the right of a word with our students, we find more information about the grammatical and collocational patterns that emerge for the word. Comparing left and right concordance lines of the same word whets students' appetites, and they are soon gripped by evolving patterns of *collocation*—that is, the tendency of words to combine with other words. The study of collocation is one of the main

**FIGURE 1**

**Extract of Concordance Lines for the Word *Made***

```
   Eighteen western governments have made a joint protest to the Burmese
 to come to London for it. Smith had made a unilateral declaration of
      I understand what you mean.” I made a list of every regret I could think
associated products similar to those made by Cooper. Before expending money
  Basso, a New York designer who has made clothes for Elizabeth Taylor and
000lb bomb. [p] The terrorists home-made device was discovered in a van just
and several hundred submissions were made either in person or in writing. [p]
also get help with interest on loans made for financing essential repairs or
 wok. This impressively solid pan is made from carbon steel with easy-care non-
      changed costs thousands. Home-made gift check whether it is genuine or
  word. Once all the words have been made, have them close their holders and
 forms of alternative treatment have made headlines. The first, based on shark
```

*Note.* Generated with the Collins Cobuild *Corpus Concordance Sampler* (n.d.).

FIGURE 2

Concordance Lines of *Made* Sorted 1L and 2L

```
    uestions. Somehow this math could be made a lot more specific, and we could b
  about the fact of whether it should be made a bit more explicit. One reason I r
cuments of the DNC that ought not to be made a matter of public record because
  , are we — have decisions already been made about the fact that this is going
 second question. The statement has been made a couple of times that parents sho
  ing that's in jeopardy, he's certainly made a, I think, a concerted effort to s
e is one area in which President Chirac made a specific point about the U.S. ro
 ho doesn't think that President Clinton made a bold move. But Chapter 1, page 1
         GOLAN: I think we as a country made a commitment to spend money on TIMS
u know now Deputy Secretary of Defense, made a key recommendation that the Defe
s though, just as the point was earlier made about the greater accessibility of
    y it. It's an eighth grade test. Ed made a good suggestion that I thought ev
 . Yes, in fact, that — in fact, I even made a suggestion for this meeting that
```

*Note.* Generated with WordSmith Tools (Scott, 1996) using data from the *Corpus of Spoken Professional American English* (2000). 1L = first word to the left; 2L = second word to the left.

applications of concordancing. Fox (1998) gives the example of *high* and *tall.* Even though they are roughly synonymous, they cannot always be used interchangeably; for example, one can say *a high building* but not *a high man.* Similarly, McCarthy (1990) gives the example of *blonde,* which is very likely to collocate with *hair* but unlikely to occur with *wallpaper* or *car.*

Stevens (1995) suggests that using concordances with students can develop cognitive and analytic skills for solving real-language problems. However, we find that learners need some training before they can make the most of concordance lines, including seeing collocational patterns. Reading a concordance line takes a little getting used to. The instinctive reaction is to try to read it in detail in the usual way, from left to right. We have found it is best to skim it initially from top to bottom, looking only

FIGURE 3

Concordance Lines of *Made* Sorted 1R and 2R

```
        GOLAN: I think we as a country made a commitment to spend money on TIMS
  lear. He believes it's important. He's made a commitment to get it done by the
 rticularly sort of concerned with this, made a commitment at the beginning of t
of the lack of effort and they now have made a commitment. But they can answer
     EINWAND: I don't think that Ed has made a compelling case to back away fro
  inced over the last hour that anyone's made a compelling case that we gain anyt
   t come to that conclusion. He has not made a conclusion of that. It's Senator
   n intelligence activity in Bosnia. We made a condition of our train-and-equip
  on who will listen with whom they have made a connection with in their freshmen
    ther participate in that process. We made a couple of determinations — sugge
   " maybe a verbatim. I don't recall who made a couple of changes in the language
 second question. The statement has been made a couple of times that parents sho
              ctions. VOICE: But we have made a deal? MYERS: We're n
    tion. And in schools where they have made a decision not to use, they shouldn
```

*Note.* Generated with WordSmith Tools (Scott, 1996) using data from the *Corpus of Spoken Professional American English* (2000). 1R = first word to the right; 2R = second word to the right.

at the central patterns and working outward from them. For example, doing this with the concordance lines for *made* in Figure 3 reveals that it collocates frequently with a *case,* a *commitment,* a *decision,* and so on.

Thompson (1995) provides some activities for practising skimming concordance lines in class and for developing strategies for guessing the general context from sample line fragments. Fox (1998) notes that "the use of concordances in the classroom is in its infancy as a language teaching technique" (p. 43), and she provides many useful examples of their application and noteworthy considerations for their use. Other ideas for using concordances in class are given by Flowerdew (1996), Johns (1997), Stevens (1991), Tribble (1997), and Tribble and Jones (1990, 1997), among others. A number of Web sites also provide online samples and sample activities (see the Appendix).

## Word Frequency Lists

Another function common to corpus software is the calculation of word frequency lists (or word lists) in any batch of texts. We find that it is important to focus on this function as it facilitates enquiry in our students. If they have learned this function, when they see a statistic from corpus linguistics, they can use the corpora available to them to compare findings across language varieties and contexts, and soon they become aware that contextual factors are paramount in analyses of corpora.

Figure 4 presents a typical activity we might do with our students. We compared the word frequencies of the following sets of data: (a) shop encounters in Ireland (8,500 words from the Limerick Corpus of Irish English (L-CIE, 2003), (b) female friends chatting (40,000 words from the L-CIE), (c) the Australian Corpus of English (1 million words of written Australian English; see *ICAME Collection of English Language Corpora,* 2000) and (d) the 10 most frequent words from the Cambridge International Corpus based on a 100,000-word sample of newspapers and magazines as presented in McCarthy (1998, pp. 122–123).

From just the first 10 words of these data sets, our students can see a divide between spoken and written language. In the spoken results, they find markers of the interactive nature of spoken English, such as *I, you, yeah* (as a response token), *like, please,* and *thanks.* Comparing the Australian written corpus results with the first 10 words from the Cambridge International Corpus, trainees find that the results are almost identical. The other important issue highlighted by this short comparison is that even though both of the first word lists are from the Irish spoken corpus (L-CIE), they are not identical. The shop data have obvious traces of context with high-frequency items, including *thanks, please,* and the discourse marker *now.*

**FIGURE 4**
**Comparison of Word Frequencies for the 10 Most Frequent Words Across Four Different Data Sets**

| Rank | Shop (L-CIE) | Friends (L-CIE) | Australian Corpus of English | Cambridge International Corpus (McCarthy, 1998) |
|---|---|---|---|---|
| | Spoken | Spoken | Written | Written |
| 1 | you | I | the | the |
| 2 | of | and | of | to |
| 3 | is | the | and | of |
| 4 | thanks | to | to | a |
| 5 | it | was | a | and |
| 6 | I | you | in | in |
| 7 | please | it | is | is |
| 8 | the | like | for | for |
| 9 | yeah | that | that | it |
| 10 | now | he | was | that |

*Note.* L-CIE = Limerick Corpus of Irish English.

