(This is a reprint of O'Keeffe and Farr (2003) "Using Language Corpora in Language Teacher Education: pedagogic, linguistic and cultural insights". TESOL Quarterly, 37(3), 389-418.)

Citation:

- O'Keeffe, A. and Farr, F. (2003) "Using Language Corpora in Language Teacher Education: pedagogic, linguistic and cultural insights". TESOL Quarterly, 37(3), 389-418.
- O'Keeffe, A. and Farr, F. (2012) "Using Language Corpora in Language Teacher Education: pedagogic, linguistic and cultural insights". D. Biber and R. Reppen, (Eds) Corpus Linguistics (Volume 4): Methods and Applications. London: Sage, 335- 365.

Using Language Corpora in Initial Teacher Education: Pedagogic Issues and Practical Applications[1]

ANNE O'KEEFFE and FIONA FARR
University of Limerick

Recent years have seen a vast increase in the amount of materials such as dictionaries and grammars which are 'corpus-based' and it is difficult to dispute the contribution of corpus linguistics to English language description. There have also been many developments in the use of corpora in the classroom in data-driven learning (Johns 1991). However, this rapid development in new technology has not been matched in teacher education provision. This paper aims to make a case for the inclusion of corpus linguistics in initial language teacher education. We argue that apart from enhancing teachers' research skills and language awareness, language corpora can aid pedagogic awareness through the use of in-house classroom corpora, and raise sociocultural awareness through the comparative investigation of large-scale commercially available corpora. We also look at the theoretical and practical considerations that need to be taken into account in the integration of language corpora in a teacher education program. We conclude that it is vital, given the pervasive nature of language corpora and their findings (especially in published materials), that future teachers have the critical evaluative skills to discern and mediate for the needs of their learners.

Applied linguists researching the field of technology and education have, for some time, referred to the technological and digital global economy in which we live (for example, Cummins, 2000; Warschauer, 2000; Chapelle, 2001, among others). Literacy is no longer just about reading and writing. Society now demands 'multi-literacies' (Warschauer, 2000), which include a high proficiency in digital and on-line competencies (see also Pennington, 2001 and Doering & Beach, 2002). Consequently, language teacher educators have a fundamental obligation to educate teachers in this

[1] The authors are grateful to the anonymous reviewers for detailed comments on an initial draft of this article. Thanks also to Susan Conrad, Gwyneth Fox and Michael McCarthy for their feedback on and encouragement with earlier versions. Any remaining inconsistencies or inaccuracies are the authors' responsibility.

respect. The initiation and implementation of many national educational policies and directives targeted at teacher education institutions are testament to such an obligation. Murray (1998) and Barnes and Murray (1999) discuss in-service and initial information and communication technology (hereafter ICT) teacher education provision for foreign language teaching in this context and suggest that 'ICT can no longer be an added extra but rather an intrinsic part of a teacher's methodological repertoire' and conclude that 'this transition must occur in the initial teacher training period to have the greatest effect' (Barnes & Murray, 1999, 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 integration (see for example, Egbert, Paulus, & Nakamichi 2002; Meskill, Mossop, DiAngelo, & Pasquale, 2002).

There are also affective benefits of successful mastery of ICT including positive attitude, increased confidence and teacher empowerment (see Egbert et al., 2002; Murray, 1998; Tammelin, 2001). Doering & Beach, (2002, p. 128) point out that 'it is primarily through active participation with technology as opposed to receiving instruction about technology that pre-service teachers learn to recognize the value of technology tools'. As with all methods, all materials and pedagogic apparatus, there are advantages and disadvantages associated with computer technology. This has not and will not detract from the need for its integration in teacher education programmes (Cummins, 2000; Chapelle, 2001) and for language teacher education programmes, this means instruction in on-line resources, a range of CALL software, and language corpora, the focus of this article.

CORPUS LINGUISTICS AND LANGUAGE TEACHING

The contribution of corpus linguistics to the description of the language we teach is difficult to dispute. 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). Being able to examine large quantities of spoken and written texts on computer has revealed language patterns and uses that had hitherto eluded intuition, and in so doing, linguists have vastly improved our dictionaries (see Fox, 1998) and grammars (see Biber, Johansson, Leech, Conrad, & Finegan, 1999 the *Longman Grammar of Spoken and Written English*, a grammar which draws on a corpus of 40 million words). In addition, numerous studies have shown us that the language presented in textbooks is often based on faulty intuition about how we 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 and finds 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 involves the 'backshift rule' for tenses in the 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 our understandings of units of fixed phrasing, collocation, and language patterning (Sinclair, 1991; Svartvik, 1991; Aston, 1995; Murison-Bowie, 1996)

