Lang. Teach. (2008), **41:4**, 563–574 © Cambridge University Press
doi:10.1017/S0261444808005247

Accessing and interpreting corpus information in the teacher education context

Michael McCarthy School of English Studies, University of Nottingham, UK
Jeannemike11@aol.com

As more and more teachers become aware of corpus resources and their applications, questions arise as to how best to prepare teachers in training appropriately to access and interpret corpus information. This is important as an element of materials evaluation as more and more types of materials and resources become corpus-informed. To date, relatively little attention has been given in teacher education programmes to the growing influence of corpora and the skills of evaluating and using corpora, but such skills will become increasingly important as the influence of corpora becomes more dominant in our profession. In this talk I argue that a shift is needed in the relationship between teachers, academics and publishers, from the teacher seen as consumer to the teacher as participant in the corpus revolution. I then consider the questions which need to be asked in evaluating corpora and the pedagogical resources which are based upon them. I look at the strengths and weaknesses of available corpus resources, and how corpus statistical output can be interpreted. Finally, I touch on the issue of reflective practice corpora within teacher education.

The title of this talk is 'Accessing and interpreting corpus information in the teacher education context'. Anyone might justifiably wonder why one needs to give a talk with this title. Surely by now, after twenty-five years of corpus linguistics playing an ever-widening role in language teaching and learning, we no longer need to advocate that knowledge of corpus linguistics and its influence should be part of teacher education? In reality we DO need to discuss the topic because, around the world, in perhaps the overwhelming proportion of teacher education programmes, there is still little systematic account taken of what has been called the 'corpus revolution' (Rundell & Stock 1992). And in the literature itself, only few scholars have confronted the issue directly, a notable example being O'Keeffe & Farr (2003).

In my travels across several continents, my common experience is that teachers have heard of corpora, but they are not quite sure what they are. They are sometimes frightened of what their use might imply: Does a teacher need to have a high level of expertise in computational linguistics or information technology in order to be part of this pedagogical revolution?

Revised version of a plenary paper presented at the QuiTE [Quality in Teacher Education] Annual Conference, held in London, 9 November, 2007.

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

Does one have to be a native speaker of a particular language in order to understand and use corpus information in that language? These are not uncommon and, perhaps, not unreasonable questions. For that reason we need to develop an overview, a broad look at what it would mean to integrate corpora (what we know about corpora and how and why we use them) into teacher education. The corpus revolution, as it has been called, could be said to represent a 'paradigm shift' (Kuhn 1962). It is one of those changes in social structure that comes along every so often and affects everything; in this case, the change was the advent of computers and their exploitation in the study of language in use.

What has shifted is a number of things: firstly, the way we look at language has changed, the way we understand language, and, perhaps most importantly, the way we understand the difference between spoken and written language, in ways that we were never able to do before. But the technological change has also caused a shift in the way that we create language teaching materials and resources, syllabuses and, to a certain extent, what happens in class. The corpus revolution has established itself, of this there is no doubt. It is now twenty years since the first corpus-based dictionary was published – the Collins-Cobuild dictionary (Collins Cobuild 1987). In 1987 this revolutionary learners' dictionary was launched and, generally, people in our profession had very little idea of how it was put together or why it was so innovative. Although there was no 'market demand' for it (the ELT publishers' mantra of the present day), the market was not clamouring for corpus-based dictionaries, and nobody was pleading for it, it appeared thanks to a visionary researcher (Professor John Sinclair of Birmingham University) and an equally visionary publisher, and it did indeed revolutionise things. When people realised how powerful this new dictionary was, that it contained information that was based on real usage, that it was not just put together in the traditional way by just absorbing all the other dictionaries and adding a few more words, it rapidly became a bestseller. This was something completely different. Now, twenty years later, no self-respecting publisher would dream of publishing a learners' dictionary that was not based on a corpus. Imagine, if you will, a major publisher at a conference boasting that they had just put out a fabulous new learners' dictionary and, best of all, that it was NOT based on a corpus. They would be laughed out of court no doubt, because corpus-based learners' dictionaries have so firmly established themselves with all the major publishers and vast numbers of users. But to a certain extent, and this is, I think, the nub of the issue, the teacher has always been seen as a CONSUMER of corpus-based materials. At major conferences, publishers typically announce new reference grammars based on a corpus, or vocabulary materials based on a corpus, and, generally speaking, the role of the teacher has been just to consume these products with awe, with a sense of admiration for the amount of research that went into them, for the technology that often accompanies them (e.g. CD-ROMs), and so on. But what we need to bear in mind alongside this, and one of the reasons we must challenge the idea of the teacher merely as the consumer of commercial products, is the shift that has taken place in teacher education generally. Our task is to see how we can create a fit between the shift in the perception of what a teacher is and how he or she develops professionally, and the paradigm shift that corpus technology has represented. What we need to consider in this paper is the teacher not as consumer but as researcher, as reflective practitioner, as someone more actively involved in their own professional development and in what happens in their classrooms. That is the challenge. How do we assist this new breed of teacher, the

