Corpus Linguistics in EFL Language Teaching: Insights From Research and Practice

Shelley Staples
University of Arizona
Laurence Anthony
Waseda University

Corpus linguistics can provide curriculum developers and teachers with theoretical foundations and guidance when deciding learning objectives, identifying materials and methods, and evaluating learner outputs. In this Exposition piece, we discuss how corpus linguistics can inform EFL language teaching in the areas of materials creation, skill development (reading, writing, speaking, and listening), and evaluation, with special attention given to lexico-grammar and vocabulary. We also provide examples from courses delivered at Waseda University to illustrate these approaches.

コーパス言語学は、カリキュラム開発者と教師に学習目標の決定、教材と教育方法の特定、学習者の成果の評価を行う際の理論的基礎とガイダンスを提供します。この解説記事では、語彙文法に注意を払いながらコーパス言語学が資料作成、スキル開発（読み書き、スピーキング、リスニング）　および評価の分野におけるEFL言語教育にどのように役立つかを解説します。　また、これらのアプローチを解説するために、早稲田大学で実施されているコースの例も紹介します。

Keywords: corpus linguistics; data-driven learning; lexico-grammar; vocabulary

https://doi.org/10.37546/JALTJJ45.2-4

Corpus linguistics (CL) is a research methodology that helps researchers, curriculum developers, teachers, and even learners to understand language use in different domains (e.g., journal articles, health care communication, conversation) through the analysis of a large, principled set of authentic "texts", called a corpus, which is sampled to represent the target language. A corpus (or 'corpora' in the plural form) is usually comprised of written or spoken texts from the target domain, but it can also be comprised of a mixture of different language modes including video and audio files. In this case, it would be called a multimodal corpus. Researchers, teachers and even learners of a foreign language can interact with corpora using special corpus software tools and gain a deep understanding of how language works in the real world. In some cases, this new knowledge may complement their existing knowledge, but in many cases, it may challenge their pre-existing ideas. Importantly, in an EFL context, the corpus linguistics methodology can empower researchers, teachers, and learners who are L2 speakers of English by providing them with data and tools that deliver insights which are not readily known even by L1 speakers.

In this paper, we will discuss how corpus linguistics can inform EFL language teaching in the areas of materials creation, skill development (reading, writing, speaking, and listening), and evaluation, with special attention given to lexico-grammar and vocabulary. To illustrate some of these ideas, we will provide real-world examples taken from courses delivered as part of the Center for English Language Education in Science and Engineering (CELESE) program at Waseda University (https://celese.jp/about). We will conclude the paper with some suggestions for important areas of future research that might inform EFL instruction.

Corpus Linguistics in EFL Materials Creation and In-Class Teaching

Corpus linguistics has profoundly changed the way in which EFL language teaching materials are created. One of the earliest examples of this trend was the use of corpora in the creation of the *Collins COBUILD English Language Dictionary,* a project headed by John Sinclair at The University of Birmingham (Sinclair, 1987). The COBUILD dictionary was unique for its time as it included frequency of use information on the words included, as well as examples taken directly from the COBUILD corpus, which allowed learners to see how words were used in authentic contexts. Today, almost all learner's dictionaries are created with the help of corpus linguistics methods, from the design of the underlying corpus and the selection of entries and examples to the inclusion of supplementary notes on grammar and usage patterns. We can see a similar trend in the creation of corpus-based reference grammars, such as the *Grammar of Spoken and Written English* (Biber et al., 2021) and the creation of corpus-informed textbooks for General English and English for Specific Purposes (ESP). A good example of the latter is the *Touchstone* series of textbooks created by McCarthy et al. (2014) that takes into account the differences between spoken and written discourse and offer students examples of the "messiness" of spoken interactions.

Teachers interested in incorporating CL into their own courses can, as a first step, evaluate textbooks for their use (or not) of corpora to inform their design, and assess the approaches taken to introduce topics to students. For example, does the textbook include authentic texts/dialogues that allow students to explore language in use (see McCarthy & O'Keefe, 2014 for a discussion of this issue), or, when introducing passive voice, is the higher frequency of the agentless passive mentioned (see Meunier & Reppen, 2015 for a discussion of passive voice presentation in corpus vs. non-corpus informed textbooks)? Going one step further, insights from corpus linguistics can also be used directly by teachers as they prepare materials for class. For example, instead of adopting the traditional approach of covering all tense-aspect pairs in a grammar class, starting with present progressive, teachers can take a more frequency-based approach which would start with simple present and not introduce progressive until later (see Biber and Reppen, 2002 for a discussion of this phenomenon in learner textbooks).

