LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

CONCEPTUAL REVIEW ARTICLE

Exploring Learner Language Through Corpora: Comparing and Interpreting Corpus Frequency Information

Dana Gablasova, Vaclav Brezina, and Tony McEnery

Lancaster University

This article contributes to the debate about the appropriate use of corpus data in language learning research. It focuses on frequencies of linguistic features in language use and their comparison across corpora. The majority of corpus-based second language acquisition studies employ a comparative design in which either one or more second language (L2) corpora are compared to a first language (L1) production corpus or two or more L2 corpora are compared to each other. This article critically examines some of the central tenets of the comparative method related to the interspeaker variation in L1 and L2 use, the representativeness and comparability of corpus data, the interpretation of difference found between corpora and the appropriate use of statistics. Using and discussing a set of five L1 spoken English corpora and three L2 English corpora (two spoken and one written), we approach these areas empirically exploring different sources of variations and methodological options that corpus-based SLA studies offer.

Keywords corpus-based SLA; comparative design; native speaker corpora; nonnative speaker corpora; corpus frequency; comparing L1 and L2 speakers

---

We wish to thank the anonymous reviewers for their valuable comments. The research presented in this article was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1, and Trinity College London. Information about access to the data used in this article is provided in Appendix S1 in the Supporting Information online.

Correspondence concerning this article should be addressed to Dana Gablasova, Department of Linguistics and English Language, Lancaster University, Lancaster LA1 4YL, UK. E-mail: d.gablasova@lancaster.ac.uk

DOI: 10.1111/lang.12226

Introduction

One of the major aims of second language acquisition (SLA) research is to describe and explain the factors that underlie the learning and use of a second (additional) language (Ädel, 2015; Ellis, 2012;). Despite the considerable volume of SLA research to date, researchers are "still a long way from understanding how second languages are learned, why many individuals have difficulty in reaching high levels of proficiency in a second language, or even what the best pedagogical approach might be" (Gass & Mackey, 2014, p. 1). Corpora and corpus techniques are becoming increasingly more integral to the investigation of SLA, complementing findings from experimental and classroom studies. Corpus approaches can provide a detailed description of interlanguage at different stages of development and thus offer insights into the sources of variation in the language use and level of proficiency reached by language learners.

This article, in particular, focuses on one type of information provided by corpora—information about the frequency of linguistic features occurring in contextualized language use. This information can provide further insights into the linguistic system developed by second language (L2) learners and users[1] (Ädel, 2015; Callies, 2015; Granger, 2012). As Leech (2011, p. 5) stated, "if asked what is the one benefit that corpora can provide and that cannot be provided by other means, I would reply 'information about frequency.'" Corpus frequencies can be analyzed at different levels of language use (e.g., lexical, grammatical, and pragmatic) and with respect to a single occurrence or the co-occurrence of linguistic features (Ädel, 2015; Granger, 2012, 2015; McEnery & Hardie, 2011; McEnery, Xiao, & Tono, 2006). Whereas frequency information from first language (L1) and expert users can help us understand the target use of language, information from L2 production can uncover linguistic patterns that point to underlying factors in SLA.

Frequency information about features in L2 production is undoubtedly a valuable source of evidence about language development and use; however, like in other SLA research, it would be difficult to interpret this evidence without a reference point. As a result, the majority of corpus-based SLA studies employ a comparative design in which a L2 corpus[2] is compared either to a corpus of language from native speakers of the target language or to another L2 corpus (Callies, 2015; Granger, 2009, 2015; McEnery et al., 2006). This research design allows researchers to contextualize and evaluate L2 corpus evidence, making the findings more informative than if a L2 corpus was considered alone (Granger, 2015). While acknowledging its advantages, researchers have also pointed to problems inherent in the contrastive approach and called for

a careful reconsideration of methodological choices involved in comparing corpora (Ädel, 2015; Gries, 2015; Gilquin & Granger, 2015; Leech, 1998; see also the discussion of comparative designs within the framework of the Contrastive Interlanguage Analysis in Granger, 2009, 2012, 2015). Drawing on and extending this debate, this article examines some of the central ideas involved in comparing and interpreting frequency information from two or more corpora. While relevant to the comparison of all corpora, the comparison of L1 and L2 corpora, one of the most common approaches in this research design, will be used to demonstrate the issues. Overall, the article seeks to highlight some of the core notions in corpus linguistics (CL), thus providing an important background for a meaningful application of CL methods in language learning research.

Corpus-Based Frequency Information As an Insight Into Language Development

Language use varies according to a number of linguistic, cognitive, and situational factors and their interaction; this variation can display different degrees of systematicity (Romaine, 2003; see also Ädel, 2015, for an overview of approaches to variation in corpus-based studies). When studying systematic, predictable, variation, different speaker-internal and speaker-external predictors are correlated with particular language outputs (frequencies of linguistic features). Findings of these studies show that speakers' language can differ as a result of group characteristics (such as gender, age, regional affiliation, socioeconomic background, cultural background, or proficiency) or individual preferences (such as personal style) (e.g. Buchholtz & Hall, 2005). Language production can also be systematically affected by situational characteristics such as genre, formality, or purpose of the communication and topic (e.g., Biber & Conrad, 2009).

As documented by numerous studies discussed in this article, corpora are effective for studying variation in both L1 and L2 language use. At this stage, it is important to define variation, the central notion of these empirical studies and an important aspect of the discussion in this article. Broadly speaking, there are two approaches to variation, the formal and the functional approach, which differ considerably in how they conceptualize a variable and in terms of their research design. The more restrictive formal approach follows the variationist tradition in sociolinguistics (e.g., Labov, 1972) and focuses on the frequency of different options (forms or variants) that speakers use to express the same meaning. For example, "Have you eaten?" and "Did you eat?" are two grammatical realizations of the same meaning. Ädel (2015,

p. 402), translating this approach to learner corpus research, sees variation as "different speakers expressing the same meaning by means of different forms ('inter-learner variation')" or "a single speaker using different linguistic forms to express the same meaning on different occasions ('intra-learner variation')." On the other hand, the functional approach, following the Biber (e.g., 1988) tradition of register variation, focuses on the frequency of linguistic features (lexico-grammatical characteristics) in texts that speakers use and that are "best suited functionally to their situations and communicative purposes" (Biber & Conrad, 2009, p. 265). For example, pragmatic or discourse markers are used in speech and writing with a variety of functions. These markers can occur almost anywhere in the text and are not in a direct competition with each other. Both types of variation have been studied in the corpus-based SLA studies and thus for purposes of this article we adopt a broad definition of variation that subsumes both the formal and the functional approach.

