Language Learning & Technology
ISSN 1094-3501
October 2017, Volume 21, Issue 3
pp. 9–27

EMERGING TECHNOLOGIES

Data-informed language learning

Robert Godwin-Jones, Virginia Commonwealth University

APA Citation: Godwin-Jones, R. (2017). Data-informed language learning. *Language Learning & Technology*, *21*(3), 9–27. Retrieved from http://llt.msu.edu/issues/october2017/emerging.pdf

Introduction

Data is being collected about us constantly—as we browse the web, make purchases online, run a red light, use an ID to enter a workplace, or carry out multiple other activities where cameras, sensors, and data capture devices are in place. The data collected may be of little practical value to us. In fact, it may in some instances be considered an invasion of privacy. On the other hand, large sets of data can be beneficial to educational institutions (learning analytics), companies (customer profiling), or government agencies (body cameras on police officers). The ubiquity of networks and the affordability of storage mechanisms has made it much easier to collect large amounts of data than in the past. In the field of language learning generally, and in second language acquisition (SLA) and computer-assisted language learning (CALL) specifically, data has played a central role, through collections of language in use (written or spoken) from both native speakers and learners.

When systematic and representative in its selection, such a collection of texts is commonly called a corpus. Corpora have been used in areas as different from one another as lexicography (real world examples in dictionaries) and air traffic safety (transcriptions of pilot–tower communications). Particular kinds of corpora serve special purposes, such as parallel corpora for machine translation (the same texts translated into a second language [L2]) or children speech data for identifying childhood developmental issues. Although data collection has been used in language learning settings for some time, it is only in recent decades that large corpora have become available, along with efficient tools for their use. Advances in natural language processing (NLP) have enabled rich tagging and annotation of corpus data, essential for their effective use in language acquisition applications. Corpus consultation by both teachers and learners has become more common, most often through the use of concordance software that shows word and expression search results in context.

The "corpus revolution in both applied linguistics and language instruction" (Boulton & Cobb, 2017, p. 388) is in part a product of the greater availability of corpora, corpus tools, and more powerful computers, but also is due to the fact that data-based learning aligns with current theories and practices of SLA (see Chambers, 2007; Flowerdew, 2015). Corpus use integrates well into constructivist and learner-centered approaches to language acquisition. It supports the mandate in contemporary communicative language instruction for the use of authentic language materials and for the development in learners of metalinguistic knowledge and learner autonomy. At the same time, the prominence of formulaic and patterned language in spoken discourse, as demonstrated by the analysis of captured corpus data, has led to an "explosion of activity" (Gablasova, Brezina, & McEnery, 2017, p. 156) in research and language instruction related to language chunks (recurring word patterns). In this column we will be looking at different dimensions of corpus use in language learning.

Authenticity and "Real" Language in Corpora

Language teachers like to share with students examples of authentic language use, whether that be popular songs, magazine articles, or television clips. This can bring variety into the classroom, motivate students,

and serve to combine culture and language. Today, assuming available online access in the classroom, it is easy to bring such materials to students, although selection will vary with target language and location. This represents a radical change from pre-Internet days, when language teachers made regular treks abroad to collect *realia* (train schedules, menus, newspapers, etc.). I remember the excitement of being able to use a small cassette recorder to capture conversations with friends, dialogs during service encounters, or, in one case, an interview with a former member of the Hitler Youth. Such materials were invaluable assets, especially when combined with vocabulary lists, transcripts, and exercises, all manually created. Although they could be made available in a language lab or similar setting, the materials were mostly used by a teacher in the classroom. Today, audio, as well as video, graphics, and text—all in digital formats—can be easily made available online. Digital formats for multimedia allow random access and easy integration into CALL programs. Also possible now is easier processing of keywords or transcripts, with, in some cases, automatic captioning and transcription available. Digital texts can be annotated (sometimes automatically) and combined with resources such as graphics and audio recordings, as well as links to online reference materials.

Having authentic language materials available in digital formats allows them to be collected and organized, in effect creating a custom small corpus. This may add significantly to the benefits for classroom use, as Braun (2006) points out:

> In the learning and teaching context small corpora, and especially small and homogeneous corpora, have a number of advantages. They provide a more systematic range of material than any individual text or sample of spoken language (such as the occasional recording of a TV show) or other 'scattered' material collections (such as paper copies of newspaper articles, etc.) which are often used in teaching contexts. Contributions by different speakers/writers to similar topics make even a small corpus less dependent on the idiosyncrasies of an individual speaker/writer and provide a greater range of expressions. (p. 31)

A corpus likely calls to mind a very large collection of texts, such as the British National Corpus (BNC), a 100 million word collection of written and spoken English from a wide range of sources. While such a large corpus provides the representativeness needed for linguistic analysis, a small pedagogically-oriented corpus can be a valuable teaching asset.

The availability of recording devices proved to be a wonderful resource for language teachers, but has also been a boon to linguists. The possibility of recording, transcribing, and analyzing large amounts of written or spoken language provides never before available evidence of how language works. The research on this data through SLA, corpus linguistics, and conversation analysis has taught us much about the nature of language in actual use, as Ellis (2017) comments: "Corpus linguistics demonstrates that language usage is pervaded by collocations and phraseological patterns, that every word has its own local grammar, and that particular language forms communicate particular functions: Lexis, syntax, and semantics are inseparable" (pp. 40–41). Usage-based research in linguistics using data from corpora has confirmed that grammar and vocabulary are linked as *lexicogrammar*, a term coined by Halliday (1961) to describe the interdependence of lexis (vocabulary) and syntax (grammar). Using concordance software on large sets of language data, it is apparent that native speakers use language patterns extensively, repeating and recycling words and expressions over and over again. This has given rise to the prominence in linguistics accorded to word groupings (frequently used combinations of words or phrases). *Phraseology*, examining the use of formulaic or pattern language, has become an important sub-field in applied linguistics. It has been estimated that from 50% to as much as 80% of language use in English is formulaic (Geluso, 2013), although that obviously can vary according to context and individual. This has proven to be a fertile area also within SLA. The *lexical approach* to language acquisition emphasizes this codependence of grammar and vocabulary (see Harwood, 2002; Lewis, 1993; Millar & Lehtinen, 2008).

Another result of findings from collections of authentic language has been to demonstrate how unnatural the spoken language represented in textbooks tends to be. Dialogues typically present polite speakers using standard grammar, engaged in regular turn-taking to carry out successfully a transaction of some kind. In

real life, exchanges often do not build logically nor do they have clearly articulated goals. Instead, it is not unusual for a conversation to contain random, off-topic utterances, frequent meaningless back-channeling (e.g., *Yes, I see*), a lot of repetition, and frequent overlapping talk. Textbook dialogs are mostly quite different:

> By contrast, the language of some coursebooks represents a 'can do' society, in which interaction is generally smooth and problem-free, the speakers co-operate with each other politely, the conversation is neat, tidy, and predictable, utterances are almost as complete as sentences, no-one interrupts anyone else or speaks at the same time as anyone else, and the questions and answers are sequenced rather in the manner of a quiz show or court-room interrogation. (Carter, 1998, p. 47)

Language collected from real usage provides a counterpoint to this model, often containing more typical, non-goal oriented language, with dialogues not always geared towards transactional goals but towards developing relationships. Gilmore (2007) comments the following: "The contrived materials of traditional textbooks have often presented learners with a meagre, and frequently distorted, sample of the target language to work with and have failed to meet many of their communicative needs" (p. 103). Students will quite likely—and logically—assume that the language of the textbook and of the classroom can be generalized to actual conversational usage. Supplementing textbooks with examples drawn from corpora can provide a valuable complement to the standardized language most commonly used by textbook authors, introducing features of spoken language rarely part of textbooks.