## CORPUS APPLICATIONS TO THE ACQUISITION OF PEDAGOGIC PRACTICE

As mentioned above, Sternberg and Horvath (1995) present three characteristics associated with the prototypical category of *expert teacher:* (a) teaching knowledge, (b) teaching efficiency, and (c) teaching insight. Within this framework, we have structured classroom corpus tasks in our programme. We present and discuss samples of these below. We draw upon the corpora for developing learning activities that attempt to develop these three areas of expertise.

### Acquiring Teaching Knowledge

Three types of knowledge are necessary for expert teaching, according to Shulman (1987). The first is content knowledge of the subject matter to be taught. Suggestions for how this knowledge, that is, knowledge of English, can be acquired with the aid of corpora are offered in the section Corpus Applications in Raising Linguistic Awareness. The second type is pedagogic knowledge, which includes skills such as classroom management and motivational strategies (e.g., effective questions, nomination, instructions, student groupings, classroom organisation, use of teaching aids, lesson planning). Finally, "content-specific teaching knowledge" (Sternberg & Horvath, 1995, p. 11) includes applying teaching knowledge in a specific sociocultural and organisational setting. This knowledge tends to be more tacit (Freeman, 1991) and therefore more elusive to acquisition, but it is nonetheless a determining feature of a distinguishable expert teacher (Sternberg & Horvath, 1995, p. 12).

We use a series of tasks that incorporate the use of classroom corpora for advancing students' pedagogic and content-specific teaching knowledge of effective questioning strategies. Using the example shown in Figure 5, trainees start by looking at questioning patterns in our classroom corpus. They investigate the correlation between a question type and its productivity (they quickly notice, e.g., how much more productive referential questions are than *yes/no* questions). They are then asked, in Task (c), to look more broadly at the placement of

**FIGURE 5**

**Sample Material Based on the Limerick Corpus of Irish English for Raising Awareness of Pedagogic Knowledge**

a) Run concordances of questions used in the classroom corpus to determine their frequencies (*wh-* questions can be extracted by searching each of the *wh-* questions individually, and *yes/no* and intonation questions can be found by searching "?")

b) Analyse and compare the productivity of each question type by running an analysis of student responses in terms of length and quality (use up to 10 examples of each question type).

c) How does each type fit in the typical initiation-response-follow-up (IRF; Sinclair & Coulthard, 1975) classroom exchange structure? Use the KWIC[a] facility to help with your analysis.

d) Compare and contrast the place of questions in the IRF model with their place in other discourse structures in two additional registers of your choice from L-CIE.

e) Investigate how questioning integrates with other strategies, for example, nomination or gesture using both the transcriptions and video recordings in a qualitative way. Pay particular attention to the contextual and pragmatic factors at play.

f) Compare data from Subcorpus X (expert teachers) with Subcorpus Y (nonexpert teachers)[b] and comment on good and bad practice in context.

g) Transcribe part of one of your teaching practice lessons where you are eliciting from students using questions. Analyse your questioning strategies and note your reflections in your teaching journals to form the basis of a comparative discussion with your peers in the coming weeks.

[a] Key word in context. Instead of viewing only short concordance lines, students can view an extended context for each occurrence of the search term. [b] We have found it useful sometimes to use data from expert and nonexpert teachers (instead of experienced versus inexperienced teachers) so that we do not establish a belief that inexperience equates with lack of expertise, and vice versa.

*questions + response + follow-up* for each question type within Sinclair and Coulthard's (1975) initiation-response-feedback model. Task (d) asks students to compare the question patterns across nonclassroom contexts so that they see how different the structure is elsewhere. For example, in casual conversation, it would be usual for a friend to ask a question and to follow up the answer with an evaluation like *very good*. This comparison brings to light how predetermined teacher-led classroom discourse can be. Task (e) focuses the students on the broader realm of classroom management by asking them to look at the combination of strategies that are employed in questioning, such as asking the question, scanning, and then nominating. By comparing questioning patterns between expert and nonexpert teachers in Task (f), the students can discern effective and ineffective practices. Task (g) initiates a longer term reflective process in which students use their own data and reflect on their own strategies.

We have found our classroom corpus to be very useful because it allows us to conduct quantitative and qualitative analysis of almost any aspect of classroom interactions. Wegerif, Mercer, and Rojas-Drummond (1999), for example, provide excellent commentary on and description of how they have applied corpus techniques to the comparative analysis of the effectiveness of different teaching approaches in a Mexican context. They empirically examine the influence that the teacher's sociocultural approach has on the development of problem-solving skills among students. In quantitative and qualitative analyses of a transcribed video and audio corpus of classroom language, they investigate the corpus of talk from two classrooms employing different teaching methodologies over a period of 1 year in relation to the problem-solving skills the students develop during this period. Verbal strategies are isolated, allowing the researchers to uncover techniques used for the "social construction of knowledge through scaffolding the pupils' engagement in independent problem-solving and reasoning" (p. 133).

## Acquiring Teaching Efficiency

Through other activities, we attempt to engender awareness of efficiency in students. The short activity shown in Figure 6, which has instruction giving as its focus, is based on the notion of *teacher modes* (see McCarthy & Walsh, 2003), whereby teachers are said to have various modes of talk in the classroom. By assessing and increasing their awareness of these modes, teachers can improve classroom competence. Here we focus on the *instructional mode,* in which teachers are giving instructions to the students. First, we ask students to generate a word list using our classroom corpus and then to isolate all the verbs within this

**FIGURE 6**

**Sample Material Based on the Limerick Corpus of Irish English for Raising Awareness of Pedagogic Efficiency**

a) Run a word frequency list for the classroom corpus and isolate all the verbs.

b) Identify which verbs are likely to be used when the teacher is in instruction mode, and run concordances of their imperative forms to test your hypothesis.

c) Search for any other key word(s) you think may be used frequently when giving instructions, for example, *Let's, can you/we, please.*

d) Isolate three instruction-giving episodes and examine their entire contexts to comment on the language, procedures, and pacing. Find examples of redundancies or inaccuracies in the teacher's instructions, and comment on the pace of delivery.

e) Rewrite the instructions in a way that you consider to be more efficient.

list. Task (b) asks students to predict which of these verbs are used in giving instructions and to check their predictions by means of concordancing. Tasks (b) and (c) focus the students on the imperative nature of instructional talk, and Tasks (d) and (e) focus qualitatively on the need to conduct instructional episodes with precision and clarity. We find that our students begin to develop the desired reflectivity and insight from this activity because it provides them with a framework within which to measure their practice.

## Acquiring Teaching Insight

Insight is the ability to solve problems in creative and effective ways. Sternberg and Horvath (1995, p. 14) give the example of teachers using analogy to help students understand difficult concepts. Instances of successful teacher insight skills can be isolated through qualitative analysis of classroom corpora with expert teachers, for example, asking questions such as "In this lesson how does the teacher effectively explain differences in use between the various conditional structures in English? Relate your answers to the teacher presentation stage of the lesson and also to subsequent student production." Even more beneficial is the remedial self-examination of novice teachers' transcripts for parts of the lesson where they encountered difficulties not anticipated during preparation. In the example shown in Figure 7, we again use our local classroom corpus to focus on a typical classroom dilemma that all novice teachers can relate to: A student asks for a detailed lexical explanation that the teacher has not anticipated.