A number of factors conditioning variation in language attainment and use can be examined using L2 corpora; as Evert (2009) states, "many linguistic questions can be operationalized as a frequency comparison." For example, recent corpus-based research on L2 development and production have studied a broad range of lexico-grammatical as well as pragmatic structures such as the use of phrasal verbs (e.g., Chen, 2013; Gilquin, 2015), pragmatic markers (Buysse, 2015; Zhang & Sabet, 2016), or the development of different aspects of vocabulary and formulaic sequences (e.g., O'Donnell, Römer, & Ellis, 2013). A common approach in these contrastive studies is to compare the frequency of the target feature in one or more L2 corpora with frequency in a L1 corpus, report the findings in terms of a difference in the frequency of occurrence and explain what factors can account for this difference (transfer from a specific L1, the difficulty of the linguistic feature, the level of L2 proficiency, etc.). For example, Gilquin (2015) compared the frequency and type of phrasal verbs used by L2 speakers of English with French L1 using part of a corpus of spoken L2 English (Louvain International Database of Spoken English Interlanguage [LINDSEI]) with a reference corpus of spoken L1 English (Louvain Corpus of Native English Conversation [LOCNEC]), linking the observed difference between the groups to the effect of French on the ability of L2 speakers to acquire and produce these verbs. The majority of corpus-based SLA studies so far have focused on linking the patterns in the interlanguage (i.e., the frequency of the target linguistic features) to the effects of transfer from L1 (cf. Ädel, 2015; Granger, 2015); however, other predictors of L2 variation have also been explored (e.g., the effect of task and L2 proficiency) and will be further discussed in this article.

Table 1 L1 corpora

Corpus	Tokens	No. of speakers	Type of interaction
CANCODE (CANC)	5,076,313	1,138[a]	
BNC – Demographic (BNC_Dem)	4,329,797	1,407	informal conversations
BNC – 2014 Spoken (BNC_SP)	4,789,185	376	
BNC64	1,575,995	64	
LOCNEC	124,935	50	informal (elicited) conversations

Note. [a]The number of individual speakers is not available. All statistics for CANCODE reported in this article use per text (conversation) frequencies. CANCODE = Cambridge and Nottingham Corpus of Discourse in English; BNC = British National Corpus; LOCNEC = Louvain Corpus of Native English Conversation.

Table 2 L2 corpora

Corpus	Tokens	No. of speakers	Type of interaction
LINDSEI	1,215,176	554	informal (elicited) conversations
TLC v. 2.1	3,296, 351	1,449	semiformal (elicited) monologues & conversations
ICLE v.1	2,576,379	3,799	academic writing

Note: LINDSEI = Louvain International Database of Spoken English Interlanguage; TLC = Trinity Lancaster Corpus; ICLE = International Corpus of Learner English.

The Corpora and Linguistic Structures Used in This Article

To demonstrate the issues involved in the interpretation and comparison of corpus frequencies, data from several corpora are analyzed or discussed in this article. The list of corpora used in this article is provided in Tables 1 (L1 corpora) and 2 (L2 corpora); an extended version, containing more information about the corpora and access to the data can be found in Appendix S1 in the Supporting Information online along with more detailed information about the corpus methods used to extract the information.

The corpora were analyzed using CQPweb (Hardie, 2012) and the SketchEngine (Kilgarriff et al., 2014). Ideally, for the sake of direct comparability, only one corpus query system should be used; however, this was not possible due to corpus accessibility restrictions. Special care was therefore taken to ensure comparability of tokenization and part-of-speech tagging when different taggers and corpus annotation schemes were involved (see Appendix S2 in the Supporting Information online).

Frequency comparison has been used widely in corpus-based research to study language use at all levels of language (lexical, syntactic, morphological, pragmatic as well as their combinations). The frequency with which speakers use linguistic features can provide us with an insight into the state of their interlanguage and is a first step in the study of what motivates the use (or avoidance) of these feature in their language. For example, a very low occurrence of passive or past tense structures in the contexts in which these are regularly used by L1 speakers may point to issues with acquisition of these structures by L2 learners. To illustrate the points discussed in this article we selected four linguistic structures to represent a range of levels of the linguistic analysis: (i) a pragmatic marker *I think,* (ii) an adjectival phrase modified by an adverb as a syntactic feature (ADV+ADJ), (iii) the simple past tense as a morphological feature, and (iv) the passive construction as a syntactic/stylistic feature. These structures are highly frequent in language and have been studied frequently as indicators of L2 proficiency (to read more about findings on L2 use based on these features, see, e.g., Baumgarten & House, 2010; Gablasova, Brezina, McEnery, & Boyd, 2015; Hinkel, 2004; Pérez-Paredes & Díez-Bedmar, 2012). These structures were selected to demonstrate that the issues discussed in this article are present at every level of linguistic analysis that involves frequency comparison; however, due to space constraints, usually only one set of results will be presented and discussed, with the remaining results, demonstrating the same point, available in Appendix S3 in the Supporting Information online. While the issues explored in this article are relevant to studies of learner language based on different kinds of corpora (e.g., corpora of written, academic or online language) corpora of spoken language will mostly be used as examples in this article.

Issues in Corpus-Based Studies of Learner Production

In this part of the article, we address core issues related to the analysis and comparison of frequencies of linguistic features in corpora. We first focus on representativeness and comparability of corpora and the range of interspeaker variation found inside the corpora; although variation in language use can be observed in both L1 and L2 production, the corpus-based approaches to the study of these groups differ. We then address the issue of how to meaningfully compare frequency information from corpora and discuss how to interpret differences found when corpora are compared and what statistical tests one may use to evaluate these differences. All of these issues have direct implications both for our ability to evaluate corpus-based findings and to generalize them to the wider population of language learners and users.

Representativeness and Comparability of Corpora

When comparing two corpora, researchers typically wish to study the effect of a particular independent variable or variables. The corpora that are compared should thus differ with respect to this variable but, ideally, be similar in other respects (Leech, 1998). This assumption underlies the ability to attribute the observed difference in corpus frequencies with a reasonable degree of certainty to the predictor variable and measure the strength (size) of its impact. In current practice, there are in particular two issues that make this seemingly straightforward assumption problematic—the representativeness and comparability of corpora.

In order to explain patterns identified in the corpora, it is essential to carefully examine the representativeness of the corpora. Corpora are samples that seek to represent language used by a certain group of speakers in a certain setting (or across a range of settings; McEnery & Hardie, 2011). If, for example, researchers wish to study the crosslinguistic effect of a particular L1 on the L2 acquisition and use, then all of the speakers in the L2 corpus should share the characteristic (i.e., the same L1 background) and typical linguistic patterns found in the corpus can then be related to this trait (McEnery et al., 2006; McEnery & Hardie, 2011). One of the main challenges in corpus-based SLA analyses is therefore to avoid introducing a systematic bias into the corpus (and the corpus-based analysis), which may happen if there are several characteristics shared by many or all speakers in the corpus that can influence their language production and act as a confounding variable. As Leech emphasized (1998, p. xix), "we have to be cautious in drawing general inferences from the results of a corpus analysis, and alert to the influence of hidden variables implicit in the way we collected and sampled the data." Due to practical issues in corpus building, such as recruiting participants, it is relatively easy to introduce such characteristics (e.g., the speakers may be from the same educational system, be interviewed by one person, or be of a particular age). For example, in addition to being L1 speakers of Chinese, all participants in the Chinese component of LINDSEI also share other characteristics: They come from one educational institution and none of them spent time in an English-speaking country. By contrast, all the German participants also come from one educational institution and have on average spent more than 9 months in an English-speaking country (Gilquin, DeCock, & Granger, 2010). Thus differences found between these two groups could be equally related to the shared L1 background or learning history of these speakers.

