FELT
Focus on ELT Journal

Focus on ELT Journal
ISSN: 2687-5381
focusoneltjournal@gmail.com
Yildiz Teknik Üniversitesi
Turquía

Using corpora for language teaching and assessment in L2 writing: A narrative review

Faruk Kaya, Ömer; Uzun, Kutay; Cang#r, Hakan
Using corpora for language teaching and assessment in L2 writing: A narrative review
Focus on ELT Journal, vol. 4, núm. 3, 2022
Yildiz Teknik Üniversitesi, Turquía
Disponible en: https://www.redalyc.org/articulo.oa?id=688974207004
DOI: https://doi.org/10.14744/felt.2022.4.3.4

Using corpora for language teaching and assessment in L2 writing: A narrative review

Ömer Faruk Kaya a ofarukkaya@trakya.edu.tr
Trakya University, Turquía
https://orcid.org/0000-0001-7329-5557
Kutay Uzun b kutayuzun@trakya.edu.tr
Trakya University, Turquía
https://orcid.org/0000-0002-8434-0832
Hakan Cangır c hcangir@ankara.edu.tr
Ankara University, Turquía
https://orcid.org/0000-0003-2589-2466

Focus on ELT Journal, vol. 4, núm. 3, 2022

Yildiz Teknik Üniversitesi, Turquía

Recepción: 09 Septiembre 2022
Revisado: 22 Octubre 2022
Aprobación: 22 Noviembre 2022
Publicación: 30 Diciembre 2022

DOI: https://doi.org/10.14744/felt.2022.4.3.4

Redalyc: https://www.redalyc.org/articulo.oa?id=688974207004

Abstract: Corpora have primarily been used in linguistic research, but they have not yet become a pedagogical mainstay of language teaching and assessment practices. Therefore, this narrative review paper aimed to inform practitioners and researchers by examining the advantages and disadvantages of data-driven learning and exploring the use of corpora in foreign language teaching, particularly in writing. Specifically, the goals of this paper include: (1) elucidating what data-driven learning is and its potential to shape the learning experience, (2) explaining and exemplifying how learner corpora can guide EFL learners with particular attention to academic writing, and (3) providing insights into the indirect uses of corpora in teaching and assessing academic writing in L2. The review has met its objectives by presenting evidence compiled from the results of corpus-related studies and references to the use of corpus in language instruction.

Keywords: Academic writing, Data-driven learning, Corpus Linguistics, Learner Corpus, Corpus-assisted writing assessment.

Introduction

In this narrative review, we attempt to provide comprehensive analysis of the current knowledge regarding the use of corpora in foreign language teaching. We start our discussion and summary of the target literature by using data-driven learning as a generic heading and then narrow down our focus to the use of corpus for language teaching and assessment. More specifically, our aim is to provide the practitioners with a smooth introduction into the field and help them gain insights into the use of current corpus tools in the foreign language classroom. Our final humble aim with this narrative review is to bring the issue of data-driven learning within the scope of language learning and assessment to light, particularly in the Turkish context, and trigger further studies combining corpus linguistics and language acquisition research with strong and practical pedagogical implications. We hope this review with its up-to-date examples and reader-friendly narration will achieve to present the concept of data-driven learning (or "corpus- assisted language learning")

from a broad perspective and be used in educational contexts by faculty (especially in introductory seminars at universities) to expose students to the related literature in their field of study.

To have an exhaustive list of related studies for our narrative review, we made a list of the prominent figures in the field (e.g., Gaëtanelle Gilquin, Sylviana Granger, Anne O'Keeffe, Peter Crosthwaite, Thomas Cobb, Pascual Pérez-Paredes, and Yukio Tono, to name but a few), and sought to transfer their insights into our summary in a logical way. Additionally, we scanned through special issues of journals (e.g., *Language teaching*, Corpora in language teaching and learning special issue; *International Journal of Applied Linguistics*, Corpus-based language teaching and learning: Applications and implications special issue) to explore the current trends in the use of corpora in language learning and teaching. The ideas regarding which corpus tools to present have been borrowed from the recently published articles since those tools are considered cutting-edge and widely used by researchers and practitioners around the world. Last but not least, we grounded our main review layout and idea organization in the book chapters dedicated particularly to data-driven learning (e.g., Crosthwaite & Cheung, 2019; Gilquin & Granger, 2022). The following sections (a) summarise data-driven learning by relating it to certain theoretical backgrounds, (b) discuss the use of corpora in language teaching by giving examples of its direct and indirect applications, (c) scrutinize the use of corpora for language assessment highlighting their versatility in teaching (academic) writing and automated scoring.

Literature Review

Data-driven learning

The widespread adoption of the internet and growing technology have promoted changes in the understanding of language education in recent years. Corpus linguistics is an innovative way of language analysis through research materials called corpora, "a collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety" (McEnery et al., 2006, p. 5). The scope of corpus linguistics is not only limited to language research. First mentioned in Johns (1991), the term data-driven learning (DDL) refers to the pedagogical application of corpus linguistics. In DDL, students analyze the language using corpus tools and follow similar procedures of linguistic analysis and, optionally, the teacher acts as a coordinator of the student-led research. In other words, learners take the role of researchers by identifying and analyzing recurrent patterns in corpora to make generalizations and test their hypotheses about language (Johns, 1997). Learner researcher or scientist approach in language teaching is not new. Cobb (1999) found that students using lexicography tools (corpus tools) to learn language performed better in transferring their vocabulary knowledge to novel contexts than those who

did not. The effectiveness of DDL is also reflected in the recent literature. In a metanalysis, Cobb and Boulton (2015) analyzed eight studies employing pre–posttests as a treatment to measure the effectiveness of using corpora in teaching. The overall effect size (Cohen's d) was reported to be 1.68, indicating that the use of corpora in language teaching is highly effective. Gilquin and Granger (2022) cite several studies exploring the effect of DDL diversified through the years (e.g., Crosthwaite, 2020; Meunier, 2020; Yao, 2019). A shift of interest has been observed, and the monopoly of the English language in these studies has been broken with growing (though still limited) interest in other languages (e.g., Yao, 2019). Another change of focus and a potential area of research promoting further investigation is the use of DDL with young learners. Studies in the literature are rare, and Crosthwaite's attempt (2020) is the first in the related literature, which can guide future researchers. Additionally, while earlier studies focused on the use of DDL, particularly for writing, approaches exploiting other skills and activities have also been developed (e.g., Meunier, 2020), though further research is needed.

