21

What can a corpus tell us about language teaching?

Winnie Cheng and Phoenix Lam

1 Corpora and language teaching

Corpora and language teaching can best be introduced by Fligelstone's (1993) three-tiered model of teaching about corpora, teaching to exploit corpora and exploiting corpora to teach. Teaching about corpora refers to teaching corpus linguistics as an academic subject, teaching to exploit corpora introduces students to different methods and tasks to exploit corpora for learning purposes and exploiting corpora to teach involves using 'a corpora-approach to inform teaching' (Huang 2018: 383). Renouf (1997) adds a fourth tier, which is teaching to establish resources, and this involves learner participation in corpus data collection, corpus design and corpus compilation.

Another way of looking at the relation between corpora and language learning is Leech's (1997) distinction between indirect and direct corpus applications in pedagogy. While indirect corpus applications mostly involve corpus-based studies informing syllabus design, material writing and creation of reference tools, such as wordlists, dictionaries and grammars, direct corpus applications involve teachers and learners working with corpora in the classroom; direct access to corpora by learners spreads along a deductive-inductive cline, and at times the inductive and deductive approaches are combined in practice.

A form of direct corpus application is "data-driven learning" (DDL), proposed by Johns (1991), in which language learners can be simultaneously active learners and language researchers accessing corpus data directly. DDL is 'a learner-focused approach which promotes learners' discovery of linguistic patterns of use and meaning by examining extensive samples of attested uses of language' (Pérez-Paredes *et al.* 2019: 145). It involves 'using the tools and techniques of corpus linguistics for pedagogical purposes' (Gilquin and Granger 2010: 359). Gilquin and Granger (2010) summarise a few advantages of the DDL approach. These include bringing authenticity to the classroom, serving a major corrective function when learners compare their own writing with a corpus of expert writing or an error-annotated learner corpus and offering discovery learning which is potentially motivating and fun. O'Sullivan (2007: 277) proposes the potential acquisition or refining of a range of micro-cognitive skills through the DDL

DOI: 10.4324/9780367076399-21

approach, including ‘predicting, observing, noticing, thinking, reasoning, analysing, interpreting, reflecting, exploring, making inferences (inductively or deductively), focusing, guessing, comparing, differentiating, theorising, hypothesising, and verifying’.

Recent research using DDL in language teaching and learning finds that it is generally effective and efficient (Boulton and Cobb 2017; Lee *et al.* 2019; Pérez-Paredes 2019) and receives positive feedback from learners (Chambers 2019). The use of corpora in DDL in English for academic purposes (EAP) has increased significantly in the past ten years (see Chapters 24, 28, 29 and 30, this volume), but few studies have examined the use of corpora by doctoral students during their thesis writing process (Charles 2018) and by secondary school students (Boulton and Pérez-Paredes 2014).

Still another way of looking at corpora and language teaching is the different ways in which teachers and learners, acting as corpus researchers, exploit corpora. Meunier and Reppen (2015), citing Tognini-Bonelli (2001), compare a corpus-driven and a corpus-based approach. The corpus in a corpus-driven approach ‘serves as an empirical basis from which researchers extract their data and detect linguistic phenomena without [too many] [*our addition*] prior assumptions and expectations (Tognini-Bonelli 2001)’, with the conclusions being ‘drawn exclusively on the basis of corpus observations’; in contrast, in a corpus-based approach, ‘linguistic information (frequencies, collocations, etc.) is extracted from a corpus to check expectations or confirm linguistic theories’ (Meunier and Reppen 2015: 499).

In addition, Meunier and Reppen (2015) describe a “corpus-informed” approach, which comprises the following features:

- The inclusion of results, conclusions and discoveries from research carried out on a variety of corpora (e.g. native or learner corpora, spoken or written, from different genres, produced by expert or novice writers/ speakers);
- The selection of what exactly should be included (e.g. structures, vocabulary, contexts of use, collocational and colligational patterns, frequency);
- The decisions linked to the presentation of the corpus information (e.g. text, graphs, concordances, data-driven approach, other);
- When the materials focus on skills, the selection of suitable texts (oral or written) as a prompt for instruction.

(Meunier and Reppen 2015: 499)

In Poole’s (2016) study conducted in a university in the United States, L2 writing was taught to develop ‘rhetorical awareness and the understanding of the interrelation between language, rhetorical purpose, and context’ (p. 101). A ’corpus-aided approach’ was adopted, where it was the teacher who prepared the corpus data and materials, and students did not ‘use a corpus program, perform a corpus query, or generate a single concordance line’ (p. 103). Instead, the teacher mediated between the corpus materials and the students and assisted in the contextualisation of findings.

As noted by Chambers (2019), in the past decade, a number of studies have reviewed the development of the field of corpora and language learning, in particular DDL. Examples are narrative reviews of different topics and methods and the history of development (e.g. Boulton 2017); two meta-analyses of empirical, quantitative studies of DDL (Cobb and Boulton 2015; Boulton and Cobb 2017); one meta-study of corpora and vocabulary acquisition (Lee *et al.* 2019); one review of the uses and spread of corpora and DDL in CALL research (Pérez-Paredes 2019); and special issues on the role of corpora in two main journals in CALL, namely *The Journal of the European Association for Computer Assisted Language Learning* (ReCALL) (2014) and *Language Learning & Technology* (LLT) (2017). Callies and Götz (2015) point out that the use of corpora and corpus linguistics tools and methods has proven to be beneficial, especially for assessing L2 proficiency; however, such use is relatively new in the area of language testing and assessment.