This process is likely to necessitate some adjustment on the part of both teacher and students. Working with corpus data takes a willingness to adapt to a form of language and content presentation that is quite different from the nicely edited and packaged textbook view. As Boulton (2009) points out, the benefits are worth the possible discomfort:

> Learners, like teachers, might find the messy nature of real language in use to be destabilising at first, preferring the teacher to have all the answers. But it would seem disingenuous to coddle learners with simplified language, disempowering them and leaving them unprepared for the realities of the authentic language we are presumably preparing them for. (p. 10)

Some may argue, and have done so, that data from collected text collections, and especially from concordances displaying corpus data, do not actually represent "real" or authentic language either, given that that language is not presented in its original context or addressed to the public for which it was written or spoken (Waters, 2009; Widdowson, 1998). As Gilmore (2007) points out, when we speak of authentic materials for language learning, "most researchers use the term to refer to cultural artefacts like books, newspapers, magazines, radio and TV broadcasts, web sites, advertising, music and so on" (p. 107). However, these kinds of discourse are often carefully prepared or even scripted and are very different in tone, nature, and interest level from the everyday conversations and routine writing typically captured in corpora. From that perspective, corpus data reflects conversation as it is actually used—it may not be contextually authentic, but it certainly is genuine (see Braun, 2006).

The supposed lack of authenticity for corpus data derives principally from how data is represented in concordances: short snippets of text with minimal contextual information. There is, however, generally the possibility, if enabled in the software, for users to view longer text selections, or possibly the entire text from which the concordance line is taken. The Contemporary Written Italian Corpus (CWIC) enables swapping between a concordance line and the whole text it came from and browsing whole texts by text type (Kennedy & Miceli, 2010). Access to the full text, along with information the system provides about text provenance, can supply the context which may authenticate the text for the learner (see Duda & Tyne, 2010). From a teaching and learning perspective, the question of whether corpus data can be judged to be authentic or not is less important than whether that data can play a useful role in language learning (Widdowson, 2000). In any case, authenticity in today's digital world is a problematic concept. Kern (2014) points out that given the oversized role of online communication today, "notions of cultural authenticity are similarly problematized by the anonymous origin and massive reappropriation of much material available

on the Internet" (pp. 340–341). This holds true as well in terms of linguistic authenticity, as global online communities create discourse in multiple "language crossings" (Kramsch, 1998, p. 70).

Corpus Use in Instruction: Direct and Indirect Approaches

An area in which access to corpora has made a significant difference is lexicography. The real world usage cleaned from corpus data provides writers of dictionaries, thesauri, and other language reference materials a rich source of usage details and examples. This indirect benefit of corpora for language learning extends to the authoring of language learning materials, such as proficiency guidelines, standardized tests, and textbooks. Textbook authors, in particular, now have the ability to tap into corpus data to bring examples of interactional and pragmatic language, as well as to expose students to more genres, registers, and regional variations available in corpora. One of the first English language corpora, from the COBUILD project, led to new learner-oriented dictionaries (Sinclair, 1987), grammars (Sinclair, 1996), and ESL textbooks (Willis & Willis, 1988)—all of which benefited greatly from using resources the COBUILD corpus made available.

One of the benefits corpora have brought to teachers is help in discovering and annotating texts for language study that align with student interest and proficiency levels. Using NLP, tools and services have been developed that automatically analyze texts for readability, based on vocabulary (lexical diversity and density) and complexity (sentence length, text structure). One of the earliest such readability indices was developed in the 1960s by Swedish scholar Björnsson. It was called LIX, the Lasbarhetsindex (readability index). Subsequently, a number of other methods for evaluating the level or complexity of texts have been developed, as recently discussed in Pilán, Vajjala, and Volodina (2016) and Xia, Kochmar, and Briscoe (2016). Web-based tools are available for analyzing texts for readability in English (TextAnalyzer or Text Readability) and in German (Lesbarkeitsindex). The AntWordProfiler (from Laurence Anthony) is a freeware online tool for profiling the vocabulary level and complexity of English language texts. Words and Phrase (from Mark Davis) creates vocabulary lists and other structured information from user-supplied texts. The Adelex Text Analyser from the University of Granada evaluates lexical difficulty of texts written in English and features special tools for analyzing characteristics of texts for Spanish L1 users. That includes an analysis of cognates that "have been classified as 'transparent' (i.e., guessable) and 'opaque' (i.e., non-guessable), in order to establish the percentage of English words which do not add difficulty to a text for a Spanish speaker, even though their frequency may be low" (Adelex Text Analyser, 2017, n.p.). Chinkina & Meurers (2016) describe Form-Focused Linguistically Aware Information Retrieval, a web and corpus query system designed to allow teachers to find linguistically appropriate texts. The system can also feed into platforms providing input enhancement, such as WERTi, that automatically highlights text segments (according to targeted lexical or grammatical structures). Such a system can be used as well by students outside the classroom, to work with L2 texts of personal interest.

Corpus access has affected classroom language instruction in a variety of ways. When introducing vocabulary or grammar, for example, one might show students the item used in sample extracts from a corpus. In discussing common errors in writing, the teacher could show examples of native speaker use, with the term shown in different sample sentences. This direct integration of corpora in instruction usually involves the use of concordance software by teachers or students. Most common as a search retrieval interface is the keyword in context (KWIC) format to display lines containing the word or expression in its context of use. Sentences are truncated, with the software displaying a pre-determined or user-chosen number of words before and after the search term. This gives students real-world examples in context, thereby providing valuable usage information. That might be, for example, words frequently accompanying the search term, such as adjective-noun combinations like *handsome + man* (collocations) or how the term is incorporated grammatically, as in a particular kind of preceding verb, like *will* or *won't + budge* (colligation). Viewing idiomatic constructions in a concordance can be especially informative, as learners are able to see both literal and non-literal uses. The frequency of occurrence of an idiomatic expression in a corpus can provide guidance on how common that construction is among native speakers. The experience working with targeted corpus data can help students with learning vocabulary (especially how to use words

appropriately) and writing (how to make better use of cohesive devices, connectors, and genre conventions). According to the overview of corpus use in language learning by Boulton & Cobb (2017), the most common uses of concordances are in L2 writing and translation. Translation studies profit immensely from the availability of parallel and multilingual corpora (see Frankenberg-Garcia, 2005; Molés-Cases & Oster, 2015).

How students access corpus data depends on the teacher and on the dynamics of the classroom. Teachers may print out or show projected on a screen a list of examples which illustrate the usage of an expression or its grammatical behavior. This might take the form of a contrastive analysis, distinguishing L2 usage from patterns in the students' L1. Commonly, corpora data are used to highlight frequent errors in students' lexis or syntax. That includes mismatches between the frequency of use by learners versus native speakers, including overuse, underuse, or misuse of particular words or constructions. Studies often target examples of lexicogrammar such as phrasal verbs (in English, Mizumoto, Chujo, & Yokota, 2016; in German, Vyatkina, 2016a; in Spanish, Salido, 2016) or constructions that are important for discourse cohesion, such as linking adverbials (Cotos, 2014), logical connectors (Cresswell, 2007), or modal particles (Belz & Vyatkina, 2005). Corpora provide particularly good insight into collocations (see Geluso, 2013; Pereira, Manguilimotan, & Matsumoto, 2016; Salido, 2016; Thomas, 2015). Working with corpora can move students beyond an exclusive concern with grammar, leading them to appreciate that naturally sounding discourse is not just a product of correct syntax and knowledge of vocabulary but also depends on how language patterns are used. Corpus use traditionally could be found most often in instruction of more advanced learners, especially in composition courses. Recent studies have, however, shown success working with intermediate or even lower proficiency learners (e.g., Boulton, 2009; Karras, 2016; Mueller & Jacobsen, 2016; Oghigian & Chujo, 2010). Mukherjee (2004, 2006) has been an important pioneer in this area.