Another way that teachers can use corpus linguistics in their materials creation is with a (learner) corpus of their students' own assignments. Learner corpora have traditionally been used by teachers (and researchers) to identify errors in learner output, but there is growing interest in using learner corpora to identify examples of both positive and negative language use. With a learner corpus at hand, teachers can apply corpus methods such as KWIC (Key-Word-In-Context) concordancing and cluster analysis to identify common patterns in learner writing and use some of the authentic examples in class to illustrate both effective and ineffective language patterns. They can also use learner corpora to introduce students to genres that are more commonly used in the classroom and in more familiar contexts than general corpora (Seidlhofer, 2002; Tribble, 2001), presenting student texts as models for discussion of language choice and effectiveness. Finally, teachers can create a corpus of the textbook materials they are asked to use and apply corpus methods such as word list and keyword list generation to identify the most frequent words to teach.

We have just noted that corpus texts can be used directly in the classroom as models of language use. Expanding on this concept, some notable scholars in the late 1980s and early 1990s (e.g., Johns, 1990) proposed introducing the principles of corpus linguistics directly to students in the form of data-driven learning (DDL). DDL is an inductive approach to language learning whereby students are provided with data from a corpus and asked to analyze and reach conclusions about common patterns of language used in the target domain. Initially, Johns adopted a 'soft' approach to DDL that relied on printed handouts of KWIC concordance outputs. Today, the dramatic increase in the power of computers has allowed for a 'hard' approach to be possible, where students query a target corpus directly either through a web-based corpus analysis tool (e.g., SketchEngine[1], CQPWeb[2]) or a desktop tool (e.g., AntConc[3], WordSmith Tools[4]). In fact, numerous large scale meta-analyses of results of 'hard' DDL have shown it to produce moderate to large gains (effect sizes) in learning, particularly in the areas of vocabulary, lexico-grammar, and writing (Boulton & Cobb, 2017; Boulton & Vyatkina, 2021). Also, while the 'soft' DDL approach showed lower effect sizes in Boulton and Cobb's (2017) meta-analysis, the reported gains are still larger than for many other types of computer assisted language learning (CALL) (c.f. Plonsky & Ziegler, 2016).

Several models have been introduced for conceptualizing the lesson arc for DDL. The "4 Is" approach proposed by Lynne Flowerdew (2009) builds off an earlier "3 Is" model (which excluded step 3) by Carter and McCarthy (1995) and includes the following steps:

1. Illustration: looking at data
2. Interaction: discussion and sharing observations and opinions
3. Intervention: optional step to provide learners with hints or clearer guides for induction
4. Induction: making one's own rule for a particular feature

Ma et al. (2021) proposes a broader model for DDL that comprises the following four steps:

1. testing students' knowledge
2. hands-on corpus search by students
3. inductive discovery by students
4. output activities

Steps 2 and 3 from Ma et al. (2021) clearly fit within steps 1-4 from Flowerdew's model, and thus the two can be usefully combined. Notably, both models can be used with the 'soft' and 'hard' approaches to DDL.

In the CELESE program at Waseda University, corpus linguistics methods are used as part of materials creation across the entire curriculum, from required courses in communication strategies (CS), academic listening comprehension (ALC), concept building and discussion (CBD), and academic reading (AR) to elective courses, such as technical writing (TW) and technical presentation (TP). All materials for the program are developed in-house, which allows for target vocabulary, grammar patterns, illustrative, dialogues, and examples to be informed by corpora from the target domain of science and engineering. An extreme version of this happens in the second part of the TW course, where students are encouraged to create their own materials in the form of a corpus of target research articles from their own specific disciplines, such as physics or mathematics. Then, in class, students are guided on how to use these materials in combination with corpus tools to inform their own writing practices.

Corpus Linguistics in EFL Skills Development

Vocabulary and Extensive Reading

Corpus linguistics provides insights on language use in authentic settings, whether those be real-world conversational settings or the fictional worlds of novels and plays. As a result, many of the findings from corpus linguistics can be applied directly to reading instruction.