The remainder of this chapter reviews the literature in three areas related to corpora and language teaching. First, corpus methods in language teaching are concerned with what corpus linguistics concepts, tools and techniques are involved in or applied to language teaching. Second, corpus evidence as teaching materials focuses on indirect applications or use of corpora in producing teaching and research resources such as dictionaries, grammars, course books and vocabulary lists. Third, corpus tasks for language teaching focus on direct applications or use of corpora by teachers, as well as learners, for a range of pedagogical purposes. The chapter ends with a discussion about future directions and areas for further research and practice.

2 Corpus methods in language teaching

Corpus methods in language teaching are related to basic corpus linguistics concepts and corpus tools and techniques that are involved in, or applied to, language teaching. They also beg the question of some key issues that teachers should consider when applying corpus methods to language teaching.

A review of recent research shows that the main corpus methods applied in language teaching are concordancing and word frequency counting. Hyland (2003) describes two uses of concordancing, namely a research tool for L2 writers to systematically investigate a specific linguistic item or phenomenon and infer underlying rules and a reference tool for L2 writers to consult to find immediate solutions to linguistic problems encountered when they are writing. Concordancing is performed by a concordancer, which is a typical tool for language learners to get access to a corpus and provide learners with a variety of language learning affordances (Flowerdew 2015) (see Chapters 9, 10 and 14, this volume).

Research in corpora and language teaching has examined different uses and purposes of concordancing. Huang (2014) discusses the use of concordancing for Chinese university students majoring in English to learn about the collocational and colligational patterns of abstract nouns. In Boulton and Cobb's (2017) meta-analysis of 64 experimental and quasi-experimental quantitative studies which examine the effectiveness of using corpus linguistics tools and techniques for second language learning or use, DDL is found to be most effective with a hands-on concordancer, with some key advantages including exposure to authentic language, identification of common patterns of language and promotion of learner autonomy. By contrast, Lee *et al.* (2019) find that purposefully curated concordance lines, rather than a hands-on concordancer, are more effective for learners and when learning materials and hands-on corpus practice are arranged concurrently. It should be noted, however, that Lee *et al.* (2019) only focus on vocabulary and measure the effect size through a quantitative approach different from that adopted in Boulton and Cobb (2017).

Learner concordance use, according to Charles (2018), is the focus of most DDL work, although most corpus software also offers other tools such as clusters, collocates, n-grams, concordance plots, wordlists and keyword lists, which tend to be under-used in

DDL. Charles (2018) highlights that different tools in *AntConc* allow learners to address different learning issues. Wordlists, n-grams and keyword lists, for example, require no user input and thus can identify potentially problematic issues which are unknown to the learners. For concordances, clusters, collocates and concordance plots, issues or problems which are already known to the learners can be addressed. Charles (2018) also explains the ways in which these tools can be used for editing purposes at the levels of content and organisation. Of all the *AntConc* tools introduced to the doctoral students, concordance was rated most highly, followed by clusters, collocates and keyword lists. It is argued that 'attention to the affordances of all available corpus tools is needed if corpus pedagogy is to realise its full potential as a valuable approach for language learning' (Charles 2018: 24).

Sha (2010) describes the characteristics of a search engine to be used in DDL, as follows:

- Should be capable of providing as many authentic usages and expressions as needed;
- Should be simple to use; the user does not have to learn complicated query syntax;
- Should guarantee a high search speed;
- Can be simultaneously used by thousands of students;
- Should require no registration or client installation;
- Should be cost-effective in the long run.

(Sha 2010: 382)

The student teachers in Ebrahimi and Faghih's (2016) study found that some corpus tools are more useful for language teachers and researchers than for learners. For *Lextutor*, the text-based concordances and n-gram phrase extractor tools are considered too technical for learners, whereas the Vocabprofile tool is considered useful for teachers for analysing learners' writing. For *AntConc*, while respondents felt that it could assist teachers in correcting learners' writing, they found it suitable only for adult learners with a high level of English proficiency.

In Yoon's (2016) study of concordancers and dictionaries as problem-solving tools for English as a second language (ESL) academic writing, a reference suite (RS) was developed. The RS is a mini web browser that allows free access to five concordancers and three types of dictionaries. The concordancers are *Corpus of Contemporary American English*, *Google* search engines, *Google Scholar* (GS), *Custom Search Engine* (CSE) and *JustTheWord* (JTW). The dictionaries are *Naver* (an online bilingual Korean-English/English-Korean dictionary), LDOCE (an online version of the *Longman Dictionary of Contemporary English*) and Thesaurus (Roget's 21st Century Thesaurus). The purpose of Yoon's (2016: 212) study was to examine 'the potential of the reference suite as a cognitive tool that extends the cognitive powers of L2 writers and mediates their problem solving while writing'. Research data were collected by getting students to record 'their writing processes using screen capture software' (p. 215) while working on the assignment, followed by 'a stimulated recall session' (p. 216). Results from the study showed that the reference suite was indeed effective as a cognitive tool, especially for helping learners with lexical and grammatical problems, though different learning goals and needs had to be taken into account to capitalise on its use.