To explain language learning, DDL approach embodies a range of learning theories, such as constructivism, the noticing hypothesis, and Vygotskyan sociocultural theories. Constructivism is a theory that supports the notion that learners build knowledge actively, mainly through inductive processes, and learning should be an independent process (Collentine, 2000). Besides, Flowerdew (2015) notes that inductive learning fosters the activation of higher- order thinking skills, such as hypothesis forming and drawing inferences. By stimulating higher-order thinking skills, it is possible to retain what is learned for a longer period and improve language skills (Corino & Onesti, 2019). Since constructs are taught without giving the rules explicitly, and learners discover the rules by themselves, Schmidt's *Noticing Hypothesis* (Schmidt, 1990, 2001) can provide a theoretical ground for DDL, as well. According to Schmidt, noticing is an essential step towards language acquisition and is facilitative of learning. Using corpus tools, a range of 'awareness-raising' and 'consciousness- raising' activities can be designed to help the learners to notice the target forms and infer the grammar rules on their own. Keyword-in-context (KWIC) function of corpus tools, for instance, presents learners with textual enhancements by highlighting the target structure in a sentence making the input more salient. To illustrate the effectiveness of using concordancers, Smart (2014) compared different approaches to grammar instruction, namely, inductive corpus- informed instruction, deductive corpus-informed instruction, and traditional grammar instruction without the use of corpora (e.g., Presentation–Practice–Production (PPP)). Though the focus of the study is limited to teaching passive voice in English, the results indicated that receiving inductive corpus-informed instruction leads to significantly improved grammatical ability.

Despite its assumed benefits and effectiveness, second language teachers may not employ DDL in their classrooms (Flowerdew, 2010). The limited spread of DDL can be attributed to the lack of clarity

regarding the theoretical background (see O'Keeffe, 2021 for an overview), limitations in the pedagogical application, practitioners' prejudices against its use (Satake, 2020), and lack of research investigating the issue from different angles (e.g., the effect of DDL on learner anxiety – Zare et al., 2022). To begin with, although noticing is claimed to be a "necessary and sufficient condition for the conversion of input to intake" (Schmidt, 1990, p. 129), SLA research explains that merely noticing a feature does not immediately lead to acquisition or intake. Exposure is another concept that is fundamental for language acquisition. Indrarathne et al. (2018) suggest that an analysis of the type (whether guided or unguided) and frequency of exposure might help determine the extent of exposure for students to notice the target linguistic feature. Therefore, observing the mental activities of learners is key to understanding the nature of such concepts as noticing and the necessary length of exposure. It has been a topic of discussion that some technologies like key stroking, voice capturing, and eye tracking might yield valuable information about the learners' cognition (see De Smet et al., 2018; Indrarathne et al., 2018; Smith, 2012). Such instruments might help increase the reliability of DDL studies on noticing and exposure. According to O'Sullivan (2007), engaging in mentally challenging activities that require a process of reflection and reasoning stimulates the learners' cognition and facilitates the development of learning processes. However, O'Keeffe (2021) notes that the lack of testing of the link between the activation of higher-order thinking skills and DDL casts doubts on the claims and leaves the statement open to interpretation. Another argument against DDL is that some learners may resist independent process-oriented learning and getting learners to explore large chunks of data might result in no discovery (O'Keeffe, 2021). To address this problem, the teacher can guide students through question prompts that can facilitate problem-solving and reasoning processes. For example, Chang and Sun (2009) found the use of question prompts to have had a beneficial effect on learners' performance and confidence to self-edit their writing, and the use of prompts might work as a pathway to independent learning. A study (Zare & Karimpour, 2022) underlining the insufficient research focus on learner psychology and approaching the issue from the students' perspective concludes that learners think DDL approach, which encourages the use of concordances in language learning, is less appealing and motivating than a traditional instruction approach. Additionally, despite the current dominance of mobile phones in learning and teaching languages, most research studies choose to explore the use of DDL approach through computers only. That is to say, the issue of mobile data-driven language learning has gone unnoticed. One of the rare studies by Pérez-Paredes et al. (2019) criticizes this lack of interest and reports their participants' positive attitudes towards the mobile-based DLL approach.

Most teachers consider themselves to lack the knowledge to use corpora, find corpus use time-consuming, and therefore do not adopt data-driven learning in their classrooms (Satake, 2020). Thus, to

encourage and inform the practitioners and teachers, this study explores the uses of corpora in both language teaching and language assessment, especially for writing skills, and it offers a rationale for using corpora. The following section provides an elaborate definition of corpus and makes a case for using it for pedagogical reasons. Underpinning this, the paper sheds light on the uses of corpora in language pedagogy, both directly and indirectly. Then, it discusses corpus-assisted language assessment and defines the ways corpora can influence language assessment. Lastly, corpora's potential to contribute to the evolution of automated essay-scoring programs is discussed. In summary, focusing primarily on writing skills, this review serves as a guide for using corpora for language assessment and language teaching, and it gives insights into the possibilities that corpora can influence the development of essay-scoring automation.