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

teacher who knows about concepts like action research, who maybe even knows a little bit about data-driven learning, who themselves are probably computer literate at a basic level or greater, and turn that 'consumer' into a more active participant in the corpus revolution? And, of course, following from this are the implications for the content and emphasis of teacher education programmes, our main preoccupation in this paper.

The teacher as consumer is where we started, a passive consumer simply using corpus-based resources. Those resources now include not only dictionaries but also vocabulary materials (e.g. McCarthy & O'Dell 2005, 2008; Barlow & Burdine 2006), grammar reference books and materials (e.g. Biber et al. 1999; Carter, Hughes & McCarthy 2000; Carter & McCarthy 2006) and major coursebook series (e.g. McCarthy, McCarten & Sandiford 2004–2006). So, what would it mean to be an ACTIVE user-consumer of such technologically-inspired resources? If the teacher were to change from the passive role and become active, what would this actually mean, what would the teacher need to do and to know, and how could teacher education programmes help?

There are two main things, I believe, which need to be done. The first is more likely to become part of teacher education itself, and that is to develop the teacher's ability to evaluate corpus-based resources. This is not a simple matter and we will look below at a couple of examples of why this is quite a difficult task. How does a teacher, however professional he or she might be, make sense of, and appraise, these new materials that are coming out based in one way or another on computer analysis of language corpora? The second area that I believe to be crucial is the emergence of the teacher as lobbyist, lobbying academics and publishers and telling those academics and the publishers what it is they want and need of this new technology. Teachers should be central stakeholders in the corpus revolution, but, so far, in the case of the development of corpora, the types of research done on corpora and outcomes in terms of resources, teachers have spoken only with muted voices, and not always been listened to. Right now, most of the clamour is coming from elsewhere, from the academics in higher educational institutions and other people working in this field. We academics get ideas, we do a computer search of our corpora, we find a nugget of information about the language and we think: 'Wouldn't it be wonderful if we put this into language teaching materials?'. But we don't as yet get much feedback from the other direction saying: 'This is what academics ought to be doing, these are the sorts of things you should be creating for us, this is the type of language we want you to look at'. Where such voices are being heard louder, I often hear remarks such as: 'We don't want just to look at English as a vast, homogenised whole. We want information about academic English, we want information about business English. My students need information about the hospitality industry and how English is used in the workplace, and so on'. These voices should be respected and heeded if the applied linguistics profession is to be a two-way street.

So the active consumer needs to be well informed, and it is this level of information that we must look at now. How do we inform teachers, and what is it that we inform them about? We shall allow ourselves the luxury now of imagining an ideal world where we have a teacher education programme that we can design and where there is a slot in the syllabus where we can offer something on corpus.