Research has also examined the use of multimodal corpora and multimodal resources and tools for DDL. Meunier (2020) presents useful examples of tools to be used for DDL, including some multimodal resources (e.g. *PlayPhrase.me, LyricsTraining*). A new development has taken place combining ‘mobile-assisted language learning (MALL)’ and DDL (Pérez-Paredes *et al.* 2019: 145). As observed by Pérez-Paredes *et al.* (2019), despite the affordances of individualisation and personalisation of MALL, an integration of MALL and DDL has not yet been widely explored. The researchers conducted an evaluation study involving the creation and use of a self-created mobile language learning app, with the objective of exploring ‘the opportunities and challenges of mobile DDL for language learners, teachers, and developers’ (p. 148). Participants were learners of English, German and Spanish aiming to achieve an A2 or a B2 proficiency level, with reference to the Common European Framework of Reference (CEFR or CEF). The app was designed to improve the learners’ writing skills by ‘offering context-driven information through word frequency and vocabulary analysis’ and to further help improve writing by ‘providing lexical alternatives’ (p. 148). The app used freely accessible natural language processing (NLP) tools, namely *Lextutor* for English (Cobb 2003), with ‘a text analyser, a vocabulary profiler, and a part-of-speech tagger’ (p. 148). After a learner’s text had been analysed, improvement could be made by exploring corpus-based reference tools, namely the *Collins Dictionary*, *Netspeak* and *Stands4*, in the areas of ‘definitions, synonyms, and example sentences’ (p. 148).

While Ballance (2016) observes a long history of using computer-generated concordances for language learning, Ballance (2017) notes that concordancing has not been widely used in mainstream language learning contexts, possibly due to difficulty in interpreting the short, truncated keyword in context (KWIC) format, insufficient teacher training in this area, limited access to technological resources and conflict between the cognitive demands of concordance use and language learning. Chen and Flowerdew (2018) describe criticisms of corpus applications in language classrooms, including the decontextualisation of truncated concordance lines and the amount of time investment.

With regard to collocations (see Chapters 9, 14 and 15, this volume), they can be a challenging area for L2 learners because they contain ‘some element of grammatical or lexical unpredictability or inflexibility’ (Nation 2001: 324). Ackerman and Chen’s (2013) study shows that the productive use of collocations is particularly challenging, and learners make mistakes concerning collocation in translation, rely on only a small number of collocations and use inappropriate synonyms. These issues may arise from L1 influence and require a high level of collocational competence for them to be addressed (Ackerman and Chen 2013).

3 Corpus evidence as teaching materials

The indirect applications of corpora in pedagogy are most common in the creation of dictionaries and grammars, and to a lesser extent in the design of course books and other supplementary materials. It has been remarked that the use of corpora in these teaching resources is so normalised that users may not actually know what a corpus is (Frankenberg-Garcia 2012) or may be unconscious of the role of corpora as a technology in motivating the paradigm change from a rule-based to an evidence-based approach to language teaching (Chambers 2019). In comparison, the influence of corpora in syllabus design and language assessment has not been well-documented thus far, though a small number of examples have suggested promising potential in these areas, such as McCarten and McCarthy (2010) (for more on the influence of corpora in syllabus design, see Chapters 25 and 26, this volume).

In dictionary making, the incorporation of corpus evidence has become mainstream since the first corpus-based dictionary, the *Collins Cobuild English Dictionary* (Sinclair 1995), was produced. Today, corpora are an indispensable item in a dictionary compiler's toolkit, and many major publishers have their own in-house corpora for this purpose. The use of corpora in dictionary-making provides important contents for entries such as frequency information, real-world examples and patterns of use (see Chapter 27, this volume). As such, corpora used in dictionary compilations are often very large in size. The *Macmillan English Dictionary*, for example, makes use of a general corpus now containing almost 1.6 billion words of written and spoken English as a basis for language description (Rundell 2020). Its online counterpart, MacmillanDictionary.com, is based on the World English Corpus, which is composed of 220 million words of written and spoken text from a variety of social and geographical contexts (Macmillan 2020).

In addition to general corpora, specialised corpora have increasingly been built and used to provide more specific empirical analyses, as well as to identify areas of interest for target audiences. The *Longman Language Activator*, produced in 1993, was the first dictionary to draw on the analysis of learner corpus data for its design. More recently, the *Cambridge Learner Corpus*, as part of the *Cambridge International Corpus*, is a 50-million-word collection of anonymised exam scripts produced by learners of English worldwide, which has been used for describing common problems in learner dictionaries compiled by the publisher (Cambridge University Press 2020). Another teaching-oriented corpus applied to dictionary making is the *Macmillan Curriculum Corpus*, which consists of 20 million words collected from school textbooks and examination syllabuses and has been used in the production of the *Macmillan School Dictionary* and *Macmillan Study Dictionary* (Rundell 2020). Corpus evidence has also been applied to the making of bilingual dictionaries and dictionaries of other languages, including the *Oxford-Hachette French-English Dictionary* and the *Frequency Dictionary of Czech* (Frekvencnı Slovnık Cestiny).