Use of corpora in language teaching

In the last decades, corpora (i.e., large electronic collections of authentic and semi-authentic texts) and corpus-analytic techniques have given valuable information about the patterns of language. A perusal of corpora can give information on various categories including the behavior of words, multi-word phrases, grammatical patterns, semantic and pragmatic features, and distribution of various patterns across genres and registers (Flowerdew, 2009). To illustrate its use in the educational context, Timmis (2010) constructed a corpus by recording a conversation with his colleagues during a dinner to create material that could serve as a language model for his students. Also, Chambers and Le Baron (2007) formed a one-million- word academic corpus of French as a language resource for learners interested in developing their academic writing skills. Whether small or big, L1 speaker data on authentic language use can inform teachers, learners, and material designers about the proper uses of the target language and its norms. The use of representative corpora for textbook design has gained attention and the recent course books by well-known publishers like Cambridge and Oxford University claim they provide a corpus-informed syllabus with more authentic lexical and grammatical content. L1 corpora and their applications in EFL research have also paved the way for corpus-driven and balanced comprehensive vocabulary lists (e.g., new general service lists), which have guided material and curriculum designers (Brezina & Gablasova, 2015).

According to Granger (2002, 2015), although investigations of L1 corpora have been beneficial to the field of language learning and teaching, the data on its own is not enough for providing an ideal model for language learners. A survey of L1 corpus data, no matter how detailed, cannot give information on learnability factors, the perceived difficulty of structures, or the language transfer effect. Complementary material to L1 speaker corpora, learner corpora is defined as "systematic computerized collections of texts produced by language learners" (Nesselhauf, 2004, p. 125). Inquiry of learner corpora helps to detect the deviations of the

learner language from L1 speaker norms or spot the differences among groups of language learners. Such contrastive analysis might provide a wealth of empirical data that can help tailor teaching materials to better suit learner needs. For example, the Italian version of *The English in Mind* series contains 'Get it Right!' sections, which provide authentic examples of typical Italian learner errors (Granger, 2015). Learner corpora used here highlight the errors and give students a chance to compare their language with other groups of language learners. Additionally, a more recent research study (Naismith et al., 2022) claims that lexical frequency information extracted from a learner corpus can help measure the lexical development of language learners regardless of the learning context.

Both learner corpora and L1 corpora have contributed to language teaching in various ways and forms. To make a distinction, the pedagogical application of corpus tools and methods can be direct or indirect. Direct applications of the corpus refer to the hands-on use of data (i.e., Data-driven learning) while indirect applications include the creating and informing of pedagogical resources like reference books (Granger, 2015). The choice depends on the availability of corpus software and websites and the level of learners. Boulton (2008) notes that at earlier levels, exploiting corpora indirectly in the language classroom seems to be a more logical choice. Although we can see confident assertions in the literature highlighting the advantages of the indirect approach, in their meta-analysis Boulton and Cobb (2017) claim the opposite, and Vyatkina (2016) concludes that students can benefit from either approach.

Given these inconclusive findings in the literature, this review looks at both the direct uses of corpora and the indirect use of corpora while examining the place of learner corpora in both approaches.

Direct use of corpora in language teaching (Data-driven learning.

Corpus consultation in language teaching and learning has been more indirect than direct (McEnery & Xiao, 2011). It is attributable to several factors such as time constraints because of the curricular pacing, teachers' motivation to use corpora in their classrooms, the skill requirements of using corpora, access to computers or internet connection, and the lack of knowledge about the uses of corpora. Adapting data analysis tools like concordancers to pedagogical settings is of great importance because they might bring innovations and creativity to language teaching, especially for writing development. The advent of corpora has affected writing skill development more than any other skill area. Writing has gained importance in second language studies, partly due to increased dependence on computers for communication and the effects of globalization (Silva & Brice, 2004). However, learners need a good inventory of resources to help them gain autonomy in developing their writing skills. According to Cobb and Boulton (2015), massive but controlled exposure to input plays a major role in the reproduction of grammar, lexical, and other patterns of language students

need for communication. Analysis of large amounts of language samples requires the use of a computer program or web-based tools, such as a concordancer. In Stockwell's (2007) definition, a concordancer is a tool for searching through the contents of the database in different modules, like keywords-in-context (KWIC) or word sketches. Manual calculations or identification of language indices are both energy and time-consuming. On top of that, it requires expertise most students cannot attain. Through concordancers, learners can effortlessly enter the target structure they want to retrieve and get a varied picture of the authentic uses of language patterns. The main advantage of this is that it only takes a few seconds to scan the data, and most modern concordancers have user-friendly interfaces. In Lee and Swales (2006) four L2 English doctoral students using corpus tools compiled a corpus of their academic writings and compared the data with expert language users. In this strongly student-led research, participants found having access to the corpus empowering and helpful. Their opinions also matched their performance, as some students reported that their writing skills improved after the experience. If learners get the notion of statistically weighted lexical preferences with the assistance of concordancers, they may have the chance to produce lexically more sophisticated and natural-sounding utterances, particularly in academic writing. This idea is also reflected in the literature. For instance, Ander and Yıldırım (2010) in a study to identify and categorize the common lexical errors that appear in Turkish Elementary level EFL students found that the most frequent errors participants made were related to word choice category, which is likely to result in poorer writing performance. Although spelling checkers and feedback tools can detect spelling mistakes, they might not detect misused vocabulary. Crosthwaite (2017) in a study of DDL- mediated error correction, reported that students used Sketch Engine for Language Learning (*SKELL*; Baisa & Suchomel, 2014) and BNCWEB (Hoffmann et al., 2008) platform for error correction and corrected their word choice errors successfully.

