FIRST PAGE PROOFS

Incorporating Corpora in Second-Language Acquisition Research

STEFAN TH. GRIES

Introduction

Q1

Over the last few decades, corpus-linguistic applications have become much more widely used in SLA, a field that traditionally had a more experimental bent to it. More and more studies coming from mostly usage-based linguistics (and less so formal/generative approaches) are now targeting various levels of linguistic analysis—lexis, syntax, semantics—using the more naturalistic data and the often highly quantitative and computational methods corpora offer to researchers. The following two sections provide very brief and selective surveys of two areas in which corpus methods have had the highest impact on SLA research, (a) the quantitative measurement of linguistic features in learner writing or speaking (often relative to comparable production by native speakers or other targets) and (b) case studies targeting specific phenomena in learner production. The final section offers a quick evaluation and sketches out a few desiderata for the field.

Characterization of Texts

One very widespread kind of application of corpus-linguistic methods in SLA involves developing measures of various linguistic characteristics of (spoken or written) learner production. Such measures can be important in two major ways: (a) to quantify proficiency levels of learners and their production and (b) to develop or validate tests, assessment tools, or even teaching material at the right levels for different kinds of learners. A trifecta of notions central to much such SLA research is *complexity*, *accuracy*, and *fluency* (*CAF*), but there has also been a lot of discussion regarding how to operationalize these ultimately multi-dimensional constructs. As Housen et al. (2012, pp. 4–5) discuss: Complexity has been understood as including both linguistic and cognitive complexity and, relevant to corpus approaches, the former refers to “the intrinsic formal or semantic-functional characteristics of L2 elements […] or to properties of (sub-)systems of L2 elements.” Accuracy “in essence refers to the extent to which an L2 learner’s performance […] deviates from a norm (usually the native speaker)” (but Housen et al. also caution that the “A” in CAF should maybe be broadened to include *appropriateness* and *acceptability*), and fluency characterizes the “ease, eloquence, ‘smoothness’ and native-likeness of speech or writing”; from these definitions, it is obvious that many quantitative corpus-linguistic operationalizations are conceivable for each of these notions.

In addition to corpora of learner production, native-speaker corpora have also been used insightfully, namely as reference points of sorts. For example, Murakami and Ellis (2022) use the native-speaker Corpus of Contemporary American English as a proxy for learners’ input and Paquot’s (2019) study of phraseological complexity (see below) used the native-speaker L2 Research corpus (approximately 7,800 articles/66m words covering L2 research between 1980 and 2014) to calculate *PMI* scores of dependency relations to be compared to learner data.

The Encyclopedia of Applied Linguistics. Edited by Carol A. Chapelle
(Section Editors: Danijela Trenkic and Angela De Bruin).
© 2025 John Wiley & Sons, Ltd. Published 2025 by John Wiley & Sons, Ltd.
DOI: 10.1002/9781405198431.wbeal20357

The second major way in which such measures are important is as predictors or control variables in studies targeting linguistic phenomena. For instance, learner proficiency is likely one of the most important predictors for SLA studies of most linguistic phenomena and the degree to which learners master them, which is why it would ideally be included in most studies.

The domains in which such measurements have been developed and studied include vocabulary, grammar/morphosyntax, and phraseology/lexicogrammar. On the level of *vocabulary*, one of the most thoroughly measured constructs is lexical complexity, a notion that encompasses multiple different dimensions such as lexical density (e.g., the ratio of content words to all words), lexical diversity (e.g., the [lack of] repetitiveness of word use), and lexical sophistication (e.g., the proportion of advanced or specialized vocabulary). One important path-breaking study is Kyle and Crossley (2015), who introduced a tool called TAALES (Tool for the Automatic Analysis of LExical Sophistication), which computes 135 indices related to 1- to 3-gram frequencies, very basic dispersion information (range), and other psycholinguistically relevant indices (including concreteness, familiarity, meaningfulness of [referents of] words, and others). What is more, they also used regression modeling with cross-validation to show that holistic lexical and speaking proficiency scores are around 48% predictable by such lexically specific predictors.

