

Quantitative research methods and study quality in learner corpus research

Magali Paquot & Luke Plonsky

FNRS – Université catholique de Louvain / Georgetown University

Abstract

This study aims to provide the first empirical assessment of quantitative research methods and study quality in learner corpus research. We systematically review quantitative primary studies referenced in the Learner Corpus Bibliography (LCB), a representative bibliography of learner corpus research maintained by the Learner Corpus Association which contained 1,276 references when the project started. Each primary study in the LCB was coded for over 50 categories representing six dimensions: (a) publication type (i.e. conference paper, book chapter, journal article), (b) research focus (e.g. lexis, grammar), (c) methodological features (e.g. keyword analysis, error analysis, use of reference corpus), (d) statistical analyses (e.g. X^2 , t-test, regression analysis), and (e) reporting practices (e.g. reliability coefficients, means). Results point to several systematic strengths as well as many flaws, such as the absence of research questions, incomplete and inconsistent reporting practices (e.g. means without standard deviations), and lack of statistical literacy (i.e. LCR studies generally overrely on tests of statistical significance, do not report effect sizes, rarely check or report whether statistical assumptions have been met, rarely use multivariate analyses). Improvements over time are however clearly noted and there are signs that, like other related disciplines, learner corpus research is slowly undergoing a methodological reform.

Keywords

Learner corpus research, study quality, quantitative research methods, reporting practices, research synthesis

1. Introduction

The origins of Learner Corpus Research (LCR) are usually dated to the late 1980s but the field did not really take root until the 1990s. As a truly interdisciplinary enterprise, LCR sits at the crossroads between a variety of disciplines including, principally, corpus linguistics and second language acquisition (Granger 2009).

Methodologically, LCR shares central tenets of corpus linguistics: (1) LCR is empirical, analyzing learners' written or spoken production; (2) LCR is based on the analysis of preferably large and principled collections of continuous and contextualized learner language texts sampled to be representative of a particular learner population; (3) LCR makes extensive use of computer software tools and programs for manipulating machine-readable learner data (e.g. searching, sorting and enriching learner texts with various metadata and linguistic analyses) and (4) LCR often combines various types of quantitative and qualitative analytical procedures (Biber & Reppen 2015a: 1).

In terms of theory, a large proportion of LCR studies share with second language acquisition (SLA) research the goal of gaining a better understanding of the human capacity to learn languages other than the first (Ortega, 2009: 2).

Since its inception, the scope of LCR, its analytical underpinning and contribution to scientific knowledge and professional practices have been discussed regularly in the scholarly literature (e.g. Dagneaux et al. 1998; Gilquin et al. 2007; Granger 1996; Granger 2004; Granger 2009; Granger 2015; Meunier 2010; Myles 2008; Paquot & Granger 2012; Pendar & Chapelle 2008). Theoretical, methodological and empirical advances have also recently been summarized and discussed in the *Cambridge Handbook of Learner Corpus Research*

(Granger et al 2015). With the exception of Stefan Gries's work (e.g. Gries & Wulff 2013; Gries 2015), however, relatively little attention has been paid so far to the state or development of LCR's quantitative research methods. In other words, it remains an empirical question whether and to what extent quantitative studies in LCR have been carried out in adherence to standards of methodological quality, an unfortunate circumstance given that "progress in any of the (...) sciences depends on sound research methods, principled data analysis, and transparent reporting practices" (Plonsky & Gass 2011: 325).

The main goals of this article are therefore to systematically describe and evaluate quantitative research and reporting practices in LCR both cumulatively and over time. Our motivation for carrying out this study is simple: by looking to the past of LCR, we hope to inform future LCR research practices and contribute to their improvement. In doing so, we build on current methodological developments in corpus linguistics and the growing body of methodological syntheses in L2 research that examine study quality across and within substantive domains (e.g., Liu & Brown 2015; Plonsky 2013).

2. Background

2.1. Quantitative designs and statistical techniques in corpus linguistics

Over the past five decades, corpus linguistics has developed into a major methodological paradigm in applied and theoretical linguistics (e.g. Biber & Reppen 2015b). This vibrant and quickly evolving field has witnessed and is still witnessing controversial but necessary discussions about the role of statistics and overall methodological sophistication in its scientific enterprise (see, e.g. Gries 2014). Many corpus linguists today adhere to the view that corpus linguistics is an "inherently distributional discipline" (Gries 2015b: 50): corpora exclusively contain (a) information on (relative) frequency of occurrence, (b) information on (relative) frequency of co-occurrence, and (c) information on dispersion. From this

perspective, it directly follows that the analysis of frequencies and distributions should best be conducted with an advanced (and cutting-edge) statistical toolbox (e.g. Biber 1988; Gries 2006a; Gries 2010; Gries 2014; Gries 2015c; Kilgariff 2005; Paquot & Bestgen 2009).

In spite of largely promising developments, a variety of methodological shortcomings has repeatedly been observed in the literature. Among the most serious shortcomings, we highlight the following issues:

- (a) Corpus linguists often report results for complete (sub-)corpora and rarely inspect by-speaker or by-text results (Brezina & Meyerhoff 2014; Gries 2006a).
- (b) Corpus linguists rarely explore homogeneity and/or variability within and between corpora (Kilgariff 2001; Gries 2006b; Gries 2014).
- (c) Many studies investigating the use of particular words/structures narrowly focus on the frequencies of these features as the main diagnostic of importance (Gries 2006a; Gries 2014).
- (d) Corpus linguists rarely provide information concerning dispersion as a supplement to frequency data (e.g. Baayen 2001; Gries 2014).
- (e) Corpus linguists often fail to report whether the assumptions of statistical tests have been checked and met (Baroni & Evert 2008; Köhler 2013; Gries 2015b).
- (f) A majority of studies ignore the repeated-measurement nature as well as the hierarchical structure of corpus data; as a result, they violate a fundamental assumption of most of the statistical methods used (e.g. chi-squared tests, correlations, generalized linear models), namely that the data points are independent of each other (Gries 2015c: 101).
- (g) Corpus linguists often apply multiple statistical tests on the same data without correcting the significance level (Gries 2015b).

- (h) Corpus linguists often mistakenly choose univariate and bivariate statistics, which fail to address the multifactorial nature of the phenomena of interest (Gries 2015b).
- (i) Many corpus linguists are still relying on a small set of off-the-shelf corpus tools and their research agenda is limited by what those tools can do (Gries 2010; Gries 2015c: 93).

While this selected list of methodological shortcomings paints a bleak picture of the field, Gries (2015b) has argued that corpus linguistics currently exhibits an overall larger degree of statistical expertise and methodological appropriateness than a decade ago (Gries, 2015b). Could this statement also be made about LCR today? Apparently not -- Gries (2015a) reports similar problems as described above with how statistical analyses are conducted and reported in LCR (see also Gries 2013a; Gries & Wulff 2013) and comments:

The somewhat obvious fact that corpus methods are by definition distributional and quantitative has not yet led to an analogous increase in the statistical sophistication of LCR – neighbouring fields such as sociolinguistics, psycholinguistics or corpus linguistics in general exhibit an overall larger degree of sophistication (Gries 2015a: 173).

However, methodological discussions in corpus linguistics have been largely anecdotal and have never been driven by a systematic and principled review of the relevant body of research. To put LCR methodological developments and statistical sophistication to the empirical test, we turn to systematic research synthesis, which stems from within and beyond applied linguistics and which provides a well-developed set of procedures for examining methodological quality.

2.2. Assessing study quality in SLA

Scientific enquiry on study quality, i.e. “the combination of (a) adherence to standards of contextually appropriate methodological rigor in research practices and (b) transparent and complete reporting of such practices” (Plonsky 2013: 657), originated in the now-thriving domain of meta-research (for a recent introduction, see Ioannidis et al. 2015). From the meta-analyst’s perspective, there are three primary motivations for assessing study quality. First, many meta-analyses weight effect sizes according to the methodological quality of the studies they were extracted from. In order to do so, numerous measures have been proposed and employed to assess the quality of quantitative empirical research (Wells & Littell 2009). Second, synthetic measures of study quality also enable meta-analysts to examine possible relationships between methodological features in primary studies and the outcomes they produce. Underlying this type of analysis is the assumption that “study results are determined conjointly by the nature of the substantive phenomenon under investigation and the nature of the methods used to study it” (Lipsey 2009, p. 150). A third impetus for examining study quality at the meta-analytic level relates to the prospective function of synthetic research. That is, by systematically examining methodological practices in a given domain, the synthesist is then able to make empirically-grounded recommendations for improving future research.

While it has quickly gained recognition as an interdisciplinary domain in its own right in other social and educational sciences, study quality has only recently been taken up as an empirical matter in SLA and applied linguistics more generally (e.g. Derrick 2016; Liu & Brown 2015; Plonsky & Gass 2011; Plonsky 2013; Plonsky 2014; Plonsky & Kim 2016; Ziegler 2016). Nevertheless, this body of research has already generated extremely useful findings. Plonsky (2013), for example, described and evaluated designs, statistical analyses, reporting practices, and outcomes (i.e. effect sizes) in a sample of 606 primary studies,

published from 1990 to 2010 in *Language Learning* and *Studies in Second Language Acquisition*. One of the most striking results is likely the persistent use of small or underpowered samples in SLA, with a median group or sample size of 19 participants. Other problematic trends observed in experimental design include relatively few randomly controlled trials and delayed posttests.

Results from Plonsky (2013) also pointed to the relatively narrow repertoire of statistical procedures used in the domain, a finding observed in individual subdomains of SLA as well, such as interaction research (Plonsky & Gass 2011) and task-based language teaching (Plonsky & Kim 2016). Overall, the most frequently used statistics are those that test for differences between group means, e.g. ANOVAs and *t*-tests. Correlation-type statistics (including multiple regression) were also found with some regularity but other multivariate statistics are scarce (see Plonsky & Oswald in press). Plonsky (2013: 668) noted that the intensive and extensive use of statistical testing (and more particularly null hypothesis testing, see Norris 2015; Plonsky 2015) taken together with the small samples typical of SLA should give us pause for their combined, debilitating effects on statistical power and therefore internal validity of study outcomes in the domain.

