Annual Review of Applied Linguistics (2019), **39**, 74–92
doi:10.1017/S0267190519000096

CAMBRIDGE UNIVERSITY PRESS

POSITION PAPER

Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use

Tony McEnery[1]*, Vaclav Brezina[1], Dana Gablasova[1] and Jayanti Banerjee[2]

[1]Lancaster University, UK and [2]Trinity College London, UK
*Corresponding author. E-mail: a.mcenery@lancaster.ac.uk

Abstract

In this article we explore the relationship between learner corpus and second language acquisition research. We begin by considering the origins of learner corpus research, noting its roots in smaller scale studies of learner language. This development of learner corpus studies is considered in the broader context of the development of corpus linguistics. We then consider the aspirations that learner corpus researchers have had to engage with second language acquisition research and explore why, to date, the interaction between the two fields has been minimal. By exploring some of the corpus building practices of learner corpus research, and the theoretical goals of second language acquisition studies, we identify reasons for this lack of interaction and make proposals for how this situation could be fruitfully addressed.

Introduction

"Corpus linguistics is essentially a technology" (Simpson-Vlach & Swales, 2001, p. 1). This quote succinctly captures the instrumental nature of the discipline as well as its strong connection to modern computational methods. While it is possible to analyze language manually, robustness of analysis of and depth of insight into attested language use can arguably be achieved only with the aid of computational technology. It is important to note at the outset of this article that corpus linguistics investigates spontaneous spoken and written language use, which we will refer to more briefly as "language use" in this article. Computational technology enables fast searches and detailed statistical analyses of data sets that comprise millions or even billions of words. For this reason, the field of corpus linguistics has seen major growth since the 1960s facilitated by advances in computational technology that can store, retrieve, and process large amounts of linguistic data. Corpus linguistics is a quantitative paradigm grounded in the empirical tradition of language analysis. It uses large quantities of observational data compiled into data sets, called corpora, to provide evidence about language use by both first language (L1) and second language (L2) speakers (for an overview, see Barlow, 2005). To exemplify the methods of corpus linguistics, let us look at a brief illustrative analysis of two near synonyms—*good* and *great*—in the Trinity Lancaster Corpus (Gablasova, Brezina, & McEnery, forthcoming). The corpus, which is

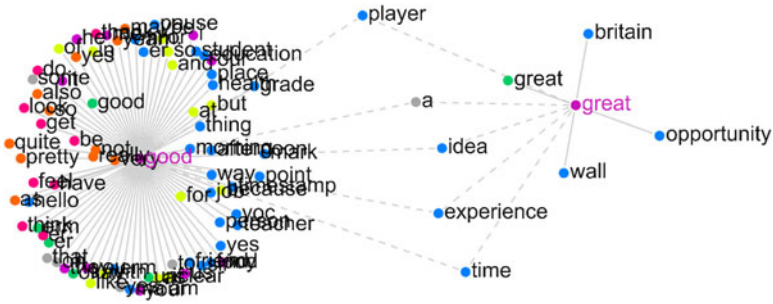

Figure 1. *good* and *great* in the *Trinity Lancaster Corpus* of L2 English

introduced in more detail later in this article (see the "Pragmatics and Learner Language" section), contains 4.1 million words of orthographically transcribed learner speech, gathered from more than 2,000 speakers with a range of L1 backgrounds across four tasks. The results of the analysis are visualized in Figure 1.

Figure 1 shows a collocation network (Brezina, 2018; Brezina, McEnery, & Wattam, 2015; Philips, 1985; Williams, 1998) displaying the associations and shared associations (collocations) of the words *good* and *great* as used by more than 1,500 different L2 speakers. These associations have been shown to have an important role to play in discourse, for example, in creating metaphorical connections between words (Brezina, 2018) or promoting social evaluation (Brezina et al., 2015; McEnery, 2005). In this case, the network functions as a visualizing device, providing a statistical summary of L2 use of these two adjectives. Different colors in the graph are used to denote different word classes of the associations (e.g., nouns are blue, adjectives are green, verbs are red, etc.). The analysis reveals that the more general evaluative adjective *good* is more frequent in L2 speech (4,812 occurrences in almost 2 million words) than the adjective *great* (590 occurrences) and has more collocates in the corpus examined (84 vs. 9). Unlike *great*, *good* is often modified by adverbs such as *really*, *pretty, so, quite,* etc. The commonality in the use of the two adjectives lies in the fact that both pre-modify nouns such as *player, idea, experience,* and *time,* displayed in the middle of the graph with links to both adjectives in question.

The reason for introducing the example is to illustrate four points that summarize the role of corpus linguistics in the study of second language acquisition (SLA). First, corpus linguistics provides access to large databases of language use that can reflect different forms of language, such as spoken and written L2. L2 language is typically compiled in what we call "learner corpora." Second, these databases are easily shared, enriched with annotations and used by a wide range of researchers for a wide range of purposes. By allowing data to be shared in this way, corpus linguistics enabled a form of open linguistics before the notion was widely popularized, reducing effort in data collection and promoting the replication and reproduction of results and providing an important source of hypothesis testing. Third, although we can observe linguistic phenomena such as the frequent co-occurrence of certain words in individual texts or speakers, only with the scale of the analysis afforded by corpus linguistics can we be confident about important recurrent patterns across many speakers and contexts. Corpus linguistics typically takes into consideration hundreds or thousands of different texts or speakers. For example, we used data from more than 1,500 speakers in producing Figure 1. To perform analysis on this scale, advanced computational

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

technology is required to search through the data, compute frequencies, and provide statistical summaries such as those shown in the graph, which was produced using #LancsBox (Brezina et al., 2015; for an example of the program in use, see Kecskes & Kirner-Ludwig, 2017). Fourth, corpus findings are based on the observation of a very large number of examples of language use. They are primarily focused on language output and can tell us relatively little (and even if so, indirectly) about the processes behind language learning and production. However, this potential limitation is also a point of possible methodological synergy and fruitful collaboration, bringing together corpus linguistics and experimental methods in SLA research (e.g., Baker & Egbert, 2019; Ellis, Römer, & O'Donnell, 2016).

It is clear that corpora are accepted as a useful source of data by the academic community and seen as commercially useful (e.g., Jamieson, 2005; McEnery & Hardie, 2011). However, corpora are proceeding relatively slowly in terms of impact on L2 research, despite the claim by Granger (2009) that "this new resource will soon be accepted as a bona fide data type in SLA research" (p. 17). The interaction that Granger predicted between learner corpus research (LCR) and SLA has yet to occur, something noted by Gries (2015) and Myles (2015) among others. And so far, there is no clear sign that the situation is changing. For example, when we look at the supplement of *The Modern Language Journal* edited by Duff and Byrnes (2019), we can use the series of papers it contains on the theme "SLA Across Disciplinary Borders: New Perspectives, Critical Questions, and Research Possibilities" as a litmus test. More than 20 years after the publication of early work on LCR (e.g., Granger, 1998b), of the 15 contributions, Slabakova (2019) uses L1 corpus data, Hall (2019) cites L1 corpus studies as corroboration for points made, and Ellis (2019) argues for the use of corpora in general. However, *learner* corpora are never specifically referenced or discussed in any of the state-of-the-art papers in that important volume.