The first thing we must get to grips with is what questions need to be asked when evaluating corpus-based resources. Step one is actually helping teachers to understand the nature of a

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

corpus, what it is and what it looks like. When I'm working with groups of teachers I usually tell them that they are already corpus users; we are, all of us, corpus users, because we use the internet. The internet is simply a huge corpus. It is millions of pages of texts, web pages, blogs, emails, online chat, and so on. It is a corpus, and every day all of us search that corpus, using the popular search engines. We all search the world's biggest corpus; we put in words or phrases or bits of words and bits of phrases and we get back thousands of answers. That is one of the basic things that corpus linguists do with their language corpora; they search large databases of language, spoken and written, and they get answers, often thousands of answers, in the form of statistics (e.g. frequency counts, collocation statistics) and they have to make sense of those answers, just as we have to make sense of the results we get from an internet search engine.

The next question is: Are the resources in question corpus-based or not? Or are they only 'corpus-informed' (see McCarthy 1998: 22f.)? Already here we enter a fine distinction of terminology. The distinction between CORPUS-BASED and CORPUS-INFORMED is one that is commonly understood in the following way: corpus-based materials are materials that try to be absolutely faithful to what the computer tells you about language use, whether you like it or not, whether it is useful or not. The late Professor John Sinclair, my great teacher and mentor, was indeed dedicated to the view that one shouldn't override what the technology is telling you: *Trust the text* was the title of one of his many memorable publications (Sinclair 2004). For example, if a spoken corpus analysis suggests that everybody, including the so-called most educated speakers, uses *there is* with a plural noun, then we might not like it but it is a fact we must learn to live with as an ineluctable change in our common usage. That's what we call being corpus-based: everything we do is based on what we get from the corpus. CORPUS-INFORMED is a more nuanced approach. It is a way of saying that one is going to do what is useful, what one's students want, what is needed, what is feasible, what is practicable, what is going to be most usable for the students. In this case one takes from the corpus what one believes will fulfil those ambitions, those needs, and the teacher's or material writer's task is essentially to mediate corpus information, filtering it for pedagogical purposes. The 'filters' might include what we know about the learning process, what the constraints of the curriculum are, what the local educational conditions, culture and traditions are, and so on. So these are two quite different positions that need to be explored in a teacher education programme. It is not just a question of simply telling teachers about the technology that is involved in corpora, and then leaving them to their own devices. There are big questions here: Is the corpus going to lead us or are we going to, as it were, lead the corpus in the directions that we need it to go?

The next major question to ask about any language-teaching resource, whether it be a dictionary, a grammar book, a set of materials or a syllabus is: What corpus was used? A publisher that says that a new dictionary is corpus-based must be prepared to say what the corpus was. Directly related to the question of what corpus was used is how big it is and how it was designed. The question 'how big is a corpus?' is one of those 'how long is a piece of string?' questions. We need to be able to judge whether the size of a corpus is suitable to the use to which it is put, and we shall consider some examples of this presently. But the primary question is, how big was the corpus used in the generation of the materials and resources under evaluation; how was it composed and put together? All of the major

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

publishers of English language teaching resources have large corpora: Pearson–Longman, Oxford, Cambridge, Macmillan, Thomson, all have corpora of some sort, but they vary in size and they vary considerably in how they are composed. This is most notably so in the balance (or lack of it) between the amounts of written and spoken data included in the corpus.

Allied to the question of size and composition is the expertise of those who exploited the corpus, the corpus users. Increasingly we see materials and other pedagogical resources emerging that claim to be corpus-based or corpus-informed, and we should rightly ask the question: Who were the people that used it to produce these materials and resources? What was their level of expertise? How much training did they themselves have in using corpora? The way you use a corpus, your way of thinking and analysing the language that is there in the corpus is by no means an easy and straightforward matter. After 25 years of researching corpora I feel I'm still on a steep learning curve, and would not call myself an expert by any means. So the expertise of corpus users is not an otiose question; the quality of the output most definitely depends on the quality of the input, and the input in this case is largely to do with framing and asking the right questions of the data to hand.