The rationale for incorporating corpus consultation in the early stages of language instruction is to demonstrate to students actual language patterns, supplementing textbook or classroom language. In the process, learners gain insight into the benefits of accessing corpus data. To maximize the impact on developing literacy in this area, it is beneficial to provide students with direct access to using a corpus. Students engaged in what corpus pioneer Johns (2002) labeled *data-driven learning* (DDL) typically use computers (their own or school-supplied) to run concordancing software. In this model, students engage with the content to uncover patterns of usage on their own, rather than searching for examples which confirm or illustrate grammar or usage rules supplied by the teacher or textbook. This inductive method features discovery learning, putting students in the driver's seat, with the goal being, in the words of Johns and King, to "cut out the middleman...the underlying assumption being that effective language learning is a form of research" (1991, p. 30). The idea is that the input enhancement provided by an expression in a KWIC display (usually bolded or color-highlighted) leads to *noticing* (Schmidt, 1990). This can provide students with insights regarding the gap between their own usage and that of native speakers. This is in line with constructivist theories of learning, as students build knowledge from personal interactions with the material (Boulton & Cobb, 2017; Flowerdew, 2015).

This learner-centered approach can contribute to learner autonomy as students gain skills and knowledge in the use of corpora for writing and reference. That is only likely to happen if learners are successful in their work with corpora in the classroom. In fact, a number of studies have pointed to problems using a hands-on approach to corpus consultation. These are summarized by Boulton and Cobb (2017); aside from general discomfort with computer tools, they point to the following issues:

> Chopped-off concordance lines may help expose patterns yet be off-putting to some and are not designed for gaining meaning as traditionally conceived via linear reading; most corpora are composed of authentic native language well beyond the comfort level of many learners; and DDL work requires substantial training, and the processes are time consuming when learners could simply be told or use pedagogically derived resources such as dictionaries. (p. 351)

Some of the issues relate to the user interface, usually KWIC, which can be confusing to students, especially

initially. The availability of more user-friendly renditions of corpus data is one of the developments that could help in moving students towards greater use of corpus consultation. Some projects have found that using printouts of concordance data, or other corpus-derived materials prepared in advance, may be a more effective and a student-friendlier approach than using computers in the classroom or lab (Huang, 2014; Smart, 2014; Vyatkina, 2016b). On the other hand, the meta-analysis by Boulton and Cobb (2017) points to somewhat better results for projects using computers.

Guided induction

An approach that has proven successful in a number of projects is guided induction (see Flowerdew, 2009; Johansson, 2009; Mizumoto et al., 2016). In this model, the teacher selects examples from the corpus data, but does not explicitly provide grammar or usage rules. In contrast to DDL, as envisioned by Johns (1997), the teacher here plays an active role. The selections from the corpus are chosen carefully to illustrate language features and to contain language and content accessible to students. As outlined in Smart (2014) and Vyatkina (2016b), this method typically involves the following steps:

1. Illustration: Students examine the corpus data.
2. Interaction: Students discuss the material and share observations and opinions with other students.
3. Intervention: As needed, the teacher provides scaffolding through additional examples, or more guidance.
4. Induction: Students discover rules for a particular feature, based on the materials studied.

Using such an approach, Smart (2014) found that students working on passive voice constructions in English outperformed students using other methods. In a study focused on students learning collocations in German, Vyatkina (2016a) found that guided induction was successful in working with students at a low proficiency level. That study also showed that working on new knowledge, not just learner errors, could be carried out with this method. Flowerdew (2015) discusses a number of projects, most using the inductive approach, and concludes that the results are "promising rather than conclusive" (p. 31). She finds that most studies are short-term and on a relatively small scale. She cites the need, as do others, for studies of longer-term benefits, which she speculates might be in the area of enhancements of metalinguistic and metacognitive knowledge (see also Mueller & Jacobsen, 2016).

This approach is possible with students using computers or tablets, but also using paper handouts. It may be, in fact, that handouts are more practical for enabling the "interaction" step listed here, as the students group together for discussion. This component of the process introduces social learning, something generally missing in corpus consultation in the classroom. This peer activity integrates corpus work more easily into the communicative-oriented classroom. Huang (2014) found that student learning was enhanced when corpus work was combined with small group discussion. Peer collaboration in this context may not always be effective, as shown in Cho (2016), if there is too substantial a gap in proficiency levels among learners. In a non-classroom setting, a social component could be implemented through online collaboration, for example, through Moodle (see Cotos, 2014). That might as well involve students communicating online with native speakers, using in real exchanges the knowledge from corpus work. In a project by Belz and Vyatkina (2005), students first worked with samples of usage (modal particles in German) before using the expressions in online exchanges. Learner diaries are another opportunity to practice use of concordance-learned constructions. They might provide a good opportunity to use features important in longer form writing such as discourse markers, transitioning expressions, or connectors.

Kennedy and Miceli (2017) describe the process of moving from seeking language patterns in corpora to re-configuring them for communication:

> In contrast to pattern-hunting, which involves open-ended questions, pattern-refining work is specific problem-solving with the corpus; it is aimed at finding models for patterns when you do know what you want to say and know one or more component words of the target pattern. (p. 94)

As indicated here, the use of corpus data in instruction may be targeted towards comprehension or towards

production. Corpus-derived data can be used as enriched input to aid in comprehension, enabling learners to focus on patterns of use. For active production of language, it is helpful to make students aware of morphological or syntactical properties, as Frankenberg-Garcia (2014) describes:

> While example sentences meant to facilitate the comprehension of a previously unknown word should contain sufficient context to enable a learner to infer what that word means, example sentences for language production should focus on which other words frequently go together with the target word (collocation) and on the grammatical preferences and constraints of the target. (p. 129)

For comprehension purposes, Frankenberg-Garcia (2014) suggests using "a definition plus examples that specifically contain contextual clues to facilitate understanding" (p. 139). For production, on the other hand, multiple examples for constructions like collocations are most useful. In any case, providing a wealth of examples is helpful for understanding and remembering, but also has the benefit of students becoming more aware of the patterned and recycled nature of real language usage.

In the guided induction method used by Frankenberg-Garcia (2014) and others, "learners were essentially spoon-fed with the right type of examples" (p. 141). In using corpora on their own, as one hopes students do, this guidance is not available. Studies have shown that using concordances can be a daunting experience, even for motivated learners. In a project involving future language teachers, Leńko-Szymańska (2014) found that as many as 14 training sessions with graduate students was not enough to make them comfortable with the use of corpora and concordance software. Due to technical and practical issues with DDL, it may be useful to consider providing a "corpus apprenticeship" to students, as outlined by Kennedy and Miceli (2001, 2010, 2017). Intermediate-level learners of Italian gained experience through a semester long course using corpora in a variety of ways, as an integral part of the course, not, as is often the case, an extra or experimental add-on. The goal was to provide students with corpus consultation literacy—in gaining proficiency in pattern hunting and defining—and to help them become familiar and comfortable with the principles of effective reference consulting. A study by Frankenberg-Garcia (2005) on learners as researchers, demonstrated the need for this kind of extended guidance, as students did not use effectively either paper-based or electronic reference materials. They particularly undervalued monolingual support options. Teachers who have queried their students in this area, or observed their behavior, will likely see these findings as familiar. Several studies have shown better results in students using concordances for look-up and writing rather than dictionaries or other conventional reference works (e.g., Boulton, 2009; Mueller & Jacobson, 2016). To prepare students for working independently with concordances and corpus query methods, it is important to show in the classroom advanced search functions such as wildcards and Boolean searches.