Corpus representativeness has direct implications for corpus comparability, the degree to which two corpora are similar. If two corpora differ in too many respects (if they represent different genres of a language or speakers), it may be difficult to ascertain, with a reasonable degree of certainty, the source of any difference observed. In the majority of SLA corpus-based studies, corpus comparability has been defined almost exclusively through its design criteria (i.e., what the corpus sets out to represent; Leech, 1998; Barlow, 2005; Gilquin & Granger, 2015) with the actual representativeness usually not validated (examined) after the data collection. However, deciding the extent to which two corpora are comparable is a complex issue that should not be reduced to the matter of how similar/different the language elicitation situation was. Even with the same design criteria, the language in corpora can differ systematically. For example, the picture description task used to elicit language in LINDSEI (a corpus of L2 spoken English) and LOCNEC (a corpus of L1 spoken English, designed to mirror LINDSEI) could be considered a highly controlled task, with the same picture described by all L1 and L2 speakers. However, when we look at transcripts from LINDSEI subcorpora from different L1 groups, it is apparent that while the language in some subcorpora is highly monologic (e.g., in the Chinese subcorpus), in others the interviewers took a more active role, with the situation resembling a conversation (e.g., the Bulgarian subcorpus).

When deciding whether two corpora can be meaningfully compared, the likelihood of occurrence of the target linguistic features also has to be considered with respect to the nature of the linguistic data in the corpus. For example, a large number of lexico-grammatical features are register dependent (e.g., Biber, 1988) and the topic of contributions can affect the occurrence of specialized words; function words, grammatical structures, and more general vocabulary items, on the other hand, can be expected to occur regardless of the topic.

Although the representativeness and comparability of corpora have major implications for the validity of corpus findings, they have so far received little direct attention in corpus-based SLA studies. It is imperative that researchers familiarize themselves with the corpora that they use. This involves considering the metadata available about the corpus and speakers included in it as well as looking at (at least a sample of) the corpus files. Particular attention should be given to the interspeaker variation in each of the corpora; if interspeaker patterns within one corpus vary considerably, this may make it difficult to interpret any difference found between the corpora. The issues related to interspeaker variation and its sources are discussed in the following section. Researchers should also consider replicating research based on a particular L2 corpus with other L2 corpora containing similar data to provide further validity to their claims.

Interspeaker Variation in Corpora of L1 and L2 Production

L2 Learners and Users

When comparing frequency information from two corpora, it is imperative to first examine the interspeaker variation found within each of the corpora in order to determine to what extent the quantitative summary (e.g., a measure of central tendency such as the mean) can be used to represent the language produced by individual users in the corpus. It is also important to reflect on the sources of the variation among the users in a corpus in order to decide whether speakers from different corpora can be meaningfully compared.

Linking variation in the language use of L2 speakers to various factors has been one of the central issues of SLA with both speaker-internal variables (such as age, proficiency, and learning history) and speaker-external variables (such as the type of instruction and the effect of task) studied extensively (e.g., Ellis, 2012). Corpus-based SLA studies have not yet developed a similar tradition, with the majority of studies focusing on L1 background as a factor in L2 acquisition (Granger, 2015). However, possibly due to a closer interaction with SLA research, there has recently been a gradual shift toward also paying more attention to other variables common in SLA research (Ädel, 2015; Callies, 2015), with a steady emergence of studies exploring factors such as the effect of different tasks on the L2 production (Ädel, 2008; Gablasova & Brezina, 2015; Gablasova et al., 2015; Tracy-Ventura & Myles, 2015). Despite these promising developments in this area, both speaker-related variables (in particular L2 proficiency) and functionally conditioned variation have not yet received sufficient attention despite the role they play in the variation in language use of L2 speakers.

One of the major variables affecting language use that has not yet received sufficient attention in corpus-based SLA studies is the level of L2 proficiency. Proficiency is a major source of variation in L2 use and it has been used as a predictor, control, and outcome variable in numerous SLA studies. Despite its significance, at present, standards for establishing and reporting speakers' proficiency in L2 corpora differ greatly, making it difficult to interpret and compare findings from different groups of L2 users. In the majority of learner corpora, L2 users' proficiency has been estimated with reference to external criteria (such as number of years spent studying English or a number of years abroad). For example, Buysse (2015, p. 66) described speakers in her corpus-based study as learners who "were at least in their second year of studying English as a major subject at university" despite the fact that the learners came from five different educational backgrounds. The practice of relying on external indicators of proficiency can result in speakers of mixed proficiency

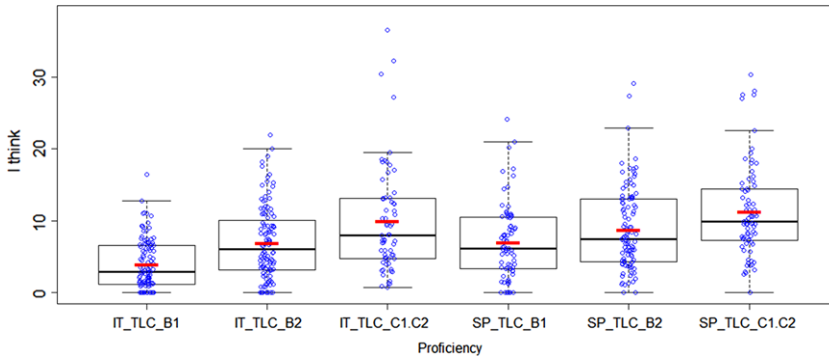

Figure 1 *I think* in TLC: Proficiency levels and L1 backgrounds. [Color figure can be viewed at wileyonlinelibrary.com]

being grouped together in one corpus designed to represent, for example, speakers of a certain L1 background. In other cases, speakers of different proficiency levels can be compared with respect to a different independent variable. For example, Callies (2015) points out that in the International Corpus of Learner English (ICLE), based on written academic language from speakers of different L1s, proficiency of speaker groups varies considerably with some groups performing at the B2 (e.g., Chinese) and others at the C1 (e.g., Russian) level of the Common European Framework of Reference for Languages (CEFR) and with yet other groups showing a "higher degree of intra-group variability" (e.g., German and Czech components; pp. 28–29). These groups of speakers are routinely compared with each other (see, e.g., Lu & Ai, 2015; Wulff, 2015) or grouped together (e.g., O'Donnell et al., 2013) despite the fact that L2 proficiency can act as a confounding variable in these studies distorting the effect of the variables that researchers wish to observe (e.g., the effect of a particular L1 as in the case of Lu & Ai, 2015).