Finally comes the question: What aspect of the resources were influenced by the corpus? With a dictionary this is usually fairly straightforward: all the words that are in the dictionary come from the corpus with perhaps others added which slipped through the corpus net (so difficult is it for any corpus to embrace the entire language). But now that the influence of corpora is spreading wider and wider we often find that it is only one part of a work which is influenced by a corpus. One current major adult EFL coursebook, for instance, advertises itself as corpus-informed but, on closer inspection, it is only the vocabulary list which comes from the corpus; the rest of it, the grammar, the skills, the texts, the reading tasks and so on are not based on the corpus. So this is a very important question to ask: Precisely what component(s) of the materials were actually informed by the corpus? Was it everything? There is at least one adult course on sale now that is very widely dependent on corpus resources, that has constructed not only its grammar syllabus, and its vocabulary syllabus but its speaking skills syllabus, its listening syllabus and its writing syllabus using information from large corpora.

What else do teachers need to know? We've talked so far about teachers having an awareness of corpora and the role that they play in the resources that are now currently available. Let us now consider in more detail some of the things that teachers need to know. We mentioned earlier the question of evaluating corpus size. This really is a case of how one wants to use the corpus. For lexicography, for the writing of dictionaries, you need hundreds of millions of words to get a cross-section of the words of any language and sufficient examples of the rarer words and expressions, since even in large corpora of millions of words, many words and expressions only occur once or twice. For example, do people really say 'raining cats and dogs' or is this something that is only heard in EFL classrooms? In fact, the Cambridge International Corpus shows the expression only occurring once per 50 million words, and none of the examples occurring in speech. What can we tell from one or two single examples? Not much. To be able to define accurately what the idiom means and to observe what patterns, if any, it forms, and what situations it is used in, a lexicographer probably needs a couple of dozen examples. To get sufficient examples of *raining cats and dogs*, one would need a corpus of at least half a billion words. And indeed, most learners' dictionaries are based on corpora of

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

at least a hundred million words, often more. The Cambridge International Corpus has now gone over one billion words, so we can see where things are heading (for further information, see <http://www.cambridge.org/elt/corpus/international_corpus.htm>). So, for general dictionaries, the corpus has to be very big: tens, even hundreds of millions of words.

Reference grammars, on the other hand, are quite different, and fewer words are needed. Large, corpus-based reference grammars of the type now commercially available can be written on the basis of a meagre hundred million words, perhaps something like the British National Corpus, which is a hundred million words of written and spoken texts (see <http://www.natcorp.ox.ac.uk/>). The reason is that grammar phenomena occur far more frequently. If you want to write a description of the preposition *of*, it is likely to occur more than 25,000 times in a one-million-word general corpus of present-day English. The search will in fact yield far too many examples, so one needs tools in the analytical software which can randomize the output and give only a random sample of the many occurrences, for instance, a sample of one thousand examples. For most of the grammar of English, apart from obscure structures like the past subjunctive, which is rare in modern-day English, a relatively smaller corpus will suffice.

For course books as well, a hundred million words are generally adequate. Specialist materials for domains such as English for academic purposes, business language, and so on, offer the opportunity to collect data in very narrow contexts. Such data are very closely circumscribed, one knows exactly where they came from, who was speaking, what they were doing. Far less in terms of quantity is needed. One business language corpus in existence at the moment is just one million words and has produced many useful insights into how business language operates (McCarthy & Handford 2004; O'Keeffe, McCarthy & Carter 2007). Thus, when we ask how big a corpus needs to be, the answer is a question: What will it be used for? And this is something that teachers need to develop an awareness of: a feeling for the appropriateness of the size of a given corpus. It is an awareness that can only come from hands-on practice and discussion, which should ideally form part of a teacher education programme.

The value of the information extracted from the corpus is the next issue that needs to be considered. How useful, how valuable is the information that we get from a corpus, whether large or a small? Let us consider some examples from a five-million-word corpus of mixed written texts from newspapers, novels, magazines, letters, and similar material. This is to underline the point I made earlier. For vocabulary analysis (as opposed to grammar) you need considerably more data because many words will only occur once or twice in even quite large corpora. The word *bindweed* is an example. It is not a terribly exciting or interesting word but if a student is concerned with English for agricultural purposes then it may be much more important. However, in our five-million-word corpus, we only find one example. We almost certainly need a more specialised corpus to get a range of contextualised occurrences. The next example word is *mercantile*; for anyone interested in business or financial language, this corpus would not be ideal: it only yields four examples. So the general, homogenous corpus is not very good for specialist usage. On the other hand, this same corpus returns 436 examples of the past perfect continuous/progressive form (*had been working*, *had been looking*, *had been travelling*, etc.). The value of this for a grammarian is quite considerable. To have 436 examples of the past perfect continuous form considerably facilitates obtaining an answer to

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

the classic teacher's question: How do we use that structure, when do we say *had been working*, *had been travelling*, etc.? This is not something that's actually easy to decide just from intuition (on this point, see Carter & McCarthy 2006: 621). So the corpus output is an extremely valuable resource.