Perhaps the most influential corpus work on reading has been in the area of vocabulary. Early pioneering researchers, such as West (1953) and later Nation (2001) and others took large corpora of general and specialized English and profiled the vocabulary used in the texts to generate lists of the most productive vocabulary in terms of frequency and dispersion. Today, many such lists exist across a huge range of target domains, such as general English, academic English, TOEIC, law, politics, sports, and many more (see https://www.newgeneralservicelist.com/ for many such lists). These lists can be used not only to evaluate the vocabulary knowledge of learners, but also used in combination with a vocabulary profiling tool, such as AntWordProfiler[5], to gauge the difficulty (and suitability) of texts for a target learner audience. In preparation for a reading class, a teacher can profile the target reading and then decide whether or not to gloss any potentially difficult vocabulary or perhaps even simplify the target text if the reading goal is fluency (see Donley & Reppen, 2001 and Huang & Liou, 2007 for example implementations of this approach). Another obvious application of such profiling is in the creation of reading materials for high-stakes entrance examinations.

The systematic profiling, glossing, and simplification of reading materials based on corpus-informed frequency lists has led to the development of modern graded reader book series, such as Oxford Bookworms (Bladon, 2014), Cambridge Young Readers (Prowse, n.d.) and many others. We also see the approach used in the creation of more specialized academic reading materials such as the *Longman Academic Reading* (Bottcher et al., 2014) and the *College Reading* series (Byrd et al., 2006), many of which are based on the Academic Word List (AWL) of Coxhead (2000). These readers have been shown to be effective as part of an extensive reading program (e.g., Huang & Liou, 2007), where the learners aim to read a large number of books over a set period with gradually increasing difficulty. Importantly, the books should always be at a level that is below the learner's current reading level so that the books are relatively easy for them to understand and can be read fluently and for enjoyment.

In the first year of the CELESE program at Waseda University, students are given a vocabulary goal of mastering the first 2000 word families of the West (1953) general service list. To support this goal, the students are provided with the complete word list that includes a pronunciation guide and authentic example sentences from the British National Corpus (BNC). In addition, all the course materials are designed to illustrate the use of these words, and other science, technology, engineering, and mathematics (STEM) related target words, in context. In the second year, the vocabulary goal switches to the Academic Word List (AWL) of Coxhead (2000). Again, all the words are provided in the form of word lists, the materials are designed to highlight these words, and the students are evaluated on the use of these words in their writing. Although CELESE does not run a formalized extensive reading program, students are encouraged to develop their reading fluency using science news articles, which are evaluated in terms of their vocabulary load.

Lexico-Grammar and Writing

Corpus tools are easily able to generate the most frequent words in a target corpus and show examples of how these words are used in context through KWIC concordances. Writing instruction, however, must go beyond vocabulary and guide learners on how to combine vocabulary with syntactic patterns to create phrases, clauses, sentences, paragraphs, and whole sections of discourse that adhere to the conventions of a particular register (Biber & Conrad, 2019) and discourse community (Swales, 1990).

The COBUILD project (Sinclair, 1987) mentioned earlier was initially designed as a lexicography project, but results soon emerged that blurred the lines between vocabulary and grammar and led to new insights on the connections between the two. This area of work was later termed lexico-grammar, but it also relates to 'pattern grammar', a term coined by Hunston and Francis (2000). One of the most notable works in lexico-grammar that is relevant to EFL language teaching is that of Willis (2003), who introduces numerous patterns that are useful for learners to know, such as the "FORGET + WH clause" that appears in the "*I forgot what I said*" and "*They always forget where the car keys are*". Hunston (2022) argues against presenting lexico-grammatical patterns to students in the form of a list, and instead recommends using awareness raising activities, such as re-writing activities and the hands-on analysis of corpus data by the learners through the data-driven learning (DDL) approach discussed earlier.

When it comes to writing, the DDL approach has been shown to be particularly effective at the tertiary level in the teaching of academic research paper writing, as discussed in detail by Anthony (2016, 2019), Charles (2007, 2014, 2018), and others. Charles, for example, describes how students collect high-quality research papers in their own discipline, convert the papers into a text-based form, and then load these papers as a corpus into the AntConc[3] corpus analysis toolkit. Once the corpus is loaded, the students can directly query the existence of common words, multi-word units, and phrases, as well as lexico-grammatical structures and discourse markers. Anthony (2016) reports on the many strengths of this approach, especially in a STEM context. Students are not only empowered to find answers to their individual language questions, but the language insights they gain are directly relevant to their learning goals, i.e., research article writing for publication in high impact journals.

Learner corpora can also be used effectively in the language classroom. Staples (2022) shows how learner corpora are used in the teaching of writing to promote asset-based approaches to language learning, with students examining lexico-grammatical patterns in a corpus of student papers from the same course context (Staples & Dilger, 2018-). Here, student papers are used as models for "allowable contributions to the genre" (Tribble, 2001, p. 381) and the students are asked to engage in questions around language choices that create more or less effective versions of a given assignment.