The corpus debate

Svartvik (1991, p. 555) points out 'the attitude to the use of corpora in linguistic research has had its ups and downs'. Many practitioners and applied linguists point to the problems of adopting corpus-based material in the language classroom (see for example Cook, 1998; Owen, 1996; Prodromou, 1997a, 1997b; Seidlhofer, 1999; Widdowson, 2000). We stress that it is important that these issues be dealt with in initial teacher education and that all teachers who use corpora or corpus findings in

the classroom be aware of these concerns. One of the core disputes centres around the 'reality' of a corpus. Sinclair (1991, p. 6), for example, making the case for the use of 'real' language in the classroom, asserts that 'one does not study all of botany by making artificial flowers'. However, Widdowson (2000) warns that just because corpus data is 'real', we should not assume that using such data in the classroom will bring with it more 'reality'. The reality that corpus findings represent is, he argues, third rather than first person reality. He asserts that problems arise when 'partial description' of 'decontextualised language' (ibid) is used to determine language prescription for the classroom. However, one could make this case for almost all of the 'authentic' classroom materials that we use with our students. As teachers we know how to adapt materials for our students and we know how to structure tasks so that materials are tailored to our local needs. After all, it is teachers who will decide whether to use corpus materials. More importantly it is teachers who will engage in the process of recontextualising any useful findings from corpus-based description and it is teachers who will mediate between corpus-based content and the needs of the actual learners in their individual classroom contexts. To do this, teachers 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, p. 31) 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' (emphasis from original text). 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 our teacher education courses (see also Conrad 2000; Chapelle 2001). In the context of teacher education for teachers who are speakers of English as a Lingua Franca, Seidlhofer's (1999, p. 240) comments in relation to corpus linguistics: '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'. Seidlhofer's comments, we feel, are equally applicable to all

teachers. In this paper, we hope to make the case for the inclusion of corpus applications and methods in initial teacher education programmes so that teachers of the future may be in a position whereby they can decide if their learners' needs will be best served by the inclusion of language corpora either as a teacher resource or a self-access application. We detail considerations and activities from our own context of teacher education at the University of Limerick, Ireland, where we have been developing and integrating corpora in our programmes for the past six years.

CORPUS APPLICATIONS FOR THE TRAINING OF TEACHERS

Sternberg and Horvath (1995) discuss three characteristics which can be used to identify what we consider to be an 'expert' teacher. They suggest that to belong to this prototypical category, which generally, though not always, comes with experience, one must be more *knowledgeable*, more *efficient* and have better *insight* than non-experts (either experienced or inexperienced). Whether or not one accepts this paradigm, it is ultimately the responsibility of initial training courses to aim to produce teachers who have at least started their journey along the road to expertise even if limited in experience. In our case, the use of corpora for this purpose came about because the training materials that we had been using for methodological skills acquisition (that is commercially-available classroom transcripts and video recordings) have two major shortcomings: 1) they have traditionally lent themselves almost exclusively to qualitative scrutiny, the conclusions of which may sometimes be elusive to and over-subjectified by inexperienced trainees, and 2) the practices of teaching must be interpreted within their contexts of realisation, much of which is lost in their reproduction and extraction for third party analysis operating in far-removed realities. In other words, socio-cultural and environmental factors which create and cast the lesson cannot easily be captured in their entirety by non-present third parties in different educational and/or cultural surrounds. This is particularly true in our Irish context as many of the training materials available commercially are either British or American produced and often mismatch the training conditions experienced by our trainees.

We have found that the acquisition of pedagogic knowledge, efficiency and insight can be encouraged through the mediated integration of classroom corpora for trainees. However, to rectify the contextual mismatch, we have been engaged in the process of building our own English Language Teaching classroom corpus for this purpose. For example, Farr (2002) reports on a study where teachers were recorded and these classroom interactions were then transcribed to form a mini-corpus which in turn was used as the basis for analysis of the correlation between question forms and productivity in the language classroom. Our classroom corpus will ultimately include four data types: transcriptions of experienced teachers operating in different socio-cultural settings from our trainees, transcriptions of experienced teachers operating in the same socio-cultural settings, transcriptions of other trainees operating in different socio-cultural settings, and transcriptions of our trainees during their on-site teaching practice sessions.

Another area of application for corpora in language teacher education which we will look at is in raising linguistic awareness (and this is very much tied up with the *knowledge* category as detailed by Sternberg and 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, is the development of a critical awareness of what corpus findings represent. As we hope to illustrated below, corpus investigations can engender enquiry in trainee teachers so that they do not readily accept corpus findings as absolute truths.

Before we take a practical look at how pedagogic, linguistic and sociolinguistic awareness can be developed and enhanced by the use of language corpora on teacher education programmes, we will first need to cover some critical level corpus software functions.

Critical Technological Expertise for Corpus Exploration

At first corpus linguistics can seem very daunting and it is important for us not to frighten our trainees 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.

Concordaning

We always begin with concordancing as it is a core tool for analysis in corpus linguistics. It is the process involved in 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 that you have searched for) in the centre of the line. There will be 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 example below shows a sample of concordance lines for the word *made* using the COBUILD Sampler[2] (freely available online, see below for URL). It provides 40 examples based on any or all of the following corpora: British books, ephemera, radio, newspapers, magazines (26 million words); American books, ephemera and radio (9 million words) and British transcribed speech (10 million words):

FIGURE 1

Extract of concordance lines for the word *made* from the COBUILD Sampler

```
   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
```

Apart from free Internet concordancing sites, there are many software packages available commercially, most of which allow the user to go back to the original source text of any one of lines or at least provide a much larger sample if required.

[2] URLs, where available, for all corpora and software mentioned in this article are listed in Appendix A below.

A key manipulation of a concordance involves *sorting* alphabetically to the left and to the right of the node word or phrase. Let us provide an example using *Wordsmith Tools* (OUP) analysing the Corpus of Spoken Professional American English (CSPAE, a two million word corpus available to buy on CD ROM made up of academic discussions, committee meetings and White House press conferences). Let us again sample the word *made*, but this time we will present the line-samples in two different sorting formats: *Figure 2* will show it sorted to the left of the node word and *Figure 3* will present the data sorted to the right. Note that 1R and 2R refer to the first and the second word to the right of the node and 1L and 2L refer to the first and second word to the left respectively.

