Farr, F. and **O'Keeffe, A**. (2019) "Using corpora to analyse language". In S. Walsh and S. Mann (Eds) *Routledge Handbook of English Language Teacher Education* London: Routledge, 268-282.

18 Using corpus approaches in English language teacher education

Fiona Farr and Anne O'Keeffe

Abstract

The aim of this chapter is to explore the ways in which corpus linguistics (CL) can facilitate teacher development in terms of content, pedagogy, technology, and research. Based on our own and other reported experiences of using CL in ELTE, we demonstrate the ways in which this approach is one of the ways teacher educators can more easily align their espoused theories (what they say they believe) with their theories in practice (what can be reasonably understood to be their beliefs based on direct observation of their practice) (Schön 1987). In other words, so that they can practise what they preach in terms of supporting novice teachers to become independent, aware, critical, inquiring, reflective practitioners. We have long argued (e.g. O'Keeffe and Farr 2003, Farr 2010) that corpora continue to play a minor part in much teacher education but are a resource with much potential (see also Römer 2006). This chapter is based on very recent advances in our corpus-based understandings of both language and educational processes and relies on a range of sources of evidence as presented in published research findings.

Introduction: ELTE and corpora?

If language teaching is complex, then language teacher education is an incredibly complex endeavour. Supporting novice teachers in their preparation for the myriad of variables and influencing factors that will play a part in the success, or otherwise, of their own future students is a task that must be approached with the care it deserves. The many questions we ask ourselves as language teacher educators when considering course design, development and delivery are: how much content (language) should be covered? What theories should be included? Where is the practicum best integrated and how much is enough? What is the nature of the received wisdom the student teachers come with and how can that be best utilised, or should it perhaps be critiqued and discarded? How can the latest technologies be

incorporated? What of teacher research - which methods, what kind of data and analytical skills are required? How can we imbue a strong commitment to ongoing professional development, so that the teacher education course is seen as the starting and not the finishing point? And what are the theoretical and practical frameworks needed to support this CPD? And then there are the philosophical stances that novice teachers have towards the nature of language, learning and teaching, also known as wisdom of practice (WoP) (Shulman 2004). How can these be honed, critically developed and best aligned with the teacher's practice? Taking Shulman's original definition of content (knowledge about language), pedagogic (knowledge about teaching) and pedagogical content knowledge (transforming content knowledge into a format suitable for teaching) (Shulman 1987, 1986), layering this with WoP (Chappell 2017: 435), and adding technological knowledge (Mishra and Koehler 2006) to give a TPACK (technological, pedagogic and content knowledge) acronym, goes some way to simplifying the complexity of the ELTE context. This is represented visually in Figure 18.1, as overlapping and integrated zones.

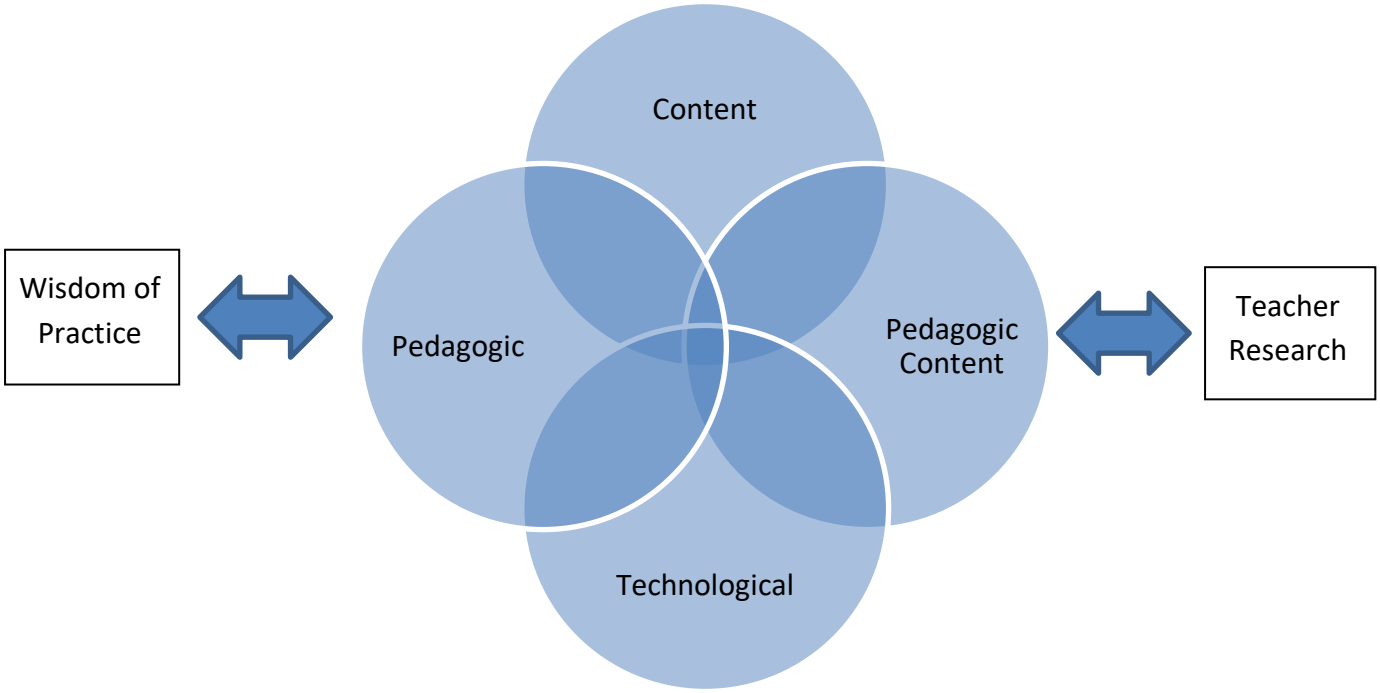

Fig 18.1: Teacher Knowledge

The aim of this chapter is to explore the ways in which corpus linguistics (CL) can facilitate teacher development in each of the spheres represented in Figure 18.1. Based on our own and other reported experiences of using CL in ELTE, we demonstrate the ways in which this approach is one of the ways teacher educators can more easily align their espoused theories (what they say they believe) with their theories in practice (what can be reasonably understood to be their beliefs based on direct observation of their practice) (Schön 1987). In other words, so that they can practise what they preach in terms of supporting novice teachers to become independent, aware, critical, inquiring, reflective practitioners. We have long argued (e.g. O'Keeffe and Farr 2003, Farr 2010) that corpora continue to play a minor part in much teacher education but are a resource with much potential (see also Römer 2006).