This article aims to better understand the apparent disconnect between learner corpus studies and SLA and to facilitate and stimulate collaboration leading to further innovative theoretical and practical (methodological) work. The road to further collaboration between corpus linguistics and SLA has already been paved by corpus studies that aim to contribute to the theory of SLA (for an overview, see Myles, 2015), and corpus linguistics has been covered in major language learning reference works (e.g., Leclercq, Edmonds, & Hilton, 2014; Mackey & Gass, 2012). We believe there is strong potential for conceptual and methodological innovation in research on L2 learning (e.g., Rebuschat, Meurers, & McEnery, 2017). However, the convergence of any two fields should be attempted with care, given the different goals that gave rise to each of the fields and the different pathways that the two disciplines have taken (Gablasova, Brezina, & McEnery, 2017b).

With the aim of contributing to productive collaboration between corpus linguistics and SLA, this article provides an in-depth critical account from the corpus linguistics perspective of the development of corpus-based research on L2 learning and use. To do so, it first provides some historical background to learner corpus studies. It then focuses on the present, discussing issues that may work against the convergence of SLA and LCR. Finally, we make some recommendations in terms of the future potential for the collaboration. We conclude by arguing that a fresh perspective on the interaction of learner corpus studies, SLA, and other disciplines can contribute to our understanding of L2 learning.

Background: Learner Corpus Studies

LCR refers to research conducted on corpora representing L2 use. These corpora usually include different groups of L2 learners and users (e.g., grouped according to

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

consideration of L1 background or proficiency in the target language) and/or L2 use from a particular linguistic setting (e.g., from a specific linguistic task). Some corpora are constructed in language learning contexts that are more complex than others, so the corpora may record multiple L1 backgrounds for learners, for example, as is the case with the Guangwai-Lancaster Chinese Learner Corpus (https://www.sketchengine.eu/guangwai-lancaster-chinese-learner-corpus/) and the Trinity Lancaster Corpus.

Learner corpus studies grew out of corpus linguistics (for an overview, see McEnery & Hardie, 2011), which in turn can be linked to precomputational studies, so-called early corpus studies, based on the manual analysis of large volumes of naturally occurring language data. Many of these studies were focused on developing resources for learning languages including French (Henmon, 1924), Spanish (Buchanan, 1929), and English (West, 1953). These were painstaking studies, based on collections of written material, manually analyzed with substantial funding from research foundations such as the Rockefeller and Carnegie (see Fries & Traver, 1940, for a comprehensive account of such studies in the early part of the 20th century).

It is tempting to point to a specific study and date and to identify that as the beginning of LCR. In reality, however, just as there were early corpus studies, there were studies that foreshadowed LCR. The idea that learner language itself was a potential object of study, to be pursued with methods analogous to corpus linguistics, was present well before landmark learner corpora were developed. Studies such as Juvonen (1989) on Finnish learners of Swedish, Cornu and Delahaye (1987) on Flemish learners of French, and Huebner (1983, 1985) on a Hmong learner of English predate, yet clearly foreshadow, LCR. Also, as mentioned by Granger (1998a), work looking at the categorization and analysis of errors made by language learners predated LCR. This work was typically based on small, paper-based corpora of attested learner language; some of the data sets may have been larger, though still paper based (see Færch, Haastrup, & Phillipson, 1984). Clearly, there were corpus-based studies of learner language, typically small in scale, before LCR started.

The starting point of LCR is usually dated to the emergence of what is commonly taken to be the first learner corpus, the International Corpus of Learner English (ICLE, Granger, Dagneux, & Meunier, 2002). ICLE was innovative in that it self-consciously drew upon the rich methodological background of corpus linguistics (Granger, Gilquin, & Meunier, 2015) and presented an advance on the smaller studies of learner language already noted, as it (a) scaled up data collection beyond what had previously been achieved by researchers such as Juvonen (1989), both in terms of volume of data collected and the number of respondents; (b) provided a machine-readable, reusable resource that was easily accessible; and (c) covered a range of L1 backgrounds for the L2 speakers in the corpus. As such, it inherited the advantages of corpus linguistics—(a) it was much easier to extract complex frequency data from the corpora; (b) it was possible to run a wide range of statistical analyses on the data (for an overview, see Gries, 2015); (c) increased advantage could be gained by using automated linguistic analysis packages on the data, enhancing the effect of (a) and (b) further; (d) the packages and measures developed that aided studies of L1 corpora, such as concordancing, collocation, frequency lists, keyword analysis, and corpus annotation, helped with the study of L2 corpora also (see McEnery and Hardie, 2011, Chap. 2, for an overview of corpus-processing tools); and (e) the corpora could be annotated with a range of automatic and manual annotations that further enhanced their utility.

ICLE, when it was published, was composed of short argumentative essays written by advanced learners of English. Each essay in ICLE is typically 500–1,000 words in

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

length. Eleven language backgrounds were represented in the first release of the corpus, with the second version of the corpus containing 16. The respondents were usually second- or third-year university students. The initial corpus was just over 2 million words in size (version 2 was 3.7 million words) and was supplemented by a further corpus, LOCNESS (Louvain Corpus of Native English Essays) to facilitate comparison to L1 speaker data. LOCNESS has three parts: (a) pre-university student essays written by British native English speakers (60,209 words), (b) essays by university students who were native speakers of British English (95,695 words), and (c) essays by American university students (168,400 words) who were native speakers of American English (for details, see http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/locness1.htm). Finally, an exploration of spoken learner language was enabled by the production of the Louvain International Database of Spoken English Interlanguage (LINDSEI) corpus; growing from a modest corpus of 50 interviews amounting to approximately 100,000 words spoken by advanced learners of English with French L1 background (see De Cock, Granger, Leech, & McEnery, 1998, for an example of the use of this early version of LINDSEI), the corpus grew until, by the time of its public release (Gilquin, De Cock, & Granger, 2010), it covered 11 different L1 backgrounds in just over 1 million words of data. A companion corpus, the Louvain Corpus of Native English Conversation (LOCNEC), contains 124,935 words of transcribed informal interviews with 50 British English native speakers.

The Louvain corpora undoubtedly represented a landmark in the development of LCR to the extent that they defined the field and stimulated learner corpus production. However, the promise that they seemed to hold for SLA research has not been fully realized, for reasons that the next section will consider.

The Present: LCR meets SLA

Since the creation of the Louvain corpora, the range of available, and privately held, learner corpora has grown exponentially. Online repositories now hold links to dozens of learner corpora, some of which, like ICLE, cover many L1 backgrounds (for a reasonably comprehensive listing of available learner corpora, see https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html).