On the level of *grammar/syntax*, syntactic complexity is also a frequent object of study. In one important study, Lu (2010) defined it as encompassing several different complexity measures at different structural levels such as mean lengths of units or mean frequencies of various kinds of units (e.g., coordination, subordination, or phrasal elements) for different units (e.g., sentence, clauses, or T-units [a main clause with its dependent clauses]). Lu's tool, the L2 Syntactic Complexity Analyzer, relied on the Stanford parser and Tregex and is designed for analyzing advanced/college-level learners' writing and returns 14 syntactic complexity indices for learners' writing samples. His evaluation based on written English data from Chinese learners suggested that some of the complexity measures distinguish between different levels of proficiency. Adding to this kind of work is Kyle and Crossley (2017), who developed an additional notion of syntactic sophistication and showed that, in their data, it outperformed traditional syntactic complexity measures by appearing to explain more than twice as much variance of holistic writing quality scores.

Finally, there is the more recently evolving field of studying *phraseology and lexicogrammar*, in part as a result of the growing popularity of Pattern Grammar and especially Construction Grammar in cognitive-/usage-based approaches to SLA (see esp. Ellis et al., 2016). The exploration of phraseological complexity has been promoted in particular by Paquot. Paquot (2019), for example, defined phraseological complexity for relational co-occurrences (such as Adj + N or V + DirObj) on the basis of phraseological diversity and phraseological sophistication. She showed that phraseological sophistication can have more predictive power for CEFR proficiency levels than traditional measures of lexical or syntactic complexity.

Investigation of Individual Linguistic Structures

An observer of the field of corpus linguistics would have to be forgiven to think that there must have been a long and fruitful collaboration between corpus linguistics and SLA, given that the field of Learner Corpus Research (LCR) has existed for around 30 years or so. However, much of the work it has produced has had less to contribute to genuine SLA questions than might have been desirable because (a) prominent LCR models such as the Integrative Contrastive Model or Contrastive Interlanguage Analysis made little connection to the many psycholinguistic mechanisms and statistical methods that have long characterized SLA research, (b) corpus compilation did often not include all the kinds of metadata needed for fine-grained studies, and (c) statistical analyses were often not sophisticated enough for the complexity of the questions studied. Nevertheless, over the last 10 years or so, LCR's evolution has made it much more relevant to SLA research, and much

FIRST PAGE PROOFS

of the current SLA interest in corpus methods might be function of this development. This section discusses a few applications of corpus-linguistic methods to SLA questions.

One particularly fruitful area of application has been that of *alternation research*. One case in point is Wulff and Gries (2015), a study of prenominal adjective order in exploring how well Chinese and German learners of English know to say "interesting little story" as opposed to "little interesting story." They annotated 3,624 instances of Adj$_1$–Adj$_2$–N sequences from native speakers and for a variety of predictors such as the lengths of the adjectives, but also semantic, phonological, and other predictors. Examples included the semantic closeness of the adjectives to the noun they precede (adjectives can be semantically close to a noun they modify if they describe properties inherent to the noun), the adjectives' independence from comparison (e.g., low values for *sophisticated* or *prestigious*, which are often used in comparisons, but high ones for *factual* or *religious*, which are not) and their nominal character (e.g., a low value for *famous* or *ugly*, which have typical adjectives endings), but a high one for *light* or *material*, which could even be used as nouns); examples of the latter included the degree to which the chosen sequence of Adj$_1$–Adj$_2$–N leads to better segment alternation scores (i.e., a better adherence to consonant-vowel [CV] structures typical of English) or rhythmic alternation scores (i.e., a better adherence to stressed–unstressed[-unstressed] structures typical of English). Their statistical analysis consists of a two-step procedure: First, they fitted one regression model to the native-speaker data and, because of its good predictive power, then used that model to predict for every learner case which order a native speaker would have produced in the linguistic situation the learner was in. They then identified where the learners made nativelike choices and where they did not and used that information as a response variable. The results indicate, to give just two examples, that (a) the Chinese learners are significantly less nativelike than the German learners when it comes to their sensitivity to the cue of the frequency difference between the two adjectives but that (b) the Chinese learners are significantly more nativelike than the German learners when it comes to producing adjective orders that result in articulatorily preferred segment alternation patterns (e.g., avoiding the clash of many consonants in the sequence *honest clever*).