In the CELESE program at Waseda University, both traditional process-writing methods and DDL are employed, with process writing being predominantly used in the first two years of undergraduate study, and DDL being the core methodology used in the rest of the program. As an example, students are exposed to a 'hard' form of DDL in the second half of the technical writing (TW) course that they take in the third year of their undergraduate studies. The DDL approach adopted at CELESE mirrors that described by Anthony (2016) and Charles (2007), with students analyzing corpora that they build themselves using the AntCorGen[6] discipline-specific corpus creation tool. Notably, the students all major in STEM subjects, which tends to reduce issues and challenges related to computer literacy that are often discussed in the literature on DDL (e.g., Adel, 2010).

Speaking and Listening Instruction

In our modern world, there is an abundance of easily available written text data that can be obtained from the Internet and used to create general and discipline specific corpora. It is perhaps no surprise, therefore, that much of today's research in corpus linguistics is focused on written language. However, corpus-based research on spoken language has been a feature of the field from the earliest days, and that interest appears to be growing with the availability of new general and specialized spoken corpora, such as the Spoken BNC (Love et al., 2017) and the British Academic Spoken Corpus (BASE) (Thompson & Nesi, 2001).

Research using spoken corpora shows a number of key features of spoken language compared with writing. These features include the use of incomplete clauses and sentences (and the related phenomenon of ellipsis), much more frequent use of ready-made chunks (i.e., lexical bundles, formulaic language), use of vague language (e.g., *thing, stuff*), use of high-frequency vocabulary, use of hesitation markers (e.g., *um*, *uh*), use of discourse markers (e.g., *well, so*), and repetition of vocabulary. For listeners, spoken corpora of conversation and other interactive discourse shows us that backchannelling and response tokens (e.g., *yeah, right*) are important cues for speakers to know their interlocutors are listening and understanding what they are saying. Within spoken discourse, we also tend to see more language associated with stance (e.g., *really, very*) due to the strong emphasis on interpersonal and pragmatic functions (Biber et al., 2021; McCarthy & McCarten, 2022; McCarthy & O'Keefe, 2014; Staples, 2015).

One way that teachers can incorporate these important features of spoken discourse into their classrooms is through the selection of corpus-informed textbooks or engagement with ready-made online materials. The *Touchstone* series (McCarthy et al., 2014) is the most prominent example of a corpus-informed textbook for conversational English. It utilizes research from the Cambridge English Corpus (https://www.cambridge.es/en/about-us/cambridge-english-corpus) and is distinctive in its inclusion of the types of spoken features discussed above. *Real Grammar* (Biber & Conrad, 2009) is a textbook that includes several units focused on spoken characteristics (e.g., discourse markers, incomplete sentences) based on the *Longman Grammar of Spoken and Written English* (Biber et al., 1999). For more academic spoken language, teachers might choose to use a corpus-informed textbook such as *Academic Interactions* (Feak et al., 2009), which is based on the Michigan Corpus of Spoken English (MICASE). This textbook provides audio samples and transcripts from the corpus for speech events like office hours and classroom discussions.

Others have developed stand-alone ready-made materials for instructors to use in classrooms. Gablasova and Brezina (2017) and Gablasova et al. (2019) describe sample materials from the Trinity Lancaster Corpus (TLC) on disagreement and active listenership. Importantly, the TLC is a learner corpus. The materials can be accessed at https://www.trinitycollege.com/about-us/research/Trinity-corpus/corpus-resources. The MICASE Handbook (Simpson-Vlach & Leicher, 2006) also contains activities based on the MICASE corpus and ideas for using MICASE for pedagogical purposes.

Data-driven learning is also possible in the speaking and listening classroom through the use of spoken corpus interfaces, particularly those that provide multimodal search results. As an example, Youglish (https://youglish.com/) searches 100 million spoken tracks to give users samples of words pronounced in context. It also allows users to search varieties of English (e.g., US, UK, Australia). The TED Corpus Search Engine (TCSE, https://yohasebe.com/tcse/) provides users with the ability to retrieve audio and transcripts from TED talks in context (Hasebe, 2015). In addition, there are various commercial learning platforms that allow learners to query multimodal corpora and view examples phrases and sentences aligned with their associated video clips.