So, what do we mean by corpus linguistics? A simple definition is: the electronic analysis of a digital collection of naturally occurring spoken (in a transcribed format) or written language, potentially with additional contextual information. This analysis allows the user to get information such as frequency of use, patterns of use, collocations, range of use across texts, plus additional statistical information. A basic example of this is the *Google NGram Viewer*, which tracks the frequency of use of a word over time in its Google Books database and displays it on a line graph. And there are other easy to use interfaces such as Just-the-Word, based on the British National Corpus (BNC), which produces a list of the most common collocates of the search word, sorted by pattern. Such sites can be effective on shorter ELTE courses as an introduction to the concepts of CL, without the need for more serious technological upskilling (see Naismith 2017 for examples of their use on intensive CELTA courses). On the other side of the cline is the full integration of CL approaches in ways that permeate all aspects of the teacher education programme, and this is usually more plausible in the context of an MA programme at university level, where time is less of an issue and the academic demands of the course are high. This potentially will allow detailed coverage of CL approaches (perhaps through following a MOOC, such as that offered by Lancaster on the FutureLearn platform), and the use of such approaches in each of the spheres identified in Figure 18.1, including as a research tool for an MA dissertation (see Farr 2008 and Zareva 2017 for such accounts in grammar courses).

Using corpus approaches: practical issues and considerations

In our 2003 article (O'Keeffe and Farr 2003), we outlined some of the key decisions that need to be made when taking a CL approach in teacher education. It is timely to revisit these here, some 15 years later, to determine what has become redundant, what remains true, and what needs to be added to that list in order to take account of changes that have taken place during the interim period, both in terms of technology, information and pedagogic approaches. In 2003, using this terminology, we discussed 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 allowing students to work 'hands on' with the data. Below, we reframe some of these discussions in light of changes, such as ease of access, more advanced technology etc., that continue to affect the ways in which we integrate corpus use in ELTE.

To build or not to build?

It is slightly amusing that we began this section with the following statement: 'Many corpora are now commercially available and some can even be purchased for under $100' (O'Keeffe and Farr 2003: 407). In fact, much of our effort at that time was spent on trying to secure funding to buy the few corpora that were available, either in disk format or through subscription. The question of 'buying' has now become all but redundant as there are a vast array of corpora readily and freely available on-line, representing many genres, varieties and registers. One prime example, is the Brigham Young collection of corpora, collated by Mark Davies (https://corpus.byu.edu/), which contains freely-available, searchable spoken and written corpora that run into billions of words and are reportedly accessed by 130,000 distinct

users each month. Scott Thornbury has also produced a very useful video tutorial available on YouTube describing how to use COCA, the Corpus of Contemporay American English, which is part of the BYC (also available at: https://scottthornbury.wordpress.com/2013/05/12/c-is-for-coca-corpus/). Murphy and Riordan (2016) provide an up-to-date broad overview of corpus types and uses. Their chapter outlines five types of corpora: general, parallel, historical, multimodal and specialised. In each section, they provide an overview of the corpus type, the key issues associated with the type as well as its applications in pedagogical contexts. So, the question of availability and cost have largely become irrelevant. This does not however, negate the need to consider whether it is desirable to build or collate one's own corpus. This can be a time consuming and complex task (see Reppen 2016 for details of how to compile a corpus), and one which is generally undertaken by teacher researchers as part of an MA or PhD thesis project if they wish to examine a specific variety or genre that is not available to them from other sources (see below).

Spoken versus Written

Direct access to spoken versus written corpora, although easier, continues to be a challenge in relative terms. But there has been a change:

> In comparison with written corpora, spoken corpora have not developed at the same rate [….] The reasons for this are largely to do with the huge costs and time involved in compilation and transcription, as well as access to recordable data. What has developed over the last 20 years, however, is an acknowledgement of the importance

> of spoken corpora in creating a fuller understanding of everyday spoken language, especially casual conversation.
>
> (Caines et al. 2016: 348)

McCarthy (1998) accounted for the dearth of spoken data in light of costs, 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. Other than quality of recording, most of these considerations have become even more problematic. Stricter and tighter ethical protocols around data gathering and legal mandates around data storage have made it almost impossible to collect spoken language representation from some user groups (for example, children or vulnerable adults). These protocols are for the protection of individuals and are to be welcomed but they have brought a new reality to many interested in the compilation of spoken data. Having said that, where access is relatively straightforward, we have seen a growth, and a resultant discernible integration of the findings from spoken language research into materials development. In his 2015 article, Alex Gilmore examined the influence that discourse studies (including corpus linguistics) has had on language descriptions and task design in published ELT materials, and highlights that 'the "corpus revolution" (Rundell and Stock 1992) has had a major and lasting impact on language learning materials in some respects' (Gilmore 2015: 511). We now have two major grammar reference books which are corpus based (Biber et al. 1999 and more recently Carter and McCarthy 2006), a range of course books, and many on-line materials. And it would be fair to say that not many, if any, publisher would dream of producing a dictionary which was not corpus based. This has meant that we are getting to a point where the spoken

language descriptions and examples that are available to learners are much more reflective of real language use, both in substantial terms and as influenced by considerations of frequency and context of use. Therefore it is important that teacher education programmes also have a stronger focus on spoken language corpora and language awareness to equip teachers with the relevant conceptual frameworks to appropriately mediate published language learning materials.