Beyond these resources, publishers and testing organizations have also developed their own learner corpora, which are substantial but typically only available in-house. For example, the Longman Learner Corpus is held in-house by the English language teaching publisher Longman and is used in the production of the usage notes for the *Longman Active Study Dictionary*, a small dictionary covering 100,000 words and phrases that is aimed at improving the vocabulary of intermediate learners of English. The corpus contains more than 10 million words of written L2 data from English language learners with a range of L1 backgrounds across a wide range of English proficiency levels. Also, the U.S.-based Educational Testing Service has made available a corpus of 12,100 essays written in English by learners from 11 different L1 backgrounds as part of the Test of English as a Foreign Language exam (Blanchard, Tetreault, Higgins, Cahill, & Chodorow, 2014), and Cambridge University Press has built the Cambridge Learner Corpus, which, at the time of writing, contains more than 50 million words of exam scripts from a range of English language examinations, produced by more than 220,000 students from 173 countries. Learner corpora have been explored extensively by LCR researchers. Tono (2018) reported that a bibliography on LCR held by the Learner Corpus Research Association contains more than 1,000 articles and books. This shows that learner corpora and LCR

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

are well established in academic settings and commercial marketplaces, though the extent to which they are used beyond the LCR community is less clear, with use by SLA researchers being vanishingly low.

Several key issues prevent interaction between LCR and SLA. Some of these issues were identified as soon as the first major body of work on ICLE, Granger (1998b), was published. Lessard (1999), in reviewing this work, rightly pointed out that (a) the "corpora used are still relatively small, which hinders the use of statistical tests"; (b) the annotation of the data was sparse, and more was needed if the data were to be exploited more effectively; and finally (c) "apart from a few theoretical constructs such as overuse and underuse with respect to native speaker usage there is a serious lack of contact with the notions of current linguistic theory" (p. 303), that is, the work presented was descriptive and did not appeal to an explanatory framework. As evidenced in Brezina and Bottini's (2019) review of a recent, edited volume comprising 18 studies by different international LCR researchers (Coccetta, Castello, & Ackerley, 2015), the situation described by Lessard (1999) appears not to have changed very much. In what follows, we briefly consider how these three issues can be addressed to facilitate collaboration between LCR and SLA.

The Quantity and Quality of the Data

While some SLA researchers seem to realize that using learner corpus data may be an excellent way of avoiding duplication of effort in research (Mackey & Gass, 2015), learner corpora are generally still relatively small when one considers the frequency of the features that might be searched for using them—even words that one may not consider to be exceptionally infrequent, such as *chill, earner, pester, sunburn,* and *vases*, occur only once in the ICLE corpus. While some corpora, notably EFCAMDAT (Alexopoulou, Michel, Murakami, & Meurers, 2017) and the Cambridge Learner Corpus (Nicholls, 2003), which comprise 83 and 50 million words each, are large enough to allow an exploration of a range of features of written learner language, many learner corpora comprise only several hundred thousand words. Yet even where corpora appear to be large, sometimes several million words, combining factors for an analysis can swiftly lead to data-sparsity issues. For a specific focused study, the data available may, in fact, be no larger than the data sets used for such studies in the pre-LCR era.

For example, the Longman Learner Corpus appears to be very large—the version we had access to for this article consisted of 8,974,424 words. However, the corpus allows for the exploration of a wide range of linguistically meaningful variables. Our version of the corpus allowed us to look at data from 18 L1 backgrounds, eight levels of proficiency, three target varieties of English, and nine different task types. While we could look at any one of these factors in a study, researchers often look at the effect of variables combining and interacting. For example, in the pre-LCR study performed by Cornu and Delahaye (1987), the authors looked at learners with an identical L1 background (Dutch), with the same target language (French), at a specific level of linguistic competence producing one type of text (an exam essay). They did not study variation according to one factor; rather, their study combined factors. Combining factors in a corpus like the Longman Learner Corpus leads to a combinatorial explosion of possible categories—we can, in principle, select learners from one of a range of L1 backgrounds, learning a specific variety of English, producing a particular type of text at one of many levels of proficiency. If we combine these factors, and an equal amount of data was available for each of the 3,888 categories created (18 L1 backgrounds × 8

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

levels of proficiency × 3 varieties of English × 9 task types), we would have the modest total of just over 2,300 words in each category; this means that any research that combined variables in this way to undertake, not unreasonably, a focused study would probably be based on less data than was available to researchers such as Cornu and Delahaye.

However, there is not an equal amount of data available for each category created, meaning that some of the research choices the corpus presents are more apparent than real—imagine that we are interested in looking at *great* and *good* in a set essay task of intermediate learners of English, of any L1 background, with British English as their target variety. There are plenty of examples: 2,832. Yet when we try to take into consideration important variables such as L1 background, proficiency level, and task, problems emerge. For example, there are only 4 and 5 examples respectively of the use of *good* and *great* for the Thai and Malay L1s due to the small size of the relevant samples (2,886 and 3,840 words, 0.03% and 0.04% of the total corpus, respectively). Neither of these would allow us to perform a quantitative analysis of collocations for the Thai and Malay L1s as discussed in the "Introduction" section here and displayed in Figure 1. When we combine variables, the amount of available data differs wildly—though the L1 Arabic speakers have 141,248 words of set essay data in the corpus, if we consider only those speakers in the corpus who reported British English, one of the three target varieties in the data (American, Australian, and British), as their L2, the size of the data shrinks to 113,562 words. If we then want to focus on intermediate-level speakers, we have only 18,945 words to work with. The situation is worse for the Czech L1 speaker data—61,711 words of data are available for Czech L1 writers of L2 English essays—all of whom have British English as their target, so all of the permutations of variables for Czech L1 writers that have the target variety as something other than British English are empty sets. There are no data in the corpus to answer any questions related to Czech learners of American English, for example, nor can one contrast the experience of Czech learners of British and American English. Likewise, the corpus contains no data for Korean writers with a target variety other than American English. These examples are illustrative of a general problem. Apparently large, well-balanced data sets, when explored through the metadata provided with them, become very unbalanced.

Data sparsity is an important point. A large corpus does not guarantee a bounty of evidence for any research question that might be put to that corpus. Inherent in the design of any corpus is a set of possibilities and limitations. Design decisions made when constructing a corpus mean that the corpus can be used to address some questions but not others, even if, as is the case in the examples given, it may appear at first glance that a corpus might allow one to investigate a specific question. Corpus builders and corpus users need to be critical in their engagement with the data they use and need to be open about the design decisions that shaped that collection of data. If they are not, corpus linguistics may be seen, on occasion, to overpromise and ultimately to disappoint.

Data-sparsity issues can quickly remove or drastically reduce in scope the possibility of a multifactorial statistical analysis, a possible explanation for Gries's (2015) finding that these statistics "have not been widely applied" in LCR (p. 177). So, scale, in a range of ways, remains an issue for LCR. Responding to that issue will not always be a question of simply gathering more data. While gathering more data may be important, sometimes it may be a question of gathering the right data. For instance, constructing a corpus in such a way as to permit questions that SLA researchers want to pose may be

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

better than simply gathering more data that are not interesting to researchers in other fields. Similarly, one also needs to accept that simply gathering more data, even if these data are well structured, may not be the answer. If a researcher is interested in a very-low-frequency feature, even an ever-larger corpus may yield insufficient examples.