Speaking and listening are essential components of the CELESE program at Waseda University. In the first year of the program, for example, the Communication Strategies (CS) course aims to develop the students' ability to speak in various academic settings, such as research labs and conferences. Similarly, the Academic Lecture Comprehension (ALC) course aims to develop the students' ability to listen and comprehend academic lectures, as well as take notes on those lectures, and summarize the main points in the form of a written or oral report. In the second year, the Concept Building and Discussion (CBD) course is designed to develop these speaking and listening skill further so that the students can confidently present and discuss the findings of mini projects that they conduct in groups or individually. Then, in the third and fourth year of the program, these skills are extended further in the Technical Presentation (TP) course, which aims to help students deliver a conference-level oral presentation about their research and respond to questions and comments about the work. Corpus-based research has been a key factor in the development of all these courses. For example, a major corpus project was initiated to understand the language used by experienced lecturers and presenters in different STEM fields (Kunioshi et al., 2016). Similarly, materials for developing successful Q&A strategies used in the TP course are based on a corpus of Q&A interactions recorded at a real conference and later transcribed.

Corpus Linguistics in EFL Evaluation

Corpus linguistics as a field is primarily concerned with describing how language is used in the real world. In the EFL classroom, however, one of the most important jobs of the teacher is evaluating the language output of the learner and assigning a grade. This raises an interesting question: What insights do corpus linguistics provide in terms of learner assessment?

In fact, corpora are the foundation of almost all automated evaluation tools used in EFL. At the most basic level, corpora of existing public domain language and local student submissions are used by plagiarism detection tools, such as Turnitin (https://www.turnitin.com/), to measure the degree to which a newly submitted student paper overlaps with existing work. The algorithms used to measure the degree of overlap are also founded on principles developed through corpus linguistics research, such as n-gram analyses (https://en.wikipedia.org/wiki/N-gram).

In the area of error detection, corpora of manually error-corrected learner writing samples are used by many automatic error detection tools to flag potential errors in writing and offer suggestions for improvement (Callies & Götz, 2015). Similarly, corpora of writing samples at different quality levels are used by testing services, such as the Educational Testing Service (ETS) (https://www.ets.org/), to automatically grade writing submitted as part of tests such as TOEIC and TOEFL. In these cases, the algorithms that compare the submitted writing with the corpus samples and assign grades can vary from simple rule-based error counting algorithms to highly complex algorithms that involve large-language models (LLMs) and deep learning.

People in the field are currently debating what aspects of LLMs (if any) are developed out of ideas from corpus linguistics. However, it is clear that some of the most important underlying principles of LLMs match ideas that were discussed in the very early days of corpus linguistics (see Firth, 1957).

In the CELESE program at Waseda University, all student reports are graded manually by teachers. However, research is in progress to determine the effectiveness of automated corpus-based grading approaches (Wang, 2022). In addition, corpus-based methods are used to support some aspects of teacher evaluation. For example, Turnitin is used to check for potential cases of plagiarism across all courses. Also, in the Concept Building and Discussion (CBD) course, students are required to highlight the use of at least three words from the Academic Word List (Coxhead 2000) in their writing, so they are encouraged to use the AntQuickTools[7] to quickly profile their work and highlight all words from the Academic Word List automatically. Teachers are also recommended to use this tool to check that students' have completed the task correctly.

Possibilities for Future Research

With the growth of technology and access to data on the Internet, a great number of corpora are now available for teachers to choose from, including general corpora (e.g., Corpus of Contemporary American English), national corpora (e.g., British National Corpus), and specialized corpora (e.g., British Academic Written English corpus, a corpus of written work by students in British universities). These may be useful for teachers who find the existing materials (e.g., textbooks) less relevant for their contexts or who want to introduce students to corpus consultation to enhance their own learning. However, as mentioned earlier, the number of large-scale spoken corpora is still relatively small. Therefore, one important area of corpus research is determining how to effectively collect, transcribe, and annotate spoken data. A related question is how to develop corpus tools that align and visualize multimodal data in an intuitive way.

Vocabulary lists have been another major outgrowth of CL research, providing students and instructors with frequently used words in specialized areas such as engineering (e.g., Basic Engineering Word List; Ward, 2009) and medicine (e.g., Medical Academic Word List; Wang, Liang, & Ge, 2008). However, the words in these lists are almost universally ranked by either their frequency of occurrence in the corpus or their dispersion across files in the corpus. Several methods have been proposed to rank words using other measures. For example, Savický and Hlavácová (2002) have proposed an average reduced frequency (ARF) measure that combines frequency and dispersion into a single number. However, the meaning of this number is effectively impossible to interpret without knowing the values on which it is based. Schmitt et al. (2021) have released knowledge-based vocabulary lists (KVL) that are based on a measure of learners' ability to produce words. However, these lists are extremely time consuming to generate and are currently only available for Chinese, German, and Spanish learner contexts. Clearly, there is a need for more research on effective word ranking measures for different purposes. In fact, the unit of analysis in all these works (i.e., the definition of a word) is also open to challenge.