The other issue I mentioned earlier was the balance of spoken versus written data in any corpus. It is only relatively recently that spoken corpora have become big enough to make an impact. Spoken corpora are expensive and extremely time-consuming to record, transcribe and compile, compared with the relative ease with which written data can be scanned or downloaded from internet sources. The COBUILD dictionary, for instance, in its first edition (though not its latest editions), was largely based on written data. There was not a great amount of spoken data in the corpus at the time, so it was essentially a dictionary of how we write, not a dictionary of how we speak, albeit a seminal and magnificent work which has been much imitated since. Nowadays thanks to miniaturised recording technology we can record speech much more easily, transcribe it and get millions of spoken words into the computer and balance things out more satisfactorily, so there is less of an excuse for only compiling written corpora. When Ronald Carter and I were preparing one of our works we had five million words of written and five million words of spoken data at our disposal. This was a balance that enabled us to see how the language is different in those two contexts.

The next issue to consider is the sources and range of texts in the corpus. If, for example, one wanted to create a truly balanced corpus of British newspapers, would one have to take into account the fact that the popular tabloids account for 78% of the market with the quality broadsheets only achieving 22% of the market? Would one have to build one's corpus accordingly, with a 3:1 bias toward the tabloids, to truly represent the daily encounters with newspaper language experienced by the British public? This is a question about the sampling of data for any corpus. What does the corpus represent? Does it represent 'quality press' sources such as the *Observer*, *Guardian* and *Independent* as many corpora do, or does it reach out to the *hoi polloi* and represent the language of the masses, so to speak? However, strong prejudices remain against the use of data from users of lower educational achievement. In collecting data for the CANCODE spoken corpus project that Ronald Carter and I directed at Nottingham University in the 1990s (the corpus was sponsored by Cambridge University Press; see McCarthy 1998 for information on its composition), we attempted first of all to move out of the confines of England. I'm not English; I'm Welsh of Irish ancestry and for years I had felt distanced from the teaching materials I used (the audio tapes, the dialogues, the grammar prescriptions, etc.) because they all seemed to be based around the usage of educated speakers from Southern England. Carter and I decided to try to break this particular mould; we recorded people in Cardiff, Dublin, Lancashire, Yorkshire, etc., with a view to modelling English from a wider range of speakers for pedagogical purposes. But, what happened? At one IATEFL international conference we were accused from the floor of including 'semi-literate speakers' in our description of spoken English grammar (Carter & McCarthy 1995), a challenge which is not easy to rebuff when feelings are so deeply entrenched. Achieving a demographic balance is not a simple matter, and people might not actually like it when you do try to achieve it. And since the days of the CANCODE project, the debate has shifted to much wider preoccupation with international varieties of English

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

(e.g. the ICE [International Corpus of English] project) and issues surrounding non-native speaker corpora (Prodromou 2008). However, understanding the importance of demographic factors and awareness and discussion of public (and professional) prejudices and expectations are two further aspects of corpus linguistics which need to be included in any programme of teacher education. These issues of data collection overlap with questions already current in teacher education concerning models of English for pedagogy, the recognition of varieties and questions of standard and non-standard usage (see Carter & Nunan 2001: 1–4).