For working with grammatical constructions, it is likely that students having learned through constructivist approaches such as guided induction with corpora, may still need grammar reference material to consult. Increasingly, grammar references and tutorials incorporate linguistic knowledge and examples taken from corpus linguistics. Researchers have also developed approaches that strive to combine deductive and inductive approaches. The Chemnitz Internet Grammar, designed for German L1 learners of English, provides access to grammar rules and offers the ability to browse the corpus and use a concordance to find grammar patterns (Schmied, 2006). Similarly, the Check My Words toolkit provides a set of online tools incorporating corpora linked to an in-house grammar guide (Milton & Cheng, 2010). Users can toggle between the two resources and can call up hints and other help functions.

Learner Corpora

There has been growing interest in recent years in learner corpora (LCs), collections of written or spoken language by language learners. Metadata are used to provide information about sampling parameters, such as the learners' L1, proficiency level, writing or speaking context, possible study abroad, and so forth. Such collected data on student interanguage has been an important component of intelligent language tutors or

intelligent computer-assisted language learning (iCALL). Data is collected on common errors, so that system developers can anticipate learners making similar mistakes and thus build in appropriate feedback. Keeping user logs allows a system to provide customized feedback and individualized guidance (see Godwin-Jones, 2017). One of the possibilities for iCALL today is to automatically generate exercises for users. Language Muse is a tool that creates customized activities based on teacher-identified problem areas, using texts from a corpus or the web (Sabatini & Andreyev, 2016).

An issue central to the collection and processing of LCs is the absence of a commonly accepted error annotation scheme. Having a standard method of analysis and tagging makes it more likely that LCs can be shareable and be integrated into different analytical and display applications. Given how time-consuming a process LC annotation can be, collaboration and crowdsourcing can be valuable project assets—options enabled by standardization. The process is made more manageable if the content of the LC is restricted, for example, by being domain-specific. Another, semi-automatic approach to building an LC, is to have learner texts and teacher corrections fed into a database in a uniform way, as was done in the CorpusScript project.

Error analysis is frequently used in LCs in classroom applications, allowing the teacher to alert students to common problem areas in lexis or syntax. Often, a LC will be paired with a L1 corpus in order to provide contrastive analysis. Cotos (2014) studied the use of linking adverbials in English, using a L1 corpus only in one group, while the other had access to both the L1 corpus and a LC of the students' written work. While both groups showed improvement, it was more pronounced in the group with access to a corpus containing their own writing. Tono, Satake, and Miura (2014) provided coded error feedback to students on EFL essays, having them use corpus consultation in revisions. Rankin and Schiftner (2011) used a local corpus to improve student use of prepositions in English through consultation and comparisons with the International Corpus of Learner English. Belz and Vyatkina (2005) used a bilingual LC to help with improving pragmatic abilities of students learning German, contrasting learner and native speaker use of modal particles to make language more natural sounding.

LCs can be particularly useful if the data enables localization of issues related to a particular group of users, such as those with the same L1, those at the same proficiency level, or those having undertaken similar tasks. As LCs contain more language variety than L1 corpora, systematic analysis of a LC can be more difficult. Meurers and Dickinson (2017) point out that this makes it unlikely that one can use, without modification, existing annotation schemes or NLP tools developed for analyzing native language. The greater volume of annotation that is available, the more useful a LC will be for different research questions:

> For transparently connecting research questions in SLA with corpus data, systematic syntactic annotation of learner language needs to make explicit which source of evidence (morphosyntax, distribution, lexical subcategorization, top-down guidance from task context, etc.) is considered in determining the different levels of annotation. Where more than one source of evidence is considered but the evidence diverges as a characteristic of learner language, each can be encoded in separate layers of annotation. The searching and interpretation of the corpus data can then systematically refer to consistently defined corpus annotation layers—an essential prerequisite for sustainable use of annotated corpora for L2 research. (p. 85)

If the LC is open and expandable, researchers can add annotation layers as needed. Lüdeling, Hirschmann, and Shadrova (2017) used the Falco corpus of German learner language because it featured an "open, extensible, and well-documented multilayer corpus architecture" (p. 122). In any case, the schema used for metadata should be explicit, with clear documentation on the models and categories used. Systemization and sharing in this area benefit all corpus researchers. MacWhinney (2017) calls for a shared platform for corpus research as an important step forward in this area. The meta-data available for categorization is essential to being able to extract and compare data from different corpora. Gablasova et al. (2017) point to the importance of considering representativeness and comparability of corpora in research and teaching projects. The article points to the need for establishing and reporting speaker proficiency levels more fully in LCs. Wisniewski (2017) provides a framework for using Common European Framework of Reference for Languages proficiency levels to standardize data and achieve uniformity in testing.

It is in the analysis of written learner language that corpora have been particularly helpful. Combining the availability of large sets of annotated learner texts with advances in NLP, there are tools available for analyzing characteristics of L2 texts such as cohesion and complexity. Coh-metrix is a tool for computing computational cohesion and coherence metrics for written and spoken texts (McNamara, Graesser, McCarthy, & Cai, 2014), and it has shown to be effective in both L1 and L2 contexts (Vyatkina, 2016d). Kyle and Crossley (2015) describe a similar Tool for the Automatic Analysis of LExical Sophistication (TAALES) that uses advanced NLP techniques to provide text scores for a variety of lexical indices, such as word frequency, range, and academic language.

LCs are commonly collected from learners of the same L1. Salido (2016) represents a departure from that model, as the LC used in the project described came from three different L1 groups. This provides some interesting comparative insights into interference from different L1s—in this case, from Japanese, English, and Spanish speakers. One of the benefits LCs can provide is tracking the development of interlanguage over time. This presupposes that there is a long-term project and that data is collected and tagged at regular intervals with the same set of learners. This can offer valuable insights into individual variation in language use and in the development of proficiency. The Kansas Developmental Learner Corpus provides an example and model (Vyatkina, 2016c). Learner writing samples were collected every 3 to 5 weeks over several semesters. The samples were annotated for multiple linguistic features and learner errors, allowing for a variety of factors to examine in analyzing the data. That data provided information about group trends as well as about individual development paths.

Small and Learner-Created Corpora

There are several advantages to small-scale corpora, particularly as used in instructed language learning. The smaller size makes search results more manageable and browsing more practical. It is more likely that a smaller corpus is able to provide full access to complete texts. Likewise, the small size may make it more feasible to integrate audio-visual data. That option is particularly attractive for a corpus designed primarily to be used in a pedagogical context. Braun (2006) provides an example of a multimodal corpus, ELISA, consisting of interviews with English native speakers talking about their professions. As with most small corpora, there was no intent to be representative, rather topicality was the main consideration for inclusion. Each interview was structured the same, allowing comparisons across groups and individual speakers with the same essential language repertoire in each case. Small corpora may feature a customized user interface, which was the case for ELISA. Users could browse the web-based interface through thematic indices, interview summaries, or specific topics covered in all interviews. The design allowed learners to gradually explore the content:

> In a first step, learners can work with just a small number of sections—preferably those with which they are familiar from reading the entire interviews—in order to formulate a hypothesis on a particular lexical, grammatical or other issue. In a second step the range of interview sections can be extended to provide a more substantial amount of data to study the phenomenon in question and to further explore the hypothesis. Thus learners can gradually be led to benefit from the advantage of a corpus (compared to just working with one text). (p. 37)

Small corpora interfaces can offer a user-friendlier format, as the pedagogical intent allows for a streamlined and potentially more intuitive set of choices, compared to corpora mainly meant to serve linguistic research. As a further way to ease students into corpus consulting, Braun (2006) suggests starting not with a concordance view, but with a sample text, and only after that working with keywords, showing how the concordance can help in comprehension.