With regard to the issue of corpus construction, note that while the number of learner corpora available is fairly large, the variety of learner corpora is still rather limited. There are currently few well-structured data sets that permit the study of variables relevant to SLA, such as proficiency level, sociolinguistic variables (e.g., age, educational attainment, social class), contextual features (including different registers and modes of communication), and L1 background. Hence, if some potential users of learner corpora approach existing corpora with a question that cannot be answered by the data, they will perfectly reasonably use different methods to pursue their question, such as elicitation; the limited nature of current learner corpora, relative to the demands that SLA researchers would reasonably place upon them, thus limits what may be addressed with such data sets. The available data are still relatively narrow. The corpora themselves thus represent a bottleneck that, while it has widened, still exists; beyond the bottleneck, SLA researchers remain concerned about the relevance of usage data to theories of learning.

This situation is particularly severe with regard to spoken interaction. The great bulk of available learner corpora are based on writing because such data "could easily be acquired by university researchers and in many cases, they were already digitised" (Díaz-Negrillo & Thompson, 2013, p. 11). While some conversational spoken learner corpora have been produced, they have been limited in terms of features such as size (e.g., the LeaP corpus, which contains only 47 short recordings of learners producing conversational English; see Gut, 2012) or task type (e.g., the LINDSEI corpus is composed only of three tasks—a set topic, free discussion, and a picture description). Additionally, many spoken learner corpora were designed with the intention of exploring the phonetics and phonology of learners' spoken language. They were compiled with minimal consideration of naturalness or conversational context because of the need for "high quality recordings of data" and a need "to control output so that a defined set of words are produced, allowing precise comparisons of how certain phonemes are realised" (Díaz-Negrillo & Thompson, 2013, p. 11). Corpora such as these, typically produced under experimental constraints, are of limited use in exploring features of spoken interaction where spontaneous interaction and speaker role are important.

While more written learner corpora are available, these are typically composed of essays produced by students. Such essays have the virtue of being easy to capture from exams or from classrooms. Yet users of a language are required to produce writing in many more registers than this. Short notes, emails, and positive and negative reviews are all forms of writing that we may suppose these learners of English may need to acquire. But these forms are absent from the existing learner corpora. So, there is a clear potential disjunction between the data current learner corpora provide and the research questions that SLA researchers may wish to address. Those questions may be more or less amenable to a corpus-based investigation. For example, consider three questions that Saville-Troike and Barto (2017, p.2) identified as the basic questions SLA seeks to address, "What exactly does the L2 learner come to know?" "How does the learner acquire this knowledge?" and "Why are some learners more successful than others?" On the basis of existing learner corpora, it is conceivable that LCR may have a role to play in addressing the first and the third questions. On the second question, however, the corpora must be all but silent—while one could conceive of a rich longitudinal data set that

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

blended corpus data with biographical, ethnographic, and perhaps even anthropological research that would cast light on this, this question, and further questions flowing from it, is something LCR is not currently equipped to address.

Annotation

Since 1998, there has been substantial progress in the area of corpus annotation. Computational techniques have now greatly expanded what a linguistic data set, such as a corpus, can be used for. Tools that linguists can use to annotate corpora are now well embedded in stand-alone packages such as #LancsBox (Brezina et al., 2015) or Sketch Engine (Kilgarriff et al., 2014). Using either of these packages, linguists can take plain text and, with no programming skills, part-of-speech tag and group variant word forms (e.g., *going* and *gone* can be identified as variants of *go*) in their data, then search it using a set of tools and statistical procedures embedded within the program. The results of such automated analysis should always be approached with caution, of course; the analyses are never 100% accurate, and their accuracy typically degrades markedly on learner data that such systems have not typically been trained to analyze (though, see Nagata, Mizumoto, Kikuchi, Kawasaki, & Funakoshi, 2018, for a study that looks at adapting automated part-of-speech tagging to learner English). Nonetheless, in principle and in practice, such annotations allow for more meaningful searches of corpus data, motivated by linguistic categories—rather than looking for a single word form that is a noun, like *car*, it is possible to search for all nouns. Annotations may also mitigate, but do not entirely remediate, the problems of data sparsity inherent in corpus analysis. For example, it is more effective to search the category of nouns for a pattern than to explore that pattern through a specific word that is a noun. There are 1,104,780 singular common nouns in the Longman Learner Corpus—plentiful data for analysts to begin an investigation of the use of nouns by learners of English. By contrast, specific nouns such as *archaeologist, bookworm, fallacy,* and *toothpick* occur once only in the corpus. Although impressive advances in automated corpus annotation have occurred since the first comprehensive review of available automated annotations was published more than two decades ago (Garside, Leech, & McEnery, 1997), it is still the case that manual annotations are also needed, for example, of learner errors (as argued by Lüdeling & Hirschmann, 2015).

Corpora and Theory

Turning to the question of LCR and SLA theory, any engagement with theory is conditioned by the data. From what has been outlined so far, it should be clear what corpus linguistics is good at: it excels in a study of the quotidian. This is the preoccupation that LCR has inherited from corpus linguistics, which in turn inherited it from the early corpus studies. This preoccupation is helpful, in that if a theory focuses on norms and seeing how those norms vary in different conditions (e.g., when L1 background varies, when some sociolinguistic variable like age varies, or context varies), then given the right corpus data, learner corpora can make a real contribution as they are designed for precisely this form of analysis. They can provide useful frequency data, which will permit the systematic study of such variation and norms. Consequently, if the principal goal of our testing of a theory or a hypothesis relies on such data, then learner corpora have a role to play, where such data are available. For examples of studies blending SLA and corpus data in this way, see Myles (2015, pp. 318–328).

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

Contrast this, however, with the aim of falsification. While norms can be established using corpus data and statistics may measure significant presence or even significant absence, falsification requires, in principle, just a single example—it is not a question of scale. Testing a theory also often requires a consideration of extreme, atypical, cases: theories are often tested by looking at the possible but exceptional. This explains, in all likelihood, why Sampson (1992) found that the types of sentences occurring in non-corpus-based research on grammar were very different from those found in corpora. Theoreticians testing theories should, of course, develop theories that account for the usual, but they will frequently test the theory in specific conditions that, by comparison to everyday language use, may appear extreme. While corpora may act as reservoirs of examples, some of which may constitute the types of boundary-testing examples required by theoreticians, the smaller such corpora are, and the fewer contexts they are gathered across, the greater the likelihood that the single telling example will not be present in the corpus. Hence an elicitation experiment may, in some circumstances, be a better way to proceed in testing theory than constructing a corpus.