A number of previous methodological syntheses have also found cause for concern in reporting practices, both across and within individual domains of L2 research. Perhaps most concerning is the omission of basic descriptive statistics. Reporting such data is necessary because it both enables consumers' interpretations of primary findings and it avails primary data to would-be meta-analysts who require such data to calculate effect sizes. Nevertheless, Plonsky (2013) found that in thirty-one percent of the primary studies in his sample one or more means are reported without a standard deviation. Likewise, *t*-tests and ANOVAs are frequently conducted without reporting the means being compared (20%) or the *t* and *F* values resulting from those tests (24%). Though encouraging when compared with results

from previous meta-analyses, reporting practices related to reliability are also in need of improvement, with one or more estimates of reliability (i.e. inter-rater reliability, instrument reliability) reported in only 45% of the sample (see Larson-Hall & Plonsky 2015; Liu & Brown 2015; Plonsky & Derrick 2016). Other reporting practices that indicate quality and transparency were found very rarely, including confidence intervals (5%), evidence of having checked statistical assumptions (17%), a priori level of statistical significance (22%) and effect sizes (26%).

Plonsky (2014) draws on the same dataset as Plonsky (2013) to examine changes over time in research and reporting practices. A comparison of articles published across the 1990s vs. the first decade of the 2000s indicates a number of methodological improvements taking place in SLA, including an increase in the number of manuscripts that state research questions (from 70% to 87%), increases in sample sizes (from a median group or sample size of 14 to 20), and increasing availability of critical data such as effect sizes (from 3% to 42%), standard deviations to accompany means (from 59% to 76%), and reliability estimates (from 38% to 50%). Substantial improvements are still needed, however, in many areas (e.g. reporting confidence intervals, power analysis, reporting *p* values consistently). With respect to statistical procedures, the range of analyses has not changed from the 1990s to the 2000s and the field continues its unfortunate reliance on null hypothesis significance testing. In addition, the number of statistical tests has increased, thus mitigating any possible increase in statistical power due to larger samples.

While there is still much room for improvement, there are reasons to be optimistic. Results from the body of methodological syntheses described here are testimony of the growing awareness of the importance of methodological issues in SLA and applied linguistics more generally. Methods and research practices are now the focus of much scientific enquiry in the field. Evidence of this ‘methodological turn’ (Byrnes 2013:825) continues to

accumulate. In addition to an increasing number of methodological syntheses, the field has also recently witnessed an increased application of several novel statistical techniques (e.g. Larson-Hall & Herrington 2010; Plonsky et al. 2015; Ross & Mackey 2015) and appreciation of replication research (e.g. Porte 2012). Furthermore, several leading journals, including *Language Learning* and *TESOL Quarterly*, have also put out new or amended journal guidelines in the last few years to better reflect current methodological standards (e.g. Mahboob et al. 2016; Norris et al. 2015). At the level of professional associations, as of 2016, the conference of the American Association for Applied Linguistics includes a strand for proposals related to Research Methods.

2.3. The Present Study

Research practices in corpus linguistics and SLA, i.e. two fields with which LCR has much in common, are being challenged by critical but healthy methodological discussions. Some criticisms have also recently been voiced against the lack of appropriate quantitative designs and sophisticated statistical testing in LCR but so far no study has extensively reviewed features of study quality in the field. After over 25 years of existence, it is also time to provide LCR with a better understanding of its substantive interests. The study is therefore driven by the following three research questions that are answered both cumulatively and over time:

1. RQ1 – To what extent and by which means have different types of learners, learning contexts and features of learner production been analyzed in LCR?
2. RQ2 – To what extent has LCR employed various methodological procedures and statistical analyses?
3. RQ3 – To what extent have data been reported thoroughly in LCR?

3. Method

The procedures used by this study to identify, extract data from, and analyze a body of LCR studies adhered closely to the growing body of methodological syntheses in applied linguistics and to current best practices in synthetic and meta-analytic techniques (see Plonsky & Oswald 2015).

3.1. Study Identification

A first step toward addressing our research questions was to identify a sample of primary LCR studies that would represent—if not comprise in its entirety—this population of research. For this purpose we turned to the Learner Corpus Bibliography (LCB; <https://www.uclouvain.be/en-cecl-lcbiblio.html>), which contains a comprehensive and current listing of references for a wide variety of empirical, theoretical, and practice-oriented publications in this area. The LCB was initially created by the *Centre for English Corpus Linguistics* (director: Prof. S. Granger) and is now maintained in the form of a Zotero collection by the Learner Corpus Association (LCA), i.e. an international association which aims to promote the field of LCR and provide an interdisciplinary forum for all the researchers and professionals who are actively involved in the field or simply want to know more about it. When the project started (February 2015), the LCB contained 1,276 references. The process of collecting primary studies from the LCB coincided with the coding process, both of which are described in detail in the following section.

3.2. Data Collection and Coding

A coding scheme was developed to enable us to systematically code each study in our sample for a range of substantive and methodological features. As a starting point in developing this

instrument, we examined the structure and individual items from previous methodological reviews (e.g., Plonsky 2013). We then considered the types of methodological practices and substantive issues addressed in LCR and how these might be fruitfully and appropriately examined at the secondary or synthetic level. Domain specific items were added that would be relevant to as large a number of studies as possible such as mode (spoken, written), size of corpora, annotation tools employed, and so forth. In line with the dual purpose of the study to both describe and evaluate research involving learner corpora, some items were associated with higher study quality (e.g., inclusion of precision and recall estimates); others were included simply to better understand practices and interests in the domain (e.g., target language; statistical analyses). In developing this instrument we were also sensitive to practicalities such as the potential for rater fatigue as well as funding for and availability of research assistants.

Once a draft was complete, both authors piloted the instrument by independently coding two studies. The results were then compared and the potential implementation was discussed leading to the revision of a number of item values and definitions. The final version of the instrument, which is shown in Table 1 and which will be available in its original format (an Excel file) upon publication on the IRIS Database (see Marsden et al. 2016), consisted of 66 items across the following six categories: (a) Study Identification, (b) Participants, Setting, and Study Design, (c) Corpus Design and Features, (d) Corpus Analytic Techniques; (e) Statistical Analyses , and (f) Data Reporting and Transparency.

Response values for most items were coded categorically and oftentimes dichotomously based on the absence (0) vs. presence (1) of a particular feature. However, a small number of items were coded as continuous or ‘open’ items, such as target language and size (in words) of the corpora being analyzed.

Table 1

Coding scheme categories and items

Category	Items
Study Identification	author(s)*; year*; journal; publication type
Participants, Setting, and Study Design	proficiency level; age; first language(s)*; target language(s)*; setting: second vs. foreign; educational context (elementary, high school, university, language institute, other); sample size*; replication study
Corpus Design and Features	cross-sectional design; longitudinal design; mode (written, spoken, multimodal); interaction, if any (learner-learner; learner-teacher); number of different L1s represented*; number of learner corpora used*; corpus size in words*; corpus size in number of texts*; use of reference corpus
Corpus Analytic Techniques	concordance lines; word lists; keywords; lexical bundles; statistical collocations; annotations and tagging (error, part of speech tagging, parsed); software*
Statistical Analyses	correlation; chi-square; <i>t</i> -test; log-likelihood; ANOVA; ANCOVA; MANOVA; MANCOVA; factor analysis; multiple regression; logistic regression; cluster analysis; mixed effects modeling; resampling/bootstrapping; corrections for multiple comparisons (e.g., Bonferroni); number of unique statistical tests*; total number of tests of statistical significance*
Data Reporting and Transparency	research questions; reliability reported; precision; recall; instrument/coding book available; predetermined alpha level; statistical assumptions checked; missing test statistics; relative/normed frequency; log likelihood/Chi-square statistics; exact <i>p</i> value; relative <i>p</i> value;

statistical power; confidence intervals; effect sizes (d , R^2); data
visualization

Note. All items coded categorically (e.g., 0 = absent, 1 = present) unless marked with * to indicate items with an open-ended response.

Primary studies were coded by three research assistants, who were provided with extensive training and supervision by the two authors throughout the data collection process. Prior to coding, references in the LCB were sorted by year of publication. In order to ensure that the sample of coded studies would be representative of—and would enable an analysis of—changes in the domain over time, the RAs were instructed to code every fifth empirical study, a process that was then repeated each time they reached the end of the list. Theoretical/position papers (e.g. Granger, 2009), methodological discussions (e.g. Pendar & Chapelle 2008), non-empirical pedagogical discussions (e.g. Meunier 2010), and literature reviews (e.g. Paquot & Granger 2012) in the LCB were excluded. Because of our interest in examining largely quantitative research practices, studies based on only qualitative data and analyses were omitted from the sample.

We had initially intended to process at least 50% of the references in the LCB within the framework of the research project. With the help of our research assistants, however, we managed to analyze 1,027 (80.5%) out of the 1,276 references found in the online bibliography (February 2015). A large proportion of references were excluded either because they were not quantitative studies (e.g. chapters introducing the field or a new learner corpus; 516; 50.24%) or were mistakenly recorded in the LCB despite not being based on learner corpus data (133; 12.95%).¹ The set of procedures described here yielded a total sample of 378 quantitative primary studies. We are aware that our sampling procedures did not provide

¹ Our research project served the field in yet another way. In collaboration with the Learner Corpus Association, we cleaned-up the online bibliography on the basis of our analyses and enriched it with pdf scans of quite a few chapters (especially the earliest publications).

us with the entire population of research in this domain; there are certainly additional studies that fall within our criteria that are not present in the sample. Nevertheless, we are reasonably confident that our sample is representative, though not exhaustive, of the population of interest. We acknowledge that it is possible that our sampling process may bias the sample toward studies that are more visible and therefore likely of higher methodological quality than what might exist in the population of LCR, which would include unpublished studies and articles published in less recognized journals. Otherwise, however, we have no reason to suspect any systematic differences between our sample and the more general population of LCR. We would also point out that, relative to other domain-specific syntheses and methodological reviews, the sample for this study is quite large. As three points of comparison in the realm of domain-specific methodological syntheses, Plonsky & Gass (2011), Plonsky & Kim (2016), and Liu & Brown (2015) respectively reviewed 174, 85, and 44 studies of interaction, task-based learner production, and written corrective feedback.