Another benefit of small corpora is the potential to engage learners with material of interest or practical benefit to them:

> Motivation can be increased by allowing learners greater involvement in creating the corpus, deciding what goes into it, or using their own productions…This helps them to see the relevance

> of what they are doing, which can also be achieved by working on language areas they know they have problems with. (Boulton, 2009, p. 9)

Studies have shown that this can indeed help motivate students and increase interest in using corpora (Cho, 2016; Cotos, 2014; Geluso & Yamaguchi, 2014; Seidlhofer, 2000). St. John (2001) found that a student using several different corpora for learning German judged their use effective as a tool for self-directing learning, despite some initial difficulties. A small group of Korean EFL students took on the role of "language detectives" (Johns, 1997, p. 101), working on improving their writing through corpus consultation, with positive results "showing initiative and motivation in their effort to try novel constructions on their own" (Yoon & Jo, 2014, p. 103). The study also showed that concordancing served to develop positive attitudes towards autonomous learning. In addition to the benefit of seeing improvement on their language ability, a motivating factor might have been working with materials with which they connected personally in some way. The SACODEYL multimodal corpus collected samples of youth language in seven different languages, on the assumption that "it [would] be interesting for students in school to listen to 'peer voices' and to young people's topics" (Braun, 2006, p. 31). One could envision such small corpora built around student-driven topics of interest, which might vary from hip-hop lyrics to sports fandom stories.

In fact, several projects have reported on student-created corpora. Chang (2014) had Korean students studying computer science and engineering compile their own corpora through student selection of papers and articles from journals in their fields. Chang reported that students appreciated the access to this additional local corpus as a complement to the Corpus of Contemporary American English (COCA) used, as they judged that corpus not to have a sufficient number of technical articles. Cotos (2014) also had students create a corpus of articles in their individual fields. She found that enabling students "to analyze their local corpus, which was something new for them, created a motivating learning environment where the corpus-based tasks were perceived as a personally relevant learning experience" (p. 217). Charles (2015) describes a similar project and discusses other examples of users building personal corpora. Millar and Lehtinen (2008) provide practical information about the process of creating a local corpus.

Small corpora lend themselves well for use in genre studies. Examples of particular forms of written or spoken discourse can be collected and commonalities examined. An analytical and descriptive process can lead into students using the language patterns and genre conventions in their own L2 creative writing. Rohrbach (2003) had students work with a corpus of tourist brochures. After studying features and language used in the brochures, students produced their own local tourist brochure. A sample student brochure (shown in Mukherjee, 2006, p. 16) demonstrates how a student successfully used structural and linguistic elements typical of such brochures. Goth et al. (2010) present a similar case study involving students writing fables with support from a corpus of fables analyzed through NLP techniques. Cambria (2011) used a multimodal corpus drawn from web new sites to train journalism students.

Chambers (2007) discusses the usefulness of small or domain-specific corpora, providing examples of corpora consisting of texts already familiar to learners. Chambers also sites an interesting byproduct of learners constructing and using corpora, namely *serendipity*—that is, encountering interesting combinations of words or uses of expressions through open-ended, exploratory browsing. Bernardini (2000) cites examples of students following idiomatic and literal uses of expressions in a concordance and engaging in spirited discussion of the differences. Kennedy and Miceli (2001) similarly refer to "treasure hunting" (p. 79) to describe the type of discovery learning which may accompany corpus consultation. This image of students having fun with language contrasts with the impression many may have of corpus use as a rather dry method using statistical analysis. An interesting approach to corpus consultation in the classroom would be to build games or other classroom activities around unusual or highly idiomatic constructions discovered while browsing or searching corpora. The CheckYourSmile project (in French) features a variety of online games, with items drawn from a domain-specific corpus (i.e., English technical terms). The project represents a collaborative corpus, with data initially entered into the system by teachers and subsequently supplemented by contributions from students. The latter are peer-reviewed and ranked before inclusion.

Access to Corpora and Tools

The vast majority of published studies using corpus consultation in teaching and learning center on students learning English, most commonly using the BNC or COCA. The meta-analysis of corpus use in language learning (up to 2014) by Boulton and Cobb (2017) lists only 2 studies not targeting English. Much of the earliest work in this area used the corpus from the COBUILD project (UK) and the Brown corpus (USA). There are corpora compiled for Indian English, New Zealand English, Australian English, and Canadian English. Information and links to these and other corpora, as well as corpus related software, are available from the Corpus-based Linguistic Links, originally created by Lee, and now maintained by Weisser. Included is information about historical corpora, such as the Middle English Compendium; specialized corpora, such as the Air Traffic Controller corpus; and LCs, such as the Chinese Learner English corpus. Particularly helpful is an annotated list of easily accessible online corpora for use in language learning, as well as links and information on creating local corpora. The BYU corpus site, created by Davies, provides access to multiple English language and other corpora, while the Ortolang site provides links to many EU corpora, such as the House of Commons Debates. The University of Louvain lists links to a variety of LCs worldwide, as does the Learner Corpora Association. Weisser lists numerous corpora in additional languages, including those with relatively few speakers, such as Welsh and Lithuanian. Manuel Barbera's site lists additional languages with corpora available from Afrikaans to West African Pigeon English. CALPER at Penn State University maintains a corpus portal with many languages represented. Recent research and teaching projects featuring corpus consultation can be found for Spanish (Salido, 2016), Japanese (Pereira et al., 2016), German (Vyatkina, 2016b), Italian (Kennedy & Miceli, 2017), and Cantonese (Wong & Lee, 2016). This shows that studies in instructional use of corpora in languages other than English is gradually increasing—a welcome development.

Multilingual corpora incorporate data from multiple languages, usually structured as a parallel corpus. One of the oldest is the Oslo Multilingual Corpus. Others are linked from Weisser's site. In recent years, there is growing interest in multimodal corpora. Some well-known corpora, such as the BNC, have at least some audio data available. More specialized corpora feature multimedia more prominently. The Basque spoken language corpus, for example, includes a collection of narratives from native Basque Euskara speakers retelling the story of a silent movie they have just watched. A number of researchers have called for more multimodal corpora, particularly given their usefulness in language learning (Boulton & Cobb, 2017). A number of multilingual and multimodal corpora are part of the MetaShare project, an open network of repositories for sharing and exchanging language data, tools, and related web services. Abuczki and Ghazaleh (2013) provide a useful overview of multimodal corpora and annotation tools. Blache et al. (2017) provide a nice walk-through of the process of collecting and annotating a large multimodal corpus. Here, even more than is the case for text-based corpora, standardization in formats is needed. Better tools for working with multimodal corpora have become available in recent years, for instance through the EU TALK project or McaWebSearch (see Cambria, 2011). Tools developed for analyzing multimedia and text in L1, such as computerized language analysis (CLAN), can be used in L2 as well (MacWhinney & Wagner, 2010).

Weisser's site has a separate section on the web as corpus resource. In fact, a number of studies have focused on this possibility (Boulton, 2010; Frankenberg-Garcia, 2005; Geluso, 2013; Sha, 2010; Ziegler et al., 2017). WebCorp Live offers a concordance interface for searching the web. Google Fights can be used to contrast frequency uses of particular expressions. WebBootCat creates a corpus from the web and features the ability to supply a set of topics, keywords, or websites, so as to enable inclusion of texts most likely to be relevant. Weisser lists a number of caveats in the use of the web as corpus, which are echoed in published reports. Not everything in English on the web is written by native speakers, resulting in an unmarked mix of native and non-native language; that applies as well, of course, to other languages. Many different genres are represented, again without information on provenance or speakers. Search engines provide minimal context. On the other hand, the web is readily available, and is a resource with which students are very familiar. It is also a resource constantly updated, so, in contrast to most corpora, it can provide up-to-date

language. Few large corpora, with exceptions such as the COCA, are regularly updated. The Wayback machine could be used to engage in diachronic analysis of online language use. Baldry (2010) points out that the web is ideal for working with terminology related to current topics such as climate change or immigration. He advocates students going beyond the limited context available from search data and considering language in the context of genre and web page layout.