FIGURE 2

Concordance lines of *made* from CSPAE 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
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
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
```

FIGURE 3

Concordance lines of *made* from CSPAE 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
```

By looking to the left and to the right of a word with our trainees, we find more information about the grammatical and collocational patterns that emerge for the word. We find that comparing left and right concordance lines of the same word whets trainees appetites and they are soon gripped by evolving patterns of *collocation*.

Collocation refers to the tendency of words to combine with other words. The study of collocation is one of the main 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, we can say 'a high building' but not a 'a high man' or 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) tells us that using concordances with students can develop cognitive and analytic skills for the purpose of solving real-language problems. However, we find that there is need for some learner training before we can make the most of concordance lines. 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 only looking at the central patterns and working outward from these. For example, if you look again at the concordance lines for *made* in Figure 3 above, you will very quickly notice 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, p. 43) notes that 'the use of concordances in the classroom is in its infancy as a language teaching technique' and she provides many useful examples of their application and noteworthy considerations for their use. Other ideas for using concordances in class are found in Flowerdew (1996), Johns (1997), Stevens (1991), Tribble (1997), Tribble and Jones (1990, 1997) among others. There are also a number of very useful websites which provide online samples and sample activities (see appendix).

Word frequency lists

Another function common to corpus software is the extremely rapid calculation of word frequency lists (or wordlists) in any batch of texts. We find that it is important to focus on this function as it facilitates enquiry in our trainees. It means that 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. Here as a typical

example of something we might do with our trainees. We compared the word frequencies of the following sets of data: 1) shop encounters in Ireland (8,500 words from the Limerick Corpus of Irish English (L-CIE); 2) female friends chatting (40,000 words from L-CIE); 3) the Australian Corpus of English (one million words of written Australian English and 4) the ten most frequent words from the Cambridge International Corpus based on a 100,000 word sample of newspaper and magazines as presented in McCarthy (1998, pp. 122-123).

FIGURE 4

Comparison of word frequencies for the ten most frequent words across four different datasets

Rank	**Shop (L-CIE)**	**Friends (L-CIE)**	**ACE**	**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

Even from just the first ten words of these datasets, our trainees can see a divide between spoken and written language. In the spoken results, we find markers of the interactive nature of spoken English such as *I, you, yeah* (as a response token), *like, please, and thanks.* When we compare the Australian written corpus results with the first ten words from the Cambridge International Corpus, we find that they are almost identical. The other important issue that this short comparison highlights is that even though both of the first wordlists are from our Irish spoken corpus (L-CIE), they are not identical. The shop data has obvious traces of context with high frequency items

including *thanks*, *please* and the discourse marker *now*. A practical exercise for trainees based on frequency information will be given below.

CORPUS APPLICATIONS TO THE ACQUISITION OF PEDAGOGIC PRACTICE

Having covered some of the basic corpus software manipulations, let us return now to how corpora can be used to enrich the acquisition of pedagogic practice. As mentioned above Sternberg and Horvath (1995) present three characteristics associated with the prototypical category, of 'expert teacher': 1) teaching knowledge, 2) teaching efficiency, and 3) teaching insight. Within this framework, we have structured classroom corpus tasks on our programme. We present and discuss samples of these below.

Acquiring Teaching Knowledge

It has been suggested that there are three types of *knowledge* necessary for expert teaching (Shulman, 1987). The first is content knowledge of the subject matter to be taught. In our case this means knowing the English language and suggestions for how this can be acquired with the aid of corpora are offered in the next section 'Corpus Applications in Raising Linguistic Awareness'. The second is pedagogic knowledge. This includes skills such as classroom management and motivational strategies (e.g. using effective questions, nomination, instructions, student groupings, classroom organisation, use of teaching aids, lesson planning etc). Finally, and importantly, there exists 'content-specific teaching knowledge' (Sternberg & Horvath 1995, p. 11), which extends to include applying teaching knowledge in a specific socio-cultural and organisational setting. This tends to be more tacit (Freeman, 1991) and therefore more elusive to acquisition but is nonetheless a determining feature of a distinguishable expert teacher (Sternberg & Horvath 1995, p. 12). The following activity is an example of how classroom corpora can be used to advance pedagogic and content-specific pedagogic knowledge of effective questioning strategies. 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 how much more productive referential questions are compared to yes/no

questions for example). They are then asked, in Task c, to look more broadly at the placement of questions + response + follow-up for each question type within Sinclair and Coulthard's (1975) Initiation- Response- Feedback model. Task d asks trainees to compare this across non-classroom contexts so that they see how different the structure is across contexts, for example, in casual conversation, it would be usual for a friend to ask a question, and then to follow up the answer with an evaluation like *very good*. This brings to light how pre-determined teacher-led classroom discourse can be. Task e focuses the trainees 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 non-expert teachers in Task f, the trainees can discern effective and ineffective practices. Task g initiates a longer term reflective process where trainees will use their own data and reflect on their own strategies.

Figure 5

Sample material based on the L-CIE for 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 ten examples of each question type).

c) How does each type fit in the typical Initiation Response Follow-up (Sinclair and Coulthard 1975) classroom exchange structure? Use the KWIC[3] 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 sub-corpus X (expert teachers) with sub-corpus Y (non-expert teachers)[4] and comment on good and bad practice in context.