This brings us to the key fault line between SLA research and LCR. SLA research has largely been theory driven and, to date, has tended to test theory through psycholinguistic and other (quasi)experimental methods. This choice of methods is well motivated, often permitting SLA researchers to make the observations they need swiftly and reliably. This approach has always been permissive of examining learner language in use but has also included other sources of data. Ellis (1994, p. 670) noted three types of data used by SLA researchers: (a) language as produced by learners of a language (whether that be natural or elicited and, if elicited, clinical or experimental), (b) metalinguistic judgments, and (c) self-report data. We might add a fourth class here, given the current prevalence in SLA research of observational data like eye tracking and brain scanning. In fact, SLA has always used a range of methods, which includes the exploration of naturally occurring learner language, though psycholinguistic approaches have been the mainstay of much SLA research and so-called social approaches have been gaining ground in recent years. By contrast, learner corpus researchers have been more exploratory and pre-theoretical in their approach to learner language and have used corpora to explore norms and differences with the field heavily driven by a preoccupation with "basic theoretical constructs such as overuse and underuse" (Lessard, 1999, p. 303), which corpus data were well adapted to explore.

Of course, there is awareness on both sides that methodological choices have been made, but the framing of the choices can often appear to be a claim for deficiency on the part of the other. Consider the following statement by Callies (2015, p. 35): "In contrast to other types of data that have traditionally been used in second language acquisition (SLA) research, learner corpora provide systematic collections of authentic, continuous and contextualised language use by foreign/second language (L2) learners stored in electronic format." While this claim is contextualized appropriately by Callies later in the article, an SLA researcher reading this passage may believe that it is being implied that elicitation studies do not do the same. On looking at LCR, such an SLA researcher may then conclude that learner corpora do not do this to the same degree that psycholinguistic experiments may do, leading Myles (2015) to claim that, in addition to the issues raised by us in this article, SLA researchers may view LCR as presenting a problem because the range of L2 languages represented in learner corpora needs to expand and the native control data need to be more plentiful and controlled, better matching the tasks being undertaken by the learners.

All these issues reduce the strength of the contribution that learner corpus studies can make to SLA. Granger (2009) rightly noted a range of ways in which learner corpus studies *could* provide additional insights into learner language. However, if the available data do not match the demands placed upon them when testing theory, then any contribution that may be made in principle will not be made in practice.

The Future: Opportunities and Challenges

This section identifies two major areas in which corpus linguistics can systematically contribute to the field of SLA. First, corpus linguistics can further permeate areas in SLA that have not benefitted from the use of large-scale quantitative analyses until now. We illustrate this through a consideration of pragmatics and spoken language studies. In this context, we also consider the most recent developments in the construction of L2 corpora and the implications of these developments for SLA research. Second, equally, corpora and corpus methods can bring further, significant innovation to the areas in SLA that have already considerably benefitted from quantitative, and specifically corpus-based, approaches. To illustrate this point, we briefly discuss studies on formulaic language use. Here, we argue that SLA research can benefit from the latest developments in corpus methods and corpus analytical tools that offer sophisticated yet user-friendly ways of identifying such patterns, built with SLA research questions and methodologies in mind.

Pragmatics and Learner Language

If there is a single area of linguistics that is most likely to show the difference in methodological outlook between learner corpus studies and SLA research, it would be pragmatics (cf. Aijmer & Rühlemann, 2015; Taguchi & Roever, 2017). The minimal availability of naturalistic spoken interaction across a range of tasks is a clear limitation for L2 pragmatics researchers. Studies of pragmatics that have been undertaken with learner corpora are heavily skewed toward what is immediately discoverable in the corpora—there is a bias toward the frequent and the lexical, as the lineage of learner corpus studies would allow one to assume. The bibliography of LCR (https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpus-bibliography.html) currently lists 1,144 LCR publications in the past three decades. A mere 111 (9.7%) explore speech and only 16 principally focus on pragmatics; in contrast, at least 85 of the papers focus on features of grammar (7.4%), while 148 (12.9%) explore questions broadly related to lexis/lexicography. The pragmatic studies focus on the investigation of discourse markers (14) and speech acts (2). Discourse markers and speech acts reflect a very narrow view of pragmatics, but it is what is tractable given the annotations and search software available, leading to what Culpeper, Mackey, and Taguchi, (2018, p. 194) rightly called "a dearth of studies of pragmatic development utilizing corpus techniques" in SLA. The consequence of this is that SLA research on pragmatics is mostly guided by small experimental studies (see Culpeper et al., 2018). Although this may be the appropriate choice in many cases, using learner corpus data when establishing norms of usage would undoubtedly be helpful but, because of the available corpus data, is not really possible in a manner that would meet the methodological standards of SLA research. For example, Culpeper et al. (2018) argued that issues with balancing and controlling variables have meant that current spoken corpora cannot be used to explore pragmatics in SLA research as it is not possible to balance variables such as "participant background

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

and context of learning" or "control situational variables such as setting, interlocutors' relationships, context of talk, and goals of interaction" (Culpeper et al., 2018, p. 194).

Constructing corpora to permit a substantial quantitative investigation of learner speech is, in comparison to producing written learner corpora, much more complex and expensive. This complexity and expense are amplified by the need to represent key variables related to the speaker and the context of communication. To address the concerns outlined above and in an explicit attempt to create a data set of interest to researchers working on spoken interaction and learner language, Trinity College London, a major international testing board, has been working with researchers at Lancaster University to build a spoken learner corpus based on oral examinations. The resulting corpus, the Trinity Lancaster Corpus (TLC; Gablasova et al., forthcoming), is of sufficient scale and complexity that will permit, for the first time, a large corpus-based investigation of the learning (or development) and use of a range of features of spoken interaction in learner language. While other corpora are also being developed of a similar scale to this corpus, notably the International Corpus Network of Asian Learners of English (ICNALE) corpus (see Ishikawa, 2013), there is currently nothing to compare to TLC in terms of scale: At the time of writing, ICNALE contains 400,000 words of transcribed dialogues from learners of English. In contrast, the TLC amounts to 4.1 million words of orthographically transcribed speech, representing more than 2,000 learners of English (who had British English as their target variety), performing up to four different speaking tasks, with three levels of language proficiency (B1, B2 and C1/C2 on the CEFR) and from 12 different first language backgrounds, including Chinese, Spanish, and Tamil. Some respondents have multiple L1s—where that is the case, it is encoded in the corpus. The corpus is balanced for gender, and every effort was made to represent L2 speakers from different age groups, ranging from 8 to 72 years of age. With this data set, it is possible to explore the language of learners of English using many relevant variables that the literature has suggested impact the process of language learning. In addition, a companion corpus is under development that consists of British English L1 speakers undertaking the same tasks as the learners, allowing a clear and controlled contrast of L1 and L2 speech.

Returning to Culpeper et al.'s (2018) critique of the use of existing spoken learner corpora for pragmatics research, TLC's metadata allow all of these variables to be controlled. Thus, by listening to the concerns of SLA researchers, the builders of learner corpora can begin to deliver on the promise of interaction between the two fields.

Formulaic Language

Over the last two decades, the study of formulaic language has attracted increasing attention from SLA (e.g., Howarth, 1998; Wray, 2013), making it one of the core topics in L2 learning research today (Gablasova, Brezina, & McEnery, 2017a). The human ability to communicate fluently in real time depends to a large extent on a knowledge of formulaic units (phraseological competence), that is, the ability to store, access, and produce prefabricated chunks of language such as multiword expressions (*would like to*) or lexico-grammatical frames (*as far as* X *is concerned*). This competence represents a key aspect of communicating in a nativelike, effortless, and error-free manner (e.g., Ellis, 2002; Ellis, Simpson-Vlach, Römer, O'Donnell, & Wulff, 2015; Erman, Forsberg Lundell, & Lewis, 2016).