A subset of 20 studies in the sample was coded by two raters in order to estimate the interrater reliability for our coding process. Overall percentage agreement was considered adequate at 88% ($\kappa = 0.76$). However, a closer look at item-level reliability based on double-coding revealed that four individual items had been systematically misinterpreted and miscoded: (a) participant age (agreement 20%); (b) inclusion of stated research questions (60%); (c) learner context (second vs. foreign language; 65%); and (d) number of statistical tests (50%). An additional research assistant with extensive experience in synthetic research then re-coded these four items in the entire sample, leading to what we consider to be a highly reliable and trustworthy final dataset.

3.3. Analysis

The analyses we conducted followed those of previous methodological syntheses and were quite straightforward. In order to answer our research questions, frequencies and percentages of all categorical response values were calculated both overall and over time, i.e. divided into three almost identical periods of eight, eight, and nine years, respectively starting with the earliest year of publication in the sample: 1991. Descriptive statistics were calculated for the small number of items that yielded continuous data (e.g., number of statistical tests). In these cases, because the data were not normally distributed, medians and interquartile ranges are presented.

4. Results

The sample for this study consisted of 378 studies, spanning a variety of outlets, years, settings, and first and target languages. As an area of empirical enquiry, LCR has also increased dramatically in recent years (Table 2). In terms of publication outlets, this body of research has been published predominantly and roughly equally in the form of journal articles and book chapters. Interestingly, however, the preferred format for publishing in this domain appears to be changing (see Figure 1). Whereas in the 1990s, most studies were published as chapters in edited volumes, LCR is now most often found in the form of journal articles. LCR studies have been published in more than sixty different journals, particularly the *International Journal of Corpus Linguistics* (k=16), *Journal of Second Language Writing* (11), *The Modern Language Journal* (8), *English for Specific Purposes* (7), *Journal of Pragmatics* (6) and *Language Learning* (6).

Table 2

Studies across time periods and outlets

	<i>k</i>	%

Time Periods		
1991-1999	33	8.7
2000-2007	110	29.1
2008-2015	235	62.2
Outlets		
Journal article	160	42.3
Book chapter	146	38.6
Conference proceeding	50	13.2
Doctoral dissertation/thesis	13	3.4
Monograph/book	9	2.4

Note. *k* = number of studies

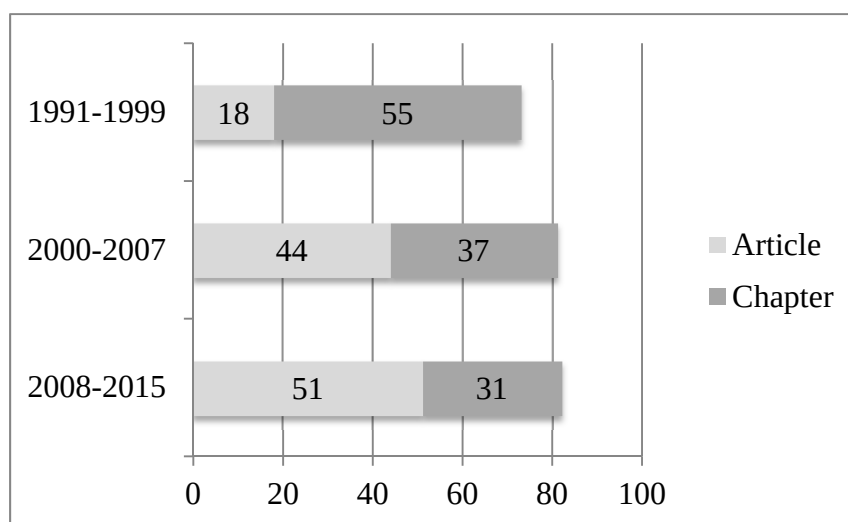


Figure 1. Change over time in primary LCR publication outlets

4.1. Learner demographics, learning contexts and study designs

One of the motivations behind research question 1 was to describe the learner demographics and sampling involved in LCR. Toward this end, we coded primary studies and calculated the frequency and percentage for a number of such variables, the results of which are presented in Figure 2. Approximately two-thirds of the L2 production data analyzed in this body of research stems from what researchers have labeled as ‘advanced’ learners. This result should perhaps be interpreted with caution. Only a small number of researchers ($k = 72$; 19%) derived this label from any kind of language test; others who described the proficiency of L2 users did so according to semesters/years of study, a general impression, or other techniques lacking an empirical foundation (see Thomas 2006).

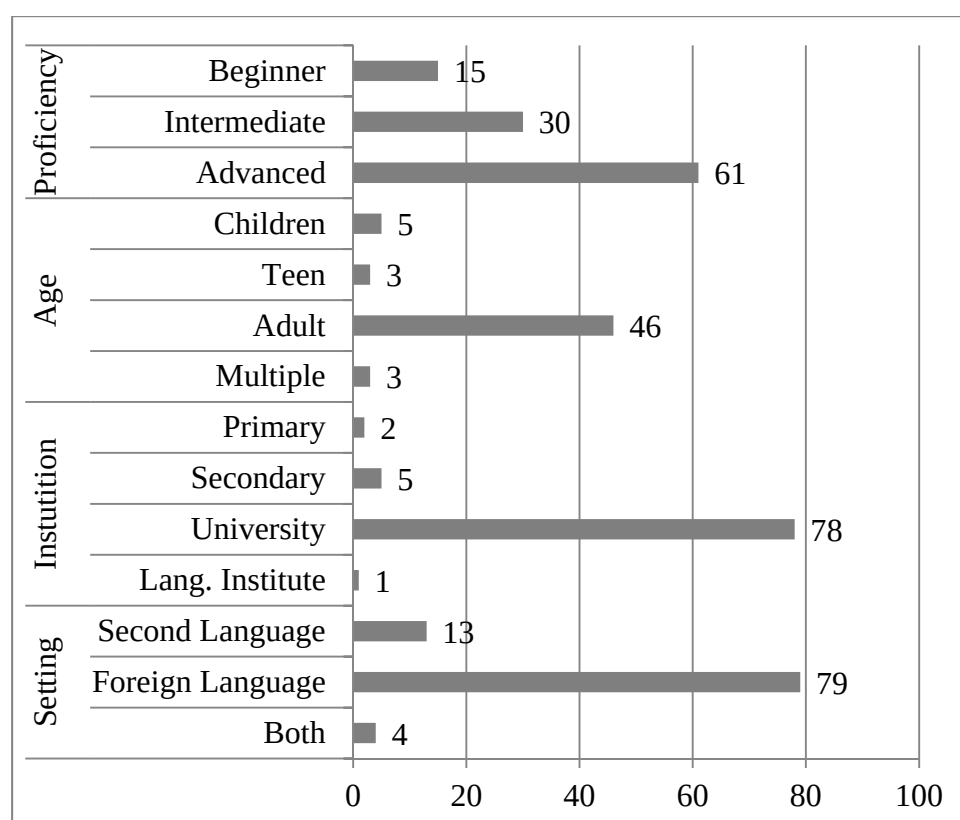


Figure 2. Learner characteristics expressed as a percentage of the total sample

Note. These categories were not mutually exclusive. Cross-sectional studies, for example, might examine corpora based on learners at different proficiency levels. At the same time, a small number of studies did not fit

into these categories; for these reasons, and due to incomplete, missing, or ambiguous reporting, some percentages here and throughout the Results do not add up to exactly 100.

Three additional learner and contextual characteristics were coded for: age, educational institution, and setting (second vs. foreign language). Although a large number of studies did not provide data for these variables, particularly on participant age, the results here reflect a very strong tendency in LCR to study learner language produced by adults studying at universities. The vast majority of corpora were also based on learners in a foreign language environment.

In an attempt to further understand the research practices and data analyzed in LCR, we coded primary studies for a number of features related to the nature and size of corpora under investigation. Approximately three-quarters of the studies in our sample analyzed written learner language. Spoken learner language (largely in the form of transcriptions) was the focus in less than a third, and a small number of studies (3%) analyzed both. However, interest in spoken language has increased substantially from the 1990s onward. Among studies based on oral corpora, a small number has analyzed language produced during interaction among learners only ($k = 19$; 5%), and a somewhat larger portion has analyzed corpora produced by learners speaking with language teachers or expert users ($k = 70$; 19%). Studies of oral interaction that include both types comprise 2% ($k = 7$) of the sample.

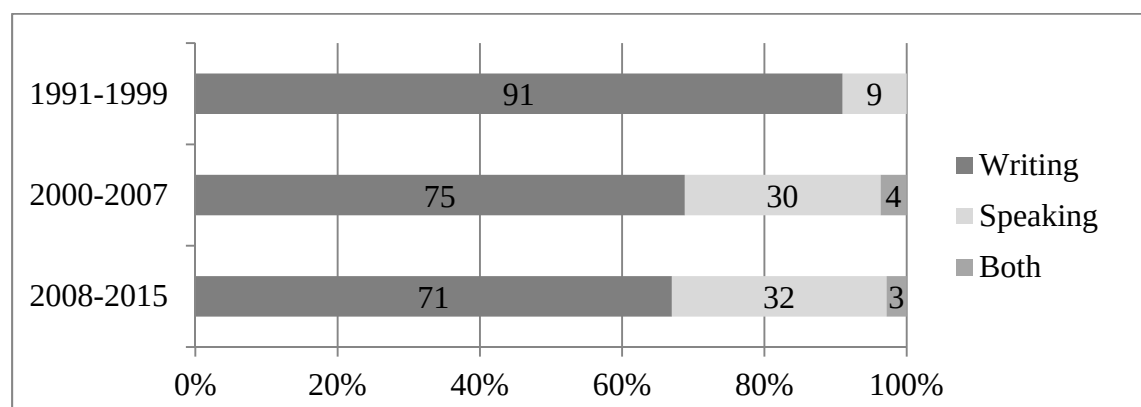


Figure 3. Mode of corpora analyzed in LCR across time

Note. Overall percentages across categories: writing = 74%; speaking = 29%, both = 3%

Research involving learner corpora has taken a variety of designs. As shown in Figure 4, 13% of our sample has examined change over time using a longitudinal design. Another, more frequent and more convenient approach to assessing development is to collect data across proficiency levels using a cross-sectional design (25%). Both of these design types have increased steadily over time, a change which likely reflects the field's interest in L2 development and/or an emphasis on designing longitudinal corpora. The majority of the studies, however, do not use proficiency or time as an independent variable.