The final question I want to deal with in this paper is one that has rightly generated an increasing amount of debate, and that is the relationship between native-speaker and non-native-speaker data in corpora. It is still very difficult to persuade publishers of English language teaching materials that it would be a good idea to base them, in part or whole, on non-native-speaker speech or writing. This is because the vast majority of teachers out there in the publishers' markets don't like the idea anyway. When market research is done, teachers often respond that they want native speaker evidence, not evidence from non-native users. The whole issue is very complex, and attitudes are not necessarily the same amongst teachers as amongst students (Timmis 2002). Although non-native user corpora ('user' in the sense of people who use English in the world, for example in business or study) are under-developed and under-exploited, LEARNER corpora (usually in the form of examination scripts or essays or classroom transcripts) have been used a great deal and are frequently a resource for evidence in error warnings in teaching materials (e.g. Carter & McCarthy 2006) or as sources for the targeting of particular language features in materials (McCarthy & O'Dell 2005). And here we have another area where corpus linguistics overlaps with extant and long-standing preoccupations in teacher education: the question of the status of errors, of how and when errors should be targeted, and the kinds of feedback which are effective (Edge 1989). Thus, although I am advocating a place for corpus studies in teacher education, that place may not only be a specific compartment or module in the programme syllabus, but should ideally inform other, already existing components of the syllabus.

I asserted at the beginning of this talk that a shift had taken place in how we view the role of the teacher in his/her professional development, that the teacher is seen now not as consumer but as a researcher or reflective practitioner, as being actively involved in their own professional development and in what happens in their classrooms. To foster the notion of teacher as researcher or reflective practitioner, I would argue that information on access to available corpora is something that should be part of teacher education programmes. This can be done in a variety of ways: corpora on CD can be purchased by institutions (these often come with built-in search engines). The British National Corpus, and the one-million-word British segment of the International Corpus of English (ICE-GB; see <www.ucl.ac.uk/english-usage/ice/avail.htm>) can both be bought on CD-ROM. Online access to corpora is the other main way in which teachers should, in my view, be helped and trained. The best example of an online corpus currently available is the Michigan Corpus of Academic Spoken English (MICASE), which is free to access online and on which powerful searches can be carried out. However, other online corpora typically offer only limited access to part of the corpus or a limited amount of output in the form of on-screen concordances, etc. On the other hand, going back to our analogy between a corpus and the internet, the

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

world-wide web itself can be used as a corpus, and a massive one at that. The Webcorp website (<www.webcorp.org.uk>) is a software interface that enables one to search the web and gives the kind of output that linguists use, concordances and collocational statistics, rather than just the output one gets from a search engine such as Google or Yahoo. The search engine is very powerful and very sensitive. One can choose what to search; for example one can search British broadsheet newspapers, and within those, the business pages only. This is the type of information that teachers should have and be helped to use.

Corpus-analytical software is the next issue. There are freeware suites available on the world-wide web. These are free downloadable software for searching language databases. They are usually not very sophisticated; that's why they're free. There are also application suites that one can buy, the most famous of all being Scott's *Wordsmith tools* (Scott 1999, 2004), though one cannot gainsay that such software suites are not always easy to use and take a considerable amount of time and frustrated energy to master. One can only hope that software suites will become simpler. Additionally, online software is available, for example the CALPER GOLD[1] software currently in its final stages of development at the Pennsylvania State University in the USA. This software suite, directly accessible online, will enable teachers to track the development and change in their learners' production over time. So if, for instance, a teacher teaches the past perfect in week seven of a course, he or she might hope to see it coming out in week eight and week nine and week ten in the students' writing if appropriate tasks are set. The CALPER GOLD software will be able to track over time emerging structures and emerging words and patterns. It presents its output in a very user-friendly way, using coloured graphs, pie charts and so on, not in the form of intimidating tables of statistics, which can be very de-motivating to teachers. So teachers can, themselves, access this kind of software, and, hopefully, without the frustrations caused by other suites. It is all Windows-based, it is all easy, and it is all done with the click of a mouse. Gone are the days when researchers had to write complex syntax commands even to get the computer to produce a simple frequency list, which was the situation that pertained when I myself began corpus research.