The contribution of corpus linguistics to the understanding of formulaic language use to date has been significant. First, corpus linguistics developed robust, quantitative

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

methods specifically aimed at identifying different types of recurrent formulaic units that consist of lexical or lexico-grammatical units (in particular, collocations) that can occur in adjacent as well as nonadjacent positions in text. It would not be possible to identify these patterns (which involve different degrees of surface variation) without the use of automatic searches of large quantities of language. Second, through the creation of corpora, the field has also provided data sets of the necessary size and structure that allow researchers to apply these sophisticated quantitative methods. As a result, the corpus-based approach to identifying formulaic units in L1 and L2 production, which relies on frequency as the primary variable, has complemented the phraseological approaches that usually require manual analysis and focus on factors such as semantic properties (e.g., semantic unity, completeness, transparency, and opaqueness) of the formulaic units (Granger & Paquot, 2008). Corpus-based analysis of formulaic language has become increasingly prominent in SLA research, having been used to describe different levels of L2 proficiency (e.g., Granger & Bestgen, 2014) as well as to compare L1 and L2 speakers of the target language (e.g., Erman et al., 2016; Schmitt, 2012). See Ellis et al. (2015) for a wider review of the use of learner corpus approaches to studying formulaic language and the role they play in SLA research.

However, although the benefits of using corpora and corpus methods in SLA studies on formulaic language use have been considerable so far, the full potential of learner corpora in the area hasn't yet been fully realized. For example, L2 learning research could benefit from corpus advances in studies of collocations, a particularly strong and dynamic domain within corpus linguistics (e.g., Brezina et al., 2015; Gries, 2013). Likewise, collocation-based analyses of learner corpora have proven to be rich sources of hypotheses about language learning and processing that can be further explored using other experimental techniques (see Durrant & Siyanova-Chanturia, 2015; Gilquin & Gries 2009). Tools such as #LancsBox (described earlier) have proved to be very efficient in analyzing and visualizing associations between words in corpora, and packages such as AntGram (Anthony, 2018) and kfNgram (Fletcher, 2012) have been developed specifically to allow the exploration of multiword sequences. SLA research thus could benefit from the principled approach corpus linguistics offers: both (a) to identify and visualize collocations and (b) to investigate the role of collocational knowledge in first and additional language (L1 and L2) learning. For example, corpus linguistics works with a large number of association measures, that is, measures used to identify the collocatability of units of language and the probability of their co-occurrence (e.g., Evert, 2008). While SLA studies have incorporated the use of association measures into their research methods, the set of measures adopted so far has been very limited, and in some cases, the rationale for their selection has not been clear (for overviews of some of the statistics used to calculate collocation and their associated rationales, see Gablasova et al., 2017a; González Fernández & Schmitt, 2015). Another major area of potential collaboration between SLA and corpus linguistics lies in exploring the different types of collocations and their relationship to psycholinguistic reality (Durrant & Siyanova-Chanturia, 2015; Gablasova et al., 2017a). Work such as this is important in establishing common ground between LCR and SLA research as it seeks to demonstrate that collocations are not merely artefacts of usage, but are also a part of the language system itself and thus have a role to play in the processes of learning.

Conclusion—The Language Learning Nexus

It is well established in SLA that multiple methods, a range of disciplinary inputs, and a range of data types are needed to explore language learning (as clearly established in,

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

e.g., Gass, Behney, & Plonsky, 2014). The same point is made by researchers in other fields, notably psychology (see Ellis, 2019). Working across disciplinary boundaries has long been acknowledged to be difficult—yet the promise of it is great. It is a promise that has been identified by LCR researchers with regard to SLA, but as this article has shown, that promise has yet to be realized.

Changing this situation requires those seeking to make a contribution to embrace the intrinsic complexity of the language learning process. We believe it is best conceived of as a nexus. Scollon (2001) viewed L1 acquisition as a nexus—SLA is the same. It has the potential to bridge cognition, sociology, educational research, psychology, neurology, and many other areas. Identifying areas with something to contribute is not, however, where the difficulty arises; it lies in how, and whether, to combine them to formulate and address a research question given that they "intersect, never perfectly, never in any finalized matrix or latticework of regular patterns" (Scollon, 2001, p. 142). In such a context, simply thinking of bridging areas is not enough—it is too passive an approach to promote interaction. What is needed is an active engagement with a dynamic nexus within which multiple disciplinary and methodological perspectives interact. It draws in subjects that have lesser or greater contributions to make in this area, but that at present barely connect though they may clearly be relevant to the theory and practice of language learning.

Given the complex, protean nature of the L2 learning nexus, clearer acknowledgment of the different goals and working practices of LCR researchers, SLA researchers, and researchers in other areas is essential. A key to this is to set aside directionality and dominance; learner corpus studies do not necessarily represent "a real improvement over the narrow empirical foundation of most SLA studies" (Granger, 1998a, p. 11) any more than SLA theory will provide LCR, which "remains rather descriptive" with "theoretical frameworks which would enable rigorous interpretation or explanation of the data" (Myles, 2015, p. 330). Within the nexus, if the empirical data needed to test a hypothesis may reasonably be furnished by traditional/existing sources of data, such as elicitation, there is no need for a wider empirical base. If the goal of a piece of research is the production of a descriptive grammar or learner dictionary, questions of theory may not be paramount. The key to bringing the fields together in the nexus is to accept that each may be pursued, quite legitimately, without regard to the others, but even where that happens, the results produced may ultimately contribute to the complex nexus formed by language learning.

We would like to end this article on a positive and practical note. Most researchers would agree that interdisciplinary and multidisciplinary research involving LCR and SLA is a desirable endpoint. Yet, as Frickel, Albert, & Prainsack, (2017, p. 8) pointed out, "decisions encouraging interdisciplinarity are often made based on faith more than on evidence." Interdisciplinarity in academia is also often discussed in abstract terms, lacking day-to-day practical suggestions for researchers in the relevant disciplines. Adopting Bronstein's (2003) model of collaboration, which is informed by the practical needs of social workers, practitioners whose success and failure have immediate consequences and who receive immediate feedback from their clients and other stakeholders, we propose the following five points of useful interaction between LCR and SLA:

1. Interdependence: Try to understand the theoretical assumptions, methodological practices, and the limitations of your own discipline and the discipline you want to collaborate with. Clearly recognized limitations can create a space for genuine

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

collaboration through the recognition of the interdependence of the two disciplines. This article has been a conscious attempt to work toward doing this.