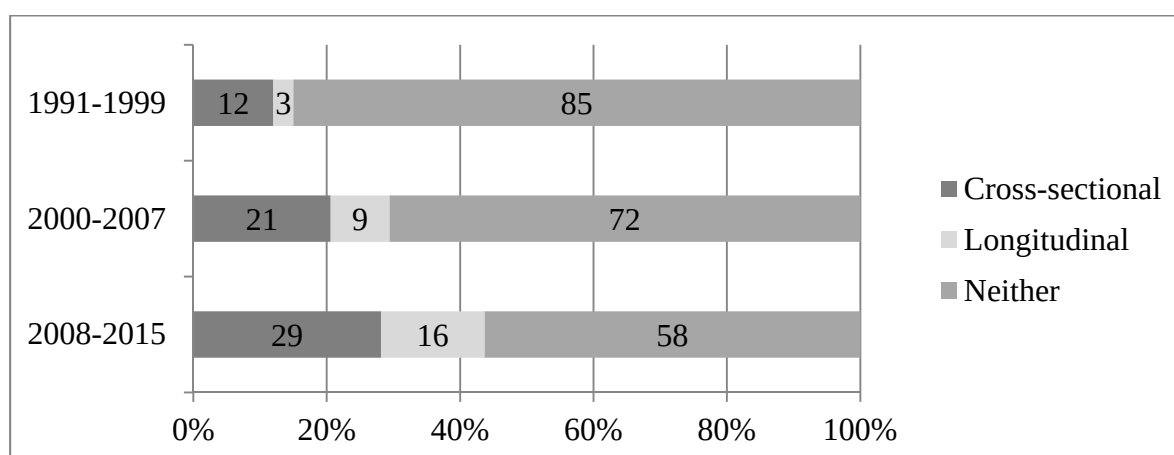


Figure 4. Major designs in LCR across time

Note. The overall percentage across categories: cross-sectional = 25% ($k=94$); longitudinal = 13% ($k = 49$); neither = 65% ($k = 244$)

As we might expect, and as depicted quite clearly in Figure 5, the vast majority of studies in LCR involve English as the target language. Almost a third (31%) of the studies analyze one or more sub-corpora of the *International Corpus of Learner English* (ICLE, Granger et al 2009) and 9% investigate learner speech in the *Louvain International Database*

of *Spoken English Interlanguage* (LINDSEI, Gilquin et al. 2010). In fact, only six other target languages were found in 1% or more of the sample, only one of which is a non-European language: Spanish, French, Italian, Korean, Dutch, and German. Note, however, the consistent trend since the 1990s to analyze foreign languages other than English, which runs parallel to the increase in the number of learner corpus compilation projects targeting other languages as evidenced in the *Learner Corpora around the World* webpage (<https://www.uclouvain.be/en-cecl-lcworld.html>).

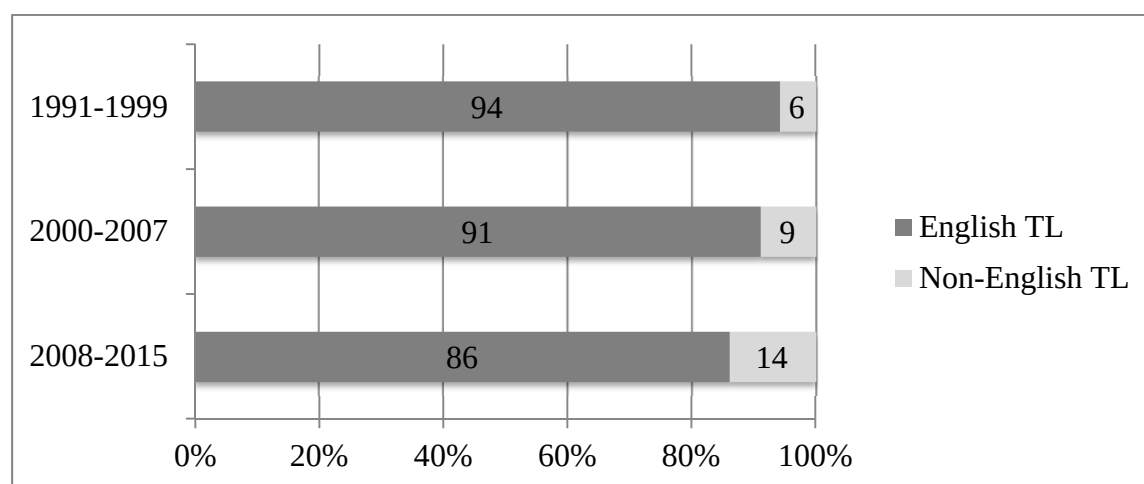


Figure 5. Percentage of studies with English vs. non-English target languages

Note. Studies in the entire sample with English as a target language totaled 333 (88%)

Another important design feature of LCR is whether or not to use a reference corpus. Overall, 44% of the sample used one reference corpus (typically LOCNESS for writing and LOCNEC for speech, two corpora of language productions by Anglophone university students, cf. Granger et al. 2009 and Gilquin et al. 2010) and 18% used two or more reference corpora (typically LOCNESS and a reference corpus of expert writing such as the British National Corpus, <http://www.natcorp.ox.ac.uk/>). Figure 6, however, reveals an interesting trend away from the use of reference corpora. More than 40% of all LCR studies in 2008-

2015 investigate learner language use without comparing results with corpus data sampled from native/expert speakers.

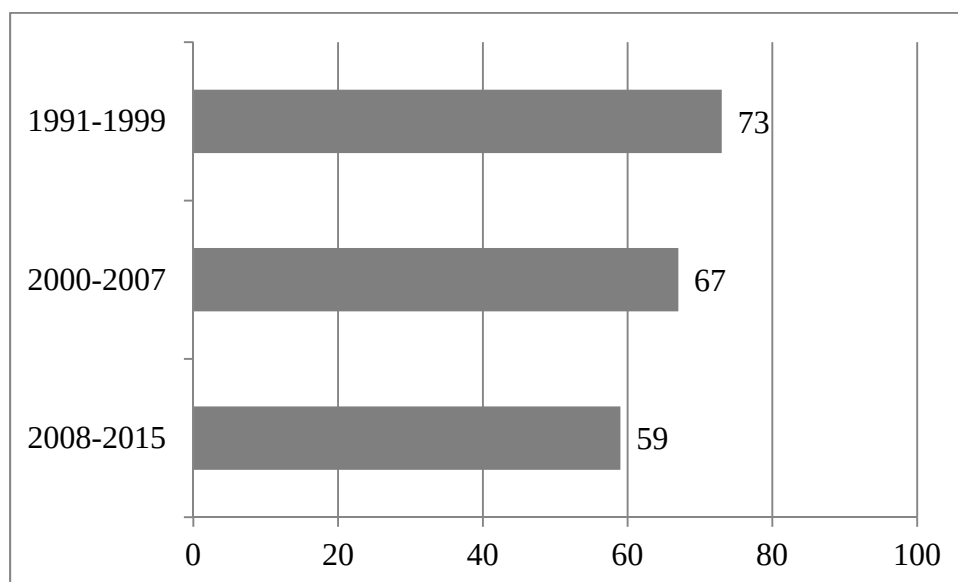


Figure 6. Percentage of studies using one or more reference corpora

Note. Overall, 63% of the sample used one or more reference corpora

This study was interested not only in the types of learners and target languages in LCR but in the size of corpora being analyzed as well. Primary studies were coded for both the number of texts and the total number of words in the learner corpora; the descriptive statistics—medians, interquartile ranges, minimum, and maximum values, which were deemed more appropriate than means and standard deviations due to strong positive skews—are presented in Table 3. The typical study in this domain analyzes a corpus of approximately 150 texts comprising about 150,000 words. These figures range widely, however, as shown in the interquartile ranges and especially in contrasting the minimum and maximum values. Looking across the three time periods of 1991-1999, 2000-2007, and 2008-2015, we see that neither of these measures of corpus size appears to be increasing (or decreasing) over time.

Table 3

Number of texts and total words in LCR overall and over time

Period	Texts		Total Words	
	Median (IQR)	Min - Max	Median (IQR)	Min - Max
1	153 (745)	6 - 2000	258,000 (391,598)	3,430 - 7,000,000
2	102 (327)	15 - 40,000	150,000 (201,903)	3,286 - 25,000,000
3	171 (490)	10 - 12,100	151,448 (369,113)	3,394 - 25,000,000
TOTAL	148 (451)	6 - 40,000	151,448 (349,943)	3,286 - 25,000,000

Moving away from the makeup of corpora in this domain and toward the techniques applied to them, different types of annotation techniques overall and over time are shown in Figure 7. No clear patterns are visible in terms of changes in annotation practices over time. Overall, however, part-of-speech tagging appears to be the most popular annotation technique (21%) followed by error tagging (12%). Very few studies in this domain (4%) work with parsed corpora.