[3] Key word in context: instead of reading the search word in short concordance lines, an extended context for each occurrence can be viewed.

[4] It is a good idea to sometimes use data from expert and non-expert experienced teachers (instead of experienced versus inexperienced teachers) so that we do not establish a belief that inexperience equates with non-expert and vice versa.

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.

We have found our classroom corpus to be very useful since quantitative and qualitative analysis of almost any aspect of classroom interactions can be conducted. Wegerif, Mercer, and Rojas-Drummond (1999), for example, provide excellent commentary 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 socio-cultural approach of the teacher has on the development of problem-solving skills among students.

Acquiring Teaching Efficiency

It is assumed that expert teachers can achieve their aims with relatively more speed and accuracy than non-experts can. An example of how awareness of efficiency can be engendered in trainees is presented in the following short activity where *instruction giving* is the focus. Here we have based the activity on the notion of teacher modes (see McCarthy & Walsh, 2003) whereby teachers are said to have various modes of talk in the classroom which can be assessed to improve classroom competence through teacher awareness. Here we focus on the *instructional mode* where teachers are giving instructions to the students. Firstly, we ask students to generate a wordlist using our classroom corpus and then to isolate all the verbs within this. Task b asks trainees to predict which of these verbs are used in giving instructions (*instructional* mode) and then they are asked to check their predictions by means of concordancing. Tasks b and c focus the trainees on the imperative nature of instructional talk while Tasks d and e focus qualitatively on the need for instructional episodes to be conducted with precision and clarity.

Figure 6

Sample material based on the L-CIE for 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 e.g. 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.

We find that our trainees get much out of this activity not least of all because it provides them with a framework within which to measure their practice. Anecdotal evidence suggests that they would not be as insightful or reflective without the structured use of our 'local' classroom corpus.

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 trainees' transcripts for parts of the lesson where they encountered difficulties which had not been anticipated during preparation. An example of this is presented in Figure 7. Here again we use our 'local' classroom corpus to focus on a typical classroom dilemma which all trainees can relate to where a students asks for a detailed lexical explanation, one which has not been anticipated by the teacher.

Figure 7

Sample material based on the L-CIE for 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.

Tasks a and b ask trainees firstly to draw on the standard dictionary resource to find the difference between the problematic words and then to use a corpus concordancer to compare their patterns in contexts of use. It is hoped to show through this activity how a dictionary definition can be greatly enhanced by concordancing because so many patterns of use can be viewed at once in many contexts. Task c leads trainees inductively back to classroom application.

CORPUS APPLICATIONS IN RAISING LINGUISTIC AWARENESS

Every teacher on an initial language teacher education programme expects to attain a high level of descriptive linguistic competence in relation to the language they are going to teach. Gabrielatos (2002/2003, p. 3) 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'. Corpora offer great potential in developing language awareness and research skills within teacher education (see Hunston, 1995; Kennedy, 1995 Coniam, 1997). Below we will share some examples of how we have used corpora to raise linguistic awareness on our teacher education programmes by using language corpora.

From our experience, a grammatically-tagged corpus (one where all of the items used have been labelled according to their word-class) is a very useful supplement to the development of critical syntactic knowledge of the English language system. For example:

1. Trainees 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 then presented with an untagged version of a text from a corpus and they try to identify the word classes, in student groupings or individually.
4. They check their answers against the tagged version of the same corpus and they carefully examine any inconsistencies and use them as the basis of a word search of a particular word to further test their hypotheses, for example, the classification of the word '*right'*, which, in various contexts can function in different ways.

This process subtly develops a sense of enquiry leading from the trainee's own research question to inductive exploration using a corpus as a problem-solving resource. Both the *ICAME Collection of English Language Corpora* and ICE GB provide a rich supply of grammatically tagged data. A tagged corpus also proves a very useful resource for the independent study of syntax as there is a ready made 'answer key' which trainees can consult.

A sample activity using concordance lines that we use to develop awareness of lexis and word classes is shown below.

FIGURE 8
Sample material based on the L-CIE for language awareness raising.