The Sketch Engine is a multifunctional tool (accessed through a web interface), which is used by lexicographers, language researchers, and teachers. Users can have 30-day trial access to the website, and it requires payment when the trial period ends. It draws its sources from various corpora and is a versatile tool offering functions such as concordancing, thesaurus, and a word sketch for language analysis. To give an instance, students can be presented with a list of definitions for the words that they commonly confuse, words such as "aspect" and "consequence" or "principle" and "principal". Then, the teacher might ask students to work out possible definitions for the target words using concordance lines. Here, students try to discover the meaning while searching through concordance lines using SKELL. As a follow-up activity, the teacher can direct students to use the thesaurus to check their answers and produce their unique sentences using those words. Corpus query tools employed in the DDL approach must be 'learner-friendly' (Lee et al., 2019, p. 747) and accessible to students with limited corpus experience. Crosthwaite and Cheung (2019) state that complex corpus query tools

can easily discourage learners from using the DDL approach. SKELL has the potential to provide learners with simple and neat query output that is more appealing and encouraging for the uptake of DDL.

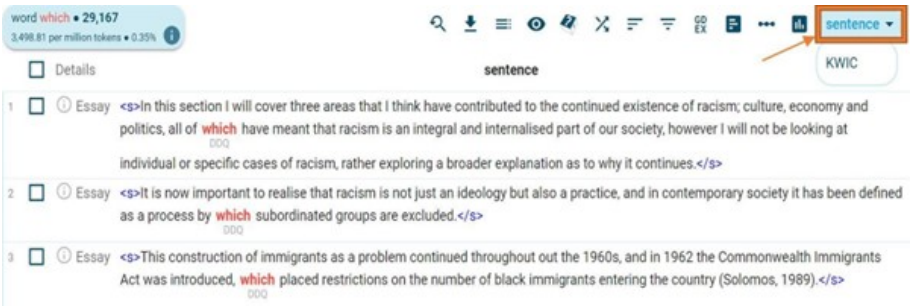

Figure 1
Concordance Function of Sketch Engine for Language Learning SKELL

Corpus tools can also be used to teach the difference between complex forms such as the relative pronouns "which" and "that". As shown in Figure 1, SKELL gives the option to display the target word in either KWIC form or sentence form. A teacher can ask students to switch to sentence form and find at least three sentences in each targeted form; "which" and "that". When students gather enough sentences, they start discussing whether clauses and phrases following "which" are necessary or not based on the corpus evidence. Then, students decide on the function of "which" and "that" by exploring the sample sentences and the patterns they appear in.

Collocations are a significant barrier for L2 learners; hence several programs have been developed to assist students in choosing the appropriate collocation (Granger, 2015). Collocaid (Frankenberg-Garcia et al., 2019), which is accessed through a free web interface, is a text editor for assisting students with the conventions of academic writing in an interactive DDL approach. Although it is still a prototype, Collocaid can answer such questions as: Is X a typical or appropriate collocate of Y? What words are conventionally used together with X? Collocaid provides options on the correct uses of collocations through multiple concordances via interactive menus (see Figure 2). British Academic Written Corpus (BAWE.), dictionaries and textbooks, crowd-sourced feedback (www.collocaid.uk), and various academic word lists form the database of Collocaid.

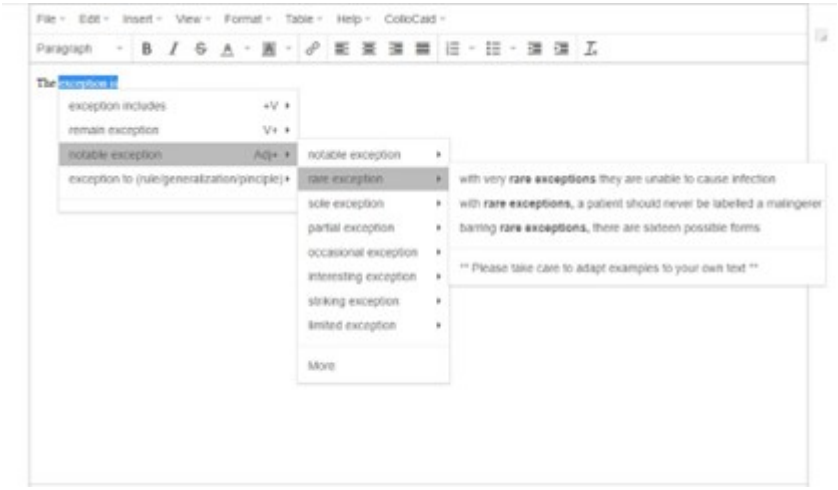

Figure 2

Excerpt from a collocations database underlying ColloCaid web-based text editor.

Collocaid can be an excellent auxiliary tool for language education because not only does it help with proofreading and editing, but it can also lead to discoveries about the words that go together. It follows a minimally intrusive way, as collocations are retrieved only on demand and in as much detail as users want.

Another free-of-charge corpus website, Just the Word (Edmonds, 2013), is a popular corpus-driven tool that demonstrates combinations of the queried word with other words as well as concordance lines highlighting the word combination patterns under observation. Its simple and user-friendly interface does not require potential users to have in-depth knowledge of corpora.

Figure 3

Just the Word; the function of the "alternatives from thesaurus" button.

When users type one or multiple words in the search box and click on the alternatives button, it can give information about the co-occurrence strength of those items (Figure 3). The strength of combinations is decided based on the frequency of occurrence. The green lines indicate the frequency of use, and the interface provides the users with various

word combination patterns (e.g., verb + noun, adverb + verb, verb + preposition, and such).