Like dictionaries, almost all grammars are to some extent corpus-based (Boulton and Pérez-Paredes 2014). Examples of major reference grammars are the *Longman Grammar of Spoken and Written English (LGSWE)* (Biber *et al.* 1999) and the *Cambridge Grammar of English* (Carter and McCarthy 2006), both of which provide descriptions of the use of English in both its written and spoken forms instead of simply focusing on the written language, as was customary in traditional grammar (see Chapters 16 and 25, this volume). Studies have shown that grammar books not based on corpus findings do not reflect authentic language use. Meunier and Reppen (2015), for example, demonstrate that crucial information on the passive voice is missing in non-corpus-informed ELT grammar materials. They argue that the description of some language features would particularly benefit from the incorporation of corpus findings, including the passive, the conditionals, relative clauses and aspect.

In language course book production, the role of corpora seems more contentious (see Chapter 26, this volume). While McCarten and McCarthy (2010: 13) remark that an ELT course book without corpus evidence is 'conspicuous' and cite a number of examples influenced by the use of corpora, including *face2face* and *Objective First Certificate*, Boulton (2010: 537) comments that corpora are often made 'invisible' in the presentation of course book content, even when the most well-known corpus-based English language course book example, the *Touchstone* series (McCarthy *et al.* 2005), is concerned. While its use of corpora may not be readily visible, the *Touchstone* series, together with the *English Vocabulary in Use* series (McCarthy and O'Dell 2008), has made the crucial move to incorporate common error warnings into the course book and to design tasks/activities based on error information identified from error-tagged learner corpora. A large degree of inconsistency, however, exists regarding the views and experiences of corpus use in course books. Although a number of empirical studies have reported that course books designed based on intuition do not mirror real-world language use, especially in relation to spoken language (e.g. Cheng and Warren 2006; Römer 2006), it remains unclear in the case of many language textbooks the extent to which and the ways in which they are corpus-informed, with only a minority of teaching materials thus far incorporating direct corpus evidence such as concordances into their design on a small scale (Boulton 2010) (see also Chapters 25 and 26, this volume). There is, however, a more promising development in this area in recent years, with the influence of two key corpus-based open resources on course book design. Both derived from the English Profile project, the English Vocabulary Profile (Capel 2010) and the English Grammar Profile (O'Keeffe and Mark 2017) are free online reference sources based on the *Cambridge Learner Corpus*. Importantly, the two resources offer valuable information concerning the typical lexical and grammatical profiles of learners at each level of the Common European Framework of Reference (CEFR or CEF), making it possible for course book writers to design evidence-based materials for language learners at different stages.

Finally, corpora have also found their way into the development of study lists, including the Academic Word List (Coxhead 2000); the Academic Keyword List (Paquot 2010); the Phrasal Expressions List (Martinez and Schmitt 2012); the Academic Collocation List (Ackerman and Chen 2013); the Academic Vocabulary List (Gardner and Davies 2014); and, most recently, the Oxford Phrasal Academic Lexicon (Oxford University Press 2020) (see Chapters 24 and 28, this volume). All of these lists have been produced with pedagogic purposes in mind and constitute examples of corpus-based teaching resources for testing and syllabus design. The Academic Collocation List, for example, was compiled based on the *Pearson International Corpus of Academic English* (PICAE) and designed for advanced learners of English. As advised by the authors, however, the list still requires teacher intervention for pedagogical use (Ackerman and Chen 2013). From the perspective of English as an international language (EIL), Flowerdew (2012) argues that teachers should be mindful of the variety of English on which teaching materials are based, as the so-called "modernisation" model of curriculum development has been criticised for the reason that 'Western models are applied by Western experts to Outer- and Expanding-Circle contexts' (Flowerdew 2012: 235). More information about the composition of corpora used in teaching materials will therefore allow teachers to make more informed decisions in this regard.

4 Corpus tasks for language teaching

The direct applications of corpora in language teaching involve the use of corpus tools and techniques in tasks for pedagogical purposes. In comparison with the indirect uses of corpora in pedagogy, the direct uses of corpora are regarded as being given a marginal treatment in the relevant literature (Leńko-Szymańska and Boulton 2015), with a slow development of the direct access to corpora by teachers and learners and of the use of corpus data in the classroom (Chambers *et al.* 2011). Outside the classroom, it has also been noted that learners' use of corpora is quite rare (Chen and Flowerdew 2018), though a number of studies have demonstrated positive responses from both trainee and in-service teachers regarding their perceptions of DDL and corpus-based instruction in language teacher education (Chen *et al.* 2019) (see Chapter 32, this volume).

Indeed, a large body of research has shown how corpus tasks can be designed to teach a variety of aspects of language, often with detailed illustrations of the tasks. The special issue of *Language Learning & Technology* devoted to corpora in language learning and teaching comprises articles which investigate the effect of DDL on different aspects of language teaching and learning, including phraseology, genre, collocation, lexico-grammatical knowledge and reading speed (Vyatkina and Boulton 2017). The most well-known corpus task is perhaps the classic DDL tasks, which involve the presentation of a concordance, either in paper or computer format, with a set of guiding questions for learners to identify patterns and make generalisations based on the patterns identified (Ballance 2016). Cotos (2014), for example, illustrates how such a discovery-oriented task based on teacher-selected examples can help in the teaching of the semantic roles, forms and syntactic distribution of linking adverbials. Another common task type is awareness-raising activities, which can be used for introducing a particular form or function. Chambers *et al.* (2011), for example, describe a task achieving this aim which focuses on the discourse functions of *right* through concordance lines to improve EFL students' conversation interactional strategies. Gablasova and Brezina (2018) also illustrate with an exercise how to heighten students' awareness of the linguistic realisations of disagreement with a transcript from corpus data rather than with concordances. Crosthwaite *et al.* (2019) provide examples of corpus pedagogic tasks, most of which involve concordance use, including gap-filling, sentence completion and frequency observation, as supplementary materials for their study, which provide useful examples for teachers of postgraduate thesis writing.