Tasks (a) and (b) first ask students to draw on the standard dictionary resource to find the difference between the problematic words and then

**FIGURE 7**

**Sample Material Based on the Limerick Corpus of Irish English for Raising Awareness of Pedagogic Insight**

Student: What's the difference between *collaborate* and *cooperate*?
Trainee: Well *collaborate* is generally used for something which is negative and *cooperate* is more positive.
Student: So can I say "I am cooperating with Maria on this project"? *Collaborate* would be wrong here?
Trainee: Well yes, no, mm I'm not too sure. What does the dictionary say? Let's check.

a) Use a dictionary to find the differences in meaning between these two words.

b) Use any large corpus from the electronic library to establish how these near-synonyms differ in terms of use and lexical patterns.

c) Redesign the part of the lesson in the extract above to make it more effective.

to use a corpus concordancer to compare their patterns in contexts of use. Through this activity, students see how concordancing can greatly enhance a dictionary definition by allowing many patterns of use, in many contexts, to be viewed at once. Task (c) leads students inductively back to classroom application.

## CORPUS APPLICATIONS IN RAISING LINGUISTIC AWARENESS

All teachers in an initial language teacher education programme expect to attain a high level of descriptive linguistic competence for the language they are going to teach. Gabrielatos (2002/2003) argues that if teachers "are to become more than 'skilled materials operators,' then teacher education needs to focus more consistently on research skills, as well as language analysis and its implications for ELT" (p. 3). Corpora offer great potential for developing language awareness and research skills within teacher education (see Coniam, 1997; Hunston, 1995; Kennedy, 1995). The following examples illustrate corpus activities from our teacher education program intended to develop students' understanding of word classes, register, and socioculturally conditioned choices of word classes in context.

From our experience, a grammatically tagged corpus (one where all of the items used have been labeled for their word class) is a very useful supplement to the development of critical knowledge of the English syntactic system. A useful sequence of activities is as follows:

1. Students are presented, either deductively or inductively, with the theory of word classes, including information on meaning, distribution, and inflection taken from a variety of grammar reference books.
2. They practise identifying the word classes in pedagogically designed texts.
3. They are presented with an untagged version of a text from a corpus and, in groups or individually, try to identify the word classes.
4. They check their answers against the tagged version of the same corpus, carefully examine any inconsistencies, and use them as the basis of a search for a particular word to further test their hypotheses. For example, they may examine the classification of the word *right,* which can function in different ways in different contexts.

This process develops a sense of enquiry, leading from the student's own research question to inductive exploration of a corpus as a problem-solving resource. Both the *ICAME Collection of English Language Corpora* (2000) and the *International Corpus of English: The British Component* (ICE-GB; 1998) contain a rich supply of grammatically tagged data. A tagged corpus also proves a very useful resource for the independent study of syntax whereby the tagging serves as a ready-made answer key that students can consult.

A sample activity with concordance lines, which we use to develop awareness of lexis and word classes, is shown in Figure 8. Such activities develop language awareness inductively and frequently lead students to form more research questions. Many student investigations, from our experience, lead to interesting comparisons across large-scale corpora available to students in our electronic library. Sometimes these mini–research projects initiate a line of enquiry that can lead to the research question for an undergraduate project or even a graduate thesis.

## Register-Specific Linguistic Choices

Although concordance-based searches and investigations can provide the basis for many insights into lexical patterns and profiles, there is also scope to explore grammatical patterns using a corpus. Figure 9 displays a task that focuses students on a grammatical item commonly presented in textbooks: question tags. This task also aims to develop a sense of questioning about corpus findings. Here the general aim is to show how results vary depending on the type of corpus used; these differences highlight the importance of contextual factors and of cross-checking findings.

**FIGURE 8**

**Sample Material Based on the Limerick Corpus of Irish English for Raising Lexical Awareness**

Below are concordance lines for the word *dead*.

a) Identify its different word classes from these examples.

b) Do any collocational patterns emerge from this evidence?

c) Divide the different examples into *positive* and *negative* meanings.

d) What synonyms could be used for the intensifier uses of *dead?*

e) Identify the examples of idioms based on the word *dead*. Use a corpus to find some more.

```
     by this time Pa would’ve been well dead 7:00 of course
       at a street corner and shoot you dead 8:48 seven
                     trees some of them dead a great many big ones which
  didn’t take enough ground to bury our dead
                 seven people were shot dead and an eighth
                          and Bernie is dead and he got him from thirty
  all the possums will be left up there dead and so it’s like er you
he pays this tribute to the poet you’re dead and so forth stanza four
    great height. chances are you’d be dead before you hit the ground
                              over your dead body huh
                        sounds sounds a dead bore so far
dleton Murray couldn’t compete with the dead brother and felt resentful
                         addressing her dead brother in her journal she said
       pretend it started off the guy’s dead but he’s sitting on the couch and
still living with her and said Stan was dead but then the telegram said
              you know i mean the guy’s dead but they this other
police believe that they were also shot dead by the same trio
       job under the table um and do it dead cheap
e hands are distraught winds waking the dead cymbalic reeds at the edge
g the ultimate shot in bowls either the dead draw or the trail of the
                     oh but that’s it’s dead easy once you get used to it
       an’t find it and we’re both at a dead end um
ts of ways especially as his mother was dead er
    everyone has a way of burying their dead
             three now and er he’s been dead for eleven hours
```

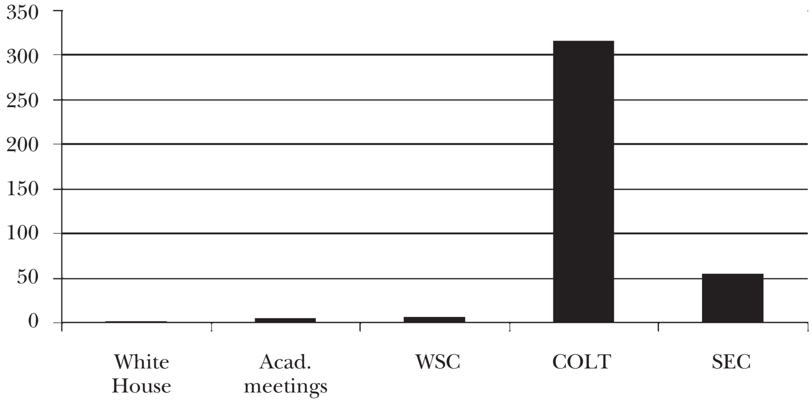
Using the example of question tags, we present findings across various corpora: the U.S. CSPAE (2000; White House press conferences and academic meetings); the Wellington Corpus of Spoken New Zealand English (WSC; see Holmes, Vine, & Johnson, n.d.); the Bergen Corpus of London Teenage Language (COLT; see University of Bergen, 2000) and the Lancaster/IBM Spoken English Corpus (SEC, n.d.). We first ask students to compare these findings from spoken corpora with those from written sources so that they see how rare question tags are in writing (in fact, they are used only in direct speech or in cases where the author addresses the reader). The spoken findings that we present show that question tags are vastly more frequent in COLT, but in Tasks (c) and (d) students see that it would be erroneous to assume that question tags are