In view of the availability of written corpora and the scarcity of large-scale spoken corpora, it is not surprising that there is relatively more research that looks at corpus-based lexico-grammar and writing instruction, and less that focuses on reading and speaking instruction. While important inroads have been made, the research on how to effectively use corpora for reading and speaking purposes is limited. In addition, and in some ways related to these limitations, most of the empirical research on the effectiveness of corpus-based instruction has focused on the use of concordance lines to inductively highlight patterns of lexico-grammar. More work to show broader contextual use of language is needed, as well as alternatives that might be more relevant for instruction beyond lexico-grammar (including genre-based instruction of writing and features of dialogic spoken discourse that rely on the unfolding of meaning over several turns). The use of corpora for spoken instruction necessarily relies on the development of multimodal corpora, which are limited and almost never found in studies of corpus-based instruction. Such developments would also align with multiliteracies frameworks for language learning, which are being adopted more broadly (see New London Group, 1996 for details of this framework).

While Chujo and Nishigaki (2004), Crosthwaite (2020), Kakiba et al. (2021), and a few others provide important first looks at what corpus-based instruction can look like in primary and secondary schools, much more work is needed to understand how corpus methods can be incorporated with other approaches commonly found in these contexts, including content-based instruction. Similarly, research on the use of corpora in post-tertiary adult learning courses is scarce, although interesting work is beginning to emerge from the teaching of teaching of Welsh to adult learners as part of the CorCenCC project (Knight et al., 2020).

Summary

In this paper, we have focused on key areas of corpus linguistics that are relevant to EFL teaching. Firstly, we introduced definitions for corpora and corpus linguistics and then discussed how corpora can be used in EFL materials creation and in-class teaching. Next, we discussed how corpus linguistics principles can be used in the teaching of vocabulary and reading, lexicogrammar and writing, and speaking/listening. We ended with thoughts on the use of corpora for evaluation/assessment and future research. In each section, we provided examples to contextualize the various approaches for EFL teaching in Japan. We hope this paper will provide teachers with the background to get started in using corpora, as well as references that they can use to gain a deeper understanding of corpus linguistics in EFL.

Notes

1. SketchEngine. https://www.sketchengine.eu/.
2. CQPWeb. https://cqpweb.lancs.ac.uk/.
3. AntConc. https://www.laurenceanthony.net/software/antconc/.
4. WordSmith Tools. https://www.lexically.net/wordsmith/.
5. AntWordProfiler. https://www.laurenceanthony.net/software/antwordprofiler/.
6. AntCorGen. https://www.laurenceanthony.net/software/antcorgen/.
7. AntQuickTools. https://www.laurenceanthony.net/software/antquicktools

Shelley Staples is Associate Professor of English Applied Linguistics/Second Language Acquisition and Teaching in the English Department at University of Arizona, United States. Her research focuses on corpus-based analyses and instruction of academic writing and speaking, with a particular emphasis on second language writing.

Laurence Anthony is Professor of Applied Linguistics and founding member of the Center for English Language Education in the Faculty of Science and Engineering, Waseda University, Japan. His main research interests are in corpus linguistics, educational technology, language data science, and English for Specific Purposes (ESP) program design and teaching methodologies.

References

Adel, A. (2010). Using corpora to teach academic writing: Challenges for the direct approach. In M. C. Campoy-Cubillo, B. Belles-Fortuño, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 18–35). Continuum.

Anthony, L. (2016). Introducing corpora and corpus tools into the technical writing classroom through Data-Driven Learning (DDL). In J. Flowerdew & T. Costley (Eds.) *Discipline specific writing* (pp. 162–180). Routledge.

Anthony, L. (2019). Tools and strategies for Data-Driven Learning (DDL). In K. Hyland & L. Wong (Eds.) *Specialised English: New directions in ESP and EAP research and practice* (pp. 179–194). Routledge. https://doi.org/10.4324/9780429492082-14

Biber, D., & Conrad, S. (2009). *Real grammar: A corpus-based approach to English*. Pearson Longman.