A wide range of studies has been concerned with the three important constructs in SLA research of CAF mentioned above. One particularly influential study involving accuracy with a high degree of statistical sophistication is Murakami (2016), who studied the use of several grammatical morphemes in English—including articles, past tense *-ed*, and plural—for 10 different learner L1 groups (operationalized by country of residence) from the EFCAMDAT corpus. Crucially, his study used a generalized linear mixed effects model (see Gries, 2021) and was, therefore, able to explore the longitudinal development that different learners exhibited and, thus, individual variation. Predictors of his model included participant-level variables (e.g., proficiency and random effects), L1-level variables (e.g., whether the L1 had an equivalent morpheme), a variable representing within-learner order of writing samples to capture within-learner longitudinal effects, but also interactions between such variables. A model selection process led to a final model showing, among other things, that there is an overall significant learning effect over time and that higher-proficiency learners are better than lower-proficiency learners. On top of these expectable results, however, there is a morpheme-by-proficiency interaction: accuracy did not increase uniformly across proficiency levels (meaning, as learners increased their proficiency level over time, their accuracy did not become better equally quickly for all grammatical morpheme: plural *s* uses become better only negligibly as a learner enters a new, higher proficiency level, but determiner use becomes significantly better. At the same time, a non-significant morpheme-by-writing sample interaction shows that the rate of longitudinal development—getting better over time in using the morphemes accurately—was comparable across the different morphemes.

Murakami's study is also noteworthy in how it tackled individual variation in more detail than many others have done. He found considerable amounts of *individual variation*—quantified as the overall standard variation of the speaker-specific adjustments—and argued that "while on average learners' longitudinal development is characterized by increased accuracy, for a great proportion of learners, accuracy decreased overall" (p. 18). While Murakami's study was ultimately on accuracy,

FIRST PAGE PROOFS

this interest in individual variation has grown considerably in general. Wulff and Gries (2021) applied predictive modeling methods to the genitive alternation (e.g., *the speech of the President* vs. *the President's speech*) and showed how it can be used to identify the learners whose overall genitive choices are least nativelike (i.e., which learner genitives are most unlike what they get as native-like input) as well as individual genitive choices that are unexpected (i.e., what might be factors that govern learner choices that were are not already aware of from how native speakers choose genitives?). Gries and Wulff (2021) also investigated the ordering of main and adverbial subordinate clauses (main clause before or after subordinate clause) and showed how differently learners (within one L1 and across different L1s) react to cues that have a strong effect on native speakers' ordering of clauses; for example, native speakers exhibit notable short-before-long effects but some learners deviate from that considerably.

A final area of application worth mentioning is that of the increased recognition of the overlap of SLA as it has traditionally been conceived of—focusing on learners that learn a non-native language in instructional, but not immersion settings—and research on *heritage speakers* (see Valdés, 2005, for one of the first discussion of this overlap). Peirce (2018) studied case- and gender-marking errors in (timed and untimed) essays of advanced heritage speakers and learners of Russian as a foreign language. Nouns, adjectives, and determiners in a sample of texts from the Russian Learner Corpus of Academic Writing were error-annotated, leading to a sample of nearly 500 errors, which was analyzed with a repeated-measures ANOVA (with error frequency [per T-unit] as the response variable). Among other things, her results show that, somewhat surprisingly, there was no significant difference between timed and untimed essays for both heritage speakers and learners but the latter exhibited much more variability in their error rates; at the same time, gender errors were much rarer than case errors, again for both speaker groups. Interestingly, in this smaller corpus study, heritage speakers performed better than L2 learners, a contrast to at least some previous findings. However, the corpus sample studied was rather small, as indicated by Peirce herself, so additional research with larger subject groups is necessary to ascertain whether these results are reliable.

Evaluation and Desiderata

Corpora have much to offer to SLA research: They provide essential frequency data, especially valuable in cognitively and psycholinguistically informed studies. Additionally, corpora offer methodological advantages, mitigating issues that commonly arise in experimental designs, such as low ecological validity and input misrepresentation. Regarding the former, by nature, experimental studies are conducted under highly controlled conditions based on carefully elicited or selected data whereas corpora, with their naturalistic and highly contextualized data, often exhibit higher ecological validity. Unlike experimental studies where participants may read one sentence at a time, corpora are naturally produced in more authentic situations. Regarding input distribution, controlled experimental designs can result in unrepresentative distributions of linguistic elements, so that even a small number of experimental stimuli can lead to observable learning effects. Moreover, unrepresentative input distributions throughout an experiment can impact the priming of structures and be influenced by factors like fatigue or habituation, distorting the results of experiments, especially in the case of learners where the pre-existing knowledge is not yet robust. Corpora studies help circumvent these issues, although they come with their own set of challenges (see below). A final advantage is that corpora allow scholars to complement experimental L2 psycholinguistic research to, hopefully, lead to converging or at least complementary evidence from a different methodological paradigm. For example, the study of how priming effects differ across different prime-target distances in a setting unaffected by unrepresentative experimental input can benefit much from corpus studies. Corpora are also generally well suited for exploratory/hypothesis-generating studies (see the suggested readings for more details).