FIGURE 9

**Sample Material Based on the Limerick Corpus of Irish English for Raising Awareness of Grammatical Patterns**

In the graph below are the results for question tags ending in *you?* from:

- two subcorpora within the Corpus of Spoken Professional American English (CSPAE): 1 million words of White House press conferences and 1 million words of academic discussions and meetings
- the Wellington Spoken Corpus (WSC) from New Zealand (1 million words)
- The Corpus of London Teenage Language (COLT) (1 million words)
- The Lancaster/IBM Spoken English Corpus (SEC) (55,000 words; these results have been normalised).[a]



Investigate the use of question tags in these and other spoken and written corpora to address the following questions:

a) Are question tags more frequent in other spoken language corpora compared to written data?

b) How are question tags used in written language?

c) Do you think question tags are used less frequently in American English?

d) What is the impact of context of use on the frequency?

e) Use any two corpora to compare findings for question tags ending in *I, he, she, it, we, they*.

f) What lessons can be learnt about care needed in selecting a corpus for your research?

[a] To make frequency results comparable, they need to be *normalised* as follows: In the SEC we found three question tags ending in *you?* This was divided by the total corpus size (55,000 words) and multiplied by 1 million, resulting in 54.5. This figure is then comparable with the other results, which are all from 1-million-word corpora.

a British phenomenon. The fact that the U.S. data came from more formal contexts than the British data affects the results. Tasks (e) and (f) focus on the need to compare data across corpora and to consider the effect of the context of the different corpora.

## Sociocultural Grammatical Choices

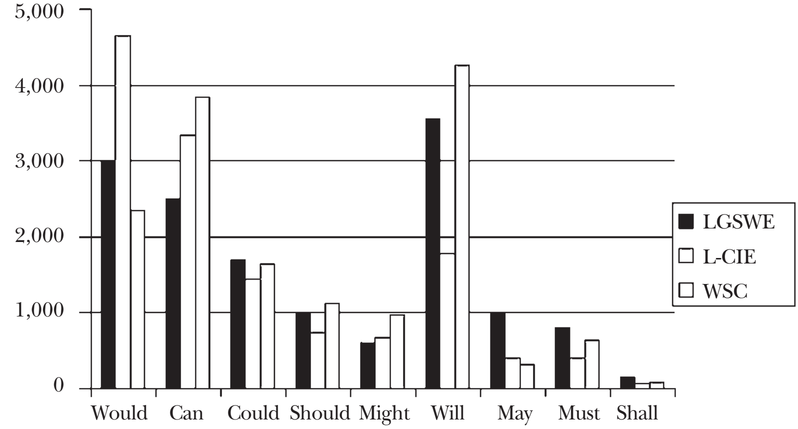
As we have stated, a central issue for the evolution of corpus use in English language classrooms around the world is the development of a critical awareness of what corpus findings represent. As illustrated above, structured corpus tasks can promote enquiry in novice teachers so that they do not readily accept corpus findings as absolutes. We feel strongly that the scrutinising of corpus findings, especially those from large-scale corpora, needs to be given overt attention. In particular, we stress the need to consider the sociocultural factors from which corpus data come, as these factors reveal much about how language is pragmatically sensitive to context. In this section, we give practical illustrations of how corpora can be of benefit in raising awareness of the sociocultural diversities that often underlie language use.

In one teacher education activity, we compare the frequency of modal verbs in British and American English, presented in the LGSWE (Biber et al., 1999), with Irish English in the L-CIE and New Zealand English in the WSC (Figure 10). One of the noticeable differences is the high occurrence of *would* in the Irish data.

The high frequency of *would* in Irish English corresponds to uses of *would* that go beyond the canonical characterisations in standard English. However, the Irish data also include *would* as a hedge consistent

**FIGURE 10**

**Distribution of Modal Verbs Across Three Corpora (per Million Words)**



*Note.* LGSWE = Longman Grammar of Spoken and Written English; L-CIE = Limerick Corpus of Irish English; WSC = Wellington Corpus of Spoken New Zealand English.

with other varieties of English; for example, in this extract from an encounter between a teacher educator and a novice teacher (following a teaching practice observation), the teacher educator uses *would* in a convergent sense instead of making a more direct statement like *You should have allowed them to work through it all* (for further discussion, see Farr & O'Keeffe, 2002):

Trainer: Do you think it would have been possible at all to just leave them work through them all? . . .
Trainee: I would say so.
Trainer: Mm.
Trainee: Given your time I would say so.
Trainer: Umhum.

However, *would* is also used in Irish English where hedging might not be expected in many other varieties of English, and its use is thus central to the sociocultural level of the interaction. Irish speakers appear to be very tentative, far beyond the demands of the interaction itself, even in situations where the propositional content of the utterance is unquestionable. For example, in the extract below from an Irish radio call-in show, a caller hedges about her hair colour, which was black but is now brown:

Caller: . . . I would have had black hair you know my hair would be brownish now . . .
Presenter: Right.

In another example, two friends reminisce, and *would* is again used for a factual statement about the location of a shop (*Swamp* refers to a chain of clothing stores):

Speaker 1: Where was it?
Speaker 2: Upper William Street. William Street. Across the road from ah. What's the name of it? Coffee place. Coffee. It would be across the road from say *Swamp* now. She used to take me in there and I used to get to drink coffee. I used to love it.

This analysis of local language use in contrast to British/American use allowed us to discover and explore a layer of tentativeness in Irish interactions, in which downtoning of indisputable facts appears to be a sociocultural norm. It is not unreasonable to expect advanced English learners, particularly as they become more proficient, to become better at recognizing such sociocultural nuances in the language they hear if teachers help them develop an interest in and ability to work with naturally occurring data. However, Rundell (1997) raises a very pertinent, broadly related question: whether imposing the "idiosyncratic linguistic features of one specific dialect of English is really an appropriate model for a majority of learners" (p. 97). In reply to his own scepticism, he points to the importance of recognizing that the specific ways in which people encode meaning reflect deeply embedded cultural characteristics. We argue that initial teacher education programs must address this level of language variation and that prospective teachers need to learn cross-corpora comparison skills in order to facilitate critical investigation of the transferability and application of corpus findings to the broader sociocultural context of their learners.

## PRACTICAL ISSUES IN USING CORPORA IN TEACHER EDUCATION

Developing these and other activities over the past years has raised a number of practical issues, including the pros and cons of building versus buying a corpus; the inclusion of spoken versus written data; the choice of small versus large corpora; the inclusion of native speaker, nonnative speaker, or learner language; and the use of paper versus online corpus work for the students.