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, D., Johansson, S., Leech, G. N., Conrad, S., & Finegan, E. (2021). *Grammar of spoken and written English*. John Benjamins. https://doi.org/10.1075/z.232

Biber, D., & Reppen, R. (2002). What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition, 24*(2), 199–208. https://doi.org/10.1017/S0272263102002048

Bladon, R. (Ed.) (2014). *Oxford bookworms library*. Oxford University Press.

Bottcher, E., Sanabria, K., Miller, J. L., Cohen, R. F., & Smith, L. C. (2014). *Longman academic reading series.* Longman.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning, 67*, 348–393. https://doi.org/10.1111/lang.12224

Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology, 25*(3), 66–89.

Byrd, P., Schuemann, C., Reid, J., Benz, C., & Folse, K. (2006). *College reading series*. Cengage.

Callies, M., & Götz, S. (2015). *Learner corpora in language testing and assessment*. John Benjamins. https://doi.org/10.1075/scl.70

Carter, R., & McCarthy, M. (1995). Grammar and the spoken language. *Applied Linguistics, 16*(2), 141–158. https://doi.org/10.1093/applin/16.2.141

Charles, M. (2007). Reconciling top-down and bottom-up approaches to graduate writing: Using a corpus to teach rhetorical functions. *Journal of English for Academic Purposes, 6*(4), 289–302. https://doi.org/10.1016/j.jeap.2007.09.009

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes, 35*, 30–40. https://doi.org/10.1016/j.esp.2013.11.004

Charles, M. (2018). Corpus-assisted editing for doctoral students: More than just concordancing. *Journal of English for Academic Purposes, 36*, 15–25. https://doi.org/10.1016/j.jeap.2018.08.003

Chujo, K., & Nishigaki, C. (2004, December). Creating e-learning material to teach essential vocabulary for young EFL learners. In *Proc. of An Interactive Workshop on Language e-Learning* (pp. 35–44). Waseda University.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238. https://doi.org/10.2307/3587951

Crosthwaite, P. (2020). *Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners*. Routledge. https://doi.org/10.4324/9780429425899

Donley, K., & Reppen, R. (2001). Using corpus tools to highlight academic vocabulary in SCLT. *TESOL Journal, 10*(2–3), 7–12.

Feak, C. B., Reinhard, S. M., & Rohlck, T. N. (2009). *Academic interactions: Communicating on campus*. Michigan ELT.

Firth, J. R. (1968). A synopsis of linguistic theory, 1930-55. In F. Palmer (Ed.), *Selected papers of J. R. Firth, 1952-59* (pp. 168–205). Indiana University Press.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics, 14*(3), 393–417. https://doi.org/10.1075/ijcl.14.3.05flo

Gablasova, D., & Brezina, V. (2017). Disagreement in L2 spoken English: From learner corpus research to corpus-based teaching materials. In V. Brezina & L. Flowerdew (Eds.), *Learner corpus research: New perspectives and applications* (pp. 69–89). Bloomsbury Publishing.

Gablasova, D., Brezina, V., & McEnery, T. (2019a). The Trinity Lancaster corpus: Applications in language teaching and materials development. In S. Götz & J. Mukherjee (Eds.), *Learner corpora and language teaching* (pp. 7–28). John Benjamins. https://doi.org/10.1075/scl.92.02gab

Hasebe, Y. (2015). Design and implementation of an online corpus of presentation transcripts of TED Talks. *Procedia – Social and Behavioral Sciences, 198,* 174–182. https://doi.org/10.1016/j.sbspro.2015.07.434

Huang, H-T., & Liou, H-C. (2007). Vocabulary learning in an automated graded reading program. *Language Learning & Technology, 11*(3), 64–82.

Hunston, S. (2022). *Corpora in applied linguistics* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108616218

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins. https://doi.org/10.1075/scl.4

Johns, T. (1990). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *CALL Austria, 10*, 14–34.

Kakiba, A., Nishigaki, C., & Oghigian, K. (2021). DDL applications to the seventh grade EFL classroom in Japan. *Bulletin of the Faculty of Education, Chiba University, 69*, 167–179.