Small versus Large

Our arguments in terms of corpus size have not changed fundamentally and our original assertion 'whether to use a large, generalized corpus or a small, specialized corpus depends on the teacher educator's particular needs' (O'Keeffe and Farr 2003: 409) still holds true. When using larger corpora, it is easier to discern repeated patterns of use and collocation, but smaller corpora are extremely useful for examining a specialised genre (Flowerdew 2004). Small corpora are useful for introducing students to corpus techniques and methods, as they can sometimes feel overwhelmed by the sheer number of 'hits' they get for a particular search item in larger corpora. This is a more salient point amid today's world of billion word corpora. Small corpora can also allow the user to more readily access contextual information, where this is available, and this can be of great benefit if you are conducting pragmatic research (e.g. investigating the use of pragmatic markers in classroom discourse or power relations within spoken or written interactions (see O' Keeffe In press). Small contextualised corpora, which include a lot of information about the participants (i.e. metadata) are also of

those interested in the corpus-based investigation of sociolinguistic variables (for example, Farr and Murphy 2009, Murphy 2010).

Learner Corpora

Through research into learner language, CL has played a key role in maintaining the continuity of focus on learner interlanguage from the early 1990s. The pioneering International Corpus of Learner English (ICLE) project (Granger 2002) brought a new intensity to the study of interlanguage because of the possibilities it opened up not just for the large-scale study of interlanguage but also for contrastive analysis (Tono and Díez-Bedmar 2014). The main focus of early learner corpus research was on learner error analysis and, as noted by Gilquin (2008: 6), this allowed for systematic and contextualised analysis of learner language and enabled researchers investigate what learners get right, what they overuse and what they under use relative to native speakers.

Looking at current trends in learner corpora, we can say that they have enjoyed a relatively strong period of growth, and with this has come an expanded portfolio of use and usefulness (see Granger and Meunier 2015 for a comprehensive overview). This has come at the same time as wider acceptance of outer circle Englishes as a continuously expanding realm (Kachru 1992). In a very recent edited volume by Brezina and Flowerdew (2017), there are discussions of learner corpora for the investigation of learner language use (disagreement, self-repetitions, disciplinary genres, phrasal verbs, and figurative language). All of these investigations and their findings are likely to inform our discussions of variables on SLA and

classroom practice in teacher education contexts. More interesting though, this volume also contains a section of three chapters on task and learner variables: the effect of task and topic on opportunity of use (Caines and Buttery 2017); the effect of proficiency and background on phrasal verb usage (Marin-Cervantes and Gablasova 2017); the effect of a study abroad period (Götz and Mukherjee 2017).

1) Another important development in learner corpora in recent years, which has important implications for ELTE, is that they are increasingly basing their calibration of level on the Common European Framework of Reference for Languages (CEFR) as opposed to the learners' year of study (O'Keeffe and Mark 2017). The *English Profile Project* (www.englishprofile.org/) offers an important strand of new CEFR-based learner corpus research which is leading to pedagogical resource outputs. They are based on the Cambridge Learner Corpus (CLC), a 55 million word collection of learner exam writing, of which over 30 million words is error coded. Apart from the ongoing body of research, the main resource outputs are listed below. These should all be of application in ELTE programmes as they will help bring more precision to materials selection, design and grading (in terms of choosing the right material for a given cohort and designing an appropriate task):The *English Vocabulary Profile* (Capel 2010) - [www.englishprofile.org/wordlists] this corpus-based resource profiles the CEFR level at which learners can use a given word and its meaning. For example, using the searchable interface, it tells us that the homonym *bark* can be used by B2 learners in the sense of the sound that a dog makes but learners cannot use it metaphorically, as in *His boss was barking at everyone in the office,* until C2 level

and the sense the hard, outside part of a tree is also not acquired until C2 level. The resource can also be used to generate lexical sets around typical pedagogical topics (clothes, food, travel, etc.) and these can be used to design vocabulary lessons around lexical sets which are linked to the given level of the learners. Using this resource in ELTE, will bring attention to the importance of being aware of teaching level-appropriate vocabulary sets and it also underscores the importance of focusing on the polysemic nature of vocabulary where the most frequent 2000 words account for the up to 83% of coverage in a given text (see O'Keeffe, McCarthy and Carter 2007).

2) *Text Inspector* (2016) is a tool which uses corpus research into vocabulary acquisition across the CEFR, including EVP (as detailed above), and allows the user to copy and paste a text into its interface so as to instantly generate a lexical profile of that text. This will prove a very useful tool for ELTE as it allows for more precision in choosing texts that are appropriately pitched in terms of vocabulary level. This will be especially important when designing assessments so that texts that contain lexical items above the level of the test are not selected. Fig 18.2 illustrates a lexical profile of a short extract from Bram Stoker's *Dracula*, by way of example.

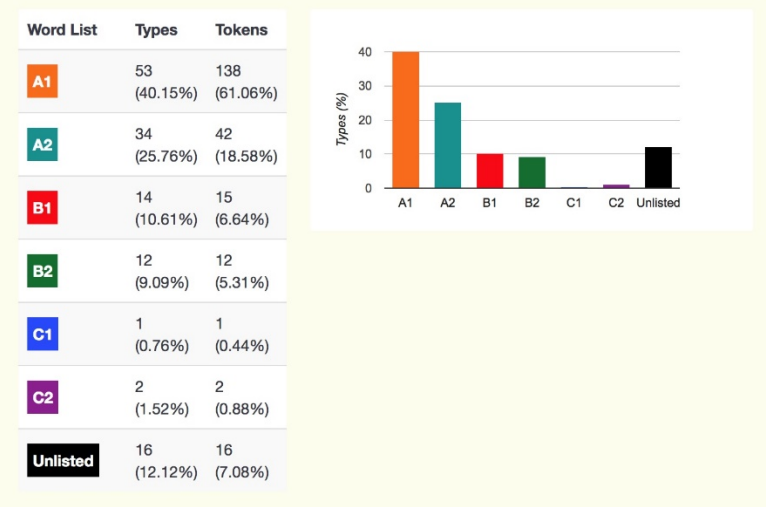

Fig 18.2 An example of how *Text Inspector* can be used to profile vocabulary of a given text, generated on www.englishprofile.org/wordlists/text-inspector (Text Inspector, 2017)