### Build Versus Buy

Many corpora are now commercially available, and some can even be purchased for under U.S.$100. As we have illustrated, having a wide variety of corpora allows for more in-depth investigation across variables such as social context and language variety. However, in some cases, the varieties or contexts one wishes to include may never have been compiled into a corpus (e.g., as was the case for us with Irish English; see also Aston, 1997; Maia, 1997), and the only solution is to build one's own corpus.

Much has been written on the principles of corpus design (see Biber et al., 1998; Crowdy, 1993, 1994; Hunston, 2002; McCarthy, 1998); however, the serious resource implications of building a corpus, especially a spoken corpus, are worth emphasizing. As an example, in building L-CIE, a 1-million-word spoken corpus, the following core costs needed to be budgeted carefully:

1. *collection of data:* Individuals need to be paid to record the data. One hour of recording includes 10,000–15,000 words (depending on the type of talk). We therefore needed to record more than 100 hours of material to ensure that we would get 1 million words. The most cost-effective means of collecting data that we have found is to pay a set

price per tape, for example, $30 per 1-hour tape, rather than costing the person-hours involved in the collection of 1 hour of data.

2. *transcription of data:* The data then need to be transcribed. The cost of transcription, which depends on the level of detail desired, was at a minimum $150 per hour of tape (i.e., around $15,000 for 1 million words).
3. *corpus research associate:* This person is responsible for the day-to-day maintenance and building of the corpus. Duties may include managing the collection and transcription of data, ensuring that recordings cover the desired distribution of contexts, cataloguing cassettes, making a database of header information, and filing speaker information and consent forms for all spoken data collected (consent forms serve as legal contracts in which the individuals who are recorded stipulate that their conversations may be used for research and possibly in the development of pedagogical materials). Ideally, this person should work full-time for the first year in the collection of a 1-million-word spoken corpus and remain part-time thereafter to maintain and update the corpus.

Written corpora are easier to compile because they do not involve recording and transcription (though if the original texts are not electronic, the time and cost of scanning must be factored in). At the same time, written corpora have the added concern of copyrights. In sum, a corpus compilation project—whether spoken or written—is not to be undertaken without considerable planning and financial resources.

## Spoken Versus Written

In general, because of the availability of data in electronic form, a written corpus is much easier to assemble than a spoken one; therefore, more written corpora exist. McCarthy (1998) accounts for the dearth of spoken corpora in light of costs (as discussed above), access to appropriate and representative speech data situations, quality of recording, time involved in transcription, and difficult decisions in relation to the level of detail to include in transcription, among other factors. However, we would argue that the efforts and resources are justifiable on the grounds of the need to reassess language interpretation and pedagogy to account for spoken as well as written norms. Some of our sample tasks have highlighted some of the many differences between findings from written versus spoken corpora, and, indeed, there are many differences within spoken corpora depending on the context and variety.

It is crucial for students to be in a position to compare corpus findings across spoken versus written corpora from as many varieties as possible.

Too often classroom descriptions of the English language are based on written norms. For this reason alone, the effort of assembling a spoken corpus is worth making. A small, specialised corpus can be assembled at a relatively low cost. For example, our classroom corpus comes from recorded data that teachers and students have donated and that we have transcribed ourselves. Though the corpus only amounts to under 100,000 words, it is rich in spoken data from our local context.

## Small Versus Large

Whether to use a large, generalized corpus or a small, specialized corpus depends on the teacher educator's particular needs. Fox (1998) remarks that "a corpus is nothing more nor less than a collection of texts input into a computer, and the number of texts will depend upon the uses that will be made of the corpus" (p. 25). To examine a relatively infrequent word and investigate generality of lexical use, a large, representative corpus is necessary so that adequate occurrences can be found from which to draw some conclusions about typical features (see, e.g., Coxhead, 2000). If, on the other hand, the object of enquiry is a word or structure that is quite common, smaller corpora may suffice, and the smaller they are, the easier they are to handle and exploit. Also, as Tribble (1997) suggests, a small corpus may be necessary if a specialized language register is involved.

Small corpora are useful for training students in corpus techniques and methods, and they often allow the user to access contextual or pragmatic information about the spoken or written text. In addition, their limits are clearer, as they cannot claim to represent an entire language, and they therefore discourage the user from overgeneralizing. Aston (1997) makes an interesting and very practical distinction between the usefulness of small and large corpora: A large corpus is necessary for developing references, but for data-driven learning (Johns, 1991) in the classroom, where the aims and needs are much more specific and localized, the smaller corpora are as good if not better. Even linguists who have traditionally favored large representative corpora exclusively now recognize the place of smaller data collections (Tribble, 1997). Of course, another advantage for the teacher educator is that such corpora are cheaper and easier to construct or buy.

At one time, a 1-million-word corpus was considered large. Today, the Bank of English (Collins Cobuild, n.d.) has more than 500 million words. What constitutes a large or a small corpus today depends on whether one is referring to a spoken or a written corpus. In very general terms we adhere to the following guidelines: For spoken corpora, anything over 1 million words is moving into the larger range; for written corpora, anything below 5 million words is quite small. However, it is often the design of the corpus as opposed to its size that determines its suitability; for example, a corpus containing only highly technical engineering language will be largely inappropriate for novice language teachers wanting to investigate the vocabulary of everyday casual conversation. Therefore, although size is an issue, it should be considered hand-in-hand with design appropriate to the long- and short-term pedagogic needs of the students. (For a full discussion of size and diversity in corpus design, see Biber et al., 1998, 1999; Coxhead, 2000; Hunston, 2002; McCarthy, 1998; Sinclair, 1991; Thomas & Short, 1996.)

## Native Speaker Versus Nonnative Speaker and Learner Corpora

The corpus developer also has to decide on the varieties of English that should be included. Prodromou (1997a), among others, raises the possibility that corpora of native speaker language may present problems for nonnative teachers. He asks, "What about the non-native speaker teacher, faced with varieties of English and cultures he or she can, by definition, never master, never own?" (p. 5). (For further discussion of native speaker ownership of the English language, see Flowerdew, 2000; Graddol, 1999; Nero, 2000; Seidlhofer, 2001; Warschauer, 2000.) One answer to this question is to have more corpora of English spoken in contexts where English is a lingua franca, such as the one described by Mauranen in this issue. Seidlhofer (1999, 2001) details a corpus development project called the Vienna-Oxford International Corpus of English (VOICE), which aims to collect approximately half a million words of spoken data from speakers who make use of ELF. Such corpora will facilitate the profiling of ELF as a robust variety that is independent of English as a native language. VOICE may, according to Seidlhofer (2001), establish "something like an index of communicative redundancy" (p. 147).