Lastly I would suggest that a good teacher education context should help teachers to build their own corpora (see Walsh 2006; O'Keeffe et al. 2007: ch.11). This, ultimately, is the most challenging example of the teacher as researcher with regard to observing the language in use. There are several ways in which teachers can be helped to engage in such research. Much can be achieved outside of class, in the teacher's own home, using the internet, downloading internet texts, recording one's friends in conversation etc. One must add the caveat here that transcription of audio files of spoken language takes a long time, roughly about 12–15 hours of transcribing for every hour of speech. However, the most exciting type of corpora that have been built so far are actually constructed in class; they are action research corpora. Walsh's EFL classroom corpus (Walsh 2006) is an example of how real class recordings can lead to new insights as regards effective teaching and understanding where and when learning opportunities arise. Corpora such as Walsh's enable teachers to see, for example, what questioning techniques work best, how students respond, how language can be used in different ways in the classroom, whether it is giving students instructions or having real

[1] CALPER: Centre for Advanced Learning Proficiency Education Research, GOLD: graphic on-line diagnostic.

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

conversations with students, and so on. This, I would argue, is the richest vein to explore in teacher education. The technology is simple; It is just a question of time and resources to transcribe data. And this is not an intractable problem; it can be addressed through data sharing. The CALPER GOLD project, mentioned earlier, will roll out during the coming year a system whereby teachers can contribute to a bank of data and if they contribute they will be able use the other data which is in that bank. This alleviates the problem of the long time it takes to build one's own corpus.

One final type of corpus, which has developed more recently, are corpora compiled from teacher education itself, for example, recordings of training sessions, of post-observation conversations between mentors and trainees, etc. Fiona Farr, of the University of Limerick in Ireland, and Camilla Vásquez and Randi Reppen, at the Northern Arizona University in the United States, have compiled this kind of corpus, which enables researchers to see how more successful and less successful teachers experience and complete their teacher training (Farr 2005, 2008; Vásquez & Reppen 2007). This is the ultimate example of reflectiveness in teacher education, where one is using a teacher education corpus to inform teacher education. In the Northern Arizona example, the researchers looked at teacher feedback sessions after training in class at one point in time and then later, after making changes to the questioning methods of mentors. The mentors changed their conversational approach to the trainees and produced more feedback and better quality feedback for and from the trainees at the second point in time.

And lest the impression be left that involvement in corpora by teachers is still a barren landscape, mention must be made of the growing number of web resources where teachers can access corpus information, either in a ready-made form (e.g. ready-made frequency lists) or where they themselves may do searches. The most notable example is Cobb's Compleat Lexical Tutor website (see <http://www.lextutor.ca/>), where corpora can be consulted, word-lists accessed, vocabulary profiles built for texts and a variety of other corpus-based resources are offered, along with useful links, all in extremely user-friendly form, and which is visited by thousands of users every day. Teachers are undoubtedly becoming more aware of the potential of corpora but the spread of awareness still has a long way to go.

A final question which arises is how corpora are developing and will develop in the future. The most important developments are in multi-modal corpora, where sound, video and transcript are linked and can be accessed via one another. When I was first involved in building spoken corpora, the only practicable medium was analogue tape recordings, and so we had to store cassettes in a big fireproof metal cupboard, and it is very difficult and time-consuming to link up the audio files with the computer data files. However, everything has changed since then; the latest corpora are digital, and the digital sound can be easily linked to the text on-screen. So the user can click a piece of text on the screen and instantly play it back in audio. Important to mention here is the work of Richard Cauldwell, who has produced a courseware facility called *Streaming Speech* (see <http://www.speechinaction.com>), in which natural speech files are digitally linked to on-screen content. Projects are also in the pipeline at various institutions to construct speech corpora which will enable users to search precisely around sound parameters and retrieve all the examples of language that have a particular sound, that one might be interested in for, say, pronunciation teaching (e.g. Campbell et al. 2007).

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

To conclude this brief overview of corpora in the context of teacher education, I suggest it may be seen as developing in three stages. The first is building awareness of what a corpus is and how it can be exploited, preferably with hands-on experience for teachers in training. The second stage is to look at the materials which are corpus-informed and to develop skills of evaluation and to encourage teachers to become more active participants in the corpus revolution, including lobbying academics and publishers to respond to practical needs. The third stage is assisting teachers to use corpora with increasing expertise and to construct their own corpora to meet their own needs and aspirations. The corpus revolution is here to stay, and teacher education cannot afford to sideline it.