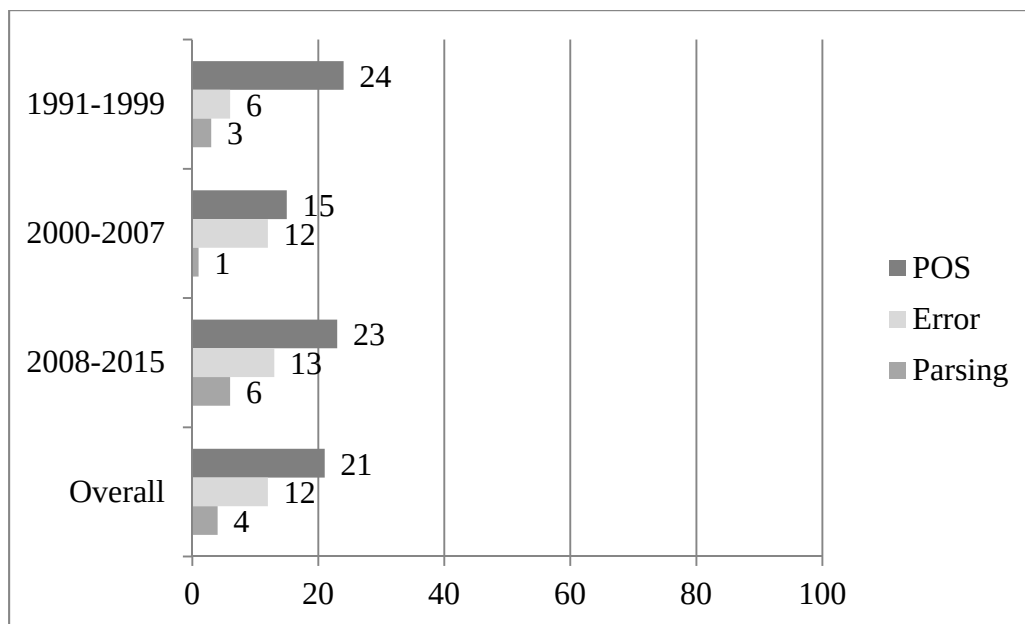


Figure 7. Percentage of studies using different annotation types

We also coded primary studies for the programs researchers used to analyze learner corpora. Unfortunately, a substantial portion of the sample (34%) did not report this information. Whatever the programs that are being used, we feel this information should be reported for the sake of transparency and replicability. Fortunately, the percentage of studies in this domain that do not report the particular program(s) used appears to be decreasing from 48% and 45% in periods 1 and 2 to 29% in period 3.

The most popular software, used by almost a third of the studies in our sample, is Wordsmith Tools. None of the other programs is used particularly often. Although the percentages are too low to merit a full consideration of changes over time for all possible programs, a few patterns stand out as noteworthy. From the 2000s on (but particularly from 2008-2015), learner corpus researchers have started to use a much wider variety of software tools including VocabProfile, Wmatrix, BNCWeb, the L2 Syntactic Complexity Analyzer, PRAAT and EXMARaLDA. Researchers' use of AntConc, Coh-Metrix, and R, the last of which can be used for linguistic as well as statistical analysis, appears to be increasing particularly sharply. We found no use of any of these programs in periods 1 and 2 but in period 3 AntConc, Coh-Metrix, and R were used in 6%, 6%, and 9% of the sample, respectively.

The software tools mentioned above are used to investigate a variety of linguistic features in learner language, including lexis, grammar, discourse and pragmatics. Figure 8 shows that LCR has always been particularly interested in lexis: between 60 and 70% of all the studies in our sample have analyzed learners' use of single words (e.g. lexical verbs) and multi-word units (e.g. verb + noun combinations, lexical bundles). Since 2000, LCR has also started to investigate pragmatic features especially in learner speech. More interestingly

perhaps, Figure 8 also reveals that a large proportion of studies were coded as including more than one type of feature. In fact, 60% of all studies coded as targeting ‘grammar’ were also coded as targeting ‘lexis’, a result which points very strongly to the special interest in the lexis-grammar interface in LCR. Among those, 39 studies (38%) were further coded as ‘discourse’ (typically studies focusing on learners’ use of connectors).

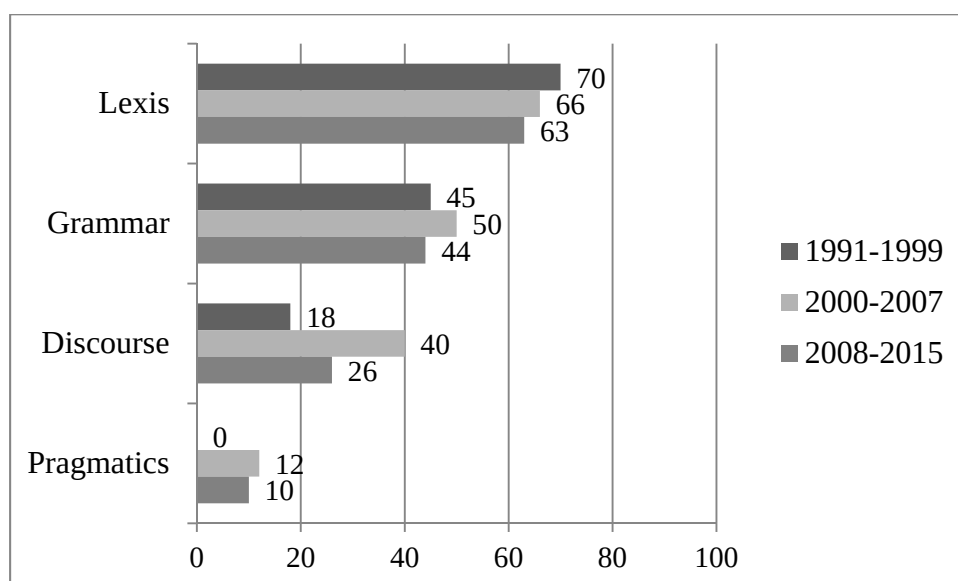


Figure 8. Target features examined in LCR

Note. Overall: Lexis = 65%, Grammar = 46%, Discourse = 30%; Pragmatics = 10%. A very small portion of the sample (2%) also targeted features of L2 pronunciation. These categories were not mutually exclusive. For example, studies that investigate the grammatical patterns and lexical preferences of words were coded as both ‘grammar’ and ‘vocabulary’. For this reason, percentages do not add up to 100.

By far, the preferred method to extract linguistic features from learner corpora has been and still is the use of concordance lines (more than 50% overall). Other relatively common methods (between 10 and 15%) are word lists, keywords and lexical bundles. No major changes over time are to be noted except for a steady increase in the use of lexical bundles.

4.2. Analyses

This study focused in particular on quantitative LCR and, as such, we were interested in describing the extent to which different statistical analyses have been used. Overall, statistical tests that involve categorical data (i.e. chi-square and log-likelihood analysis), means-based statistics (i.e., *t*-test, ANOVA) and correlations are found in a large majority of the studies that rely on inferential statistics in our sample (86 %). Logistic and multiple regression analyses are present in about 5% of the studies in this domain, and multivariate stats such as cluster analysis are found only scarcely.

As we have seen elsewhere in the results of this study, that general picture has changed over time. In the 1990s, categorical analyses were found in most studies of learner corpora that made use of statistics, and multivariate analyses were never or almost never used. Since then, the use of categorical analyses has perhaps a bit surprisingly still increased but less so than that of means-based tests. Regression and multivariate analyses are still rare but their use is nevertheless on the increase.

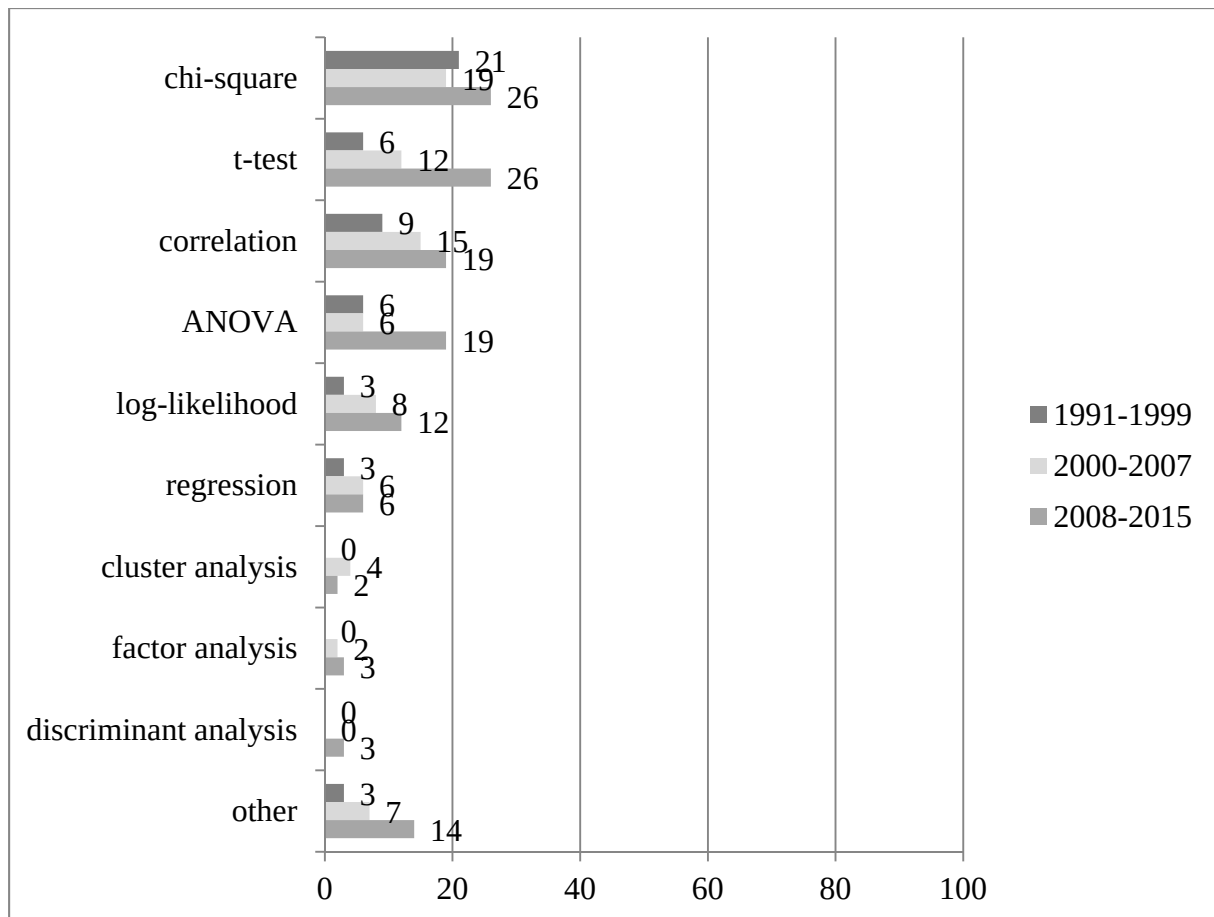


Figure 9. Statistical analyses in LCR

Note. Overall: chi-square = 23%; t-test = 20%; correlation = 17%; ANOVA = 14%; log-likelihood = 10%; regression (logistic or multiple, including VARBRUL) = 6%; cluster analysis = 2%; factor analysis (including multidimensional analysis) = 2%; discriminant analysis = 2%. The “other” category includes statistics found in less than 2% of the overall sample such as ANCOVA, MANOVA, mixed effects modeling, resampling/bootstrapping; regression includes counts for logistic and multiple regression; frequencies for t-test and ANOVA include counts for their (rare) non-parametric counterparts

Two additional and complementary analyses were conducted to better understand the use of statistical procedures in LCR. We first calculated the variety of unique tests per study, overall and over time. Not all quantitative research engages in statistical testing. In fact, as we show in Figure 10, this is the case in approximately 40% of our sample. Over time, however,

the percentage of such studies has been divided by more than two (from 73 to 31). The percentage of studies with a single unique statistical test (e.g., one or more chi-squares) has doubled from the 1990s to the most recent period in our sample. Moreover, the use of multiple unique statistical tests in the same study has more than tripled over the same period (e.g., one or more chi-squares *and* one or more ANOVAs), jumping from 9 to 31% of the samples in periods 1 and 3, respectively.