Learner corpora are collections of texts produced by writers or speakers while they are learners. Granger (1998) advances theoretical and practical arguments for the place of learner corpora in the language classroom for studying phenomena such as interlanguage, fossilization, patterns of error, and cross-linguistic similarities and differences. Biber and Reppen (1998), Granger and Tribble (1998), and Milton (1998) outline useful procedures for using corpora as a supplementary tool for language learners, whereby students compare and analyse native speaker and learner data as a means of improving their language. Future teacher education may benefit from a recent large-scale international corpus project focusing on the written English of learners from many different

L1 backgrounds—the International Corpus of Learner English (ICLE; see Granger, 1996, 1998, 1999; Granger, Hung, & Petch-Tyson, 2002). In 1995, a corpus of spoken learner English, The Louvain International Database of Spoken English Interlanguage, was set up to complement the ICLE project (see De Cock, 1998a, 1998b, 2000). Reder, Harris, and Setzler (this issue) describe a multimedia corpus of low-level learner language, with applications for second language acquisition studies as well as teacher education.

## Handouts Versus Hands On

A very practical but important decision to make when using corpus evidence for pedagogic purposes is whether to prepare and print out the data for students in class or to give students access to the data on the computer. Of course, the latter assumes the ready availability of adequate levels of technology and support. In institutions where technological support may be a concern, students may be able to use the many online self-instructional options available (see the Appendix).

Leech (1997) outlines the advantages of both the paper-based and the computer-based approaches as follows: Prepared printouts, which allow wider access to the data by more students, are most effective in lowering the affective filter of technophobic students and save class time as the teacher does the preliminary work prior to the lesson. On the other hand, using the computer in class can promote a more learner-centred approach, provides an open-ended supply of data, and allows for more tailored and customised learning; it also teaches strategies for learning with corpora beyond the classroom. Johns (1991), in describing the data-driven approach, strongly advocates the hands-on use of corpora by students because it makes the whole experience the epitome of induction. An additional argument for having students engage in concordancing is that it aims to give them control over their learning and build their competence by giving them access to the facts of linguistic performance (see Stevens, 1995). If practical reasons make it impossible for students to use computers themselves, Willis (1998) outlines at length the procedures that can be adopted for the use of paper-based concordances in the classroom.

Educators who are familiar with inductive instruction will appreciate its effectiveness but will also recognise the increased time investment required. In shorter teacher education courses, already under time pressure, inductive instruction may not be a luxury one can afford. In our teacher education programmes, we have balanced both approaches and have found that starting with printouts and working up to computer use promotes a more progressive, inductive approach, which students tend to prefer. They need to understand the theoretical and practical applications before they become sidetracked or overwhelmed by the technology. Furthermore, using both instructional modes in teacher education programmes provides a richer variety of experience and presents students with more options for their own future teaching.

## CONCLUSION

In this article we have outlined practical and theoretical aspects related to the integration of language corpora as an electronic resource in initial teacher education. Without doubt, language corpora will continue to develop as an influence in language pedagogy. Many instructional materials, including software, dictionaries, and grammars, have been corpus based in recent years. For this reason alone, all teachers should learn about corpora. Beyond this, however, the more teachers know about corpora and how to use them, the more they will be empowered to (a) evaluate corpus-based materials more objectively and (b) question publishers and academics about specific details of the corpora they use. Native and nonnative teachers need to learn to manipulate language corpora for their own pedagogic ends and to evaluate findings that are presented as facts. Corpus-using teachers will be better placed for the sociocultural mediation and pedagogic recontextualization of these resources and findings in their language classrooms of the future. At the same time, much work remains for teacher educators in further developing methodological principles for the use of corpora and empirically evaluating corpus-based approaches and their effect on learning.

### ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers for detailed comments on an initial draft of this article. We also thank Susan Conrad, Gwyneth Fox, and Michael McCarthy for their feedback and encouragement on earlier versions.

### THE AUTHORS

Anne O'Keeffe is course leader in EFL/TEFL at Mary Immaculate College, University of Limerick. Her research centres on small spoken language corpora and particularly on how they can be used to explore sociocultural nuance. Along with her colleagues, she is involved in building the Limerick Corpus of Irish English, and she has recently completed a PhD on the discourse of radio call-in.

Fiona Farr is director for the MA in English language teaching at the University of Limerick. Her research interest is in the application of spoken language corpora and

language varieties. She is part of a research group building the Limerick Corpus of Irish English and is completing a PhD on the discourse of language teacher education.

## REFERENCES

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. (1997). Small and large corpora in language learning. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC '97: Practical applications in language corpora* (pp. 51–62). Lodz, Poland: Lodz University Press.

Barnes, A., & Murray, L. (1999). Developing the pedagogical ICT competence of modern foreign languages teacher trainees. Situation: All change and plus ça change. *Journal of IT for Teacher Education, 8,* 165–180.

Baynham, M. (1991). Speech reporting as discourse strategy: Some issues of acquisition and use. *Australian Review of Applied Linguistics, 14,* 87–114.

Baynham, M. (1996). Direct speech: What's it doing in non-narrative discourse? *Journal of Pragmatics, 25,* 61–81.

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.* Essex, England: Longman.

Biber, D., & Reppen, R. (1998). Comparing native and learner perspectives on English grammar: A study of complement clauses. In S. Granger (Ed.), *Learner English on computer* (pp. 145–158). London: Longman.

Boxer, D., & Pickering L. (1995). Problems in the presentation of speech acts in ELT materials: The case of complaints. *ELT Journal, 49,* 99–158.

Carter, R. (1998). Orders of reality: CANCODE, communication and culture. *ELT Journal, 52,* 43–56.

Carter, R., & McCarthy, M. J. (1995). Grammar and the spoken language. *Applied Linguistics, 16,* 141–58.

Chapelle, C. A. (2001). ELT, technology and change. In A. Pulverness (Ed.), *IATEFL 2001 Brighton conference selections* (pp. 9–18). Kent, England: International Association of Teachers of English as a Foreign Language.

Collins Cobuild. (n.d.). *Corpus concordance sampler.* Retrieved May 30, 2003, from http://titania.cobuild.collins.co.uk/form.html

Coniam, D. (1997). A practical introduction to corpora in a teacher training language awareness programme. *Language Awareness, 6,* 199–207.

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 Carter. *ELT Journal, 52,* 57–63.

*Corpus of spoken professional American English* [CD-ROM]. (2000). Houston, TX: Athelstan. (Available from http://www.athel.com/cspa.html)

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34,* 213–238.

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing, 8,* 259–265.

Crowdy, S. (1994). Spoken corpus transcription. *Literary and Linguistic Computing, 9,* 25–28.

Cummins, J. (2000). Academic language learning, transformative pedagogy, and information technology: Towards a critical balance. *TESOL Quarterly, 34,* 537–547.

De Cock, S. (1998a). Corpora of learner speech and writing and ELT. In A. Usoniene (Ed.), *Proceedings from the International Conference on Germanic and Baltic Linguistic Studies and Translation* (pp. 56–66). Vilnius, Lithuania: Homo Liber.

De Cock, S. (1998b). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics, 3,* 59–80.

De Cock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing. In C. Mair & M. Hundt (Eds.), *Corpus linguistics and linguistic theory: Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau 1999* (pp. 51–68). Amsterdam: Rodopi.