FIRST PAGE PROOFS

FIRST PAGE PROOFS

As mentioned, despite, or in fact because of, their advantages, corpora also come with challenges. On the one hand, it has to be acknowledged that learner corpora are probably not the most naturalistic and ecologically valid corpora. They come with many non-natural constraints, especially when composed largely of timed argumentative-essay data on the usual suspects of “controversial topics.” Even to the extent that such data are ecologically valid, ecological validity also implies a degree of noise and heterogeneity that goes far beyond that of carefully controlled experiments. A related challenge of corpus data is that their distributional characteristics—often unbalanced/Zipfian with frequencies of words that decrease as a power function of their rank in the frequency table (Ellis et al., 2016) —can make the required statistical analyses quite complex. Thus, it is probably fair to say that the choice between experimental and observational corpus data, while ultimately of course affected most strongly by a study’s objectives, also often involves the question of when a researcher prefers to exercise control: before the data collection by virtue of a priori experimental design or after the data collection by virtue of complex statistical control of noise variables and confounds. Finally, if background information about speakers (L1, cognitive scores quantifying some form or aspects of intelligence, aptitude, or working memory, or motivation-related variables, etc.) is lacking, as is often the case, this makes it hard to fully account for how cognitive/sociodemographic factors affect learners’ acquisition of the L2 [see Paquot et al. (2023) for a recent survey regarding the kinds of metadata researchers want/need].

As a result of such considerations, there are many desiderata for the field. On the resource side, we need more and bigger learner corpora (and maybe especially more longitudinal corpora to better track development). We need more spoken corpora (to have to rely less on written essays), and all of this we need a larger variety of target languages, as well as more diverse representations of L1 learner backgrounds. Ideally, we would have corpora that contain both L1 and L2 production data for all speakers represented in them, as in ZAEBUC (see Palfreyman & Habash, 2021). Just as importantly, we need those corpora to be widely and *fully* available: there is a lamentable trend for some corpus compilers not to make corpora accessible, or accessible only via web interfaces (e.g., the Trinity Lancaster Corpus of L2 data or new British National Corpus of L1 data) which creates huge obstacles to innovative research. This is a paradoxical development at a time when open access and open science are otherwise booming. We also need such corpora to contain much more systematically collected metadata providing rich learner information (see examples mentioned above) and task information (to control analyses for things like task complexity, communicative stress, aspects of the prompts, etc.). This is especially important for heritage language corpora to be able to capture, for instance, in which social contexts heritage speakers use which language (see Kisselev, 2021); see Paquot et al. (2021) for a study that includes a reanalysis of an earlier study, which led to a considerable re-evaluation of earlier results.

On the methods side, there is a pressing need for enhanced statistical and computational expertise. Concerning the former and to do justice to the complexity and noisiness of observational data, we need multifactorial methods that can accommodate subject and item variation on multiple levels, along with multivariate exploratory tools to capture non-linear trends in data. However, we also need better-executed statistics. To mention just one example, many widely cited studies benchmarking measurement tools against proficiency levels do not adequately model proficiency as an ordinal response variable. Regarding the latter, the field would benefit from more programming expertise. While tools such as TAALES (Kyle & Crossley, 2015) or L2SCA (Lu, 2010) are valuable, it is important to recognize that each ready-made tool has its limitations. Researchers should have the flexibility to use a wider variety of tools or develop their own to avoid getting locked into what specific tools offer. By addressing these needs, corpus-linguistics, itself a rapidly evolving field, is poised to gain an even greater relevance in various aspects of SLA research.

wbeal0033.pub2
wbeal0258.pub3

SEE ALSO: Corpus Linguistics: Overview; Corpus Linguistics: Quantitative Methods

References

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

Gries, S. T. (2021). (Generalized linear) Mixed-effects modeling: A learner corpus example. *Language Learning*, *71*(3), 757–798.

Gries, S. T., & Wulff, S. (2021). Examining individual variation in learner production data: A few programmatic pointers for corpus-based analyses using the example of adverbial clause ordering. *Applied Psycholinguistics*, *42*(2), 279–299.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Koiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–20). John Benjamins.