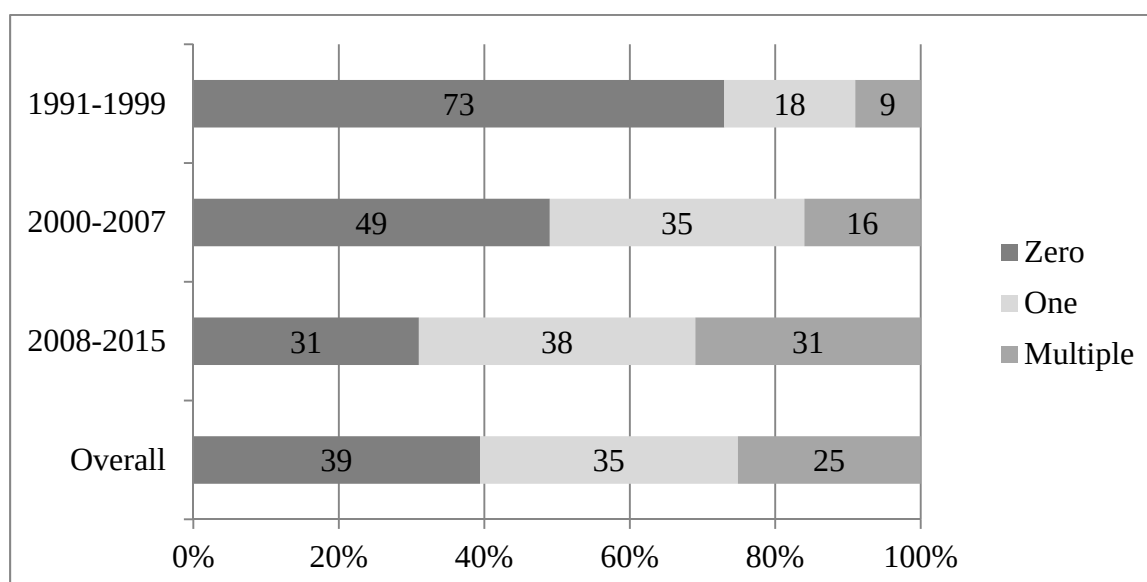


Figure 10. Percentage of studies using zero, one, or multiple unique statistical tests

The place and extent of statistical testing in LCR can also be measured by examining the total number of statistical tests conducted per study. Some studies, for instance, conducted numerous tests of statistical significance, a practice likely to produce an inflated rate of false positive findings (i.e., Type I errors; see, e.g., Plonsky 2015). It is typical for studies in this domain to run (or at least to report) nine tests. This level of statistical testing varies considerably across studies, as shown in the interquartile ranges and minimum/maximum values, but we see little change taking place here over time. In other words, although LCR has

moved over time toward using a wider variety of tests per study (see Table 4), the total number of statistical tests has increased only slightly.

Table 4

Number of tests of statistical significance in LCR overall and over time

Period	Median (IQR)	Min – Max
1	7 (12)	1 – 128
2	9 (27)	1 – 210
3	9 (18)	1 – 388
TOTAL	9 (19)	1 – 388

Note. Based only on those studies with at least one test

4.3. Reporting Practices

Our last set of results, shown in Figure 11, are based on an examination of data reporting practices, many of which are associated with enhanced transparency, replicability, and meta-analyzability. The overall most frequent practices in this category are the inclusion of research questions in written reports (57%) and the use of visual representations of data (46%), both of which appear to be increasing over time.

Four practices associated particularly with the internal validity and reliability of LCR data were investigated: reporting of estimates of (a) precision, (b) recall, (c) intrarater reliability, and (d) interrater reliability. We recognize that such estimates are not relevant to all types of studies. Nevertheless, an exceedingly small portion of researchers in this area have reported the extent to which computer- and hand-coded data are re-coded similarly by another researcher or by the same researcher him/herself. Fortunately, however, all four

categories exhibit increases over time, providing evidence for greater care in more recent years for the quality of coding and data among learner corpus researchers.

In the realm of data reporting related to statistical testing, the overall results present a fairly bleak picture, consisting of high rates of both underreported and missing data. For instance, only 8% of the studies conducting one or more statistical tests reported to have checked the assumptions, and only 36% reported the exact p value associated with the tests that were conducted. In close to a third of the studies in our sample we found omissions of standard deviations and/or the coefficients associated with statistical tests such as t or F . Looking over time, however, we do see signs of improvements in the reporting of these values. Nevertheless, power analyses, confidence, and d values were almost completely absent in LCR, with little to no sign of improvements taking place over time.

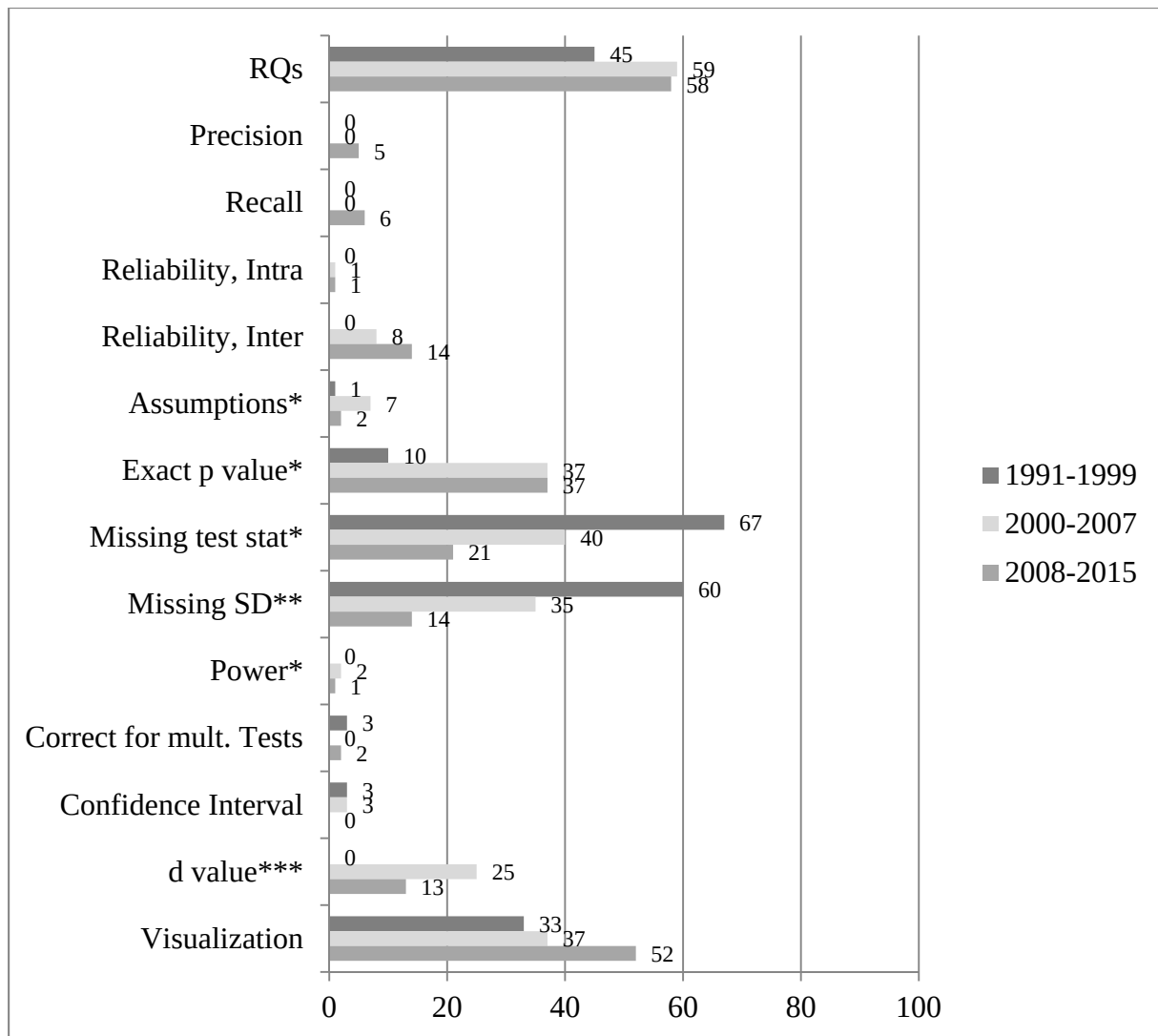


Figure 11. Percentage of studies with different data reporting practices and signs of transparency

Note. Overall findings for these reporting practices are as follows: Research questions = 57%; Precision = 3%; Recall = 3%; Intrarater reliability = 1%; Interrater reliability = 11%; Assumptions* = 8%; Exact *p* value* = 36%; Missing test statistic* = 28%; Missing SD** = 35%; Power* = 1%; Correction for multiple tests (e.g., Bonferroni) = 2%; Confidence intervals = 1%; *d* value*** = 7%; Visualization = 46%.

* Percentage based on number of studies with one or more statistical tests

** Percentage based on number of studies with one or more means reported

*** Percentage based on number of studies comparing pairs of mean scores

5. Discussion

This study applied a synthetic approach to examine the substantive interests, study designs, statistical techniques and reporting practices in LCR both cumulatively and over time. In the discussion that follows, we address the results pertinent to answering our research questions and situate them in the broader context of methodological discussions in L2 research and corpus linguistics. We also use our results to make empirically grounded suggestions for building on and improving primary research in LCR.