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
```

Such activities develop language awareness inductively and frequently lead trainees to form more research questions. Many trainee 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. For example, one undergraduate trainee who became intrigued by the high frequency of *like* in casual conversations between friends looked at the patterns of speech reporting (*I'm like…; he goes…* etc.) in these conversations compared with those used in textbooks for her BA dissertation.

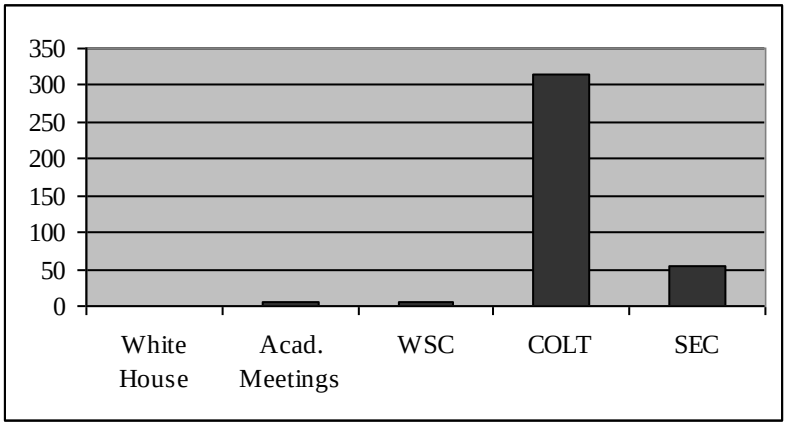
While concordance-based searches and investigations can provide the basis for many insights into lexical patterns and profiles, there is also scope to explore grammatical patterns using a corpus. Below is a task which focuses trainees on a grammatical item commonly presented in textbooks: i.e. question tags. This 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 you use, these differences highlight the importance of contextual factors and how essential it is to cross-check findings. Using the example of question tags, we present finding across various corpora: the American CSPAE (White House press conferences and academic meetings); the New Zealand Wellington Spoken Corpus (WSC); the British Corpus of London Teenagers (COLT) and Lancaster/IBM Spoken English Corpus (SEC). We first ask trainees to compare these findings from spoken corpora with those from written sources so that they see how rare they are in writing – in fact they are only used in direct speech or where the author addresses the reader and this is rare. The spoken findings that we present show that question tags are vastly more frequent on the London teenage data but it would be erroneous to assume that this means that they are a British phenomenon. We ask trainees to consider this in Tasks c and d. The context of the American data, for

example, is much more formal than the British data and so this has an impact on the results. Tasks e and f focus on the need to compare data across corpora and to consider the contextual origins of the data that they produce.

FIGURE 9
Sample material based on the L-CIE for language raising awareness.

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

- **Two sub-corpora within the CSPAE: one million words of White House press conferences and one million words of academic discussions and meetings.**
- **The Wellington Spoken Corpus (WSC) from New Zealand (one million words).**
- **The Corpus of London Teenage Language (COLT) (one million words).**
- **The Lancaster/IBM Spoken English Corpus (SEC) - 55,000 words (these results have been normalised)[5].**



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 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?**

[5] Note that to make frequency results comparable, they need to be 'normalised'. In this case, the Lancaster/IBM Spoken English Corpus (SEC) is only 55,000 words. In it we found 3 question tags ending in *you?* This was divided by the total corpus size (55,000) and multiplied by 1,000,000 to give 54.5. This figure is then comparable with the other results, which are all from one million word corpora.

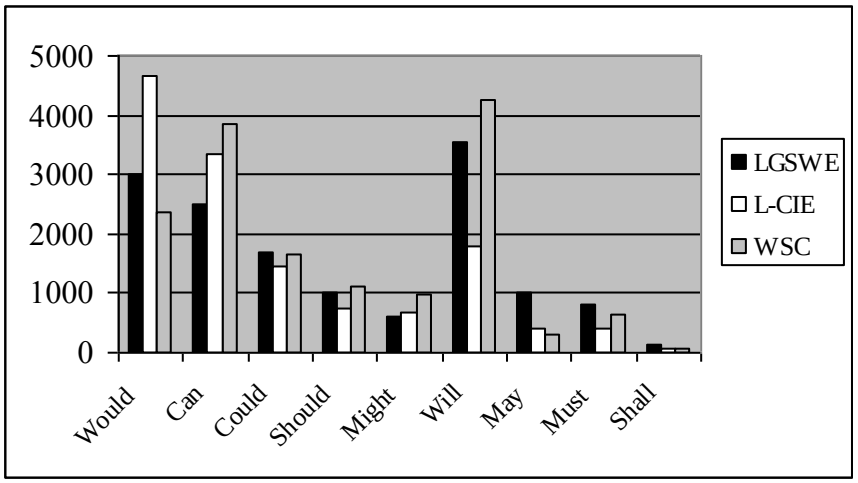
e) **use any two corpora to compare findings for question tags ending in *I, he, she, it, we, they?***

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

CORPUS APPLICATIONS IN RAISING SOCIOCULTURAL AWARENESS

As we have stated already, central to the evolution of corpus use in English language classrooms around the world is the development of critical awareness of what corpus findings represent. As we have illustrated above, structured corpus tasks can promote enquiry in trainee teachers so that they do not readily accept corpus findings as absolutes. We feel strongly that the scrutinising of corpus findings needs to be given overt attention, especially when dealing with large-scale corpora. In particular, we stress the need to take into consideration the sociocultural factors from which corpus data comes as this can tell us much about how language is pragmatically sensitive to context. In this section, we aim to give practical illustration of how corpora can be of benefit in raising awareness as to the sociocultural diversities that often belie corpus findings. In the following example, we compare the frequency of modal verbs presented in Biber et al., (1999) (LGSWE), with L-CIE (Irish English) and the Wellington Spoken Corpus (WSC) (New Zealand English).

FIGURE 10
The distribution of modal verbs across the Longman corpus, L-CIE and WSC (results per million words)



One of the noticeable differences is the high occurrence of *would* in the Irish data. We find that the Irish English use of *would* yields a range of uses of the modal verb *would*

that go beyond the 'canonical' characterisations in 'standard' English. The higher frequency we attributed in part to hedging (a common function of *would*), for example in this extract from an encounter between a trainer and a trainee (following a teaching practice observation), we can see how there is a convergent use of *would,* instead of a more direct statement like, *You should have allowed them to work through it all* (see Farr and 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.

The surplus functions of *would* in Irish English, which goes beyond descriptions in standard grammars, are central to the socio-cultural 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, we see a caller hedging about her hair colour.

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

In another example we see two friends reminiscing and *would* is again used for something that is factual (*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 explore an extra layer of tentativeness in Irish interactions, where downtoning of indisputable facts appears to be a sociocultural norm. In the context of teaching English at higher levels, it is not unreasonable to expect learners, particularly as they become more proficient, to become better at recognizing such socio-cultural nuances in the language they hear. Working with naturally-occurring data can facilitate this. However, Rundell (1997) raises a very pertinent question broadly related to this: whether imposing the 'idiosyncratic linguistic features of one specific dialect of

English is really an appropriate model for a majority of learners' (Rundell, 1997, 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 to do so trainees need to be imbued with cross-corpora comparison skills so as to facilitate critical investigation of the transferability and application of corpus findings to the broader socio-cultural context of their learners.