Doering, A., & Beach, R. (2002). Preservice English teachers acquiring literacy practices through technology tools. *Language Learning and Technology, 6,* 127–146. Retrieved June 2, 2003, from http://llt.msu.edu/vol6num3/doering/default.html

Egbert, J., Paulus, T. M., & Nakamichi, Y. (2002). The impact of CALL instruction on classroom computer use: A foundation for rethinking technology in teacher education. *Language Learning and Technology, 6,* 108–126. Retrieved June 2, 2003, from http://llt.msu.edu/vol6num3/pdf/egbert.pdf

Farr, F. (2002). Classroom interrogations—how productive? *Teacher Trainer, 16,* 19–23.

Farr, F., & O'Keeffe, A. (2002). *Would* as a hedging device in an Irish context: An intra-varietal comparison of institutionalised spoken interaction. In R. Reppen, S. Fitzpatrick, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 25–48). Amsterdam: Benjamins.

Flowerdew, J. (1996). Concordancing in language learning. In M. Pennington (Ed.), *The power of CALL* (pp. 97–113). Houston, TX: Athelstan.

Flowerdew, J. (2000). Discourse community, legitimate peripheral participation, and the nonnative-English-speaking scholar. *TESOL Quarterly, 34,* 127–150.

Fox, G. (1998). Using corpus data in the classroom. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 25–43). Cambridge: Cambridge University Press.

Freeman, D. (1991). "To make the tacit explicit": Teacher education, emerging discourses, and conceptions of teaching. *Teaching and Teacher Education, 7,* 439–454.

Gabrielatos, C. (2002/2003). Grammar, grammars and intuitions in ELT: A second opinion. *IATEFL Issues, 170,* 2–3.

Graddol, D. (1999). The decline of the native speaker. *AILA Review, 13,* 57–68.

Granger, S. (1996). Learner English around the world. In S. Greenbaum (Ed.), *Comparing English world-wide* (pp. 13–24). Oxford: Clarendon Press.

Granger, S. (Ed.). (1998). *Learner English on computer.* London: Longman.

Granger, S. (1999). Use of tenses by advanced EFL learners: Evidence from an error-tagged computer corpus. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora—studies in honour of Stig Johansson* (pp. 191–202). Amsterdam: Rodopi.

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

Granger, S., & Tribble, C. (1998). Learner corpus data in the foreign language classroom: Form-focused instruction and data-driven learning. In S. Granger (Ed.), *Learner English on computer* (pp. 199–209). London: Longman.

Holmes, J. (1988). Doubt and certainty in ESL textbooks. *Applied Linguistics, 9,* 21–44.

Holmes, J., Vine, B., & Johnson, G. (n.d.). *The Wellington corpus of spoken New Zealand English.* Retrieved May 30, 2003, from http://www.vuw.ac.nz/lals/wgtn_crps_spkn_NZE.htm

Hughes, R., & McCarthy, M. J. (1998). From sentence to discourse: Discourse grammar and English language teaching. *TESOL Quarterly, 32,* 263–287.

Hunston, S. (1995). Grammar in teacher education: The role of a corpus. *Language Awareness, 4,* 15–31.

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

*ICAME collection of English language corpora* [CD-ROM]. (2000). Bergen, Norway: Norwegian Computing Centre for the Humanities. (Available from http://www.hit.uib.no/icame.html)

*International corpus of English: The British component (ICE-GB)* [CD-ROM]. (1998). (Available from http://www.ucl.ac.uk/english-usage/ice-gb/)

Johns, T. (1991). Should you be persuaded—two samples of data driven learning materials. *English Language Research Journal, 4,* 1–16.

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). London: Longman.

Kennedy, C. (1995). Wish you were here: "Little" texts and language awareness. *Language Awareness, 4,* 161–172.

Kettermann, B. (1995). Concordancing in English language teaching. *TELL and CALL, 4,* 4–15.

*Lancaster/IBM Spoken English Corpus (SEC) tag-set.* (n.d.). Retrieved May 30, 2003, from http://www.comp.leeds.ac.uk/amalgam/tagsets/sec.html

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.

*Limerick Corpus of Irish English.* (2003). Retrieved July 23, 2003, from http://www.mic.ul.ie/lcie

Maia, B. (1997). Do-it-yourself corpora . . . with a little help from your friends. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC '97: Practical applications in language corpora* (pp. 403–410). Lodz, Poland: Lodz University Press.

McCarthy, M. J. (1990). *Vocabulary.* Oxford: Oxford University Press.

McCarthy, M. J. (1998). *Spoken language and applied linguistics.* Cambridge: Cambridge University Press.

McCarthy, M. J. (2001). *Issues in applied linguistics.* Cambridge: Cambridge University Press.

McCarthy, M. J., & Walsh, S. (2003). Discourse. In D. Nunan (Ed.), *Classroom-based language teaching methodology* (pp. 173–195). New York: McGraw-Hill

Meskill, C. J., Mossop, S., DiAngelo, R., & Pasquale, K. (2002). Expert and novice teachers talking technology: Precepts, concepts and misconcepts. *Language Learning and Technology, 6,* 46–57. Retrieved June 2, 2003, from http://llt.msu.edu/vol6num3/meskill/default.html

Milton, J. (1998). Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In S. Granger (Ed.), *Learner English on computer* (pp. 186–198). London: Longman.

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

Murray, L. (1998). CALL and Web training with teacher self-empowerment: A departmental and long-term approach. *Computers and Education, 31,*17–23.

Nero, S. J. (2000). The changing faces of English: A Caribbean perspective. *TESOL Quarterly, 34,* 483–510.

Owen, C. (1996). Do concordances need to be consulted? *ELT Journal, 50,* 219–224.

Pennington, M. (2001). Writing minds and talking fingers: Doing literacy in an electronic age. In *CALL in the 21st century* [CD-ROM]. Kent, England: International Association of Teachers of English as a Foreign Language.

Prodromou, L. (1997a). Corpora: The real thing? *English Teaching Professional, 5,* 2–6.

Prodromou, L. (1997b). From corpus to octopus. *IATEFL Newsletter, 137,* 18–21.

Rundell, M. (1997). Understatement and indirectness in English: From corpus evidence to classroom practice. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *PALC '97: Practical applications in language corpora* (pp. 90–98). Lodz, Poland: Lodz University Press.

Scott, M. (1996). WordSmith Tools (Version 3.0) [Computer software]. Oxford: Oxford University Press. (Available from http://www.liv.ac.uk/~ms2928/)

Seidlhofer, B. (1999). Double standards: teacher education in the expanding circle. *World Englishes, 18,* 233–45.

Seidlhofer, B. (2001). Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics, 11,* 133–158.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 19,* 4–14.

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

Sinclair, J., & Coulthard, M. (1975*). Towards an analysis of discourse: The English used by teachers and pupils.* Oxford: Oxford University Press.

Sternberg, R. J., & Horvath, J. A. (1995). A prototype view of expert teaching. *Educational Researcher, 24,* 9–17.

Stevens, V. (1991). Classroom concordancing: Vocabulary materials derived from relevant, authentic text. *ESP Journal, 10,* 35–46.