5.1. The ‘what’ of learner corpus research

Samples in LCR are extremely limited in diversity. Almost everything we know about learner language is based on (mostly argumentative) essays produced in university settings by young adult learners of English as a Foreign Language (cf. Granger 2015). Oversampling of such a restricted range of convenience sample demographics is a common problem throughout L2 research (e.g. DeKeyser et al. 2010; Ortega 2009). As pointed out by Plonsky & Kim (2016: 90), however, “the consequence of systematic oversight is potentially severe: massive limitations to both theoretical advancement and to the potential of our research to inform practice”.

There are several other findings related to sampling worth underlining. Approximately three-quarters of the studies in our sample analyzed written learner language. Spoken learner language was the focus in less than a third, and a small number of studies (3%) analyzed both. Although there was a sharp increase (from 9% to 30%) in the number of studies that analyzed spoken learner language from 1991-1999 to 2000-2007, this number did not increase again from 2000-2007 to 2008-2015. As such, LCR differs markedly from other L2 research domains such as task-based language teaching (TBLT) where there is a much stronger focus on spoken learner language (Plonsky & Kim 2016). There is certainly scope for more research

on corpora of spoken L2, perhaps focusing more particularly on relatively neglected features such as pragmatics and pronunciation.

Another limitation of sampling practices in LCR relates to the thorny issue of assessing proficiency in learner corpora (cf. Carlsen 2012). Most learner corpora have been compiled on the basis of external criteria (i.e. learners' characteristics and task settings) (see Boyd et al 2014 for an exception) and year of instruction or institutional status is certainly the most common criterion when assembling learner texts according to proficiency level (Granger 2004:130; see Hulstijn et al. 2010 for similar remarks on the field of SLA). This is problematic for at least two main reasons. First, learner corpora assembled on the basis of year of instruction or institutional status in a variety of settings do not "guarantee the same level of proficiency, due to factors such as the relative distance between the L1 and L2, or the amount of exposure to English in the two countries" (Carlsen 2012: 168). Second, group-based methods have been shown not to be valid ways of classifying learner corpus texts by proficiency level due to intra-group heterogeneity (Present-Thomas et al. 2013; see also Granger et al. 2009: 11-12). There is today an urgent need for more text-based or internal methods to assess proficiency level in LCR. As already argued by Carlsen (2012), learner corpus researchers may benefit from insights and practices from the professional field of language testing where achieving high rater agreement is considered essential for the very validity of test results. Increased consistency in proficiency assessment will be beneficial to the field in a number of ways: it will not only help reducing a potential source of error in LCR but it will also make it possible for other researchers to re-examine original findings, replicate analyzes and generalize the results to other learner populations.

On a more positive note, we note the increasing number of studies that have adopted a cross-sectional design to compare the production of learners at different proficiency levels (albeit, as discussed above, with the downside that proficiency levels are not often clearly

defined). Similarly, more longitudinal research is being conducted to chart the development of L2 capacities. The latter is a particularly most welcome development as the need for the longitudinal study of learner language development has repeatedly been noted in L2 research (e.g. Ortega & Byrnes 2008). In this endeavour, however, we should thrive to compile longitudinal corpora with (a) more data points and (b) spanning longer time periods if we want LCR findings to contribute to advancing SLA theories and research programs (Ortega & Iberri-Shea 2005).

In her reappraisal of Contrastive Interlanguage Analysis (CIA), Granger (2015: 11) noted that the “more popular branch of CIA has involved a comparison of learner data with native data”. The present study, however, reveals that, with a steadily decreasing number of studies using one or more reference corpora, the field seems to be moving away from looking at L1-L2 differences and more towards describing L2 production data as a worthwhile endeavor in and of itself. Support for this shift can be traced to some of the earliest theoretical models proposed in SLA, including Bley-Vroman’s (1988) Fundamental Difference Hypothesis and of course Selinker’s (1972) Interlanguage Hypothesis. More recently, however, related notions have returned to the forefront in discussions related to the “monolingual/native speaker bias” and the “multilingual turn” (e.g., Ortega 2013), movements which we may be seeing reflected in the realm of LCR as well.

Finally, it is worth highlighting the direct influence of the broader field of corpus linguistics on LCR when it comes to its substantive interests. Unlike in other linguistic traditions, the lexis-grammar interface and phraseology have always been promoted by the corpus linguistic enterprise (e.g. Sinclair 1991; Römer 2009) and LCR makes no exception. Like corpus linguists, learner corpus researchers are still relying on a limited set of off-the-shelf corpus tools (cf. Gries 2015c). Unlike corpus linguists, however, they apparently make very little use of morpho-syntactic or grammatical annotation. One likely reason for this lack

of use is that lemmatizers, part-of-speech taggers and parsers have usually been trained on the basis of native speaker corpora and there is no guarantee that they will perform as accurately when confronted with learner data (cf. Granger 2004). Pilot studies aimed at testing the reliability of annotation tools in terms of precision and recall are therefore necessary before any research can be conducted on POS-tagged or parsed learner corpora. Unfortunately, however, estimates of precision and recall are markedly absent in much of the domain as well (see Granger 1997 for an exception).

5.2. The ‘how’ of learner corpus research

Many of our results for research and reporting practices in LCR reflect those observed elsewhere in recent years in L2 research and corpus linguistics (e.g. Gries 2015c; Plonsky 2013). The analyses in LCR were found to be less than optimal. The main issue is one that has surfaced several times in recent years in broader methodological and statistical discussions: an overreliance on null hypothesis significance testing (cf. Norris 2015; Plonsky 2015).

There are several problems associated with this approach. We will not rehash the arguments in their entirety, but a few points are worth reiterating here if only briefly. By focusing almost entirely on the *p* values associated with tests of statistical significance, researchers are likely to interpret a lack of statistical significance as indicating no difference or relationship between variables (Godfroid & Spino 2015; Plonsky 2015c). Due to the relatively large samples commonly found in corpus-based studies, the low power leading to such (Type II) errors is perhaps not such a major concern. More worrisome is perhaps that the dichotomous view often associated with *p* values often blinds researchers to more informative results regarding the extent or magnitude of different relationships (Type M error; see Gelman & Weakliem 2009). For example, it is much more useful to know how strongly L2 proficiency is associated with the use of formulaic language or to know the extent to which

the use of personal pronouns differs across genres than to simply know that a statistically significant correlation or difference exists between them. In addition to these concerns, interpretations of quantitative data based on *p* values frequently lead researchers to ignore variance in statistical results as observed, which is often expressed in the form of confidence intervals (see comment on reporting practices below). By doing so, we overlook the variability across and within texts, which is critical to understanding learner language, knowledge, and development.

Despite these substantial concerns, we also observed positive developments in the choice of the statistics used. First, there has been a sharp decrease in the use of categorical tests across time, with 82% of the studies in our sample relying on (typically) chi-square tests in 1991-1999, 48% in 2000-2007 and 21% in 2008-2015. This is encouraging because it bears testimony of learner corpus researchers' increasing interest in learner variability and individual behaviour, which has prompted them to turn away from total frequencies per corpus to frequencies per text/speaker. As a result, analyses comparing means (typically *t*-tests and ANOVAs) appear to be occupying an increasingly large share of the analyses over time, from 12% (1991-1999) to 40% (2008-2015). Unlike in other domains in L2 research, however, ANOVAs and *t*-tests do not have the lion's share (Plonsky 2014). And perhaps rightly so.

As put by Gries (2015b: 64), "the recognition that corpus-linguistic statistics has to go multifactorial is maybe the most important recommendation for the field's future development." We agree entirely. It is critical that learner corpus researchers recognize the inherently multivariate nature of all linguistic phenomena. In the realm of theory, this approach will allow for a more holistic view of the many factors influencing language use and development. Methodologically speaking, and perhaps most pertinent to the present study, such an approach will require that learner corpus researchers engage with statistical methods

that are better tailored to deal with the internal complexity of corpus data (e.g. nested variables, repeated-measurements) and capture the context-constrained and multifactorial nature of (foreign) language development (Gries 2015c; Gries & Wulff 2013: 330). Furthermore, by simultaneously analyzing relationships between and among multiple independent and possibly dependent variables, multifactorial statistics reduce the number of statistical tests that are needed and thereby preserve statistical power. A small number of authors make use of binary, logistic, and multiple regression analyses ($k = 21$) but more comprehensive techniques such as resampling, bootstrapping, factor analysis, cluster analysis, and mixed-effect modeling are still extremely scarce. And very few learner corpus researchers have generally embraced the use of multivariate statistics, largely mirroring what is found in other domains within applied linguistics.

A final comment regarding the change in analyses observed over time is warranted. The percentage of studies employing no statistical test dropped sharply from 73% (1991-1999) to 31% (2008-2015) while the percentage of studies using one or more than one increased substantially (from 18% to 38% and from 9% to 31% respectively). This result replicated previous reports in L2 research and corpus linguistics of an increase in the use of inferential statistics as opposed to presenting only descriptive statistics (e.g. Gass 2009). Like in corpus linguistics, however, we note two patterns in need of improvement: (a) assumptions of statistical tests are not often checked and/or reported, and (b) multiple statistical tests are still often applied on the same data without correcting for the alpha level (Gries 2015b; Köhler 2013).

In general, however, reporting practices in the field display change over time in the right direction. But progress is slow and developments are still very much needed. For example, the increasing percentage of studies that include research questions shows that LCR is becoming less descriptive and perhaps also somewhat more theoretical, as would be

expected from a developing field (cf. Gass 2009: 10). This being said, however, research questions were only found in 58% of the studies in the third period (2008-2015) and we believe that this percentage should still increase in the future. There are different types of research questions (i.e. descriptive, comparative or relationship-based) but a valid quantitative study should always answer research questions in a scientifically rigorous manner (cf. dissertation.laerd.com 2012).