PRACTICAL ISSUES AND CONSIDERATIONS

Here we look at some practical issues and considerations that we have had to face over the last six years of developing corpora and corpus applications within our teacher education programmes. We examine the pros and cons of building versus buying a corpus; spoken versus written data; small versus large corpora; native speaker versus learner or non-native corpora; and using handouts versus having students work 'hands on' with the data.

1. Build versus Buy

Many corpora are now commercially available and some can even be purchased for under $100. As we have already illustrated, having a wide variety of corpora allows for more in-depth investigation across variables such as written/spoken language, context, variety and so on. One may also decide to build one's own corpus, for a number of reasons, such as the lack of availability of a specific language (or variety, for example L-CIE) (see Aston, 1997; Maia, 1997).The first step in building a corpus is to design a framework for the data you are going to collect. Much has been written on the principles of corpus design (see Crowdy, 1993, 1994; Biber et al., 1998; McCarthy, 1998; Hunston 2002). Because of the availability of data in electronic form, a written corpus is much easier to assemble than a spoken one. One needs to be aware of the serious resource implications of building a spoken corpus. From our experience of building L-CIE, a one million word spoken corpus, the following core costs need to be budgeted for: a) *collection of data:* that is paying individuals to record the data. Keep in mind that there is between 10,000 and 15,000 words per one

hour of recording (depending on the type of talk). One therefore needs to record over one hundred hours of material to ensure getting one million words; b) *transcription of data*: the data then needs to be transcribed. The cost of transcription depends on the level of detail desired. At a minimum it will cost around $150 per hour of tape (that is around $15,000 for one million words) and c) *a corpus administrator*: with this amount of data being collected and processed, it is essential to have an administrator for your project.

2. Spoken versus Written

Sinclair refers the current state of 'superfluity' of corpora and real language data (1997, p. 27), for example, the BNC (100 million words) and the Bank of English (over 500 million words). However, large corpora consist mainly of written British and American data. McCarthy (1998) accounts for the dearth of spoken data in light of costs (as discussed above), access to appropriate and representative speech data situations, quality of recording, time involved in transcription, difficult decisions in relation to level of detail to include in transcription, and so on. However, it can be argued that such exertion and funding is perfectly justifiable on the grounds of needing to re-assess language interpretation and pedagogy to account for spoken as well as written norms. As some of our earlier tasks will have highlighted, there are 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 trainees to be in a position to compare corpus findings, as we have argued, and to check results across spoken versus written corpora from as many varieties as possible. Too often our classroom descriptions of the English language are based on written norms. For this reason alone, the effort of assembling a spoken corpus is worth it. It is also worth noting that a small specialised corpus can be assembled at a relatively low cost. For example our classroom corpus comes for recorded data which teachers and trainees have 'donated' and which we have transcribed ourselves. Though it only amounts to under 100,000 words, it is rich in spoken data from our local context.

3. Small versus Large

Whether we use a small, specialized corpus or a larger generalized corpus really depends on our particular needs. Fox (1998, p. 25) 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'. If we are to examine a relatively infrequent word and are interested in generality of lexical use, then we need to use a larger more representative corpus in order to find adequate occurrences from which to draw some conclusions about typical features (see for example Coxhead, 2000). If, on the other hand, we need to find 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, we may need to use a small corpus if we are dealing with a very specialized language register, such as that described by Aston (1997). Small corpora are useful for training students into corpus techniques and methods, and they also allow the user to access more readily contextual or pragmatic information about the spoken or written text. Of course the ultimate advantage for the trainer/teacher is that they are cheap and easy to construct (or buy), and their limits are clearer as they can claim only to represent themselves and therefore discourage the user from over-generalizing. Aston (1997) makes an interesting and very practical distinction between the usefulness of small and large corpora - if we want to use corpora for developing materials and references, then we need a large corpus, 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). In terms of what constitutes a large or a small corpus it really depends on whether one is referring to a spoken or written corpus and whether one is seeking representation and range in the data contained therein (for a full discussion of these and other issues and examples of corpus design see: Sinclair, 1991; Thomas & Short, 1996; Biber et al., 1998; McCarthy, 1998; Biber et al., 1999; Coxhead, 2000; Hunston, 2002) In very general terms we adhere to the following guidelines: for spoken corpora anything over one million words is considered to be moving into the 'larger' range, for written anything below five million is quite small. Saying this, it is often the design of the corpus as opposed to its size which determines its suitability, for example, a corpus containing only highly technical engineering language will be largely inappropriate for language teacher trainees wanting to

investigate the vocabulary of everyday casual conversation. Therefore while size is an issue, it should be considered hand in hand with design appropriate to the long and short-term pedagogic needs of the trainees for any given purpose.

4. Native Speaker versus Learner/Non-native Speaker Corpora