3) The *English Grammar Profile* (O'Keeffe and Mark 2017) - [www.englishprofile.org/english-grammar-profile] this resource profiles the grammatical competence of learners at each level of the CEFR. It outlines over 1,200 reference level descriptors, or 'can do statements' across the six competence levels. This resource marks a major shift of focus from learner error to learner competence. In the context of ELTE, this will prove an important reference point for syllabus and materials design in relation to grammar. It shows how learners acquire forms developmentally and how, as they grow in proficiency, they often continue to acquire new functions and pragmatic functions of syntax that they may have learnt at an early stage. An example of this is the Past Simple, which learners acquire at A1 level, in a limited way, but even at B2, they are still acquiring new uses, such as examples 1 and 2 below, which show learners at B2 using the Past Simple after *If* to show politeness:

[Extracts from the Cambridge Learner Corpus taken from the English Grammar Profile open online resource: www.englishprofile.org/english-grammar-profile/egp-online]

*It would be great if **you sent** me a free copy of the next edition of The Student Tourist Guidebook (B2, Korean).*
*I would be grateful if **you gave** me this job (B2, Greek).*

Hand Outs versus Hands On

The discussion here relates to the direct versus indirect (mediated or moderated) access to language corpora, the major difference being the time and level of technical and analytical skills needed by the user for direct access in a data-driven learning (DDL) context (as discussed further below). As Boulton (2012: 152) underlines 'one of the most apparent obstacles to DDL is the use of the technology itself - the computer with its query software and interfaces for accessing electronic corpora'. Extensive training in corpus literacy is generally required for direct access on a computer or other electronic device (see for example, Farr 2008, Heather and Helt 2012, Zareva 2017) to be able to do any meaningful analysis of the data. This needs to be considered in the context of the teacher education programme. Within our own institutional contexts, at MA level, we do integrate full direct access for our student teachers, and on some, but not all, of our shorter TESOL programmes at undergraduate level. Where direct access is not a viable option, 'the obvious question is whether the computer can be removed from the equation without losing the benefits of the overall approach' (Boulton 2012: 152), and in our own practice we use learning materials that are corpus-based or corpus-informed, and at the very least the student teachers interact with a simple on-line corpus interface to get a sense of what is possible. Bouton's study

compares a hands-on and hands-off approach with groups of language learners and found that, with some caveats, the paper-based option does represent a 'viable option'. We do remain true to our original contention in 2003 that even when adopting a direct access approach, it is usually very important to begin with some paper-based introductions in order to help develop conceptual frameworks before student teachers get distracted by the technology, overwhelmed by the volume of search return items, or diverted into statistical analyses that may be outside of their comfort zone. This sentiment is echoed in the research of Ebrahimi and Faghih (2017) on the reflections of MA TEFL students on their experience with using corpora for the first time.

Developing Technological Knowledge

Zareva (2017) reports on a questionnaire-based study of 21 TESOL trainee teachers enrolled on an MA Applied Linguistics, Teacher Education and Teacher Development programme in which she explored a number of research questions relating to the use of corpus linguistics within the programme. One of the key questions focused on the level of preparedness of the graduate students. Based on their self-reports within the surveys, only 43 per cent (n=9) knew what a corpus was and only 29 per cent (n=6) had ever used a corpus. To introduce a corpus literacy component to 'the already dense content of an English grammar course' (ibid: 76) meant using five class periods over the first ten weeks of the semester. Feedback showed that teacher trainees felt this was very worthwhile. In fact, when asked whether they would recommend removing the corpus component from the grammar modules, they unanimously agreed that it should not be removed. While they felt that the preparatory input prepared them well, they had feedback on what needs to be improved. Two of the three of these related to

technological knowledge: 1) dealing with technical difficulties (how to navigate a corpus, how to use wildcards in searches, restricting and filtering searches and getting to know the corpus interface), and 2) needing more hands-on practice in class. These concerns are consistent with other studies. Farr (2008), Heather and Helt (2012) and Ebrahimi and Faghih (2017) all cite user-related struggles with the technical side of using corpora, such as dealing with the interface, figuring out how it all works, interpreting findings. Interestingly, Zavera (2017) notes, the area that trainees found most challenging in the use of corpora within their English grammar module related to coming up with a project idea. In one sense, this is heartening as it is more a cognitive than a technical challenge. However, it points to an inter-related issue, namely the need for an understanding of the iterative process involved when using corpus interfaces to interrogate language. While technical processes and challenges can be explained and abated, an individual faced with a screen of concordance lines or a frequency list still has to find their own way in terms of following a querying pathway. In other words, the interactive processes of going from frequency lists to concordances, to sorting and resorting to follow up hypotheses about patterns of use is also a core skill. While on the one hand it requires a critical level of technical skill, it also requires an iterative thought process.