To facilitate the direct use of corpora in language teaching, a number of corpus tools and interfaces have been made available. Of the resources which are publicly accessible, *Lextutor* is one which has been used in many studies reporting the direct use of corpora by learners in EAP writing classrooms (Chen and Flowerdew 2018). As a contemporary pedagogic resource, *Lextutor* consists of a collection of corpora with a built-in concordancer accompanied by language learning activities. It is, however, not only used for teaching academic writing to more advanced learners, as two corpora on *Lextutor*, the 1K and 2K graded corpora, are considered particularly useful by student teachers, but its "multi concordance" program is also found to be valuable for teaching near synonyms (Ebrahimi and Faghih 2016). Another interface which provides access to data from a number of corpora as well as corpus-based resources is corpus.byu.edu (Davies 2020). Other corpus-based teaching resources include the Academic word highlighter and Check My Words, which have been recommended for teachers specifically targeting the four skills (Timmis 2015). For the enhancement of language awareness, the *Scottish Corpus of Texts & Speech* (SCOT) allows the study of interactional features by native speakers in non-standard varieties of English. Anderson and Corbett (2009) give two specific examples of how material from the corpus can be used in a classroom setting to teach aspects of evaluative language and linguistic forms used for particular pragmatic functions in standard and local English as a lingua franca (ELF) varieties. Examples of pedagogic corpora in other languages include the SACODEYL corpus and the BACKBONE corpora, both of which contain interview data of native speakers of a number of European languages and transcripts, coupled by learning resources in the form of exercises.

For more advanced learners, especially in the fields of EAP and ESP, the creation of corpora either by teachers or learners based on one's own data has been advocated. Smith (2020) describes how the online corpus tool Sketch Engine can be used to generate corpora from teaching materials such as lecture notes, PowerPoint slides and test papers to create personal vocabulary portfolios for students. Chen *et al.* (2019) point to the availability of existing software tools for teachers and learners to automatically build large-scale, discipline-specific corpora. With the development of big data, an increasing number of studies also explore the use of search engines as a corpus tool, particularly concerning the use of Google searches (Sha 2010; Yoon 2016).

When corpus tasks are designed for pedagogical purposes, a number of issues need to be taken into account. Chambers *et al.* (2011), for example, suggest that teachers should consider which type of corpus to use and what information to add or remove from corpus texts to enhance readability. Clearly these decisions have to be made in relation to the level of learners. Boulton (2010), for instance, argues for the elimination of technical issues concerning DDL tasks by using paper-based corpus materials with less advanced language learners, as such materials are generally more teacher-led and may provide more guidance and support that these learners need. Flowerdew (2015), by contrast, shows that postgraduate students preparing for their thesis writing prefer a "high-tech" search engine interface and hands-on tasks to printed concordance output. Response to corpus-based tasks may also vary according to learning style. As suggested by Bridle (2019), inductive learners are more likely to rate corpus-based activities more positively, while reflective learners are not receptive to corpus consultation and consider it time-consuming and overwhelming. Pérez-Paredes (2010) also proposes that corpora should be pedagogically annotated and corpus texts should be carefully chosen based on the learners' learning context and proficiency level. For do-it-yourself (DIY) corpora, teachers may need to seek further support in the forms of 'refresher sessions, drop-in clinics, and online on-demand courses' (Charles 2014: 39). With thoughtful decisions and appropriate strategies, direct corpus use in pedagogy can be successfully implemented so that it is motivational and effective to language teaching.

5 Bridging corpus linguistics and language teaching

Bridging corpus linguistics and language teaching involves describing future directions and areas for further research and practice. The literature review clearly demonstrates that different approaches to the use of corpora in language teaching have been advocated and applied in various educational contexts. It also shows a range of corpus methods; that is, corpus linguistics concepts, tools and techniques, applied in language teaching, involving a variety of disciplines in different classroom settings. The review also presents a variety of corpus evidence as teaching materials involving primarily indirect applications or use of corpora in producing teaching and research resources, such as dictionaries, grammars, course books, and vocabulary lists, as well as a range of corpus tasks for language teaching involving primarily direct applications or use of corpora by teachers and learners for a range of pedagogical purposes.

Despite the fact that corpus linguistics and DDL have a relatively long history in the field, with many of the publications deriving from the COBUILD project in the late 1980s and 1990s specifically aimed at learners, as Römer (2006) notes, corpus linguistics and its applications have yet to become mainstream in language teacher education programmes or in language teaching. Römer (2006: 122) concludes that there is 'strong resistance towards corpora from students, teachers and materials writers'. Similarly, more recently, Boulton (2017) comments that DDL is still 'a marginal practice' (p. 1).