While there has been an increase in calls for a more rigorous treatment of proficiency in corpus-based SLA studies (e.g., Ädel, 2015; Callies, 2015; Granger, 2015), the majority of corpus-based SLA studies do not strive to examine more closely the proficiency of the speakers/texts in the corpus and, to report it, if possible, on a standardized scale (e.g., in terms of CEFR or major exam scores such as the International English Language Testing System [IELTS]). To illustrate the possible impact of not distinguishing more carefully among speakers of different proficiency levels in corpus-based studies let us look at the following example. Figure 1 shows the frequency of the occurrence

of *I think* (normalized to 1,000 words) across three different proficiency levels (B1, B2, and C1 and above of CEFR) of L1 Spanish and Italian speakers of English participating in two dialogic tasks in the Trinity Lancaster Corpus (TLC) of spoken L2 English ($N_{Italian,\ B1} = 111$; $N_{Italian,\ B2} = 116$; $N_{Italian,\ C1+} = 65$; $N_{Spanish,\ B1} = 77$; $N_{Spanish,\ B2} = 110$, $N_{Spanish,\ C1+} = 75$).

Figure 1[3] shows both, a relatively large interspeaker variation but also a developing positive trend with the increase in the use of the stance marker related to growing proficiency for both speaker groups. For both groups, there was a statistically significant difference between B1 and B2 (Italian L1 speakers: $t(204.23) = -5.21$, $p < 0.001$; Spanish L1 speakers: $t(171.51) = -2.07$; $p = 0.040$) and B2 and C1+ levels (Italian L1 speakers: $t(96.79) = -2.94$, $p = 0.004$; Spanish L1 speakers: $t(143.88) = -2.81$; $p = 0.006$). This difference would have remained obscured if speakers from the B1 to the C1+ levels were aggregated as has often happened in previous corpus-based SLA studies (cf. Callies, 2015). The results for the three other features (passive construction, adverbial modification, and simple past), placed in Appendix S3 in the Supporting Information online due to space constraints, show similar trends with respect to interspeaker variation that appear to some extent reduced by grouping speakers more closely according to their proficiency level. Even when proficiency does not seem to reduce (and explain) some of the interspeaker differences, it is important to examine it and, if the evidence supports this, exclude it as a possible source of variation.

The closer look at proficiency as one variable that affects interspeaker variation in language use demonstrates the value of examining or controlling variables that can affect the frequency with which L2 users produce a certain linguistic structure or expression. While corpus research of L1 production has devoted considerable attention to variation in language use in relation to the situational or functional characteristics of communication (e.g., Biber & Conrad, 2009), this factor has not yet been systematically examined in corpus studies of L2 production despite the fact that several corpora are well positioned for such research as they record production from the same users across several tasks (e.g., LINDSEI and TLC). Several corpus-based SLA studies have already provided evidence of intraspeaker variation in language use across different tasks and speaker roles (Gablasova & Brezina, 2015; Gablasova et al., 2015; Tracy-Ventura & Myles, 2015). Overall, a greater attention to both speaker- and context-related factors affecting higher or lower frequency of production of target linguistic features can contribute significantly to a better description and understanding of the linguistic system developed by the L2 users. As their range and size grows, L2 corpora increasingly lend

themselves to a systematic study of these variables, thus offering a richer description of L2 production and L2 speakers than the often-used dichotomy of learners contrasted with native speakers, as criticized by Cook (1999, 2016).

L1 Speakers

In SLA research, whether based on corpora or other data sources, the primary focus is understandably on L2 users. However, in order to evaluate to what extent L2 use resembles or differs from L1 use, language produced by native speakers needs to be closely scrutinized as well. This becomes even more imperative if the production of L1 and L2 speakers is to be compared directly.

While corpus research on L1 use has devoted considerable attention to functional variation (e.g., according to the setting and purpose of communication), similar attention has not been given to individual variation related to factors such as gender, age, or level of education. This is likely connected to the assumption that adult native speakers are relatively similar to one another in terms of language processing and use as a result of being exposed to and using the target language since birth/early years (cf. Davies, 2003, 2011; Lardiere, 2013). As a result of this shared experience with language, native speakers "are often thought to process language in ways that are efficient and accurate, and with little inter-individual variation" (Birdsong & Gertken, 2013, p. 122). In corpus-based SLA studies, the references to, for example, "full proficiency" and "native-like grammar" as the aim of learning (Ädel, 2015, p. 404) and pedagogical suggestions based on deviations from the trends found in L1 corpora suggest that many of these studies still appear to treat L1 speakers as a homogeneous group who all possess the full extent of linguistic knowledge. However, as Davies argues, the idea that "all these native speakers are at C2, the highest level on the Council of Europe Reference Scale, makes no sense. Some perhaps are, but they are unusual" (Davies, 2011, p. 306). Similarly, Birdsong and Gertken (2013, p. 122) argue that "recent research is at odds" with the "assumption" of the homogeneous nature of language mastery among native speakers. This research indicates that variation among native speakers is observed not only with respect to lexical choice (Bucholtz & Hall, 2005) but also in morpho-syntactic attainment and choices, which have traditionally been believed to be relatively constant among native speakers (see Dąbrowska, 2012, for a review of recent evidence; Andringa, 2014). Interestingly, while language mastery has been a major variable in (corpus-based) SLA research involving L2 speakers, it is rarely measured or discussed with respect to L1 speakers.