Developing language awareness (content knowledge)

In the aforementioned Zareva (2017) study, it is notable that one of the key themes from the teacher trainees who were surveyed on the use of corpus linguistics as part of their grammar modules was that they found that it heightened their awareness of grammar and that they saw the use of CL within this module as going hand-in-hand with acquiring a deeper cognitive

understanding of language as a system. One informant summed this up as follows (Zareva 2017: 75): 'It was a huge benefit to "see" the grammar we talked about in class being used in real-life language'. The same informant also noted the link to future practice through this statement: 'it was useful in understanding descriptive grammar and its relevance to ESL teaching' and s/he found an integrative benefit to the use of CL in the ELTE programme, 'it also offered a quantitative side to the class that helped reinforce the lectures.' These findings underscore those of Farr (2008). One of her informants reported that the use of CL on their MA programme helped to uncover 'functions and structures in language not found in grammar books' (ibid: 36). Farr also notes that the fact that corpora represent 'real language use in context' was viewed as an asset in their awareness of language process.

1) What is heartening about Zavera's (2017) study is that participants cited 43 advantages in all to learning how to use CL in their programme and when these are collapsed into three overarching themes in terms of their gain in expertise they bode very well for not just present gain in language awareness and content knowledge but to future application and continuing professional development within the practice:Language use and ESL teaching (74 per cent): they cited the many applications that they could envisage for using CL in their professional practice, including, *inter alia,* grammar teaching, academic vocabulary, vocabulary and collocations, developing writing skills, finding authentic examples, contrasting different language usage;
2) Research (16 per cent): for instance, they envisaged other opportunities to apply their new knowledge in other areas of their graduate studies;

3) Professional empowerment (10 per cent): often at a personal level, informants cited a sense of excitement and empowerment at about being able use their new skills to follow up their language queries and interests, summed up by the comment, ‘it opens up a whole new world to me’ (Zareva 2017: 75).

Developing pedagogical knowledge

In addition to the more well-established use of corpus approaches in language awareness contexts, there are a number of ways in which small, local or personally-constructed corpora can help facilitate improved pedagogic awareness of practice. Earlier research has described how the collection and analysis of corpora from the teacher education context can help us to understand it better (for example, the TP feedback context reported in Farr 2011, or the construction of teacher identity in personal narratives and shared discussions in online modes reported in Farr and Riordan 2015), and hence make informed decisions about how the process of teacher education can be made more effective. Such research has focused on how teacher educators can reflect on and improved their practices. The focus of the present discussion will be on how corpus-based evidence can be used to inform and improve the pedagogic practices of student and novice teachers undertaking a teacher education course. We situate this in the context of reflective practice (RP) as a framework for professional development, specifically as part of the practicum component of the course. Based originally on the work of Dewey and Schon (Farrell 2012), much has been written on the usefulness of RP in teacher education, and it has established itself firmly as a developmental tool (for example, Mann and Walsh 2017). One of the difficulties associated with it however is that it can be overly introflexive (looking inwards at oneself) and intuitive, and calls have been

made to use more evidence as a basis for change and progression, where necessary (Akbari 2007, Farr and Riordan 2017, Mann and Walsh 2013). To that end, we have implemented the use of the evidence-based PENSER (translates to think/reflect, from French) framework (Farr and Farrell 2017) at our home university. This RP framework allows the student teacher to use and analysis corpus-stored evidence of classroom interactions, compared with other genres, where relevant, to inform their reflections and plans for change. The PENSER model proposes the following five-stage framework for novice teachers to assist their development during the practicum:

1. **Problem identification**: individual challenges are identified and articulated through observing and reflecting on personal practice.
2. **Embracing**: these challenges are accepted as issues in need of further investigation, understanding and improvement and there is a commitment to these actions.
3. **Noticing**: experienced teachers are observed to facilitate a better understanding of the challenge and to provide examples of practice to be considered for assimilation.
4. **Solving**: a solution for personal change to practice is proposed, planned and implemented.
5. **Exploring and Researching**: the solution is investigated and critically evaluated to determine if the challenge has been appropriately overcome or if further engagement is needed.

At each stage in this process, appropriate corpora of classroom data can provide a source of evidence for the decisions and focus of the RP undertaken. Such corpora can be transcribed spoken language, or multi-modal to include video recordings (see Adolphs and Carter 2013 for details on building and using multimodal corpora). And while it would be far too time-consuming for an individual to build corpora of all of their own practice sessions, institutions can build up a library of such resources over time and make them available to their student teachers each year. At our home institutions we now have a range of classroom data in corpus format, and we build on this each year with contributions from students, teachers and researchers. Farrell (2015) provides good examples of how to explore individual awareness and use of regional and 'standard' language among cohorts of novice and experienced teachers as part of their professional development.

developing wisdom of practice

- Shulman (2004) developed the concept of teachers' Wisdom of Practice (WoP) to refer to the philosophical stances that frame their evaluation of what is going on in the classroom. In other words, their beliefs, their values, and their opinions, about language, how it is taught and how it is learned. This fits well with the work on teacher cognition that has triggered much understanding in the field of ELTE (for example, Borg 2006, Li Li 2016). Chappell (2017: 434) argues that 'finding ways to support teachers in articulating, interrogating, and developing their WoP is a powerful way to assist them in better understanding and developing their teaching practices'. He proposes a social and collaborative framework to develop teachers' WoPs (ibid: 441).

It involves the three stages of:Articulation: revealing the true nature of theoretical and philosophical stances

- Interrogation: comparison with other theoretical stances and comparison with actual classroom practice
- Change implementation: integrating the required innovations in practice to fit with the new theoretical orientations.

The use of corpus-data of classroom interactions, and other teacher-student interactions, can provide compelling evidence for stages two and three in this framework. Schön (1987) differentiates between espoused theories (what teachers say they believe) and theories in practice (what can reasonably be assumed to be a teacher's beliefs based on observing their practice), and using evidence from practice is one convincing way through which teachers come to realise any divergences or discrepancies between both, and initiate desired changes. In simple terms, this is about getting the walk to match the talk and vice versa.