References

Barlow, M. & S. Burdine (2006). *American phrasal verbs* (CorpusLAB Series). Houston, TX: Athelstan.

Biber, D., S. Johansson, G. Leech, S. Conrad & E. Finegan (1999). *Longman grammar of spoken and written English*. London: Longman.

Campbell, D., C. McDonnell, M. Meinardi & B. Richardson (2007). The need for a speech corpus. *ReCALL* 19.1, 3–20.

Carter, R. A., R. Hughes & M. J. McCarthy (2000). *Exploring grammar in context*. Cambridge: Cambridge University Press.

Carter, R. A. & M. J. McCarthy (1995). Grammar and the spoken language. *Applied Linguistics* 16.2, 141–158.

Carter, R. A. & M. J. McCarthy (2006). *Cambridge grammar of English*. Cambridge: Cambridge University Press.

Carter, R. A. & D. Nunan (2001). *The Cambridge guide to teaching English to speakers of other languages*. Cambridge: Cambridge University Press.

Collins Cobuild (1987). *Collins Cobuild English language dictionary*. London: Collins ELT.

Edge, J. (1989). *Mistakes and correction*. London: Longman.

Farr, F. (2005). Relational strategies in the discourse of professional performance review in an Irish academic environment: The case of language teacher education. In K. P. Schneider & A. Barron (eds.), *Variational pragmatics: The case of English in Ireland*. Berlin: Mouton de Gruyter, 203–234.

Farr, F. (2008). Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness* 17.1, 25–43.

Kuhn, T. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

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

McCarthy, M. J. & M. Handford (2004). 'Invisible to us': A preliminary corpus-based study of spoken business English. In U. Connor & T. Upton (eds.), *Discourse in the professions: Perspectives from corpus linguistics*. Amsterdam: John Benjamins, 167–201.

McCarthy, M. J., J. McCarten & H. Sandiford (2004–2006). *Touchstone: Student books Levels 1–4*. Cambridge: Cambridge University Press.

McCarthy, M. J. & F. O'Dell (2005). *English collocations in use*. Cambridge: Cambridge University Press.

McCarthy, M. J. & F. O'Dell (2008). *Academic vocabulary in use*. Cambridge: Cambridge University Press.

O'Keeffe, A. & F. Farr (2003). Using language corpora in language teacher education: Pedagogic, linguistic and cultural insights. *TESOL Quarterly* 37.3, 389–418.

O'Keeffe, A., M. J. McCarthy & R. A. Carter (2007). *From corpus to classroom*. Cambridge: Cambridge University Press.

Prodromou, L. (2008). *English as a Lingua Franca: A corpus-based analysis*. London: Continuum International.

Rundell, M. & P. Stock (1992). The corpus revolution. *English Today* 30, 9–17; 31, 21–38; 32, 45–51.

Scott, M. (1999, 2004). *Wordsmith tools. Software, Versions 3 and 4*. Oxford: Oxford University Press.

Sinclair, J. McH. (2004). *Trust the text: Language, corpus and discourse*. London: Routledge.

Timmis, I. (2002). Native-speaker norms and International English: A classroom view. *ELT Journal* 56.3, 240–249.

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press

Vásquez, C. & R. Reppen (2007). Transforming practice: Changing patterns of participation in post-observation meetings. *Language Awareness* 16.3, 153–172.
Walsh, S. (2006). *Investigating classroom discourse*. London: Routledge.

MICHAEL MCCARTHY is Emeritus Professor of Applied Linguistics at the University of Nottingham, UK, Adjunct Professor of Applied Linguistics at the Pennsylvania State University, USA, and Adjunct Professor of Applied Linguistics at the University of Limerick, Ireland. He is co-director (with Ronald Carter) of the five-million word CANCODE spoken English corpus project and the (co-)author and (co-)editor of more than 40 books and 80 academic articles.

https://doi.org/10.1017/S0261444808005247 Published online by Cambridge University Press