The issue of native speaker versus learner/non-native speaker corpora is one of growing focus. The question of whether a corpus should include 'non-native' speakers is a fraught one since the *native* versus *non-native* distinction itself is problematic. Prodromou, among other, raises issues such as the undermining effect of corpora for non-native speakers of English (Prodromou, 1997a). 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 (p. 239) (see also Graddol, 1999; Flowerdew, 2000; Nero, 2000; Warschauer, 2000; Seidlhofer, 2001). Seidlhofer (1999) provides the term 'speakers of English as a Lingua Franca' (ELF) in reference to a corpus she is building. Seidlhofer (2001) details an innovative corpus development called The Vienna-Oxford International Corpus of English (VOICE) which aims to collect around half a million words of spoken data from speakers whose first language is not English, but who make use of ELF. This corpus will facilitate the profiling of ELF as something robust and independent of English as a native language. The corpus may, according to Seidlhofer (2001, p. 147), establish 'something like an index of communicative redundancy' which may have pedagogical application.

Learner corpora are a separate issue and it is important not to confuse them with 'non-native' speaker data. As many have argued, there are millions of people globally who are so-called 'non-native' speakers of English who are also highly competent users of the language. Granger (1998) advances theoretical and practical arguments for the place of learner corpora (i.e. those comprising samples of language from learners) in the language classroom for the purposes of studying phenomena such as second language acquisition processes, interlanguage, fossilization, patterns of error, cross linguistic studies etc. Biber and Reppen (1998), Granger and Tribble (1998), and Milton (1998), outline useful procedures for using corpora as a supplementary tool for non-native speakers, whereby native and non-native speaker data are compared and

analysed by students for the purposes of language advancement. Trainees will be interested to find out more about a large-scale international corpus project focusing on the written English of learners from many different first language backgrounds which has been compiled in recent years to form the International Corpus of Learner English (ICLE) (see for example 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* (LINDSEI) was set up to complement the ICLE project (see De Cock, 1998a, 1998b, 2000).

5. Hand Outs versus Hands On

A very practical, but important decision to be taken when using corpus evidence for pedagogic purposes relates to whether we prepare and print out the data to be used by our trainees in class or whether we allow our students to have access to the data on the computer. Of course the latter assumes the ready availability of adequate levels of technology (previously outlined) and support. In institutions where the technological support may be a concern there are many on-line self instructional options available (see Appendix A). Leech (1997) outlines the advantages of both the paper-based and computer-based approaches as follows: prepared printouts 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 preliminary work is done by the teacher prior to the lesson. 'Hands on' the computer in class promotes a more learner-centred approach, provides an open-ended supply of data, and allows for more tailored and customised learning. Others, such as Johns (1991), in describing the data-driven approach, strongly advocate the hands on use of corpora as this, Johns argues, is what makes the whole experience the epitome of induction. One of the arguments for engaging in concordancing in Data-driven Learning (DDL) is that it will give users control over their learning and build their competence by giving them access to the facts of linguistic performance (see Stevens, 1995), whereby the instructor provides the evidence which allows discovery of the 'facts' about the language from real examples. It may be discerned however, perhaps for practical reasons, that concordances are a useful resource to supplement class materials rather than opting for DDL. Willis (1998) outlines at length the procedures that can be adopted for the use of paper-based concordances in the classroom.

Those of us familiar with inductive instruction will appreciate its effectiveness but will also recognise the increased time investment required and on shorter training courses, already under time pressure, this may not be a luxury one can afford to entertain. In our teacher education provision we have managed to balance both approaches and have found that starting with printouts and working up to computer use promotes a more progressive, inductive approach, which trainees 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 on training programmes provides a richer variety of experience and presents trainees with more options for their own future teaching environments.

CONCLUSION

In this paper we have attempted to outline the 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 and develop as an influence in language pedagogy. Many instructional materials, in the form of CD-ROMS, software, dictionaries, grammars, etc. have been corpus-based in recent years and if only for this reason all teachers should know about corpora. We argue strongly that the more teachers know about corpora and the more they can use them, the more they will be empowered to (a) evaluate publishers' projects more objectively, and (b) put pressure on publishers and academics to come clean about the corpora they use in their products (e.g. how much written how much spoken, etc. – issues largely fudged at present). We need to educate teachers who can manipulate language corpora for their own pedagogic ends, scrutinise and evaluate findings that are presented as 'facts', whether native or non-native speakers of English, so that they will be better placed for the socio-cultural mediation and pedagogic recontextualization of these resources and findings in their language classrooms of the future. On a final note, as practitioners who have been involved in the use of corpora we are very much aware of the need to continue to develop methodological principles in relation to their use, and more essentially to empirically evaluate such approaches and their effect on learning.

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 Proceedings of the first annual conference* (pp. 51-62). Łodz: Łodz 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. http://rice.edn.deakin.edu.au/archives/jitte/anindex.htm

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., & 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.