In addition to using a search engine to query the web, there are a variety of online tools that provide access to traditional corpora. The Compleat Lexical Tutor is one of the best-known, providing tools for data-driven self-learning. A number of corpora are available for analysis through the online interface. It is also possible to enter one's own text. Weisser's site lists other online concordances and related tools. A Windows software package widely used is WordSmith Tools. Another commercial online corpus query system which provides useful resources for DDL is SketchEngine. It sports a variety of useful features for use in instruction, including creation of *word sketches*, one-page corpus-derived summaries of a word's grammatical and collocational behavior. Also available is the Sketch Engine for Language Learning (SkELL), an online interface to check how a particular phrase or a word is used. A stripped down free version, NoSketchEngine, is available. Several recent studies feature use of SketchEngine (e.g., Frankenberg-Garcia, 2014; Gablasova et al., 2017). A popular free concordancer, often used in language learning projects, is AntConc. It is multi-platform and fully Unicode compliant, so will work with all languages. A user guide is available, as is a Google support group. Recent teaching projects using AntConc include Chang (2014), Thomas (2015), and Li (2016). In this issue, Leńko-Szymańska (2017) and Ackerley (2017) also use AntConc. The resource pages in the book by Leńko-Szymańska and Boulton (2015) give a list and URLs for additional tools, as well as corpora and related language learning and teaching resources (pp. 300–303).

Outlook and Conclusion

The interest in the use of corpus-related resources in language learning is evinced by the recent spate of special journal issues on the topic in *CALICO Journal* (2009), *ReCALL* (2014), and *Language Learning* (2017). The meta-analyses by Boulton and Cobb (2017), Mizumoto and Chujo (2015), and Chambers (2007) provide information on the current state of work in the field. There are several common themes that are discussed in these articles, as well as in other recent work in the field. One of those is a call for greater use of corpora in different levels of instruction. While there are studies describing use of corpora in secondary school, they are relatively few. Introducing corpus-based work early on exposes students to real world language, a variety of genres and registers, and provides an introduction to a technique of importance in second-language learning. Longer-term studies of students starting to use corpora early would be valuable in uncovering the usefulness of that process.

More longitudinal testing and delayed post-tests are often mentioned as desirable directions for research. Part of the information that would be valuable to have is what degree and kind of exposure to DDL is most beneficial. That information might include how many examples from concordancing make a difference in student uptake (Geluso & Yamaguch, 2014). Boulton and Cobb (2017) point out that corpus-informed teaching projects have yielded generally positive results, but that they have not been undertaken in all areas: "We reach the somewhat surprising and possibly encouraging conclusion that DDL works pretty well in almost any context where it has been extensively tried" (p. 386). There are not many examples, for instance, of corpus usage in enhancing speaking skills; Geluso and Yamaguch (2014) is one of the few examples. Also rarely studied are corpus consultation outside the classroom (Chang, 2014). There are also few studies on the use of multimodal corpora (Boulton & Cobb, 2017).

Corpus work in children's L1 usage has been successful in capturing data on language use that includes non-verbal and environmental factors, as demonstrated in the CHILDES project (see MacWhinney, 2000; Monaghan & Rowland, 2017). Ellis (2017) calls for this kind of dense data in L2 LCs, incorporating techniques from conversation analysis and L1 research:

> We need large dense longitudinal corpora of L2 use, with audio, video, transcriptions, and multiple layers of annotation, for data sharing in open archives. We need these in sufficient dense mass that we can chart learners' usage history and their development. We need them in sufficient detail that we can get down to the fine detail of conversation analyses of the moment. (p. 49)

Having such data allows for a more accurate ecological perspective on spoken language, bringing together cognitive, linguistic, and social dimensions.

Boulton and Cobb (2017) point out that a number of recent studies come from the Middle East and Asia. The geographical spread of DDL would be interesting to study from the perspective of cultural learning styles. In that direction also, would be more studies including non-Western languages. There have been several studies that deal with learning styles in general (including deductive vs. inductive approaches), but it would be beneficial to have more studies of that kind. This is an area where qualitative analysis and individual case studies would be informative.

There are not as many studies as one might expect in using corpora in language for special purposes, although there are examples such as the studies by Rodgers, Chambers, and Le Baron-Earle (2011) for biotechnology and Cambria (2011) for journalism, as well as many for English for academic purposes (see Römer, 2011). This seems like an area in which more specialized corpora would be developed, for example, in technical fields. It has been shown, not surprisingly, that students in areas such as computer science and engineering have shown to be more comfortable with corpus consultation than students in other fields (see Boulton, 2009). They tend to have had prior experience with turning data into information—something that is not the case for students in all disciplines.

More studies comparing the use of corpora for comprehension versus production would be of interest (see Frankenberg-Garcia, 2014). That might include optimal methods and tools for accessing corpus data depending on the task. We now have alternatives to KWIC which provide user-friendlier interfaces. The Just the Word online tool, for example, offers a clean and simple query option for the BNC. Graphic representations, such as collocation networks, move in that direction. GraphColl and its sequel LancBox are tools that, among other functions, create a visualization of collocations from user-defined corpora (Brezina, McEnery, & Wattam, 2015; Gablasova et al., 2017). This kind of graphic analysis could be done as well with the R statistical programming language (Stefan, 2009), but that requires coding, which GraphColl and LancBox do not. Graphic representations of data based on treebanks (parsed text from corpus data) might be informative for students. An easily implemented graphic representation can be done creating a word cloud of a collocation or colligation using a tool such as Wordle. This is an area, in fact, where students creating mindmaps and other graphic representations might serve a dual purpose of creating helpful mnemonic devices and engaging in language play.

Data-informed language learning is such a potentially powerful, engaging, and enabling approach that more experimentation in its use at all levels of instruction is needed. More work with query and display interfaces would be welcome. Students (and teachers) today are used to more attractive and intuitive (and mobile-friendly) interfaces than is normally the case for concordance software, whether it be web-based or desktop. Frankenberg-Garcia (2014) points out how different corpus data is for students from the carefully prepared language models they are used to:

> Concordancing programs are not particularly user-friendly and raw corpus data is far more difficult to understand than the edited materials language learners are accustomed to using. With the technological advances that we see today, however, there is no reason why we should not attempt to fill in the existing gap that lies between the polished, albeit limited, linguistic information neatly systematized in dictionaries and the countless other linguistic facts that can be gleaned from corpora, but which only experienced corpus users are able to access. (p. 141)

The vast majority of DDL studies are from researchers in the field, not from teachers. That is only likely to change if user interfaces to corpora improve.

References

Abuczki, Á., & Ghazaleh, E. B. (2013). An overview of multimodal corpora, annotation tools, and schemes. *Argumentum*, *9*, 86–98.