Knight, D., Morris, S., Fitzpatrick, T., Rayson, P., Spasić, I., Thomas, E-M., Lovell, A., Morris, J., Evas, J., Stonelake, M., Arman, L., Davies, J., Ezeani, I., Neale, S., Needs, J., Piao, S., Rees, M., Watkins, G., Williams, L., Muralidaran, V., Tovey-Walsh, B., Anthony, L., Cobb, T., Deuchar, M., Donnelly, K., McCarthy, M., & Scannell, K. (2020). *CorCenCC: Corpws Cenedlaethol Cymraeg Cyfoes – the National Corpus of Contemporary Welsh*. Cardiff University. http://doi.org/10.17035/d.2020.0119878310

Kunioshi, N., Noguchi, J., Tojo, K., & Hayashi, H. (2016). Supporting English-medium pedagogy through an online corpus of science and engineering lectures. *European Journal of Engineering Education, 41*(3), 293–303. https://doi.org/10.1080/03043797.2015.1056104

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics, 22*(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Ma, Q., Tang, J., & Lin, S. (2021). The development of corpus-based language pedagogy for TESOL teachers: A two-step training approach facilitated by online collaboration. *Computer Assisted Language Learning, 35*(9), 2731–2760. https://doi.org/10.1080/09588221.2021.1895225

McCarthy, M., & McCarten, J. (2022). Corpora for teaching social conversation. In R. R. Jablonkai & E. Csomay (Eds.), *The Routledge handbook of corpora and English language teaching and learning* (pp. 102–115). Routledge. https://doi.org/10.4324/9781003002901-9

McCarthy, M., McCarten, J., & Sandiford, H. (2014). *Touchstone* (2nd ed.). Cambridge University Press.

McCarthy, M., & O'Keefe, A. (2014). Spoken grammar. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language* (4th ed., pp. 271–287). National Geographic/Cengage.

Meunier, F., & Reppen, R. (2015). Corpus vs. non-corpus informed pedagogical materials: Grammar as the focus. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 498–514). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.028

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review, 66*(1), 60–92. https://doi.org/10.17763/haer.66.1.17370n67v22j160u

Plonsky, L., & Ziegler, N. (2016). The CALL-SLA interface: Insights from a second-order synthesis. *Language, Learning and Technology, 20*(2), 17–37.

Prowse, P. (Ed.) (n.d.). *Cambridge young English readers*. Cambridge University Press.

Savický, P., & Hlavácová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics, 9*(3), 215–231. https://doi.org/10.1076/jqul.9.3.215.14124

Schmitt, N., Dun, K., O'Sullivan, B., Anthony, L., & Kremmel, B. (2021). Introducing knowledge-based vocabulary lists (KVL). *TESOL Journal, 12*(4). https://doi.org/10.1002/tesj.622

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learning-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213–234). John Benjamins. https://doi.org/10.1075/lllt.6.14sei

Simpson-Vlach, R., & Leicher, C. (2006). *The MICASE handbook: A resource for users of the Michigan corpus of academic spoken English*. Michigan ELT. https://doi.org/10.3998/mpub.101203

Sinclair, J. (Ed.). (1987). *COBUILD dictionary* (1st ed.)*.* Collins ELT.

Staples, S. (2015). *The discourse of nurse-patient interactions: Contrasting the communicative styles of U.S. and international nurses*. John Benjamins. https://doi.org/10.1075/scl.72

Staples, S. (2022, July 15). *Learner corpora and data-driven learning: moving toward an asset-based approach* [Plenary session]. 15th Teaching and Language Corpora (TaLC) Conference, Limerick, Ireland.

Staples, S., & Dilger, B. (2018-). *Corpus and repository of writing* [Learner Corpus Articulated With Repository]. https://crow.corporaproject.org.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge University Press.

Thompson, P., & Nesi, H. (2001). Research in progress, the British Academic Spoken English (BASE) corpus project. *Language Teaching Research, 5*(3), 263. https://doi.org/10.1191/136216801680223443

Tribble, C. (2001). Small corpora and teaching writing. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small corpus studies and ELT: Theory and practice* (pp. 381–408). John Benjamins. https://doi.org/10.1075/scl.5.22tri

Wang, J., Liang, S-I., & Ge, G-C. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*(4), 442–458. https://doi.org/10.1016/j.esp.2008.05.003

Wang, Q. (2022). The use of semantic similarity tools in automated content scoring of fact-based essays written by EFL learners. *Education and Information Technologies, 27*(9), 13021–13049. https://doi.org/10.1007/s10639-022-11179-1

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes, 28*(3), 170–182. https://doi.org/10.1016/j.esp.2009.04.001

West, M. (1953). *A general service list of English words*. Longman, Green and Co.

Willis, D. (2003). *Rules, patterns and words: Grammar and lexis in English language teaching*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733000