2. Newly created professional activities: Create opportunities for regular interactions between researchers from the two fields. These include professional symposia, conferences, and journals that are not restricted to a narrow audience. Although specialized journals such as *The International Journal of Learner Corpus Research* (*IJLCR*) and book series such as *Task-Based Language Teaching* (*TBLT*) play an important role for each research community, especially for methodologically oriented contributions, interdisciplinary and outward-looking venues for research should be promoted. It was in this spirit that the workshop sponsored by *Language Learning* reported in Rebuschat et al. (2017) was held.
3. Flexibility: Reach productive compromises in the face of disagreement. Don't limit the scope of your analyses a priori but be ready to embrace and encompass the perspective of the other field, if it proves useful. Take one another's methods and results seriously. It is in this spirit that we propose a nexus as the crucible within which LCR and SLA should interact.
4. Collective ownership of goals: Think of language learning and SLA from the perspective of both the process (SLA) and the product (LCR). Jointly design, define, develop, and achieve research goals. It was thinking such as this that led us to design the TLC with the goals of both LCR and SLA research in mind.
5. Reflection on process: Regularly reflect on the collaboration process as defined in 1–4 above. Adjust practices accordingly and incorporate feedback.

The last point is important. We view this article, and our efforts to work toward Bronstein's points 1–4 as the beginning, not the end, of the process of interaction between LCR, SLA, and other research areas interested in language learning. The promise of linking LCR and SLA is great—yet nobody should think it is easy. They should, however, think it is a goal worth striving for.

Author ORCIDs. Tony McEnery, 0000-0002-8425-6403.

Acknowledgements. The work presented in this article was supported by Trinity College London and the UK Economic and Social Research Council (grants ES/R008906/1, EP/P001559/1 and ES/S013679/1).

References

Aijmer, K., & Rühlemann, C. (Eds.). (2015). *Corpus pragmatics: A handbook*. Cambridge, UK: Cambridge University Press.

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning*, *67*(S1), 180–208. doi:10.1111/lang.12232

Anthony, L. (2018). *AntGram* (Version 1.0.0). Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/software

Baker, P., & Egbert, J. (Eds.). (2019). *Triangulating corpus linguistics with other linguistic research methods*. New York, NY: Routledge.

Barlow, M. (2005). Computer-based analyses of learner language. In R. Ellis & G. Barkhuizen (Eds.), *Analysing learner language* (pp. 335–369). Oxford, UK: Oxford University Press.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2014). ETS Corpus of non-native written English LDC2014T06. Web download file. Philadelphia, PA: Linguistic Data Consortium.

Brezina, V. (2018). Collocation graphs and networks: Selected applications. In P. Cantos-Gómez & M. Almela-Sánchez (Eds.), *Lexical collocation analysis: Advances and applications* (pp. 59–83). Cham, Switzerland: Springer International.

https://doi.org/10.1017/S0267190519000096 Published online by Cambridge University Press

Brezina, V., & Bottini, R. (2019). Review of Castello, Erik, Katherine Ackerley & Francesca Coccetta, Eds. (2015) Studies in learner corpus linguistics. Research and applications for foreign language teaching and assessment. *International Journal of Learner Corpus Research*, *5*(1), 113–117.

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. doi:10.1075/ijcl.20.2.01bre

Bronstein, L. R. (2003). A model for interdisciplinary collaboration. *Social Work*, *48*(3), 297–306.

Buchanan, M. A. (1929). *A graded Spanish word book*. Publications of the American and Canadian Committees on Modern Languages. Toronto, Canada: University of Toronto Press.

Callies, M. (2015). Learner corpus methodology. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 35–56). Cambridge, UK: Cambridge University Press.

Coccetta, F., Castello, E., & Ackerley, K. (2015). *Studies in learner corpus linguistics: Research and applications for foreign language teaching and assessment*. Bern, Switzerland: Peter Lang.

Cornu, A. M., & Delahaye, M. (1987). Variability in interlanguage reconsidered: LSP vs. Non-LSP IL talk. *English for Specific Purposes*, *6*(2), 145–151.

Culpeper, J., Mackey, A., & Taguchi, N. (2018). *Second language pragmatics: From theory to research*. London, UK: Routledge.

De Cock, S., Granger, S., Leech, G. N., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 67–79). London, UK: Longman.

Díaz-Negrillo, A., & Thompson, P. (2013). Learner corpora: Looking towards the future. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (pp. 9–30). Amsterdam, The Netherlands: John Benjamins.

Duff, P. A., & Byrnes, H. (Eds.). (2019). SLA across disciplinary borders: New perspectives, critical questions, and research possibilities. *The Modern Language Journal*, *103*(S1), 3–5. doi:10.1111/modl.12537

Durrant, P., & Siyanova-Chanturia, A. (2015). Learner corpora and psycholinguistics. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 57–77). Cambridge, UK: Cambridge University Press.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188. doi:10.1017/S0272263102002024

Ellis, N. C. (2019). Essentials of a theory of language cognition. *The Modern Language Journal*, *103*(S1), 39–60. doi:10.1111/modl.12532

Ellis, N. C., Römer, U., & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: Cognitive and corpus investigations*. New York, NY: Wiley.

Ellis, N. C., Simpson-Vlach, R., Römer, U., O'Donnell, M. B., & Wulff, S. (2015). Learner corpora and formulaic language in second language acquisition research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 357–378). Cambridge, UK: Cambridge University Press.

Ellis, R. (1994). *The study of second language acquisition*. Oxford, UK: Oxford University Press.

Erman, B., Forsberg Lundell, F. F., & Lewis, M. (2016). Formulaic language in advanced second language acquisition and use. In K. Hyltenstam (Ed.), *Advanced proficiency and exceptional ability in second languages* (pp. 111–148). Boston, MA: De Gruyter Mouton.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (pp. 1212–1248). Berlin, Germany: De Gruyter Mouton.

Færch, C., Haastrup, K., & Phillipson, R. (1984). *Learner language and language learning*. Clevedon, UK: Multilingual Matters.

Fletcher, W. H. (2012). *kfNgram*. Retrieved from http://www.kwicfinder.com/kfNgram/kfNgramHelp.html

Frickel, S., Albert, M., & Prainsack, B. (2017). Introduction: Investigating interdisciplinarities. In S. Frickel, M. Albert, & B. Prainsack (Eds.), *Investigating interdisciplinary collaboration: Theory and practice across disciplines* (pp. 5–26). New Brunswick, NJ: Rutgers University Press.

Fries, C. C., & Traver, A. (1940). *English word lists: A study of their adaptability for instruction*. Washington, DC: American Council of Education.

Gablasova, D., Brezina, V., & McEnery, T. (2017a). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*(S1), 155–179. doi:10.1111/lang.12225

Gablasova, D., Brezina, V., & McEnery, T. (2017b). Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning*, *67*(S1), 130–154. doi:10.1111/lang.12226

Gablasova, D., Brezina, V., & McEnery, T. (forthcoming). The Trinity Lancaster Corpus: Development, description and application. *International Journal of Learner Corpus Research*.

Garside, R., Leech, G. N., & McEnery, T. (1997). *Corpus annotation: Linguistic information from computer text corpora*. London, UK: Longman.

Gass, S., Behney, J., & Plonsky, L. (2014). *Second language acquisition: An introductory course*. New York, NY: Routledge.