Ackerley, K. (2017). Effects of corpus-based instruction on phraseology in learner English. *Language Learning & Technology*, *21*(3), 195–216. Retrieved from http://llt.msu.edu/issues/october2017/ackerley.pdf

Adelex Text Analyser. (2017). Retrieved from http://www.ugr.es/~inped/ada/ada.php?ada=ughnvs3q3jtrhgqkq8pcf99uc0&lng=english

Baldry, A. P. (2010). A web-as-multimodal corpus approach to lexical studies based on intercultural and scalar principles. In M. Jaén, F. Valverde, & M. Pérez (Eds.), *Exploring new paths in language pedagogy. Lexis and corpus based language teaching*, (pp. 173–190). London, UK: Equinox Publishing.

Belz, J., & Vyatkina, N. (2005). Learner corpus analysis and the development of L2 pragmatic competence in networked inter-cultural language study: The case of German modal particles. *Canadian Modern Language Review*, *62*(1), 17–48.

Bernardini, S. (2000). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 225–234). New York, NY: Peter Lang.

Blache, P., Bertrand, R., Ferré, G., Pallaud, B., Prévot, L., & Rauzy, S. (2017). The corpus of interactional data: A large multimodal annotated resource. In N. Ide & J. Pustejovsky (Eds.), *Handbook of linguistic annotation* (pp. 1323–1356). Amsterdam, Netherlands: Springer.

Boulton, A. (2009). Data-driven learning: Reasonable fears and rational reassurance. *Indian Journal of Applied Linguistics*, *35*(1), 81–106.

Boulton, A. (2010) Learning outcomes from corpus consultation. In M. Jaén, F. Valverde, & M. Pérez (Eds.), *Exploring new paths in language pedagogy. Lexis and corpus based language teaching*, (pp. 129–144). London, UK: Equinox Publishing.

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

Braun, S. (2006). ELISA—A pedagogically enriched corpus for language learning purposes. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (pp. 25–47). Frankfurt, Germany: Peter Lang.

Brezina, V., McEnery, T., & Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, *20*(2), 139–173.

Cambria, M. (2011). Websearching and corpus construction of online news sites in ESP: Government leaders on show at G8 Summits. *ESP Across Cultures*, *8*, 7–22.

Carter, R. (1998). Orders of reality: CANCODE, communication, and culture. *ELT Journal*, *52*(1), 43–56.

Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 3–16). Amsterdam, Netherlands: Rodopi.

Chang, J. Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL*, *26*(2), 243–259.

Charles, M. (2015). Same task, different corpus. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data–driven learning* (pp. 131–154). Amsterdam, Netherlands: John Benjamins.

Chinkina, M., & Meurers, D. (2016). Linguistically aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 188–198). San Diego, CA: Association for Computational Linguistics. Retrieved from http://m–mitchell.com/NAACL-2016/BEA/pdf/BEA1121.pdf

Cho, H. (2016). Task dependency effects of collaboration in learners' corpus consultation: An exploratory case study. *ReCALL*, *28*(1), 44–61.

Cotos, E. (2014). Enhancing writing pedagogy with learner corpus data. *ReCALL*, *26*(2), 202–224.

Cresswell, A. (2007). Getting to "know" connectors? Evaluating data-driven learning in a writing skills course. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 267–288). Amsterdam, Netherlands: Rodopi

Duda, R., & Tyne, H. (2010). Authenticity and autonomy in language learning. *Bulletin Suisse de Linguistique Appliquée*, *92*, 86–106.

Ellis, N. C. (2017). Cognition, corpora, and computing: Triangulating research in usage-based language learning. *Language Learning*, *67*(S1), 40–65.

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics*, *14*(3), 393–417.

Flowerdew, L. (2015). Data-driven learning and language learning theories. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15–36). Amsterdam, Netherlands: John Benjamins.

Frankenberg-Garcia, A. (2005). A peek into what today's language learners as researchers actually do. *International Journal of Lexicography*, *18*(3), 335–355.

Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL*, *26*(2), 128–146.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*(S1), 155–179.

Geluso, J. (2013). Phraseology and frequency of occurrence on the web: native speakers' perceptions of Google-informed second language writing. *Computer Assisted Language Learning*, *26*(2), 144–157.

Geluso, J., & Yamaguchi, A. (2014). Discovering formulaic language through data-driven learning: Student attitudes and efficacy. *ReCALL*, *26*(2), 225–242.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language teaching*, *40*(2), 97–118.

Godwin-Jones, R. (2017). Scaling up and zooming in: Big data and personalization in language learning. *Language Learning & Technology*, *21*(1), 4–15. Retrieved from http://llt.msu.edu/issues/february2017/emerging.pdf

Goth, J., Baikadi, A., Ha, E., Rowe, J., Mott, B., & Lester, J. (2010). Exploring individual differences in student writing with a narrative composition support environment. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing processes and authoring aids* (pp. 56–64). San Diego, CA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W10-0408

Halliday, M. A. K. (1961). Categories of the theory of grammar. *Word*, *17*, 241–292.

Harwood, N. (2002). Taking a lexical approach to teaching: Principles and problems. *International Journal of Applied Linguistics*, *12*(2), 139–155.

Huang, Z. (2014). The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing. *ReCALL*, *26*(2), 163–183.

Johansson, S. (2009). Some thoughts on corpora and second-language acquisition. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 33–44). Amsterdam, Netherlands: John Benjamins.

Johns, T. (1997). Contexts: The background, development, and trialing of a concordance based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). New York, NY: Longman.

Johns, T. (2002). Data-driven learning: The perpetual challenge. *Language and Computers*, *42*(1), 107–117.

Johns, T., & King, P. (1991). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. *Concordancing: English Language Research Journal*, *4*, 27–45.

Karras, J. N. (2016). The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL*, *28*(2), 166–186.

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology*, *5*(3), 77–90. Retrieved from http://llt.msu.edu/vol5num3/pdf/kennedy.pdf

Kennedy, C., & Miceli, T. (2010). Corpus-assisted creative writing: Introducing intermediate Italian learners to a corpus as a reference resource. *Language Learning & Technology*, *14*(1), 28–44. Retrieved from http://llt.msu.edu/vol14num1/kennedymiceli.pdf

Kennedy, C., & Miceli, T. (2017). Cultivating effective corpus use by language learners. *Computer Assisted Language Learning*, *30*(1–2), 91–114.

Kern, R. (2014). Technology as Pharmakon: The promise and perils of the internet for foreign language education. *Modern Language Journal*, *98*(1), 340–357.

Kramsch, C. (1998). *Language and culture*. Oxford, UK: Oxford University Press.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*, 757–786.

Leńko-Szymańska, A. (2014). Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education. *ReCALL*, *26*(2), 260–278.

Leńko-Szymańska, A. (2017). Training teachers in data-driven learning: Tackling the challenge. *Language Learning & Technology*, *21*(3), 217–241. Retrieved from http://llt.msu.edu/issues/october2017/lenko-szymanska.pdf

Leńko-Szymańska, A., & Boulton, A. (2015). *Multiple affordances of language corpora for data-driven learning*. Amsterdam, Netherlands: John Benjamins.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward*. Hove, UK: Language Teaching Publications.

Li, W. (2016). “Paper” and “We” - The referring preference in academic abstracts. In *Proceedings of the Fifth Northeast Asia International Symposium on Language, Literature, and Translation* (pp. 117–122). Atlanta, GA: American Scholars Press. Retrieved from https://biblio.ugent.be/publication/8508663/file/8508678#page=117

Lüdeling, A., Hirschmann, H., & Shadrova, A. (2017). Linguistic models, acquisition theories, and learner corpora: Morphological productivity in SLA research exemplified by complex verbs in German. *Language Learning*, *67*(S1), 96–129.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Mahwah, NJ: Erlbaum.

MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning*, *67*(S1), 254–275.

MacWhinney, B., & Wagner, J. (2010). Transcribing, searching, and data sharing: The CLAN software and the TalkBank data repository. *Gesprachsforschung: Online-Zeitschrift zur Verbalen Interaktion*, *11*, 154–173.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, MA: Cambridge University Press.