To demonstrate the value of exploring interspeaker differences in L1 corpora, we examined the use of the four linguistic features by L1 speakers in five

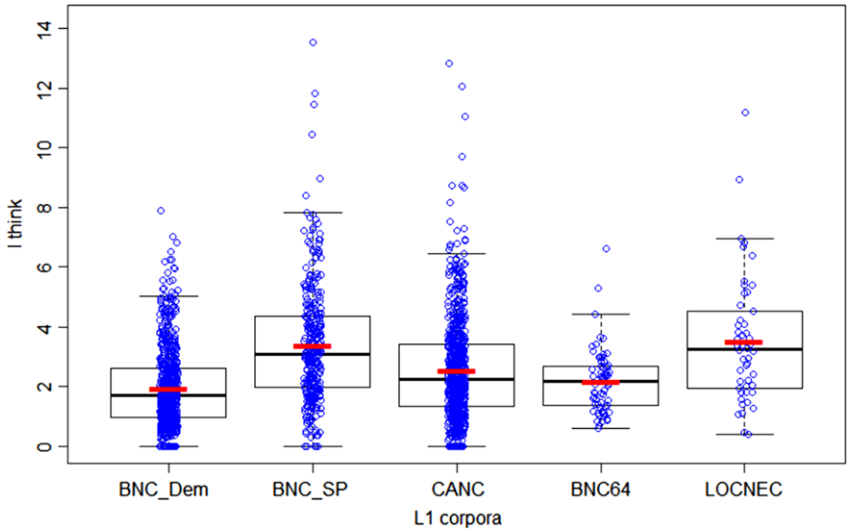

Figure 2 *I think* in five native speaker corpora. [Color figure can be viewed at wileyonlinelibrary.com]

corpora of informal spoken British English: BNC_DEM, BNC_Spoken, CANCODE, BNC64, and LOCNEC. Figure 2 offers a visual representation of the frequency (normalized to 1,000 words) with which speakers used *I think* in their conversations. The results for the remaining three linguistic features, demonstrating the same point, have been placed in Appendix S3 in the Supporting Information online.

The boxplots in Figure 2 point to a considerable interspeaker (and, in one case, interconversational) variation in the use of *I think;* (in CANCODE the points represent conversations and not individual speakers). Similar variation can be seen for the other three features, whether related to morphological, syntactic, stylistic. or pragmatic level of language description. It is of particular interest that a large degree of variation is also found among speakers in LOCNEC, a corpus that could be considered as the most homogeneous in terms of the speaker characteristics (all speakers are students at the same British university, aged 18–30), interviewed by the same person, and performing the same set of linguistic tasks. To examine interspeaker variation further, let us look at two examples from LOCNEC taken from the picture description task in which the L1 speakers were asked to describe the same set of four pictures:

Speaker 1: well there's a: obviously an artist erm .. maybe he's amateur maybe he's professional he's asked someone if they would pose for him so he can

paint their picture paint their portrait .. and . perhaps he's a professional portrait painter . erm [. . .] obviously she's not very pleased with it . she's rather displeased so erm . she's got him to repaint it or paint her again . and I think he's used a bit of artistic licence there [LOCNEC, E07]

Speaker 2: the story right .. erm . this bloke he's got his friend round . to: do a picture of her [. . .] and .. he he just gets <XX> he won't let her see it till it's finished and when it's finished she looks at it and she thinks oh my god is that supposed to be me and has a go at him and he gets all insulted and all those <X> even though [. . .] it looks like . the mirror image of her so he thinks .. oh well . even though it's a good painting .. I've offended her because she doesn't like the way she looks [. . .] so he makes up . a totally false impression of her .1. and erm .. then she shows it off to all her mates and she's really made up with it [LOCNEC, E10]

As can be seen from the extracts, the two speakers opted for a different degree of formality in their narratives with implications for their lexical and syntactic choices. Compared to the first speaker, the second speaker used more informal language (indicated by underlined phrases) characterized by colloquial expression ("oh well," "oh my god"), informal lexical expressions (e.g., "bloke," "have a go at"), a range of phrasal verbs ("show off," "make up") and intensifiers ("totally").

There are two immediate implications for corpus-based SLA studies following this illustration. First, even if the primary focus in SLA studies is on L2 production, researchers should carefully examine and report the interspeaker variation in L1 production with particular emphasis on factors to which it can be attributed (e.g., sociolinguistic variables, levels of literacy). It is important to note at this point that, while interspeaker variation can be observed in both L1 and L2 speakers, the range and sources of this variation may be different (e.g., L2 speakers are likely to be more heterogeneous and a certain degree of variation in their interlanguage can be attributed to language errors that are not common to L1 speakers). However, the variation that is present has a direct implication for how we think about "typicality" or "standards" of L1 production and what is considered acceptable in language use. The second point is related to the variation illustrated above in the production of both L1 and L2 speakers. Information about the range of interspeaker variation plays an important role in our understanding of how language is acquired by individual speakers or groups of speakers; yet, this information is often lost in research designs that aggregate speakers into groups without taking such variation into account. These designs are still very common in corpus-based SLA studies

(e.g., Buysse, 2015; Gilquin, 2015) and may obscure interesting patterns in L1 and L2 use. This issue is discussed in more detail in the section on the use of statistics.

Comparison of the Corpora and Interpreting the Difference

In corpus-based SLA studies, the ability to interpret similarities/differences found between corpora is essential for drawing valid conclusions about speakers' L2 ability. However, at present it is somewhat difficult to meaningfully interpret the difference between two corpora, because there is little understanding of what constitutes a small or large difference between any corpora. There has been only limited empirical investigation into differences/similarities between corpora (Gries, 2006; Kilgarriff, 2001) leading Kilgarriff (2001, p. 97) to conclude that "[CL] lacks strategies for describing and comparing corpora." The commonly accepted practice has been to establish whether the difference in frequency of occurrence of the target linguistic features is statistically significant; if the difference is statistically significant, this is usually taken as proof that speakers' use of target features in two corpora is systematically different.

To explore the range of cross-corpus differences, we compared L1 corpora that represent language collected in similar situations (although some of them at different points in time, see Appendix S1 in the Supporting Information online). The differences between frequencies from corpora sampling the same type of language will provide a baseline of a typical amount of variation between different corpora, enabling us to better evaluate differences found between L1 and L2 corpora. When looking at large enough samples (corpora) of L1 production, controlled for register, we hypothesize that we should not expect to see large differences between these samples. In this case, we looked at differences/similarities between four corpora representing informal British speech with respect to the occurrence of the four linguistic features described above. The descriptive statistics (mean relative frequency per 1,000 words [*M*], *SD*, and interquartile range [*IQR*][4]) can be seen in Table 3.

In addition, for each of the four variables, Figure 3 displays 95% confidence intervals (CIs) shown as error bars around the mean relative frequency (a circle in the graph) in each corpus; the error bars indicate a statistically established interval within which the mean lies in 95% of cases of random samples taken from the same population. In practical terms, largely overlapping CIs imply that the difference between the means is not statistically significant; on the other hand, nonoverlapping or slightly overlapping CIs show statistically significant differences between the corpora. Clearly, in most cases the 95% CIs in Figure 3 do not overlap, which runs counter to our hypothesis. We may view this

Table 3 Linguistic features: Mean frequency per 1,000, standard deviations and *IQR*s

	I think		Adverb + adjective	
Corpus	*M (SD)*	*IQR*	*M (SD)*	*IQR*
BNC_Dem	1.94 (1.36)	0.95–2.61	4.96 (2.47)	3.26–6.18
BNC_SP	3.39 (2.01)	1.98–4.36	8.80 (2.79)	6.73–10.44
CANC	2.55 (1.74)	1.35–3.43	8.01 (3.33)	5.99–9.49
BNC64	2.19 (1.09)	1.39–2.69	6.39 (2.06)	5.21–7.17
	Past tense		Passives	
BNC_Dem	33.73 (12.69)	24.74–41.53	2.26 (1.46)	1.41–2.83
BNC_SP	21.50 (7.66)	16.47–26.37	3.66 (1.39)	2.81–4.33
CANC	32.96 (18.06)	18.26–43.60	4.43 (2.23)	3.01–5.44
BNC64	36.66 (9.03)	29.70–42.15	3.93 (0.80)	3.45–4.54

finding as surprising considering the fact that the corpora were designed to represent the same type of language—informal British conversation produced by L1 speakers.