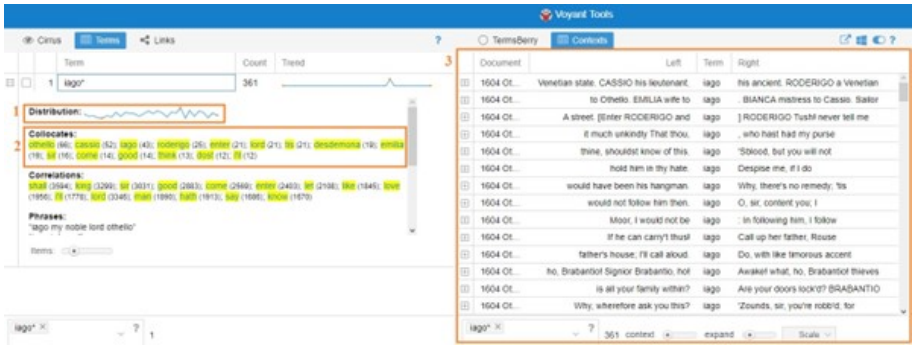

Figure 4

Excerpt for the use of grid tools in Voyant Tools. (Key: 1. Distribution of target term. 2. Collocates. 3.Contexts Tool (KWIC)).

Another versatile web interface that can be directly exploited in a language teaching environment is Voyant Tools (Sinclair & Rockwell, 2016). It is a free, web-based natural language processing (NLP) toolkit, which uses corpus methods to extract information, display measures of frequency, and analyze texts. What differentiates Voyant Tools from other text analysis tools is that it can display multiple visualizations or grid tools simultaneously. As represented in Figure 4 collocations and KWIC related to the word "Iago" are retrieved together. "Iago" is a character from Shakespeare's play, and this feature of the Voyant tool can come in handy for introducing a new character, setting the scene, and getting students to brainstorm about the topic. Users can either integrate the pre-loaded corpora (mainly literary texts) in the system into their instruction or create their own (learner) corpus and build their instructional strategies on this specialized corpus or a more representative corpus of L1 academic English like BAWE.

Rather than using a web interface, if an instructor wants to employ stand-alone software and use it in language teaching, AntConc (Anthony, 2022) could be a good option. The software is open-source and widely used by both researchers and language instructors around the world. With Antconc, one can investigate lexical and collocational frequency, create and compare wordlists, explore word clusters and n-grams either in an L1 corpus or a learner corpus (student writings for instance), create word clouds (see Figure 5) to help learners brainstorm on a particular subject before writing, and compare corpora to detect keywords (e.g., L2 English learner corpus vs. BAWE). Those features have the potential to guide language instructors while designing materials (an indirect way of DDL) and help them design in-class activities through which they can present new grammar structures in naturally occurring contexts or introduce academic registers to novice writers.

Figure 5
Excerpt for the use of word cloud in AntConc taken from Laurance Anthony's twitter thread

Indirect use of corpora in language teaching

While direct uses of corpora (data-driven learning or hands-on experience of corpora) play an important role in helping to decide how to teach, indirect uses of corpora are more concerned with informing teachers on what to teach. In the 1980s, Collins Birmingham University International Language Database (COBUILD) project laid the foundations for the development of corpus-based language teaching materials. Since then, corpora have been an invaluable tool for various areas of reference publishing, namely dictionaries, reference grammar, teaching material development, and syllabus design. After the pioneering work of *Collins COBUILD English Language Dictionary*, the dictionaries in the following three decades have made use of corpus data in such a way that as Hunston (2002, as cited in McEnery & Xiao, 2011) stated "even people who have never heard of a corpus are using the product of corpus-based investigation" (p. 96). Thanks to the conveniences and advantages brought by the corpus method, lexicographers can now reach valid and empirically based information on language use and its frequency of occurrence. Today, many other popular dictionaries (e.g., *Longman Dictionary of Contemporary English, Macmillan Dictionary Online*, and *Oxford Collocations Dictionary for Students of English* are corpus-based in one way or another.

According to Granger (2015), the impact of corpora on pedagogical grammars is less noticeable than on dictionaries. Distinguishing between the common and uncommon language choices of L1 users and the relative uses of those choices in context is important for both teachers and learners. Given that grammars are commonly used as reference books for understanding language forms, they should provide reliable and genuine instances of language that are up to date. There are various reasons to use corpora as a reference for the creation of grammar books. It is discussed in

McEnery and Xiao (2005) that non-corpus consulted grammar is prone to contain biases, and corpus consultation can enhance the quality of grammatical descriptions. It is difficult to reach and store large chunks of language samples without the help of corpus methods, thereby writers may write grammar descriptions intuitively. Thanks to corpus tools and corpora, grammars now take their source from a more expansive database of authentic language samples. *Longman Grammar of Written and Spoken English* and *Cambridge Grammar of English* are popular examples of corpus-influenced reference grammars.

The prominence of authentic and updated language examples in corpora has also found uses in syllabi design, especially those focused on communicative competence (Hymes, 1972) and vocabulary learning. Corpus data on L1 speakers give valuable insights into the patterns that learners are likely to encounter in authentic communicative situations. Information gathered from large L1 corpora might help dictionaries include more detailed descriptions of phrases and vocabulary, which might reduce misrepresentations. *Lexical syllabus* (Willis, 1990) for example, is organized around a mini corpus of pragmatically useful everyday words, and it draws heavily on spoken and written text in the target language. It relies on the provision of frequency information and authenticity of language made possible by corpus linguistics. There are also recent research studies (e.g., Cangır, 2021) claiming that we should combine corpus- extracted objective frequency values with L1 (or advanced L2) users' frequency judgments to have corpus-driven and pedagogically more convenient language teaching materials. Adding to the examples mentioned about the indirect corpus application, corpora have influenced the field of language assessment as well. Using corpora in assessment and its potential to inform test scoring methods, both automated and human, will be discussed in the next section.