Meurers, D., & Dickinson, M. (2017). Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning*, *67*(S1), 66–95.

Millar, N., & Lehtinen, B. (2008). DIY local learner corpora: Bridging gaps between theory and practice. *JALT CALL Journal*, *4*(2), 61–72.

Milton, J., & Cheng, V. S. (2010). A toolkit to assist L2 learners become independent writers. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing processes and authoring aids* (pp. 33–41). San Diego, CA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W/W10/W10-04.pdf#page=43

Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies*, *22*, 1–18.

Mizumoto, A., Chujo, K., & Yokota, K. (2016). Development of a scale to measure learners’ perceived preferences and benefits of data-driven learning. *ReCALL*, *28*(2), 227–246.

Molés-Cases, T., & Oster, U. (2015). Webquests in translator training. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 201–224). Amsterdam, Netherlands: John Benjamins.

Monaghan, P., & Rowland, C. F. (2017). Combining language corpora with experimental and computational approaches for language acquisition research. *Language Learning*, *67*(S1), 14–39.

Mueller, C. M., & Jacobsen, N. D. (2016). A comparison of the effectiveness of EFL students’ use of dictionaries and an online corpus for the enhancement of revision skills. *ReCALL*, *28*(1), 3–21.

Mukherjee, J. (2004). Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. *Language and Computers*, *52*(1), 239–250.

Mukherjee, J. (2006). Corpus linguistics and language pedagogy: The state of the art—and beyond. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (pp. 5–24). Frankfurt, Germany: Peter Lang.

Oghigian, K., & Chujo, K. (2010). An effective way to use corpus exercises to learn grammar basics in English. *Language Education in Asia*, *1*(1), 200–214.

Pereira, L., Manguilimotan, E., & Matsumoto, Y. (2016). Leveraging a large learner corpus for automatic suggestion of collocations for learners of Japanese as a second language. *CALICO Journal*, *33*(3), 311–333.

Pilán, I., Vajjala, S., & Volodina, E. (2016). A readable read: Automatic assessment of language learning materials based on linguistic complexity. *International Journal of Computational Linguistics and Applications*, *7(*1), 143–159. Retrieved from http://www.ijcla.bahripublications.com/2016-1/IJCLA-2016-1-pp-143-159-preprint.pdf

Rankin, T., & Schiftner, B. (2011). Marginal prepositions in learner English: Applying local corpus data. *International Journal of Corpus Linguistics*, *16*(3), 412–434.

Rodgers, O., Chambers, A., & Le Baron-Earle, F. (2011). Corpora in the LSP classroom: A learner-centred corpus of French for biotechnologists. *International Journal of Corpus Linguistics*, *16*(3), 391–411.

Rohrbach, J. (2003). 'Don't miss out on Göttingen's nightlife': Genreproduktion im Englischunterricht. *Praxis des neusprachlichen Unterrichts*, *50*, 381–389.

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, *31*, 205–225.

Sabatini, N. M. J. B. J., & Andreyev, K. B. S. (2016). Language Muse: Automated linguistic activity generation for English language learners. In *The 54th annual meeting of the Association for Computational Linguistics: Proceedings of system demonstrations* (pp. 79–84). Berlin, Germany: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P/P16/P16–4.pdf#page=91

Salido, M. G. (2016). Error analysis of support verb constructions in written Spanish learner corpora. *Modern Language Journal*, *100*(1), 362–376.

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129–158.

Schmied, J. (2006). Corpus linguistics and grammar learning: Tutor versus learner perspectives. In S. Braun, K. Kohn, & J. Mukherjee (Eds.), *Corpus technology and language pedagogy: New resources, new tools, new methods* (pp. 87–106). Frankfurt, Germany: Peter Lang.

Seidlhofer, B. (2000). Operationalizing intertextuality: Using learner corpora for learning. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 207–223). Frankfurt, Germany: Peter Lang.

Sha, G. (2010). Using Google as a super corpus to drive written language learning: A comparison with the British National Corpus. *Computer Assisted Language Learning*, *23*(5), 377–393.

Sinclair, J. (1987). *Collins COBUILD English language dictionary*. London, UK: Harper Collins.

Sinclair, J. (Ed.). (1996). *Collins COBUILD grammar patterns*. London, UK: Harper Collins.

Smart, J. (2014). The role of guided induction in paper-based data-driven learning. *ReCALL*, *26*(2), 184–201.

St. John, E. S. (2001). A case for using a parallel corpus and concordancer for beginners of a foreign language. *Language Learning & Technology*, *5*(3), 185–203. Retrieved from http://llt.msu.edu/vol5num3/pdf/stjohn.pdf

Stefan, G. (2009). *Quantitative corpus linguistics with R: A practical introduction*. London, UK: Routledge.

Thomas, J. (2015). Stealing a march on collocation. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 85–108). Amsterdam, Netherlands: John Benjamins.

Tono, Y., Satake, Y., & Miura, A. (2014). The effects of using corpora on revision tasks in L2 writing with coded error feedback. *ReCALL*, *26*(2), 147–162.

Vyatkina, N. (2016a). Data-driven learning for beginners: The case of German verb-preposition collocations. *ReCALL*, *28*(2), 207–226.

Vyatkina, N. (2016b). Data-driven learning of collocations: Learner performance, proficiency, and perceptions. *Language Learning & Technology*, *20*(3), 159–179. Retrieved from http://llt.msu.edu/issues/october2016/vyatkina.pdf

Vyatkina, N. (2016c). The Kansas Developmental Learner corpus (KANDEL). *International Journal of Learner Corpus Research*, *2*(1), 101–119.

Vyatkina, N. (2016d). What can multilingual discourse-annotated corpora do for language learning and teaching? In P. Furko, D. Csilla, L. Degand, & B. Webber (Eds.), *TextLink - Structuring Discourse in Multilingual Europe, Second Action Conference: Conference handbook* (pp. 21–24). Debrecen, Hungary: Debrecen University Press.

Waters, A. (2009). Ideology in applied linguistics for language teaching. *Applied Linguistics*, *30*(1), 138–143.

Widdowson, H. G. (1998). Context, community, and authentic language. *TESOL Quarterly*, *32*(4), 705–716.

Widdowson, H. G. (2000). On the limitations of linguistics applied. *Applied Linguistics*, *21*(1), 3–25.

Willis, J., & Willis, D. (1988). *Collins COBUILD English course*. London, UK: Harper Collins.

Wisniewski, K. (2017). Empirical learner language and the levels of the Common European Framework of Reference. *Language Learning*, *67*(S1), 233–254.

Wong, T. S., & Lee, J. S. (2016). Corpus-based learning of Cantonese for Mandarin speakers. *ReCALL*, *28*(2), 187–206.

Xia, M., Kochmar, E., & Briscoe, T. (2016). Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, (pp. 12–22). San Diego, CA: Association for Computational Linguistics. Retrieved from http://www.anthology.aclweb.org/W/W16/W16-0502.pdf

Yoon, H., & Jo, J. W. (2014). Direct and indirect access to corpora: An exploratory case study comparing students' error correction and learning strategy use in L2 writing. *Language Learning & Technology*, *18*(1), 96–117. Retrieved from http://llt.msu.edu/issues/february2014/yoonjo.pdf

Ziegler, N., Meurers, D., Rebuschat, P., Ruiz, S., Moreno-Vega, J. L., Chinkina, M., Li, W., & Grey, S. (2017). Interdisciplinary research at the intersection of CALL, NLP, and SLA: Methodological implications from an input enhancement project. *Language Learning*, *67*(S1), 210–232.