Developing Pedagogical-Content Knowledge

Pedagogical-content knowledge is about manipulating, mediating and developing content, English language in this case, so that it becomes useful for learners and learning contexts. This can happen in two ways. Indirectly, when corpus research findings 'get through' (Gilmore 2015: 511-512) in the form of dictionaries, reference grammars, innovative approaches to textbook design or electronic resources as detailed above. There are not many corpus-based examples of all of these resources available to teachers, and although they have not been directly involved in their development, by preparing to teach and by mediating the

materials for their students they can develop their own awareness of how corpus-based findings relevantly inform and direct pedagogic materials and how they might best be exploited for this purpose. Teachers can become more aware of notions of frequency of use, context-differentiated language use, and collocational language use, when using materials that are based on evidence from corpus-based research.

In a more direct way, teachers can engage in their own materials' development, based on corpora. The use of smaller, often locally produced/gathered, specialised corpora is quite well-established to teach academic or professional English (for example, Flowerdew 2004). Teachers are almost always de facto materials' developers to some extend for their own students. The use of corpus materials with learners fits well with the SLA concept of 'noticing'. According to Mishan and Timmis (2015: 40):

> It is little surprise really that this field that depended on the 'noticing' of patterns (lexical, grammatical, discourse, pragmatic etc.) in corpus data conceived pedagogical applications that were a perfect 'fit' with this SLA theory. An early application was data-driven learning (DDL) (e.g. Johns 1991), probably the closest pedagogically to CL research itself, which saw learners mining raw data (in the form of concordances) in order to expose grammatical patterns from which they could infer rules of use.

Mishan and Timmis provide much more detail on the principles and use of DDL in Chapter 5 of the same volume on Materials Development in TESOL. In fact, a word search for 'corpus' in that volume produced 71 occurrences, an indication of the growing importance of corpus-data in the field of materials' development. And, as discussed earlier, while DDL

models, which allow learners to have direct access to corpora through computer interfaces may not suit all learners or contexts, teacher mediation of corpus materials can be equally effective (Boulton 2012).

Teacher As Researcher

As recently as August 2017, in the ELT Journal, Peter Medgyes returns to the relationship between teacher and research in a context where he states that, 'It is a fact that a very low proportion of practising teachers are in the habit of reading ELT-related research papers' (Medgyes 2017: 491), and goes on to discuss various reasons why this is the case, including the disparate nature of the two activities. From another angle, Borg (2009) identified a number of reasons why so few teachers are directly engaged in research activity (either reading or doing research), the most frequently-cited being time constraints and limited access to books and journal articles. Those who did engage, did so primarily as part of a professional programme of study. Like Paran (2017: 501) we strongly believe that 'intuitions and beliefs are not reliable when complex issues such as teaching and learning are concerned. This is where a research-oriented or evidence-based approach comes into play'. Teacher education programmes are precisely the appropriate place to engender this spirit of enquiry, or research, if research is taken in its broad meaning. It is here that prospective teachers learn the tools of research, such as appropriate methodologies, how to survey the relevant literature and make sense of it, how to collect and analyse data and how to disseminate and report findings to others, or simply to implement changes to one's local practice. And it is in this realm of a teacher education programme that there is a place for corpus-based methodologies and principles. They can be used to systematically explore practice in a more formal manner

than suggested in the previous two sections on pedagogic knowledge, or to explore language through a research lens. In addition, corpora provide the kind of data and evidence which are central to teacher development (Mann and Walsh 2017).

Hockly (2017: 364) suggests that

> the term 'ELT researcher' is a broad one, taking in experienced academic researchers who are au fait with qualitative or quantitative research methods on the one hand, to 'teachers as researchers' on the other, that is practising English language teachers interested in carrying out more informal or ad hoc action research based on their classroom practice.

Whichever is the case, collecting or collating a corpus for research purposes is a very legitimate way to explore language in use from a variety of contexts. Data is everywhere and it is now easier than at any other time in the past to build a corpus by scraping the internet. A couple of hours of effort can easily yield a sizeable corpus ready for analysis and investigation to answer a specified set of relevant research questions. There are however a range of complex ethical considerations around digital or online data gathering and use. Consent and anonymity are not as straight-forward as they might be in face-to-face contexts. The Association of Internet Researchers (AoIR) provide ethical guidelines which will help researchers steer clear of unethical pitfalls, inadvertent or otherwise.

Looking to the future

Returning again to our (2003) paper, we note our call at that time for the need to educate teachers who can manipulate language corpora for their own pedagogic ends, scrutinise and evaluate findings that are presented as 'facts', 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. In terms looking to the future, we can say that so much amelioration has occurred technologically to allow this to happen. Corpora and software are freely available, online supporting training materials for their use abound, and so on. The conditions seem ripe for a future where the integration of corpus linguistics in ELTE will be a given. Yet, despite all of these affordances, we are mindful of Ebrahimi and Faghih's (2017: 121) caveat in relation to current practice, that, 'the reality is that they [corpora] are still rarely used by language teachers'. Collating ELTE practice-based evaluative studies of the use of corpus linguistics on MA programmes play a key role in addressing why this is so. The work of Farr (2008), Heather and Helt (2012), Ebrahimi and Faghih (2017) and Zareva (2017), among others, is crucial to informing the way forward because they put corpora to the test in ELTE and invariably expose both the highs and lows of their use. Evidence from these studies shows that student teachers are aware of the challenges in terms of technical skills, time and application but they invariably seem to enjoy the use of corpora and to see the benefit of having corpus linguistics skills as part of their 'toolkit' (Timmis 2015).

The future of ELTE seems firmly set in the socialisation of teachers in a way that is dependent on a set of understandings (Freeman 2016), and the central argument of this chapter is that using corpus-based evidence in qualitative or quantitative ways can help to develop these

understandings in ways that are relevant, local and even personal to the individual. We have briefly illustrated the ways in which such an approach can enhance the various types of 'teacher knowledge' outlined in the introduction. We are not suggesting that this is the only valid approach, but it is one that we have found extremely effective when working with prospective teachers on our programmes, and one that can complement other tools and methodologies with ease.