Gilquin, G., De Cock, S., & Granger, S. (2010). The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM. Louvain-La-Neuve, Belgium: Presses Universitaires de Louvain.

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1–26. doi:10.1515/CLLT.2009.001

González Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? The effects of frequency and amount of exposure. *International Journal of Applied Linguistics*, *166*(1), 94–126. doi:10.1075/itl.166.1.03fer

Granger, S. (1998a). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3–18). London, UK: Routledge.

Granger, S. (Ed.). (1998b). *Learner English on computer*. London, UK: Routledge.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13–32). Amsterdam, The Netherlands: John Benjamins.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229–252. doi:10.1515/iral-2014-0011

Granger, S., Dagneux, E., & Meunier, F. (2002). *The international corpus of learner English*. Louvain, Belgium: Université Catholique de Louvain.

Granger, S., Gilquin, G., & Meunier, F. (2015). Introduction: learner corpus research - past, present and future. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 2–5). Cambridge, UK: Cambridge University Press.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & F. Meunier (Eds.), *Phraseology. An interdisciplinary perspective* (pp. 27–50). Amsterdam, The Netherlands: John Benjamins.

Gries, S. T. (2013). 50-something years of work on collocations. What is or should be next …. *International Journal of Corpus Linguistics*, *18*(1), 137–166. doi:10.1075/ijcl.18.1.09gri

Gries, S. T. (2015). Statistics for learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 160–181). Cambridge, UK: Cambridge University Press.

Gut, U. (2012). The LeaP corpus: A multilingual corpus of spoken learner German and learner English. In T. Schmidt & K. Wörner (Eds.), *Multilingual corpora and multilingual corpus analysis* (pp. 3–23). Amsterdam, The Netherlands: John Benjamins.

Hall, J. K. (2019). The contributions of conversation analysis and interactional linguistics to a usage-based understanding of language: Expanding the transdisciplinary framework. *The Modern Language Journal*, *103*(S1), 80–94. doi:10.1111/modl.12535

Henmon, V. A. C. (1924). *A French word book based on a count of 400,000 running words*. Bureau of Educational Research Bulletin, No 3. Madison, WI: University of Wisconsin.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24–44. doi:10.1093/applin/19.1.24

Huebner, T. (1983). *A longitudinal analysis of the acquisition of English*. Ann Arbor, MI: Karoma.

Huebner, T. (1985). System and variability in interlanguage syntax. *Language Learning*, *35*(2), 141–163. doi:10.1111/j.1467-1770.1985.tb01022.x

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner Corpus Studies in Asia and the World*, *1*, 91–118. doi:10.24546/81006678

Jamieson, J. (2005). Trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, *25*, 228–242. doi:10.1017/S0267190505000127

Juvonen, P. (1989). Repair in second-language instruction. *Nordic Journal of Linguistics*, *12*(2), 183–204. doi:10.1017/S0332586500002043

Kecskes, I., & Kirner-Ludwig, M. (2017). "It would never happen in my country I must say": A corpus-pragmatic study on Asian English learners. *Corpus Pragmatics*, *1*(2), 91–134. doi: 10.1007/s41701-017-0007-x

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, M., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7–36. doi: 10.1007/s40607-014-0009-9

Leclercq, P., Edmonds, A., & Hilton, H. (Eds.). (2014). *Measuring L2 proficiency: Perspectives from SLA* (Vol. 78). Bristol, UK: Multilingual Matters.

Lessard, G. (1999). Learner English on computer [Review of the book *Learner English on computer*, by S. Granger (Ed.)]. *Computational Linguistics*, *25*(2), 302–303.

Longman active study dictionary (5th ed.). (2010) London: Pearson Education Limited.

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 135–157). Cambridge, UK: Cambridge University Press.

Mackey, A., & Gass, S. M. (2012). *Research methods in second language acquisition: A practical guide*. Chichester, UK: Wiley-Backwell.

Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design*. 2nd ed. New York, NY: Routledge.

McEnery, T. (2005). *Swearing in English: Bad language, purity, and power from 1586 to the present*. London, UK: Routledge.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge, UK: Cambridge University Press.

Myles, F. (2015). Second language acquisition theory and learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 309–331). Cambridge, UK: Cambridge University Press.

Nagata, R., Mizumoto, T., Kikuchi, Y., Kawasaki, Y., & Funakoshi, K. (2018). A POS tagging model adapted to learner English. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop in Noisy User-Generated Text* (pp. 39–48). Brussels, Belgium: Association for Computational Linguistics.

Nicholls, D. (2003). The Cambridge learner corpus - error coding and analysis for lexicography and ELT. In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the Corpus Linguistics 2003 conference, 28–31 March. Technical Papers, Vol. 16* (pp. 572–581). Lancaster, UK: UCREL, Lancaster University.

Philips, M. (1985). *Aspects of text structure: An investigation of the lexical organisation of text*. Amsterdam, the Netherlands: North-Holland.

Rebuschat, P., Meurers, D., & McEnery, T. (2017). Language learning research at the intersection of experimental, computational, and corpus-based approaches. *Language Learning*, *67*(S1), 6–13. doi:10.1111/lang.12243

Sampson, G. (1992). Probabilistic parsing. In J. Svartvik (Ed.), *Directions in corpus linguistics. Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991* (pp. 425–447). Berlin, Germany: De Gruyter Mouton.

Saville-Troike, M., & Barto, K. (2017). *Introducing second language acquisition*. 3rd ed. Cambridge, UK: Cambridge University Press.

Schmitt, N. (2012). Formulaic language and collocation. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–10). New York, NY: Blackwell.

Scollon, R. (2001). *Mediated discourse: The nexus of practice*. New York, NY: Routledge.

Simpson-Vlach, R., & Swales, J. (Eds.). (2001). *Corpus linguistics in North America: Selections from the 1999 symposium*. Ann Arbor, MI: University of Michigan Press.

Slabakova, R. (2019). "L" stands for language. *The Modern Language Journal*, *103*, 152–160. doi:10.1111/modl.12528

Taguchi, N., & Roever, C. (2017). *Second language pragmatics*. Oxford, UK: Oxford University Press.

Tono, Y. (2018). Corpus approaches to L2 learner profiling research. In Y. Leung, J. Katchen, S. Hwang, & Y. Chen (Eds.), *Reconceptualizing English language teaching and learning in the 21st century: A special*

monograph in memory of Professor Kai-Chong Cheung (pp. 392–409). Taipei, Taiwan: Crane Publishing Co., Ltd.

West, M. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London, UK: Longman.

Williams, G. (1998). Collocational networks: Interlocking patterns of lexis in a corpus of plant biology research articles. *International Journal of Corpus Linguistics*, *3*(1), 151–171. doi: 10.1075/ijcl.3.1.07wil

Wray, A. (2013). Formulaic language. *Language Teaching*, *46*(3), 316–334. doi:10.1017/S0261444813000013

Cite this article: McEnery T, Brezina V, Gablasova D, Banerjee J (2019). Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyze Language Use. *Annual Review of Applied Linguistics* **39**, 74–92. https://doi.org/10.1017/S0267190519000096