Focusing on *I think* as an example, as shown in Table 3, the relative frequencies with which the target structure appeared in each corpus can vary, sometimes quite considerably, with the most striking difference found between BNC_SP and BNC_Dem, both sampling interactive informal spoken English and constructed as mirror corpora (albeit with a 30-year difference between them). All or most of the differences for *I think* were statistically significant depending on the statistical tests used. In CL, either a statistical test such as the chi-squared is used (a traditional approach) or a more precise test that takes into account interspeaker variation can be employed (*t* test, analysis of variance [ANOVA], or their nonparametric counterparts). With a series of pairwise chi-square tests, all differences in the use of *I think* in the four corpora turned out to be statistically significant at $p < 0.05$. A series of pairwise *t* tests[5] showed that five out of six comparisons were statistically significant at the $p < .05$ level with the effect size (Cohen's *d*) ranging from little effect (0.19) to large (0.91) effect. Similar patterns were obtained for the other three variables as shown by CIs in Figure 3.

These findings allow us to demonstrate two points. First, corpora sampling similar native speaker language can differ considerably. The differences between the means are in the magnitude of several units (occurrences per 1,000 words).[6] In addition, large *SD*s (relative to the means) and wide *IQR*s suggest

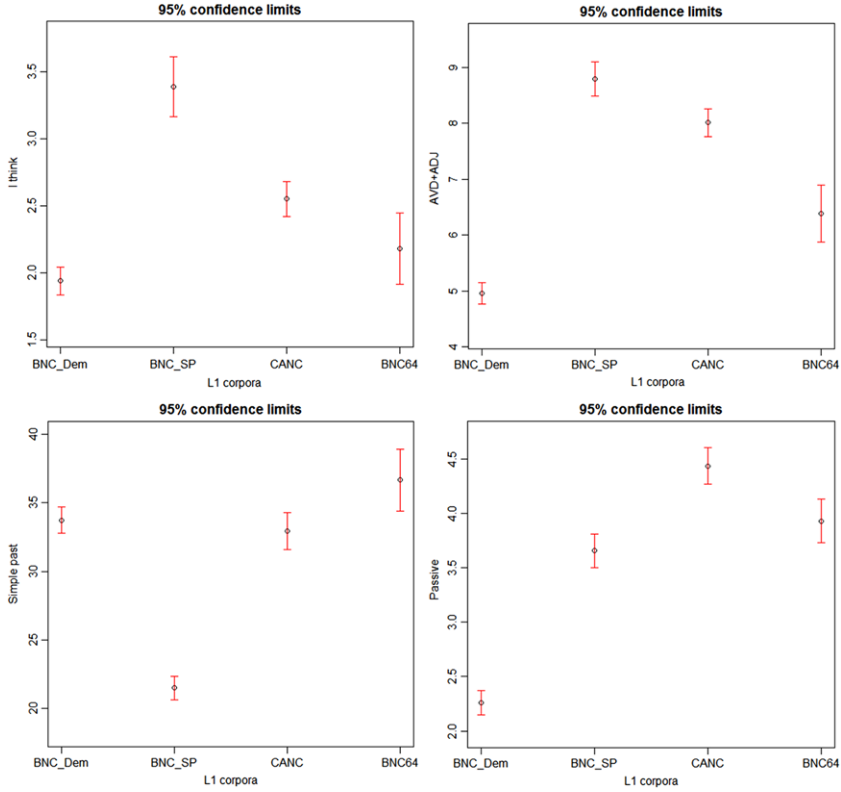

Figure 3 Linguistic features: 95% confidence intervals. [Color figure can be viewed at wileyonlinelibrary.com]

a substantial amount of variation inside each corpus (possible reasons for this variation have been discussed above). Second, assuming that all the assumptions of the tests are met and based on the 95% CIs and statistical significance testing, we have to conclude that the corpora represent different populations (uses of language) with respect to the four linguistic features. Arguably, the variation can be somewhat curbed by choosing corpora that are more controlled across a range of features (e.g., with respect to speaker characteristics, time period of data collection) and in which speakers perform exactly the same tasks. However, as we saw in Figure 2, even when using LOCNEC, in which speakers participated in the same four tasks, the variation between native speakers was large and it is likely that two L1 corpora with the same tasks would differ to a considerable degree.

Overall, drawing implications for comparative research design, we have demonstrated that not all differences between corpora can be automatically attributed to the difference between L1 and L2 use, because L1 corpora are very likely to differ from each other as well. This fact should be taken into consideration when evaluating the results of cross-corpora frequency comparison, as well as when drawing pedagogical implications based on these findings.

Variation Within and Between Corpora: The Meaningful Use of Statistics

The previous section explored the interpretation of a difference between corpora with respect to variation. Considering the large amount of within- and between-corpus variation that is part of linguistic reality, several points should be made regarding the meaningful use of statistics in corpus-based SLA studies.

First, our thinking about typical performance as documented by corpus evidence should shift from considering a single value per corpus (such as the mean frequency) to representing linguistic production in a corpus by a range of values. This can be achieved through a more systematic application of statistical procedures that take interspeaker variation into account (e.g., they use and report means, *SD*s, *IQR*s, etc.). Visualizations (such as the boxplots in Figures 1 and 2) that provide an insight into the distribution of the occurrences of target linguistic features should also be used. Needless to say, interspeaker variation should be carefully documented for both L1 and L2 speakers, because this represents an essential piece of information about the use of language. One of the major advantages of corpora is the scale of the evidence provided and the ability to see whether a certain usage pattern is systematic (e.g., whether we see it repeat itself again and again across a number of speakers). This can only be observed if we distinguish individual speakers or texts in the analysis and report within-corpus variation in the research report.