Related topics

The language of the classroom/classroom discourse; Teacher knowledge and development; From language as system to language as discourse

KEY READINGS

Ebrahimi, A. and Faghih, E. (2017) 'Integrating Corpus Linguistics Into Online Language Teacher Education Programs', *ReCALL,* 29(1): 120-135. This paper reports on the intensive use of an online CL course within an MA TEFL in Iraq and gives insight into the benefits and challenges of this approach based on the reflections of 50 participants. It gives an interesting perspective on the challenges relating to using corpus linguistics in ELTE in a context where access to IT and internet is not a given or to use the authors' term, in technology-poor educational settings.

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.

This research thematically surveys the responses of 25 MA TESOL students to the use of corpora in their ELTE programme over a two-semester phase. The longer duration of this period of use is perhaps an important variable.

Farr, F. and Murray, L., eds. (2016) *The Routledge Handbook of Language Learning and Technology,* London and New York: Routledge. Part IV: Corpora and data-driven learning.

This section offers six chapters which provide practical and illustrated advice on the integration of corpus linguistics in language teaching contexts, most of which is also directly relevant to ELTE contexts.

Timmis, I. (2015) *Corpus Linguistics for ELT: Research and Practice*, London: Routledge.

This book offers a very practical curation of the essentials of using CL for ELT research and practice. It includes step-by-step guidance on how to use CL to query language in use as well as illustrating some very practical applications for ELT practice, for example how to use CL in English for Specific Purposes classes and materials design.

Zareva, A. (2017). Incorporating corpus literacy skills into TESOL teacher training. *English Language Teaching Journal,* 71(1): 69-79. Another paper which surveys the incorporation of CL in an MA programme. This offers insight into the technical and conceptual challenges that were involved for the MA student but is hearteningly positive in its overall appraisal.

REFERENCES

Adolphs, S. and Carter, R. (2013) *Spoken Corpus Linguistics: From Monomodal to Multimodal,* New York and London: Routledge.

Akbari, R. (2007) 'Reflections On Reflection: A Critical Appraisal Of Reflective Practices In L2 Teacher Education', *System,* 35: 2, 192-207.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English,* London/New York: Longman.

Borg, S. (2006) *Teacher Cognition and Language Education: Research and Practice.* London: Continuum.

Borg, S. (2009) 'English Language Teachers' Conceptions Of Research', *Applied Linguistics,* 30: 3, 358-388.

Boulton, A. (2012) 'Hands on / hands off: Alternative approaches to data-driven learning', in Boulton, A. and Thomas, J. (eds.), *Imput, Process and Product: Developments in teachiung and languagew corpora*, Blansko: Masaryk University Press, 152-168.

Brezina, V. and Flowerdew, L. (2017) *Learner Corpus Research: New Perspectives and Approaches,* Oxford: Bloomsbury.

Caines, A. and Buttery, P. (2017) 'The effect of task and topic on opportunity of use in learner corpora', in Brezina, V. and Flowerdew, L. (eds.), *Learner Corpus Research: New Perspectives and Applications*, London: Bloomsbury.

Caines, A., McCarthy, M. and O'Keeffe, A. (2016) 'Spoken language corpora and pedagogical applications', in Farr, F. and Murray, L. (eds.), *The Routledge Handbook of Language Learning and Technology*, London and New York: Routledge, 348-361.

Capel, A. (2010) 'A1 – B2 Vocabulary: Insights and issues arising from the English Profile Wordlists project', *English Grammar Profile Journal,* 1: 1, 1-11.

Carter, R. and McCarthy, M. (2006) *Cambridge Grammar of English: A Comprehensive Guide to Spoken and Written Grammar and Usage,* Cambridge: Cambridge University Press.

Chappell, P. (2017) 'Interrogating your wisdom of practice to improve classroom practices', *ELT Journal,* 71: 4, 433-444.

Ebrahimi, A. and Faghih, E. (2017) 'Integrating corpus linguistics into online language teacher education programs', *ReCALL,* 29: 1, 120-135.

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.

Farr, F. (2010) 'How can corpora be used in teacher education?', in O'Keeffe, A. and McCarthy, M. (eds.), *Routledge Handbook of Corpus Linguistics*, London and New York: Routledge, 620-632.

Farr, F. and Farrell, A. (2017) 'PENSER: A data-informed reflective practice framework for novice teachers', *The European Journal of Applied Linguistics and TEFL,* 6: 2, 85-103.

Farr, F. and Murphy, B. (2009) 'Religious references in contemporary Irish-English: 'for the love of God almighty….I'm a holy terror for turf'', *Intercultural Pragmatics,* 6: 4, 535-560.

Farr, F. and Murray, L., eds. (2016) *Routledge Handbook of Language Learning and Technology,* London and New York: Routledge.

Farr, F. and Riordan, E. (2015) 'Tracing the reflective practices of student teachers in online modes', *ReCALL,* 27: 1, 104-123.

Farr, F. and Riordan, E. (2017) 'Prospective and Practising Teachers Look Backwards at the Theory-Practice Divide Through Blogs and E-Portfolios.', in Farrell, T. S. C. (ed.) *TESOL Voices: Insider Accounts of Classroom Life. Preservice Teacher Education*, Virginia: TESOL, 13-26.

Farrell, A. (2015) 'In the classroom', in Farr, F. (ed.) *Practice in TESOL*, Edinburgh: Edinburgh University Press, 89-110.

Farrell, T. S. C. (2012) 'Reflecting on reflective practice: (re)visiting Dewey and Schön', *TESOL Journal,* 3: 1, 7-16.

Flowerdew, L. (2004) 'The argument for using specialized corpora to understand academic and professional language', in Connor, U. and Upton, T. (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics* Amsterdam: John Benjamins, 11-33.

Freeman, D. (2016) *Educating Second Language Teachers,* Oxford: Oxford University Press.