Kisselev, O. (2021). Corpus-based methodologies in the study of heritage languages. In S. Motrul & M. Polinsky (Eds.), *The Cambridge handbook of heritage languages and linguistics* (pp. 520–544). Cambridge University Press.

Kyle, K., & Crossley, S. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786.

Kyle, K., & Crossley, S. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing*, *34*(4), 513–535.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496.

Murakami, A. (2016). Modeling systematicity and individuality in nonlinear second language development: The case of English grammatical morphemes. *Language Learning*, *66*(4), 834–871.

Murakami, A., & Ellis, N. C. (2022). Effects of availability, contingency, and formulaicity on the accuracy of English grammatical morphemes in second language writing. *Language Learning*, *72*, 899–940.

Palfreyman, D., & Habash, N. (2021). *Zayed Arabic-English bilingual undergraduate corpus (ZAEBUC)*. Zayed University/New York University. http://www.zaebuc.org (accessed September 8, 2023)

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145.

Paquot, M., König, A., Stemle, E., & Frey, J.-C. (2023). *Core metadata schema for learner corpora*. https://doi.org/10.14428/DVN/4CDX3P. Open Data @ UCLouvain.

Paquot, M., Naets, H., & Gries, S. T. (2021). Using syntactic co-occurrences to trace phraseological development in learner writing: Verb + object structures in LONGDALE. In B. Le Bruyn & M. Paquot (Eds.), *Learner corpora and second language acquisition research* (pp. 122–147). Cambridge University Press.

Peirce, G. (2018). Representational and processing constraints on the acquisition of case and gender by heritage and l2 learners of Russian: A corpus study. *Heritage Language Journal*, *15*(1), 95–111.

Valdés, G. (2005). Bilingualism, heritage language learners, and SLA research: Opportunities lost or seized? *The Modern Language Journal*, *89*(3), 410–426.

Wulff, S., & Gries, S. T. (2015). Prenominal adjective order preferences in Chinese and German L2 English: A multifactorial corpus study. *Linguistic Approaches to Bilingualism*, *5*(1), 122–150.

Wulff, S., & Gries, S. T. (2021). Explaining individual variation in learner corpus research: Some methodological suggestions. In B. Le Bruyn & M. Paquot (Eds.), *Learner corpora and second language acquisition research* (pp. 191–213). Cambridge University Press.

Suggested Readings

Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English second language speakers. *Language Learning*, *60*(3), 573–605.

Deshors, S. C., & Gries, S. T. (2023). Using corpora in research on second language psycholinguistics. In A. Godfroid & H. Hopp (Eds.), *The Routledge handbook of second language acquisition and psycholinguistics* (pp. 164–177). Routledge.

FIRST PAGE PROOFS

Gries, S. T. (2019). Priming of syntactic alternations by learners of English: An analysis of sentence-completion and collostructional results. In J. A. Egbert & P. Baker (Eds.), *Using corpus methods to triangulate linguistic analysis* (pp. 219–238). Routledge.

Q4 Gries, S. T. (to appear). Against level-3-only analyses in corpus linguistics. *ICAME Journal*.

Le Bruyn, B., & Paquot, M. (Eds.). *Learner corpora and second language acquisition research*. Cambridge University Press.

Lu, X. (2023). *Corpus linguistics and second language acquisition: Perspectives, issues, and findings*. Routledge.

Tracy-Ventura, N., & Paquot, M. (Eds.). (2021). *The Routledge handbook of second language acquisition and corpora*. Routledge.

Wulff, S., & Gries, S. T. (2019). Particle placement in learner English: Measuring effects of context, first language, and individual variation. *Language Learning*, *69*(4), 873–910. Q5

FIRST PAGE PROOFS

The abstract and keywords will not be included in the PDF or any printed version of your article, but are necessary for publication on Wiley's online publishing platform to increase the discoverability of your article.
If the abstract and keywords are not present below, please take this opportunity to add them now.
The abstract should be a short paragraph up to 200 words in length and keywords between 5 to 10 words.

Abstract: This article discusses the use of corpus-linguistic methods in second-language acquisition research. It focuses on measurement applications, specific linguistic case studies, and an evaluation coupled with an outlook over desiderata for the future.

Q6 **Keywords:** corpora; second-language acquisition; learner corpus research; quantitative methods

FIRST PAGE PROOFS