Second, the role of absent evidence (i.e., the zero occurrences in some speakers' production) needs to be carefully considered. In many cases, the mean frequency obscures the fact that many (L1 as well as L2) speakers do not produce the target structure (see, e.g., Figure 2). Taking into account that occurrence will also be dependent on the type of linguistic feature and its optionality, this phenomenon has been noted in several cases in L2. For example, Callies (2015) observed highly uneven production with many L2 speakers not producing the target feature in his studies of emphatic *do,* demonstrative clefts, first-person pronouns, and *it-* and *there*-sentences. While numerically ordered on the same scale, zero occurrences may signal two very

different states in terms of speakers' linguistic knowledge—one a matter of *ability* to produce the feature (e.g., lack of mastery or activation), the other the matter of *choice* (the speakers decide not to produce it or do not have an opportunity to produce it) (Granger, 2012). The corpus data therefore needs to be critically evaluated for the type of evidence (and absence of it) they represent. The information about the distribution of linguistic features in individual speaker production is an invaluable guide in this process (see point one).

Third, we need to be very careful when comparing corpora and interpreting the differences found. We saw that many differences that we observed between corpora were statistically significant.[7] Although the measures of statistical significance (statistical tests) can be useful tools for evaluating variation, that is, for answering the question of whether the variation between corpora is larger than the variation inside them and whether this variation may be due to chance, they are by no means a substitute for the critical evaluation of the observed differences among speakers/texts and between corpora and the size of this difference (the so-called effect size such as Cohen's *d* reported in the previous section). It should also be noted that in order to get useful information about statistical significance, we need to select tests that are appropriate for our type of data. For example, using a family of tests for categorical data such as the chi-squared or the log-likelihood (likelihood ratio) test is still a common practice in corpus-based SLA studies (e.g., Ädel & Erman, 2012; Buysse, 2015; Gilquin, 2015), despite the fact that there is compelling evidence that points to severe limitations of this practice (Bestgen, 2014; Brezina & Meyerhoff, 2014; Lijffijt et al., 2016). As shown by Brezina and Meyerhoff (2014), with this family of tests we can get statistically significant results between random groups of speakers because the elementary assumption of these tests (i.e., the independence of observations) is severely violated.[8] The *p* values are thus not reliable; in fact, they are close to meaningless. A better option in these situations is to use a test that takes individual texts/speakers as observations (Gilquin & Granger, 2015), such as the *t* test, ANOVA, or their robust or nonparametric counterparts.

Conclusion

This article examined some of the core issues in SLA studies based on corpus frequencies, which have implications for the validity of the results and their generalizability to a wider population of L2 learners and users. The article argues that critical reflection on the nature of corpus evidence is crucial for understanding the results of corpus frequency comparison. In particular,

researchers should carefully asses the amount of interspeaker variation and investigate its sources to decide whether there is a systematic difference in the language use of two or more groups of speakers or whether the difference is an artifact of the corpus design or the analytical tool used.

Language production is a dynamic process and speakers enter interactions with a plurality of aims; both L1 and L2 speakers also vary in their mastery of the language. Recording and describing variation among speakers is one of the most valuable contributions of CL to the study of SLA, complementing the evidence from more controlled studies that employ sophisticated techniques to decrease natural heterogeneity and variation of language used in the context of meaningful communication. Findings from corpus-based studies can place this evidence into a broader perspective and thus contribute to strengthening of the theoretical base of SLA.

Final revised version accepted 19 November 2016

Notes

1 Throughout this article, the term "L2 user" will be used to refer both to those who still actively learn L2 as well as those who use it, as it is impossible to make a clear distinction between the two categories in the data we used.
2 There are different terminological traditions when referring to corpora representing L1 and L2 language (corpora of L1 have been referred to as "native-speaker corpora" and corpora of L2 have been referred to as "learner" or "nonnative speaker" corpora). We will refer to the corpora from L1 speakers as "L1 corpora" and to corpora from L2 speakers as "L2 corpora."
3 In the graph (boxplot), each data point (circle) represents an individual speaker. The long line represents the median, the short red line the mean, the box represents middle 50% values in the distribution (*IQR*) and the whiskers extend to the minimum and the maximum values disregarding the outliers (1.5 times *IQR*).
4 *IQR* represents the range of the middle 50% of values in a distribution; these values tend to be more stable than the extremes (minimum and maximum values) reported in a simple range.
5 The Welsh version of the *t* test was used as a robust option, because the homoskedasticity assumption was violated.
6 For example, in their study, Fung and Carter (2007) established the cutoff distance between the L1 and L2 corpora at 1.14 occurrences of the linguistic structure per 1,000 words. If L2 speakers differed from L1 speakers by more than this, they were considered to under- or overuse the particular linguistic feature. We can see that comparisons between several corpora in Table 3 would satisfy this criterion.
7 A technical note: When assessing the significance of the differences at the .05 alpha level ($p < .05$), which is customary in linguistic and social science research, we

only corrected for family-wise type 1 error if multiple comparisons were intended to be regarded as valid at the same time as, for example, in post hoc tests; when pairwise comparisons between corpora were used to establish how often these would turn out significant in real corpus-based studies where typically only two corpora are compared, no such correction was deemed necessary.

8 Briefly, when used for a comparison of the frequencies of a linguistic feature in two (or more) corpora, these tests treat every word in the corpus as an observation analogous to a participant in a study who was asked a yes/no question. Thus in a corpus of one million words, the chi-squared test works with the assumption that we have one million participants who are independent of each other. When we find that the target word or phrase occurred 5,000 times in the corpus, the test works with the information that 5,000 of your participants answered "yes" to the question while another 995,000 said "no." Language, however, does not work like this because individual words are not independent; on the contrary, individual words form an interconnected chain of meaning relationships as they combine in sentences and texts.

References

Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness?. *Language and Computers*, *66*, 35–53. doi:10.1163/9789401206204_003

Ädel, A. (2015). Variability in learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Cambridge handbook of learner corpus research* (pp. 379–400). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139649414.018

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, *31*, 81–92. doi:10.1016/j.esp.2011.08.004

Andringa, S. (2014). The use of native speaker norms in critical period hypothesis research. *Studies in Second Language Acquisition*, *36*, 565–596. doi:10.1017/S0272263113000600

Barlow, M. (2005). Computer-based analyses of learner language. In R. Ellis & G. Barkhuizen (Eds.), *Analysing learner language* (pp. 335–358). Oxford, UK: Oxford University Press.