Use of corpora for language assessment

Testing or assessment can be defined as "the systematic gathering of language-related behavior to make inferences about language ability and capacity for language use on other occasions" (Chapelle & Plakans, 2013, p. 241). In recent years, a good number of survey articles have shown that there is a growing interest in using corpus linguistics to inform the development and validation of language assessment (Cushing, 2017). The arguments about the benefits of using corpora in reference publishing are of equal relevance to language assessment. Language assessment, like reference publishing, benefits from the capacity of corpus linguistics for comparative analysis of language. That is, the availability of large chunks of language data on both learners and L1 speakers may help distinguish between language users at various levels of proficiency. The information provided by the comparative analysis of L1 and L2 corpora might aid the construction of test items that are more consistent with the proficiency levels of L2 learners. Empirical evidence on learner language can also inform reference level descriptions and consequently influence

rating scales. Learner corpora even had an impact on the Common European Framework (CEFR) for languages (Council of Europe, 2001), which is a highly influential construct in language assessment. Tono's (2019) attempt to adapt the CEFR to the Japanese context could be given as a good example of using corpora (i.e., objective means) to decide benchmarks for language levels. A common problem with many rating scales is that they are created intuitively and cannot capture some aspects of language. Römer (2022) argues that in the rating scales of internationally recognized tests (e.g., IELTS, TOEFL iBT, Cambridge English: Advanced) descriptors of speaking proficiency do not adequately reflect authentic use of spoken English. Corpora can be used for verifying or updating rating scales. Römer (2022) suggests that implementing a phraseological approach (An approach to corpus analysis) in rating scale development can make speaking assessment more consistent with the authentic spoken language. Research on corpora can improve the detection of learner errors since it contains information about word usage and the use of grammatical patterns. In terms of tools for detecting errors, learner corpus research has long envisaged automatic approaches (e.g., Granger, 1994; Granger & Meunier, 1994). Analysis of annotated learner corpora can potentially highlight both interlingual and developmental errors. One advantage of corpus linguistics over conventional ways of error analysis is that it allows for a more systematic and exhaustive analysis of the underuse, overuse, and misuse of patterns. In a frequency-based corpus analysis, Huang (2015) documented and classified lexical bundle errors according to their structural characteristics and discourse functions (e.g., referential expressions and discursive organizers). He found that agreement errors (e.g., subject-verb agreement and antecedent-pronoun agreement) account for the majority of the errors in the essays of Chinese EFL English learners. Difficulties faced by learners can give clues about what to select as a test item or add as a distractor to a question since needs analysis is an important part of teaching. Corpora have also influenced the making of NLP algorithms for detecting and correcting errors. E-rater® scoring engine by ETS (Attali & Burstein, 2006), for example, is an NLP-influenced feedback tool that can draw a writing proficiency profile of learners and correct their errors in categories like grammar, spelling, organization, and style.

Moving beyond errors, NLP techniques embedded in corpus software packages such as parsing, part-of-speech (POS) tagging, keyword extraction, and frequency displayers have paved the way for automated language analysis. Some publicly available noteworthy web-based NLP tools are, L2 Syntactic Complexity Analyzer (L2SCA) (Lu, 2010), Web-based Lexical Complexity Analyzer (Ai & Lu, 2010), Coh-Metrix 3.0 (Graesser et al., 2004) and The Compleat Lexical Tutor (Cobb, n.d.), to name a few. In addition to the web-based NLP tools, software such as UAM Corpus Tool by O'Donnell (2016) and software (e.g., Antconc) presented by Anthony (2022), and Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (*TAASSC*; Kyle, 2016) are free to download and use (Uzun, 2022). Some patterns in

language are difficult to identify and manual annotation of certain linguistic phenomena takes a long time, so researchers often opt for indices that are easier and more time- efficient to calculate. This results in a gap in both language research and test development, as some important measures of complexity and proficiency predictors remain uncovered. Researchers using the above-mentioned software can quantify several syntactic sophistication features, lexical complexity, cohesion, and discourse variables (Uzun, 2022). For example, by using TAASSC, Kyle and Crossley (2018) measured the proficiency of L2 writers using phrasal complexity indices and found that higher-graded essays include a higher proportion of nominal subjects containing a wider range of dependents. Another important finding by Kyle and Crossley (2018) is that fine-grained indices of phrasal complexity (e.g., number of subjects per clause) are better predictors of writing quality when compared to traditional syntactic indices (e.g., mean length of clause). This information can be useful for determining the specification of the content of tests and tasks. Finding relationships between complexity measures and L2 proficiency might also influence the development of automated and human scoring practices. Statistical calculations on word frequency or other complexity indices might work as a counterbalance to human rater intuition and increase the reliability of the scoring. Jarvis (2017) used learner corpus and statistical models to evaluate the perceptions of human raters on lexical diversity. At a minimal level, the use of corpora can serve as a consistency checker, in which human grading is compared to statistical results on target features.