To date, a range of issues related to corpora and language teaching have been observed, and recommendations for future directions and areas for further research and practice have been made by corpus linguistics researchers and educational practitioners. According to Burton (2012), non-native speaker corpora may play a role in future course book production. Leńko-Szymańska and Boulton (2015) note that future directions in the field may include treating the internet as a corpus and using search engines such as Google as a concordancer. They compare the indirect and direct uses of corpora, remarking that 'the direct uses of corpora in language teaching are treated rather marginally in the literature in the field' (p. 3). Götz and Mukherjee (2019) note that to date, there is still little contribution of learner corpora to second language acquistion (SLA) and foreign language teaching. Crosthwaite (2020) observes that the use of DDL with pre-tertiary learners is rare, with a small number of studies found in the high school setting and none in a primary school context.

Chen and Flowerdew (2018) report on a critical review of research and practice in DDL in the EAP/academic writing classroom since 2000, based on 37 empirical studies. They conclude that the field of DDL in EAP/academic writing is 'still in its infancy' (p. 356) and make five recommendations for future research and practice, as follows:

> i. More descriptions of different approaches in different geographical and classroom contexts;
> ii. More such studies need to be carried out describing how corpora and more traditional teaching approaches could be combined in overall writing instruction;
> iii. Researchers and practitioners need to come up with ways of helping students to become more autonomous in their use of corpora;
> iv. [More resources][*our addition*] to show writers how to use corpus tools to identify problems in their writing without the aid of the teacher;
> v. More research into specialist training in DDL for EAP practitioners.
>
> (Chen and Flowerdew 2018: 356–7)

Based on the conclusion of the mobile DDL evaluation study, Pérez-Paredes *et al.* (2019) make suggestions in the areas of task design, such as 'the addition of further built-in tools and adaptation to hardware constraints'; specialised learner training; the creation of 'more fleshed-out tools'; and future studies to investigate 'the potential of combining DDL and MALL' (p. 145).

The concrete recommendations and suggestions presented here show the significant benefits corpus use can offer to language teaching. With continuous technological advancement and pedagogical development through the concerted effort of corpus linguists and language teachers, the research-practice gap between corpus linguistics and language teaching will be narrowed, enabling the two areas to be fruitfully bridged.

Acknowledgements

The authors are members of the English Department's Research Centre for Professional Communication in English (RCPCE), the mission of which is to pursue applied research and consultancy so as to deepen the understanding of professional communication and

to better serve the communicative needs of professional communities. The work described in this chapter is intended to fulfil part of this mission.

The authors are very grateful for the reviewers' insightful and helpful suggestions.

Further reading

Cheng, W. (2012) *Exploring Corpus Linguistics. Language in Action*, London: New York: Routledge. (This book provides a practice guide to core theories and concepts in corpus linguistics with classroom examples, corpus-based analyses and tasks for learners and teachers.)

Reppen, R. (2010) *Using Corpora in the Language Classroom*, Cambridge: Cambridge University Press. (This book describes corpus-based materials, online corpus resources and example activities for use in the classroom.)

Timmis, I. (2015) *Corpus Linguistics for ELT: Research and Practice*, London: New York: Routledge. (This book introduces corpus linguistics to English language teachers and helps make it a regular part of a teacher's toolkit. It is designed to help ELT teachers be familiar with basic concepts in corpus linguistics and using corpora through three kinds of activities: corpus search, corpus question and discussion.)

References

Anderson, W. and Corbett, J. (2009) 'Teaching English as a Friendly Language: Lessons from the SCOTS Corpus', *ELT Journal* 64(4): 414–23.

Ackerman, K. and Chen, Y.-H. (2013) 'Developing the Academic Collocation List (ACL) – A Corpus-Driven and Expert-Judged Approach', *Journal of English for Academic Purposes* 12(4): 235–47.

Ballance, O. (2016) 'Analysing Concordancing: A Simple or Multifaceted Construct?', *Computer Assisted Language Learning* 29(7): 1205–19.

Ballance, O. (2017) 'Pedagogical Models of Concordance Use: Correlations between Concordance User Preferences', *Computer Assisted Language Learning* 30(3–4): 259–83.

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

Boulton, A. (2010) 'Data-Driven Learning: Taking the Computer out of the Equation', *Language Learning* 60(3): 534–72.

Boulton, A. (2017) 'Research Timeline: Corpora in Language Teaching and Learning', *Language Teaching* 50(4): 483–506.

Boulton, A. and Cobb, T. (2017) 'Corpus Use in Language Learning: A Meta-Analysis', *Language Learning* 67(2): 348–93.

Boulton, A. and Pérez-Paredes, P. (eds) (2014) 'Researching Uses of Corpora for Language Teaching and Learning', *ReCALL* 26(2): 121–7.

Bridle, M. (2019) 'Learner Use of a Corpus as a Reference Tool in Error Correction: Factors Influencing Consultation and Success', *Journal of English for Academic Purposes* 37: 52–69.

Burton, G. (2012) 'Corpora and Coursebooks: Destined to Be Strangers Forever?', *Corpora* 7(1): 91–108.

Callies, M. and Götz, S. (eds) (2015) *Learner Corpora in Language Testing and Assessment*, Amsterdam: John Benjamins.