Gilmore, A. (2015) 'Research into practice: The influence of discourse studies on language descriptions and task design in published ELT materials', *Language Teaching,* 48: 4, 506-530.

Gilquin, G. (2008) 'Combining contrastive analysis and interlanguage analysis to apprehend transfer: detection, explanation, evaluation', in Gilquin, G., Papp, S. and Díez-Bedmar, M. B. (eds.), *Linking Up Contrastive and Learner Corpus Research* Amsterdam: Rodopi, 3-33.

Götz, S. and Mukherjee, J. (2017) 'Investigating the effect of the study abroad variable on learner output: A pseudo-longitudinal study on spoken German learner English', in Brezina, V. and Flowerdew, L. (eds.), *Learner Corpus Research: New Perspectives and Applications*, London: Bloomsbury.

Granger, S. (2002) 'A bird's-eye view of learner corpus research', in Granger, S., Hung, J. and Petch-Tyson, S. (eds.), *Computer learner corpora, second language acquisition and foreign language teaching.*, Amsterdam: John Benjamins, 3-33.

Granger, S. and Meunier, F., eds. (2015) *The Cambridge Handbook of Learner Corpus Research,* Cambridge: Cambridge University Press.

Heather, J. and Helt, M. (2012) 'Evaluating corpus literacy training for pre-service language teachers: Six case studies', *Journal of Technology and Teacher Education,* 20: 4, 415-440.

Hockly, N. (2017) 'Researching with technology in ELT', *ELT Journal,* 71: 3, 364-372.

Inspector, T. (2016) 'Online lexis analysis tool at textinspector.com [accessed 8.12.2017]'.

Johns, T. (1991) 'Should you be persuaded - two samples of data-driven learning materials', *English Language Research Journal,* 4, 1-16.

Kachru, B. (1992) 'World Englishes: approaches, issues and resources', *Language Teaching,* 25: 1, 1-14.

Mann, S. and Walsh, S. (2013) 'RP or 'RIP': A critical perspective on reflective practice', *Applied Linguistics Review,* 4: 2, 291-315.

Mann, S. and Walsh, S. (2017) *Reflective Practice in English Language Teaching,* New York and London: Routledge.

Marin-Cervantes, I. and Gablasova, D. (2017) 'Phrasal verbs in spoken L2 English: The effect of L2 proficiency and L1 background', in Brezina, V. and Flowerdew, L. (eds.), *Learner Corpus Research: New Perspectives and Applications*, London: Bloomsbury.

McCarthy, M. J. (1998) *Spoken Language and Applied Linguistics,* Cambridge: Cambridge University Press

Medgyes, P. (2017) 'The (ir)relevance of academic research for the language teacher', *ELT Journal,* 71: 4, 491-498.

Mishan, F. and Timmis, I. (2015) *Materials Development for TESOL,* Edinburgh: Edinburgh University Press.

Mishra, P. and Koehler, M. J. (2006) 'Technological Pedagogical Content Knowledge: A framework for teacher knowledge', *Teachers College Record,* 108: 6, 1017-1054.

Murphy, B. (2010) *Corpus and Sociolinguistics: Investigating age and gender in female talk,* Amsterdam and Philadelphia: John Benjamins.

Murphy, B. and Riordan, E. (2016) 'Corpus types and uses', in Farr, F. and Murray, L. (eds.), *The Routledge Handbook of Language Learning and Technology*, London and New York: Routledge, 388-403.

Naismith, B. (2017) 'Integrating corpus tools on intensive CELTA courses', *ELT Journal,* 71: 3, 273-283.

O' Keeffe, A. (In press) 'Corpus-based function-to-form approaches', in Jucker, A. H., Schneider, K. P. and Bublitz, W. (eds.), *Methods on Pragmatics*, Berlin: Mouton de Gruyter.

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 Mark, G. (2017) 'The English Grammar Profile of learner competence: Methodology and key findings', *International Journal of Corpus Linguistics,* 22: 4, 457-489.

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

Paran, A. (2017) ''Only connect': researchers abd teachers in dialogue', *ELT Journal,* 71: 4, 499-508.

Reppen, R. (2016) 'Designing and building corpora for language learning', in Farr, F. and Murray, L. (eds.), *The Routledge Handbook of Language Learning and Technology*, London and New York: Routledge, 404-412.

Römer, U. (2006) 'Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments', *Zeitschrift für Anglistik und Amerikanistik,* 54: 2. Special Issue: 'The Scope and Limits of Corpus Linguistics - Empiricism in the Description and Analysis of English' (ed. Volker Gast). XX, 121-134.

Rundell, M. and Stock, P. (1992) 'The corpus revolution', *English Today,* 30, 9-14.

Schön, D. (1987) *Educating the Reflective Practitioner: Toward a design for teaching and learning in the professionals,* San Francisco: Jossey-Bass.

Shulman, L. S. (1986) 'Those who understand: Knowledge growth in teaching ', *Educational Researcher,* 15: 2, 4-14.

Shulman, L. S. (1987) ‘Knowledge and teaching: foundations of the new reform’, *Harvard Educational Review,* 57, 1-22.

Shulman, L. S. (2004) *The Wisdon of Practice: Essays on Teaching, Learning, and Learning to Teach,* San Francisco, CA: Jossey-Bass.

Timmis, I. (2015) *Corpus Linguistics for ELT: Reserarch and Practice,* New York and London: Routledge.

Text Inspector (2017) *Online lexis analysis tool at textinspector.com* [accessed 08.12.2017]

Tono, Y. and Díez-Bedmar, M. B. (2014) ‘Focus on learner writing at the beginning and intermediate stages: the ICCI corpus’, *International Journal of Corpus Linguistics,* 19: 2, 163-177.

Zareva, A. (2017) ‘Incorporating corpus literacy skills into TESOL teacher training’, *ELT Journal,* 71: 1, 69-79.