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

Svartvik, J. (1991). What can real spoken data teach teachers of English? In J. A. Alatis (Ed.), *Linguistics and language pedagogy: The state of the art* (pp. 555–565). Washington, DC: Georgetown University Press.

Tammelin, M. (2001). Empowering the language teacher through ICT training and media education: Case HSEBA. In *CALL in the 21st century* [CD-ROM]. Kent, England: International Association of Teachers of English as a Foreign Language.

Thomas, J., & Short, M. (Eds.). (1996). *Using corpora for language research.* New York: Longman.

Thompson, G. (1995). *Collins Cobuild concordance sampler 3: Reporting.* London: HarperCollins.

Tribble, C. (1997). Improvising corpora for ELT: Quick and dirty ways of developing corpora for language teaching. In B. Lewandowska-Tomaszczyk & J. Melia (Eds.), *PALC '97: Practical applications in language corpora* (pp. 106–117). Lodz, Poland: Lodz University Press. Retrieved June 2, 2003, from http://web.bham.ac.uk/johnstf/palc.htm

Tribble, C. (2000). Practical uses of 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, England: International Association of Teachers of English as a Foreign Language.

Tribble, C., & Jones, G. (1990). *Concordances in the classroom.* London: Longman.

Tribble, C., & Jones, G. (1997). *Concordances in the classroom: Using corpora in language education.* Houston, TX: Athelstan.

University of Bergen. (2000). *Bergen corpus of London teenage language.* Retrieved May 30, 2003, from http://www.hit.uib.no/colt/

Warschauer, M. (2000). The changing global economy and the future of English teaching. *TESOL Quarterly, 34,* 511–535.

Wegerif, R., Mercer, N., & Rojas-Drummond, S. (1999). Language for the social construction of knowledge: Comparing classroom talk in Mexican preschools. *Language and Education, 13,* 133–150.

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics, 21,* 3–25.

Willis, J. (1998). Concordances in the classroom without a computer: Assembling and exploiting concordances of common words. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 44–66). Cambridge: Cambridge University Press.

# APPENDIX

## Web Sites and Software for Corpus Linguistics

### Corpora

*American National Corpus*

http://americannationalcorpus.org/

*Australian Corpus of English*

Available in *ICAME Collection of English Language Corpora* (2000).

Peters, P., & Smith, A. (n.d.). *Manual of information to accompany the Australian Corpus of English (ACE).* http://khnt.hit.uib.no/icame/manuals/ace/INDEX.HTM

*Bergen Corpus of London Teenage Language (COLT)*

http://www.hit.uib.no/colt/; available in *ICAME Collection of English Language Corpora*

*British National Corpus*

http://info.ox.ac.uk/bnc/

*Corpus of Spoken Professional American English (CSPAE)*

http://www.athel.com/cspa.html

*ICAME Collection of English Language Corpora*

http://www.hit.uib.no/icame.html

*English-Norwegian Parallel Corpus*

http://www.hd.uib.no/enpc.html

*English-Swedish Parallel Corpus*

http://www.englund.lu.se/research/corpus/corpus/webtexts.html

*International Corpus of English: The British Component (ICE-GB)*

http://www.ucl.ac.uk/english-usage/ice-gb/

*International Corpus of Learner English*

http://www.abo.fi/fak/hf/enge/icle.htm

Granger, S. (n.d.). *International corpus of learner English—ICLE.* http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm

*IViE Corpus*

Grabe, E., & Slater, A. (2002). *The prosodically transcribed IViE corpus on-line.* http://www.phon.ox.ac.uk/~esther/ivyweb/search_trans.html

*Lancaster/IBM Spoken Corpus of English (SEC)*

http://www.comp.leeds.ac.uk/amalgam/tagsets/sec.html; available in *ICAME Collection of English Language Corpora*

*Limerick Corpus of Irish English (L-CIE)*

http://www.mic.ul.ie/lcie

*Longman Spoken American Corpus*

http://www.longman-elt.com/dictionaries/corpus/lcaspoke.html

*Longman Learners' Corpus*

http://www.longman-elt.com/dictionaries/corpus/lclearn.html

*Louvain International Database of Spoken English Interlanguage (LINDSEI)*

http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Lindsei/lindsei.htm

*Michigan Corpus of Academic Spoken English (MICASE)*

http://www.hti.umich.edu/m/micase/

*Tractor*

http://www.tractor.de/faq.htm

*Wellington Corpus of Spoken New Zealand English*

http://www.vuw.ac.nz/lals/wgtn_crps_spkn_NZE.htm; available in *ICAME Collection of English Language Corpora*

## Concordancing

*Software*

Collins Cobuild. Corpus concordance sampler. (n.d.). http://titania.cobuild.collins.co.uk/form.html

Conc (Version 1.80). (2000). Dallas, TX: SIL International. (Available from http://www.sil.org/computing/conc/)

MonoConc Pro. (2000). Houston, TX: Athelstan. (Available from http://www.athel.com/mono.html)

Multiconcord. (1998). Hants, England: CFL Software Development. (Available from http://web.bham.ac.uk/johnstf/lingua.htm)

Scott, M. (1996). WordSmith Tools (Version 3.0). Oxford: Oxford University Press. (Available from http://www.liv.ac.uk/~ms2928/ and http://www4.oup.co.uk/isbn/0-19-459286-3)

UltraFind. (1997). London: Ultradesign Technologies. (Available from http://www.ultradesign.com/ultrafind/ultrafind.html)

Watt, R. J. C. (2002). Concordance (Version 3.0). (Available from http://www.rjcw.freeserve.co.uk/)

*Suggestions for Classroom Use of Concordancing*

*Hot Potatoes home page.* (n.d.). http://web.uvic.ca/hrd/halfbaked

Ruthven-Stuart, P. (2000). *How to use concordances in teaching English: Some suggestions.* http://www.nsknet.or.jp/~peterr-s/concordancing/usingconcs.html

Ruthven-Stuart, P. (2000). *Online concordance quizzes.* http://www.nsknet.or.jp/~peterr-s/concordancing/onlineconcquiz/online_conc_quizzes.html

## Corpus Linguistics

*Corpus linguistics pages.* (1998). http://info.ox.ac.uk/bnc/corpora.html#Corpus

Barlow, M. (n.d.). *Parallel corpora.* http://www.ruf.rice.edu/~barlow/para.html

*The Fifth Teaching and Language Corpora Conference.* (2003). http://www.sslmit.unibo.it/talc/

*The Tuscan Word Centre.* (2003). http://www.twc.it/

*University of Birmingham Centre for Corpus Linguistics.* http://clg1.bham.ac.uk/

## Corpus Linguistics Tutorials

Aston University, School of Languages and European Studies. (2000). *Introduction to text analysis.* http://www.les.aston.ac.uk/txtintro.html

Ball, C. N. (1997). *Tutorial: Concordances and corpora.* http://www.georgetown.edu/cball/corpora/tutorial.html

University of Essex, W-3 Corpora Project. (2000). *The W3-corpora site.* http://clwww.essex.ac.uk/w3c/