Cambridge University Press (2020) *Cambridge Learner Corpus*. Available at: https://www.cambridge.org/elt/corpus/learner_corpus2.htm.

Capel, A. (2010) 'Insights and Issues Arising from the English Profile Wordlists Project', *Research Notes* 41: 2–7. Cambridge: Cambridge ESOL.

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

Chambers, A. (2019) ‘Towards the Corpus Revolution: Bridging the Research-Practice Gap’, *Language Teaching* 52(4): 460–75.

Chambers, A., Farr, F. and O’Riordan, S. (2011) ‘Language Teachers with Corpora in Mind: From Starting Steps to Walking Tall’, *Language Learning Journal* 39(1): 85–104.

Charles, M. (2014) ‘Getting the Corpus Habit: EAP Students’ Long-Term Use of Personal Corpora’, *English for Specific Purposes* 35(1): 30–40.

Charles, M. (2018) ‘*Corpus*-Assisted Editing: More than Just Concordancing’, *Journal of English for Academic Purposes* 36: 15–25.

Chen, M. and Flowerdew, J. (2018) ‘A Critical Review of Research and Practice in Data-Driven Learning (DDL) in the Academic Writing Classroom’, *International Journal of Corpus Linguistics* 23(3): 335–69.

Chen, M., Flowerdew, J. and Anthony, L. (2019) ‘Introducing In-Service English Language Teachers to Data-Driven Learning for Academic Writing’, *System* 87: 102–48.

Cheng, W. and Warren, M. (2006) ‘I Would Say Be Very Careful of…: Opine Markers in an Intercultural Business Corpus of Spoken English’, in J. Bamford and M. Bondi (eds) *Managing Interaction in Professional Discourse. Intercultural and Interdiscoursal Perspectives*, Rome: Officina Edizioni, pp. 46–58.

Cobb, T. (2003) ‘Analyzing Late Interlanguage with Learner Corpora: Québec Replications of Three European Studies’, *The Canadian Modern Language Review* 59(3): 393–423.

Cobb, T. and Boulton, A. (2015) ‘Classroom Applications of Corpus Analysis’, in D. Biber and R. Reppen (eds) *Cambridge Handbook of English Corpus Linguistics*, Cambridge: Cambridge University Press, pp. 478–97.

Cotos, E. (2014) ‘Enhancing Writing Pedagogy with Learner Corpus Data’, *ReCALL* 26(2): 202–24.

Coxhead, A. (2000) ‘A New Academic Word List’, *TESOL Quarterly* 34(2): 213–38.

Crosthwaite, P. (2020) ‘Data-Driven Learning and Younger Learners: Introduction to the Volume’, in P. Crosthwaite (ed.) *Data-Driven Learning for the Next Generation: Corpora and DDL for Pre-Tertiary Learners*, London: Routledge, pp. 1–10.

Crosthwaite, P., Wong, L. and Cheung, J. (2019) ‘Characterising Postgraduate Students’ Corpus Query and Usage Patterns for Disciplinary Data-Driven Learning’, *ReCALL* 31(3): 255–75.

Davies, M. (2020) *corpus.byu.edu*, Available at: https://corpus.byu.edu/overview.asp.

Ebrahimi, A. and Faghih, E. (2016) ‘Integrating Corpus Linguistics into Online Language Teacher Education Programs’, *ReCALL* 29(1): 120–35.

Fligelstone, S. (1993) ‘Some Reflections on the Question of Teaching, from a Corpus Linguistics Perspective’, *ICAME Journal* 17: 97–110.

Flowerdew, J. (2012) ‘Corpora in Language Teaching from the Perspective of English as an International Language’, in L. Alsagoff, S. L. McKay, G. Hu and W. R. Renandya (eds) *Principles and Practices for Teaching English as an International Language*, London; New York: Routledge, pp. 226–43.

Flowerdew, L. (2015) ‘Using Corpus-Based Research and Online Academic Corpora to Inform Writing of the Discussion Section of a Thesis’, *Journal of English for Academic Purposes* 20: 58–68.

Frankenberg-Garcia, A. (2012) ‘Raising Teachers’ Awareness of Corpora’, *Language Teaching* 45(4): 475–89.

Gablasova, D. and Brezina, V. (2018) ‘Disagreement in L2 Spoken English: From Learner Corpus Research to Corpus-Based Teaching Materials’, in V. Brezina and L. Flowerdew (eds) *Learner Corpus Research: New Perspectives and Applications*, London: Bloomsbury Academic, pp. 69–89.

Gardner, D. and Davies, M. (2014) ‘A New Academic Vocabulary List’, *Applied Linguistics* 35(3): 305–27.

Gilquin, G. and Granger, S. (2010) ‘How Can Data-Driven Learning Be Used in Language Teaching?’, in A. O’Keeffe and M. McCarthy (eds) *The Routledge Handbook of Corpus Linguistics*, London: Routledge, pp. 359–70.

Götz, S. and Mukherjee, J. (eds) (2019) *Learner Corpora and Language Teaching*, Amsterdam: John Benjamins Publishing Company.

Huang, L.-S. (2018) ‘Taking Stock of Corpus-Based Instruction in Teaching English as an International Language’, *RELC Journal* 49(3): 381–401.