Millions of learners are taking tests every year and the manual scoring of those tests is prone to be influenced by biases, fatigue, and inconsistencies. The machine learning approach to scoring has become important with the rising interest in high-stakes tests. The motivation behind this can be explained by the ability of automated essay scoring to provide reliable and accurate scoring of large volumes of test responses. The development and evaluation of automated essay-scoring systems (AES) have greatly benefited from the use of learner corpora (Higgins et al., 2015). For example, learner corpora and NLP tools can aid the system training and calibration of scoring engines (see Jarvis, 2017; Zechner et al., 2009). Some corpus- influenced scoring systems can be listed as; Pearson's Intelligent Essay Assessor™ (*IEA*; Landauer et al., 2000), e-rater® by ETS, Project Essay Grade (Page, 2003), and IntelliMetric® (Rudner et al., 2006) by Vantage Learning. MyAccess! ® (Vantage learning, 2007) using IntelliMetric, WriteToLearn® using IEA, and the Criterion® online writing evaluation software using e-rater can be counted as the adaptions of scoring engines to classroom use (Higgins et al., 2015).

Having its spark in the well-acknowledged scoring engines mentioned above and the studies in the automated scoring literature, our corpus-driven and NLP-enhanced project *Automated Grading of L2 Writing Using Corpus Linguistics and NLP Methods* aims to design a reliable automated essay-scoring algorithm. To achieve that goal, our team is investigating the predicting power of lexical sophistication and lexical

errors in L2 English writing performance. We are working on various mixed-effects models exploiting parameters from TAALES (Kyle & Crossley, 2015) and GAMET (Crossley et al., 2019) so far. Our preliminary results tentatively indicate that the number of words, a bidirectional lexical association measure (delta-p), concreteness ratings of the lexical items, frequency profiles, errors, and overall vocabulary knowledge of the participants can predict the overall writing performance of L2 English users (. = 350) to a moderate extent (*R.* =.45). To be more precise, longer texts, stronger delta-p, better overall vocabulary knowledge are associated with higher writing scores. On the other hand, higher academic lexical frequency, higher concreteness ratings, and higher error counts are associated with lower writing scores. The findings of the research and detected algorithms will be used to develop a model to predict writing performance in L2 English and design an automated grading software for L2 writing.

Conclusion

In this narrative review, we aim to provide a summary of the field of data-driven learning by approaching the issue from the perspective of teaching and assessment. We are well aware of the bias and subjectivity this type of review paper brings, and thus we accept that this summary is just another attempt to illuminate the use of corpora in language learning and our account of the phenomenon is likely to have its limitation. More studies like these should be conducted to have a more comprehensive understanding.

Potential uses of corpora are varied: they include data-driven learning, teaching material development, syllabus design, language testing, and many NLP applications. Given the growing popularity of learner corpora in language research, the present review has focused on the use of corpora in language pedagogy, focusing mainly on academic writing skills. The idea of using corpora in language teaching is promising yet not widely embraced by language teachers and not a mainstream application in their teaching practices. The overview of pedagogical applications of corpus findings and review of publications shared in this narrative review paper can be useful for raising consciousness on the use of corpora in various dimensions of language pedagogy. Corpora provide creative ways of designing and presenting activities and tasks that reflect the authentic language, as well as aiding the development of reliable teaching and assessment materials. The purpose of this narrative review article was to inform the readers about the potential of corpora in both direct (e.g., use of concordances to explore lexical patterns in academic writing) and indirect exploitation of corpora (e.g., use of corpora to create word lists) and to motivate teachers to use them. As mentioned in the earlier parts of this paper, narrative reviews are valuable pieces in that they have the potential to guide novice readers in the field and help shape future scientific endeavours. We hope this review will lead prospective researchers in the field in the right direction by giving them a brief

overview of the salient aspects of the target field. Finally, reviews like these will encourage practitioners to employ corpus tools more in their classes; the pedagogical use of corpora will reach a wider audience, and the use of corpora will become common practice.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References

Ai, H., & Lu, X. (2010, June 8–12). A web-based system for automatic measurement of lexical complexity. Paper presented at the 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA.

Ander, S., & Yıldırım, Ö. (2010). Lexical errors in elementary level EFL learners' compositions. Procedia - Social and Behavioral Sciences, 2(2), 5299-5303. https://doi.org/10.1016/j.sbspro.2010.03.864

Anthony, L. (2022). AntConc (Version 4.1.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. The Journal of Technology, Learning and Assessment, 4(3), 1-30.

Baisa, V., & Suchomel, V. (2014). SkELL: Web interface for English language learning. In A. Horák, & P. Rychlý (Eds.), Proceedings of recent advances in Slavonic natural language processing (pp. 63-70). NPL Publishing Consultants.

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

Boulton, A. (2008). DDL: Reaching the parts other teaching can't reach? In A. Frankenburg-García (Eds.), Proceedings of the 8th Teaching and Language Corpora Conference (pp. 38- 44). Associaçao de Estudos e de Investigaçao Científica do ISLA-Lisboa.

Boulton, A., & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. Language Learning, 67(2), 348-393. https://doi.org/10.1111/lang.12224

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the New General Service List. Applied Linguistics, 36(1), 1-22. https://doi: 10.1093/applin/amt018

Cangır, H. (2021). Objective and subjective collocational frequency Association strength measures and EFL teacher intuitions. Pedagogical Linguistics, 2(1), 64-91. https://doi.org/10.1075/pl.20014.can

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

Chambers, A., & Le Baron, F. (2007). Chambers-le Baron corpus of research articles in French. Oxford Text Archive, http://hdl.handle.net/20.500.12024/2527.