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: 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: IATEFL .

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

Cook, G. (1998). The uses of reality: a reply to Ronald Carter. *ELT Journal, 52*, 57-63.

Conrad, S. (2000). Will corpus linguistics revolurtionize grammar teaching in the 21st century? *TESOL Quarterly, 34*, 548-560.

Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics, 22*, 75-95.

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) 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. (1998b) 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: Homo Liber.

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. http://llt.msu.edu/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. http://llt.msu.edu/default.html

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: John 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.

Freeman, D. (1994). Knowing into doing: teacher education and the problem of transfer. In D. Li, D. Mahoney, & J Richards (Eds.), *Exploring second language teacher development* (pp. 1-20). Hong Kong: City Polytechnic of Hong Kong.

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., & 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.

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

Greenbaum, S., & Nelson G. (1996). The International Corpus of English (ICE) project. *World Englishes, 15,* 3-15.

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

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.

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.

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.

Maia, B. (1997). Do-it-yourself corpora…with a little help from your friends. In B. Lewandowska-Tomaszczyk and P. J. Melia (Eds.) *PALC '97 Proceedings of the first annual conference* (pp. 403-410). Łodz: Łodz 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. http://llt.msu.edu/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. *CALL in the 21st Century.* CD ROM, Kent: IATEFL.

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: form corpus evidence to classroom practice. In B. Lewandowska-Tomaszczyk, & P. J. Melia (Eds.), *PALC '97 Proceedings of the first annual conference* (pp.90-98). Łodz: Łodz University Press.

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

Seildhofer, 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. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. (1997). Corpus evidence in language description. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 27-39). London: Longman.

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? *CAELL Journal*, *6*, 2-10. – should this be CALL??

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. *CALL in the 21st century*. CD ROM. Kent: IATEFL.

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

Thompson, G. (1995). *Collins cobuild concordance sampler 3: reporting.* London: Harper Collins Publishers.

Tribble, C. (1997). Improvising corpora for ELT: quick and dirty ways of developing corpora for language teaching. In B. Lewandowska-Tomaszczyk, & J. Melia (Eds.), *Proceedings of the first international conference on practical applications in language corpora* (pp.106-117). Łodz: Łodz University Press 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: IATEFL.

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.

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 applied linguistics. *Applied Linguistics, 21*, 2-25.

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 A: USEFUL WEBSITES

1) Corpora

American National Corpus
http://americannationalcorpus.org/

Australian Corpus of English
(available on ICAME CD-ROM)
http://khnt.hit.uib.no/icame/manuals/ace/INDEX.HTM
http://www.hit.uib.no/icame.html

British National Corpus
http://info.ox.ac.uk/bnc/

Corpus of London Teenage Language (COLT)
(available on ICAME CD-ROM)
http://www.hit.uib.no/colt/
http://www.hit.uib.no/icame.html

Corpus Linguistics Page
http://info.ox.ac.uk/bnc/corpora.html#Corpus

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

International Corpus of English – Great Britain (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
http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm

IVIE On-line Corpus
http://www.phon.ox.ac.uk/~esther/ivyweb/search_trans.html

Lancaster/IBM Spoken Corpus of English (SEC)
(available on ICAME CD-ROM)
http://www.comp.leeds.ac.uk/amalgam/tagsets/sec.html
http://www.hit.uib.no/icame.html

Limerick Corpus of Irish English (L-CIE)
http://www.mic.ul.ie/ivacs/

Longman Corpus of Spoken American English
http://www.longman-elt.com/dictionaries/corpus/lcaspoke.html

Longman Learners' Corpus
http://www.longman-elt.com/dictionaries/corpus/lclearn.html

Michigan Corpus of Academic Spoken English (MICASE)
http://www.hti.umich.edu/m/micase/

Mike Scott's Webpage (info on wordsmith tools)
http://www.liv.ac.uk/~ms2928/index.htm

Mike Barlow's parallel corpus page
http://www.ruf.rice.edu/~barlow/para.html

Parallel corpus research at Lund University
http://www.englund.lu.se/research/corpus/corpus/webtexts.html

Teaching and Language Corpora (TALC)
http://www.sslmit.unibo.it/talc/

The English-Norwegian parallel corpus project
http://www.hd.uib.no/enpc.html

The Louvain International Database of Spoken English Interlanguage (LINDSEI)
http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Lindsei/lindsei.htm

The Tuscan Word Centre
http://www.twc.it/

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

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

Wellington Spoken Corpus
(available on ICAME CD-ROM)
http://www.vuw.ac.nz/lals/wgtn_crps_spkn_NZE.htm
http://www.hit.uib.no/icame.html

2) Concordancing software and sites

Cobuild Concordance Sampler (The Bank of English)

http://titania.cobuild.collins.co.uk/form.html

WordSmith v3.0

http://www.liv.ac.uk/~ms2928/

http://www4.oup.co.uk/isbn/0-19-459286-3

Mono-Conc Pro

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

Concordance

http://www.rjcw.freeserve.co.uk/

Ultra Find

http://www.ultradesign.com/

Conc 1.80

http://www.sil.org/computing/conc/

Multiconcord: the Lingua Multilingual Parallel Concordancer

http://web.bham.ac.uk/johnstf/lingua.htm

Suggestions for classroom use of concordancing

http://www.nsknet.or.jp/~peterr-s/concordancing/usingconcs.html

http://web.uvic.ca/hrd/halfbaked
http://www.nsknet.or.jp/~peterr-s/concordancing/onlineconcquiz/online_conc_quizzes.html

3) Corpus linguistics tutorial sites

http://www.georgetown.edu/cball/corpora/tutorial.html

http://clwww.essex.ac.uk/w3c/

http://www.les.aston.ac.uk/txtintro.html