Huang, Z. (2014). 'The Effects of Paper-Based DDL on the Acquisition of Lexico-Grammatical Patterns in L2 Writing,' *ReCALL* 26(2): 163–83.

Hyland, K. (2003) *Second Language Writing*, Cambridge: Cambridge University Press.

Johns, T. (1991) 'From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-Driven Learning', *English Language Research Journal* 4: 27–45.

Lee, H., Warschauer M. and Lee J. H. (2019) 'The Effects of Corpus Use on Second Language Vocabulary Learning: A Multilevel Meta-Analysis', *Applied Linguistics* 40(5): 721–53.

Leech, G. (1997) 'Teaching and Language Corpora: A Convergence', in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds) *Teaching and Language Corpora*, Harlow: Addison Wesley Longman, pp. 11–23.

Leńko-Szymańska, A. and Boulton, A. (eds) (2015) *Multiple Affordances of Language Corpora for Data-Driven Learning*, Amsterdam: John Benjamins.

Macmillan (2020) *Corpus*, Available at: https://www.macmillandictionary.com/corpus.html.

Martinez, R. and Schmitt, N. (2012) 'A Phrasal Expressions List', *Applied Linguistics* 33(3): 299–320.

McCarten, J. and McCarthy, M. J. (2010) 'Bridging the Gap between Corpus and Course Book: The Case of Conversation Strategies', in A. Chambers and F. Mishan (eds) *Perspectives on Language Learning Materials Development*, Bern: Peter Lang, pp. 11–32.

McCarthy, M. J., McCarten, J. and Sandiford, H. (2005) *Touchstone 2a: Student's Book*, Cambridge: Cambridge University Press.

McCarthy, M. J. and O'Dell, F. (2008) *English Vocabulary in Use*, Cambridge: Cambridge University Press.

Meunier, F. (2020) 'A Case for Constructive Alignment in DDL: Rethinking Outcomes, Practices and Assessment in (Data-Driven) Language Learning', in P. Crosthwaite (ed.) *Data-Driven Learning for the Next Generation. Corpora and DDL for Pre-Tertiary Learners*, London: Routledge, pp. 1–18.

Meunier, F. and Reppen, R. (2015) 'Corpus versus Non-Corpus-Informed Pedagogical Materials: Grammar as the Focus', in D. Biber (ed.) *The Cambridge Handbook of English Corpus Linguistics*, Cambridge: Cambridge University Press, pp. 498–514.

Nation, I. S. P. (2001) *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

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): 457–89.

O'Sullivan, I. (2007) 'Enhancing a Process-Oriented Approach to Literacy and Language Learning: The Role of Corpus Consultation Literacy', *ReCALL* 19(3): 269–86.

Oxford University Press (2020) *OPAL (Oxford Phrasal Academic Lexicon)*, Available at: https://www.oxfordlearnersdictionaries.com/wordlists/opal.

Paquot, M. (2010) *Academic Vocabulary in Learner Writing: From Extraction to Analysis*, London and New-York: Continuum, pp. 56–8.

Pérez-Paredes, P. (2010) 'Corpus Linguistics and Language Education in Perspective: Appropriation and the Possibilities Scenario', in T. Harris and M. M., Jaén (eds) *Corpus Linguistics in Language Teaching*, Bern: Peter Lang, pp. 53–73.

Pérez-Paredes, P. (2019) 'A Systematic Review of the Uses and Spread of Corpora and Data-Driven Learning in CALL Research during 2011–2015', *Computer Assisted Language Learning*, doi: 10.1080/09588221.2019.1667832

Pérez-Paredes, P., Ordoñana Guillamón, C., Van de Vyver, J., Meurice, A., Aguado, P., Conole, G. and Hernández, P. (2019) 'Mobile Data-Driven Language Learning: Affordances and Learners' Perception', *System* 84: 145–59.

Poole, R. (2016) 'A Corpus-Aided Approach for the Teaching and Learning of Rhetoric in an Undergraduate Composition Course for L2 Writers', *Journal of English for Academic Purposes* 21: 99–109.

Renouf, A. (1997) 'Teaching Corpus Linguistics to Teachers of English', in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds) *Teaching and Language Corpora*, London: Longman, pp. 255–66.

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): 121–34.

Rundell, M. (2020) *From Corpus to Dictionary*, Available at: http://www.macmillandictionaries.com/features/from-corpus-to-dictionary/.

Sha, G. (2010) 'Using Google as a Super Corpus to Drive Written Language Learning: A Comparison with the British National Corpus', *Computer Assisted Language Learning* 23(5): 377–93.

Sinclair, J. (ed) (1995) *Collins COBUILD English Language Dictionary*, 2nd edn, London: Collins.

Smith, S. (2020) 'DIY Corpora for Accounting and Finance Vocabulary Learning', *English for Specific Purposes* 57: 1–12.

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

Tognini-Bonelli, E. (2001) *Corpus Linguistics at Work*, Amsterdam: John Benjamins.

Vyatkina, N. and Boulton, A. (ed) (2017) Special Issue on Corpora in Language Learning and Teaching, *Language Learning & Technology* 21(3).

Yoon, C. (2016) 'Concordancers and Dictionaries as Problem-Solving Tools for ESL Academic Writing', *Language Learning & Technology* 20(1): 209–29.