Chang, W. L., & Sun, Y. C. (2009). Scaffolding and web concordancers as support for language learning. Computer Assisted Language Learning, 22(4), 283-302. https://doi.org/10.1080/09588220903184518

Chapelle, C. A., & Plakans, L. (2013). Assessment and testing: Overview. In C. A. Chapelle (Ed.), The encyclopedia of applied linguistics (pp. 240-244). Blackwell/Wiley. https://doi.org/10.1002/9781405198431.wbeal0603

Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. Educational Technology Research and Development, 47(3), 15-31. https ://doi.org/10.1007/BF02299631

Cobb, T. (n.d.). Compleat Lex Tutor v.8.5 [Software]. Accessed 17 July 2022 at https://www.lextutor.ca

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), The Cambridge handbook of English corpus linguistics (pp. 478-497). Cambridge University Press. https://doi.org/1 0.1017/CBO9781139764377.027

Collentine, J. (2000). Insights into the construction of grammatical knowledge provided by user-behavior tracking technologies. Language Learning & Technology, 3(2), 44-57. https://doi.org/10125/25072

Corino, E., & Onesti, C. (2019). Data-Driven Learning: A Scaffolding Methodology for CLIL and LSP Teaching and Learning. Frontiers in Education, 4(7), 1-12. https://doi.org/10.3389/ feduc.2019.00007

Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. Cambridge University Press. Accessed 16 July 2022 at https://rm.coe.int/1680459f97

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing. Journal of Writing Research, 11(2), 251- 270. https://d oi.org/10.17239/jowr-2019.11.02.01

Crosthwaite, P. (2017). Retesting the limits of data-driven learning: feedback and error correction, Computer Assisted Language Learning, 30(6), 447-473. https://doi.org/10.1080/09588221.2017.1312462

Crosthwaite, P. (2020). Data-driven learning for the next generation: Corpora and DDL for pre-tertiary learners.

Crosthwaite, P., & Cheung, L. (2019). Learning the Language of Dentistry: Disciplinary Corpora in the Teaching of English for Specific Academic Purposes. John Benjamins. https://doi.org/10.1075/scl.93

Cushing, S. T. (2017). Corpus linguistics in language testing research. Language Testing, 34(4), 441-449. https://doi.org/10.1177/0265532217713044

De Smet, M. J. R., Leijten, M., & Van Waes, L. (2018). Exploring the process of reading during writing using eye tracking and keystroke logging. Written Communication, 35(4), 411447. https://doi.org/10.1177/07410883187 88070

Edmonds, P. (2013). Just The Word. Accessed 17 July 2022 at http://www.ju st-the-word.com/

Flowerdew, J. (2009). Corpora in Language Teaching. In M. H. Long & C. J. Doughty (Eds.), The handbook of language teaching (pp. 327-350). Wiley-Blackwell. https://doi.org/10.1002/9781444315783.ch19

Flowerdew, L. (2010). Using corpora for writing instruction. In A. O'Keeffe, & M. McCarthy (Eds.). The Routledge handbook of corpus linguistics (pp.

444-457). Routledge. https://www.routledgehandbooks.com/doi/10.4324/9780203856949.ch32

Flowerdew, L. (2015). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska & A. Boulton (Eds.), Multiple affordances of language corpora for data-driven learning (pp. 15–36). John Benjamins. https://doi.org/10.1075/scl.69.02flo

Frankenberg-Garcia, A., Rees, G., Lew, R., Roberts, J., Sharma, N., & Butcher, P. (2019). ColloCaid: a tool to help academic English writers find the words they need. In F. Meunier (Eds.), CALL and complexity – short papers from EUROCALL 2019 (pp.144–150). https://doi.org/10.14705/rpnet.2019.38.1000

Gilquin, G., & Granger, S. (2022). 'Using data-driven learning in language teaching'. In A. O'Keeffe, & M. McCarthy (Eds.) The Routledge handbook of corpus linguistics. Second Edition (pp. 430-442). Routledge. https://doi.org/10.4324/9780367076399-30

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. Behavior Research Methods, Instruments, & Computers, 36(2), 193-202. https://doi.org/10.3758/BF03195564

Granger, S. (1994). The Learner Corpus: A revolution in applied linguistics. English Today, 10(3), 25-33. https://doi.org/10.1017/S0266078400007665

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), Computer learner corpora, second language acquisition and foreign language teaching (pp. 3–33). John Benjamins. https://doi.org/10.1075/lllt.6.04gra

Granger, S. (2015). The contribution of learner corpora to reference and instructional materials design. The Cambridge Handbook of Learner Corpus Research, 485-510. https://doi.org/10.1017/cbo9781139649414.022

Granger, S., & Meunier, F. (1994). Towards a grammar checker for learners of English. In U. Fries, & G. Tottie (Eds.) Creating and using English language corpora (pp. 79-91). Rodopi.

Higgins, D., Ramineni, C., & Zechner, K. (2015). Learner corpora and automated scoring. In S. Granger, G. Gilquin, & F. Meunier (Eds.), Cambridge handbook of learner corpus research (pp. 567–586). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.026

Hoffmann, S., Evert, S., Smith, N., Lee, D., & Berglund-Prytz, Y. (2008). Corpus linguistics with BNCweb-a practical guide (Vol. 6). Peter Lang.

Huang, K. (2015). More does not mean better: Frequency and accuracy analysis of lexical bundles in Chinese EFL learners' essay writing. System, 53, 13-23. https://doi.org/10.1016/j.system. 2015.06.011

Hymes, D. (1972). On communicative competence. In J. Pride, & J. Holmes (Eds.), Sociolinguistics (pp. 269-285).

Notes

1 It can be accessed free-of-charge for research and teaching purposes. (https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written- english-corpus-bawe/)

Notas de autor

a M.A Student, Social Sciences, Trakya University, Türkiye, ofarukkaya@trakya.edu.tr

b Assoc. Prof. Dr., Department of Foreign Languages Education, Trakya University, Türkiye, kutayuzun@trakya.edu.tr

c Lect., Dr., School of Foreign Languages, Ankara University, Türkiye, hcangir@ankara.edu.tr