Baumgarten, N., & House, J. (2010). I think and I don't know in English as lingua franca and native English discourse. *Journal of Pragmatics*, *42*, 1184–1200. doi:10.1016/j.pragma.2009.09.018

Bestgen, Y. (2014). Inadequacy of the chi-squared test to examine vocabulary differences between corpora. *Literary and Linguistic Computing*, *29*, 164–170. doi:10.1093/llc/fqt020

Biber, D. (1988). *Variation across speech and writing*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511621024

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511814358

Birdsong, D., & Gertken, L. M. (2013). In faint praise of folly: A critical review of native/non-native speaker comparisons, with examples from native and bilingual processing of French complex syntax. *Language, Interaction and Acquisition/ Langage, Interaction et Acquisition*, *4*, 107–133. doi:10.1075/lia.4.2.01bir

Brezina, V., & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, *19*, 1–28. doi:10.1075/ijcl.19.1.01bre

Bucholtz, M., & Hall, K. (2005). Identity and interaction: A sociocultural linguistic approach. *Discourse Studies*, *7*, 585–614. doi:10.1177/1461445605054407

Buysse, L. (2015). "Well it's not very ideal . . . " The pragmatic marker well in learner English. *Intercultural Pragmatics*, *12*, 59–89. doi:10.1515/ip-2015-0003

Callies, M. (2015). Using learner corpora in language testing and assessment: Current practice and future challenges. In E. K. Castello, K. Ackerley, & F. Coccetta (Eds.), *Studies in learner corpus linguistics: Research and applications for foreign language teaching and assessment* (pp. 21–35). Frankfurt, Germany: Peter Lang. doi:10.3726/978-3-0351-0736-4

Chen, M. (2013). Overuse or underuse: A corpus study of English phrasal verb use by Chinese, British and American university students. *International Journal of Corpus Linguistics*, *18*, 418–442. doi:10.1075/ijcl.18.3.07che

Cook, V. (1999). Going beyond the native speaker in language teaching. *TESOL Quarterly*, *33*, 185–209.

Cook, V. (2016). Where is the native speaker now?. *TESOL Quarterly*, *50*, 186–189. doi:10.2307/3587717

Dąbrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, *2*, 219–253. doi:10.1075/lab.2.3.01dab

Davies, A. (2003). *The native speaker: Myth and reality*. Bristol, UK: Multilingual Matters.

Davies, A. (2011). Does language testing need the native speaker? *Language Assessment Quarterly*, *8*, 291–308. doi:10.1080/15434303.2011.570827

Ellis, R. (2012). *The study of second language acquisition*. Oxford, UK: Oxford University Press.

Evert, S. (2009). *Rethinking corpus frequencies*. Presentation at the ICAME 30 Conference, Lancaster, UK.

Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics*, *28*, 410–439. doi:10.1093/applin/amm030

Gablasova, D., & Brezina, V. (2015). Does speaker role affect the choice of epistemic adverbials in L2 speech? Evidence from the Trinity Lancaster Corpus. In J.

Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics 2015* (pp. 117–136). Dordrecht, Netherlands: Springer. doi:10.1007/978-3-319-17948-3_6

Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2015). Epistemic stance in spoken L2 English: The effect of task type and speaker style. *Applied Linguistics*. Advance online publication. doi:10.1093/applin/amv055

Gass, S. M., & Mackey, A. (2014). *The Routledge handbook of second language acquisition*. Abingdon, UK: Routledge. doi:10.4324/9780203808184

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching. In K. Aijmer (Ed.), *Corpora and language teaching* (pp. 13–332). Amsterdam: John Benjamins. doi:10.1075/scl.33.04gra

Granger, S. (2012). How to use foreign and second language learner corpora. In A. Mackey & S. M. Gass (Eds.). *Research methods in second language acquisition: A practical guide* (pp. 7–29). Oxford, UK: Wiley-Blackwell. doi:10.1002/9781444347340.ch2

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*, 7–24. doi:10.1075/ijlcr.1.1.01gra

Gries, S. Th. (2006). Exploring variability within and between corpora: Some methodological considerations. *Corpora*, *1*, 109–151. doi:10.3366/cor.2006.1.2.109

Gries, S. Th. (2015). Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics*, *16*, 93–117. doi:10.1177/1606822X14556606

Gilquin, G. (2015). The use of phrasal verbs by French-speaking EFL learners. A constructional and collostructional corpus-based approach. *Corpus Linguistics and Linguistic Theory*, *11*, 51–88. doi:10.1515/cllt-2014-0005

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

Gilquin, G., & Granger, S. (2015). Learner language. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 418–436). Cambridge, UK: Cambridge University Press.

Hardie, A. (2012). CQPweb—Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, *17*, 380–409. doi:10.1075/ijcl.17.3.04har

Hinkel, E. (2004). Tense, aspect and the passive voice in L1 and L2 academic texts. *Language Teaching Research*, *8*, 5–29. doi:10.1191/1362168804lr132oa

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, *6*, 97–133. doi:10.1075/ijcl.6.1.05kil

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíč ek, M., Kovář , V., Michelfeit, J., et al. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*, 7–36. doi:10.1007/s40607-014-0009-9

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.

Lardiere, D. (2013). Nativelike and non-nativelike attainment. In J. Herschensohn & M. Young-Scholten (Eds.), *The Cambridge handbook of second language acquisition* (pp. 670–691). Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9781139051729.038

Leech, G. (1998). Preface. In S. Granger & G. Leech (Eds.), *Learner English on computer* (pp. xiv–xx). Abingdon, UK: Routledge. doi:10.4324/9781315841342

Leech, G. (2011). Frequency, corpora and language learning. In F. Meunier, S. De Cock, & G. Gilquin (Eds.), *A taste for corpora. In honour of Sylviane Granger* (pp. 7–31). Amsterdam: John Benjamins.

Lijffijt, J., Nevalainen, T., Säily, T., Papapetrou, P., Puolamäki, K., & Mannila, H. (2016). Significance testing of word frequencies in corpora. *Digital Scholarship in the Humanities*, *31*, 374–397. doi:10.1093/llc/fqu064

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing*, *29*, 16–27. doi:10.1016/j.jslw.2015.06.003

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge, UK: Cambridge University Press. doi:10.1017/CBO9780511981395

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London: Taylor & Francis.

O'Donnell, M. B., Römer, U., & Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics*, *18*, 83–108. doi:10.1075/ijcl.18.1.07odo

Pérez-Paredes, P., & Díez-Bedmar, M. B. (2012). The use of intensifying adverbs in learner writing. In Y. Tono, Y. Kawaguchi, & M. Minegishi (Eds.), *Developmental and crosslinguistic perspectives in learner corpus research* (pp. 105–123). Amsterdam: John Benjamins. doi:10.1075/tufs.4.11per

Romaine, S. (2003). Variation. In C. J. Doughty & M. H. Long (Eds), *The handbook of second language acquisition* (pp. 410–435). Malden, MA: Blackwell. doi:10.1002/9780470756492.ch14

Tracy-Ventura, N., & Myles, F. (2015). The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research*, *1*, 58–95. doi:10.1075/ijlcr.1.1.03tra

Wulff, S. (2015). What learner corpus research can contribute to multilingualism research. *International Journal of Bilingualism*. Advance online publication. doi:10.1177/1367006915608970

Zhang, G. Q., & Sabet, P. G. (2016). Elastic "I think": Stretching over L1 and L2. *Applied Linguistics*, *37*, 334–353. doi:10.1093/applin/amu020

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1. Data.
Appendix S2. Corpus Tools and Search Terms.
Appendix S3. Additional Results.