As found across a number of other domains of applied linguistics (see Larson-Hall & Plonsky 2015), our study revealed unacceptably high rates of both underreported and missing data. Omissions of standard deviations, for example, are immediately problematic for interpretations made at the primary study level. Missing standard deviations, which are used to calculate effect sizes such as Cohen's *d*, may also indicate a lack of concern for the replicability and meta-analyzability of results. We are speculating here. However, only 2% (*k*=8) of our sample referred to itself as a replication (see Granger & Bestgen forthcoming, for a discussion of the importance of replication studies in LCR). Furthermore, despite the widespread use of meta-analysis elsewhere in applied linguistics (see Plonsky & Oswald 2015) and occasionally in other subdomains of corpus linguistics (e.g., Durrant 2014), we are not aware of one single meta-analysis in LCR.

Finally, among the few studies that make use of morpho-syntactically or grammatically annotated data, very few have reported precision and recall rates. Similarly, an extremely small proportion of learner corpus researchers have reported the extent to which hand-coded data are re-coded similarly by another researcher or by the same researcher himself/herself. Fortunately, however, at least the category of inter-rater reliability estimates exhibits a marked increase over time, providing evidence for greater care in more recent years for the quality and reliability of coding and data among learner corpus researchers. Because of this positive change in estimates of inter-rater reliability, which generally constitutes a better

type of evidence than intra-rater agreement, we are not particularly concerned with the lack of the latter.

5.3. Recommendations for future research

Our focus throughout much of this synthesis has been retrospective. However, the purpose of this study was not only to look back but ahead as well. Based on the findings of the current study, we present in this section a set of recommendations for areas in need of particular attention. Many of these comments echo suggestions made previously as part of methodological discussions in L2 research and corpus linguistics (see Granger 1997; Granger 2015; Gries 2013; Gries 2015a; Gries 2015b; Mahboob et al. 2016; Norris et al 2015; Plonsky 2013; Plonsky 2014). For the sake of clarity and economy, our recommendations appear in the form of a list.

1. Substantive areas in need of further attention
 - a. Pragmatics
 - b. Pronunciation
2. Learner demographics, learning contexts and study design
 - a. Investigate a greater variety of learner demographics (e.g. younger and older learners, beginning learners, a wider variety of settings especially outside university, more target languages).
 - b. Investigate a greater variety of learner production (i.e. speech in its various forms, more varied genres and tasks).
 - c. Resort to text-based methods to assess proficiency.
 - d. Carry out more cross-sectional and longitudinal studies.
3. Analyses

- a. Do not let software tools limit your research questions.
 - b. Do not aggregate learner data and use case-by-variable datasets.
 - c. Check the assumptions of statistical tests.
 - d. Conduct fewer tests of statistical significance and correct for the alpha level.
 - e. Be skeptical of p values.
 - f. Calculate, report and interpret descriptive statistics, including effect sizes and confidence intervals.
 - g. Consider multivariate statistics.
4. Data reporting and transparency
- a. Formulate research questions.
 - b. Identify each software tool used, report the settings employed and describe each methodological step.
 - c. Report precision and recall rates for any automatic annotation tool (e.g. POS-tagger, parser) used.
 - d. Report more thoroughly descriptive statistics, including standard deviations will all means.
 - e. Report exact p values with all statistical tests if $p > .001$.
 - f. Report and interpret (intra-rater, inter-rater and instrument) reliability estimates.

6. Conclusion

LCR is a recent but burgeoning scientific enterprise which has largely borrowed its methodological toolbox and conceptual framework from second language acquisition and corpus linguistics, i.e. two disciplines which do not go back much earlier than the 1960s or 1970s (cf. Gass 2009; Biber & Reppen 2015a) and have lately been engaged in the process of

methodological reform (e.g. Plonsky 2013; 2014; Norris et al. 2015; Mackey & Marsden 2016). As previously described in the fields of SLA and corpus linguistics, this study has found quantitative studies in LCR to have increased steadily in number and variety since its inception. It has also revealed a number of (albeit perhaps slower) methodological improvements taking place. However, a significant number of weaknesses related to sampling practices, data analyses and reporting practices were also observed in the sample of primary studies. Consequently, we feel the findings of this study warrant close attention. It is our hope that the field of LCR will, at the very least, reflect on and continue to investigate not only the substance of the field (the *what*—language use, development) but also the means by which we arrive at our understanding of such constructs (i.e., the *how*). We would also encourage action on this front at both institutional, societal, and individual levels.

In order to continue its path toward the kind of methodological reform that this paper has shown to be needed, LCR should (a) develop and enforce methodological standards for quantitative research and (b) improve statistical literacy. Researchers, reviewers and editors will find directions in a variety of resources including textbooks on research methods and statistics in SLA and (corpus) linguistics (e.g. Gries 2013b; Levshina 2015; Mackey & Gass 2011; Plonsky 2015b), publications by the American Psychological Association (e.g. APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) and journal guidelines (e.g. Norris et al. 2015; Mahboob et al 2016). However, in view of its multidisciplinary nature, there is arguably a need for specific recommendations for LCR. Of course, much of the responsibility for improving research and reporting practices will fall, by default, on journal editors and reviewers. With this great task in mind, we recommend that the editors of the new *International Journal of Learner Corpus Research* (IJLCR) consult with trusted individuals and sources and establish a set of guidelines for publication of LCR for use both in IJLCR as well as for researchers in this

domain publishing in other venues. We would also call upon the leadership of the Learner Corpus Association to consider designating a task force to construct methodological standards for LCR, spanning the whole research process from corpus collection to manuscript writing.

The suggestions we have laid out are optimistic and ambitious. However, we feel that they are necessary in order to improve the field of LCR. This is necessary not only because a strong methodological foundation is needed to build theory and practice but also because, ultimately, “[r]espect for the field of [LCR] can come only through sound scientific progress” (Gass et al 1998: 407). We hope that the findings of this paper—and our recommendations for moving forward—will help enable the field of LCR to produce more scientific progress and thereby move toward a more robust impact in the realms of theory and practice.

Acknowledgment

We thank our MA students, Geoffrey Bodart, Dan Brown, Deirdre Derrick, and Sophie Longueville for their invaluable help with the coding of the LCR references. The first author gratefully acknowledges the financial support of the Belgian Fund for Research (FNRS) and the Language & Communication Institute (UCLouvain). Our gratitude also goes to the reviewers who helped shape this manuscript into its present form. All remaining errors remain ours.

References

- APA Publications and Communications Board Working Group on Journal Article Reporting Standards. 2008. “Reporting Standards for Research in Psychology”, *American Psychologist* 63(9), 839-851.
- Baayen, R.H. 2001. *Word Frequency Distributions*. Kluwer Academic Publishers.

- Baroni, M. & Evert, S. 2008. "Statistical methods for corpus exploitation". In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics. An International Handbook*, article 36. Mouton de Gruyter, Berlin, 777-803.
- Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
- Biber, D. & Reppen, R. 2015a. "Introduction". In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press, 1-8.
- Biber, D. & Reppen, R. 2015a (Eds.). *The Cambridge Handbook of Corpus Linguistics*. Cambridge: Cambridge University Press.
- Bley-Vroman, R. 1988. "The fundamental character of foreign language learning", in W. Rutherford & M. Sharwood Smith (Eds.), *Grammar and Second Language Teaching: A Book of Readings*. Rowley, MA: Newbury House, 19-30.
- Boyd, A., Hana, J., Nicolas, L., Meurers, D., Wisniewski, K., Abel, A., Schöne, K., Štindlová, B. & Vettori, C. 2014. "The MERLIN corpus: Learner Language and the CEFR". Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, May 26-31, 2014.
- 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), 1-28.
- Byrnes, H. 2013. "Notes from the editor", *The Modern Language Journal*, 97(4): 825-827.
- Carlsen, C. 2012. "Proficiency Level—a Fuzzy Variable in Computer Learner Corpora", *Applied Linguistics*, 33(2), 161–183.

- Dagneaux, E., Denness, S., & Granger, S. 1998. « Computer-aided error analysis”, *System: An International Journal of Educational Technology and Applied Linguistics*, 26(2), 163–174.
- DeKeyser, R., Alfi-Shabtay, I. & Ravid, D. 2010. “Cross-linguistic evidence for the nature of age effects in second language acquisition”, *Applied Psycholinguistics*, 31, 413-438.
- Derrick, D. J. 2016. “Instrument reporting practices in second language research”, *TESOL Quarterly*, 50, 132-153.
- dissertation.laerd.com 2012. “Types of quantitative research question”, Laerd Dissertation Online. Available at <http://dissertation.laerd.com/types-of-quantitative-research-question.php> [last accessed 20 May 2016]
- Durrant, P. 2014. “Corpus frequency and second language learners’ knowledge of collocations: A meta-analysis”, *International Journal of Corpus Linguistics*, 19, 443–477.
- Gass, S. 2009. “A historical survey of SLA research”, in T. K. Bhatia & W. C. Ritchie (Eds.), *The New Handbook of Second Language Acquisition*. Bingley, England: Emerald, 3-27.
- Gass, S., Fleck, C., Leder, N. & Svetics, I. 1998. “Ahistoricity revisited. Does SLA have a history?”, *Studies in Second Language Acquisition* 20, 407-421.
- Gelman, A., & Weakliem, D. 2009. “Of beauty, sex, and power: Too little attention has been paid to the statistical challenges in estimating small effects”, *American Scientist*, 97, 310–316.
- Gilquin, G., Granger, S., & Paquot, M. 2007. « Learner corpora: The missing link in EAP pedagogy”, *Journal of English for Academic Purposes*, 6(4), 319–335.

- Gilquin, G., De Cock, S., & Granger, S. 2010. *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Presses universitaires de Louvain: Louvain-La-Neuve.
- Godfroid, A., & Spino, L. 2015. "Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence", *Language Learning*, 65, 896-928.
- Granger, S. 1996. "From CA to CIA and back: an integrated approach to computerized bilingual and learner corpora", in K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in Contrast. Text-based Cross-linguistic Studies*. Lund: Lund University Press, 37-51.
- Granger, S. 1997. "Automated retrieval of passives from native and learner corpora: precision and recall", *Journal of English Linguistics* 25(4): 365-374.
- Granger, S. 2004. "Computer learner corpus research: Current status and future prospects", in U. Connor & T. Upton (Eds.), *Applied Corpus Linguistics: a Multidimensional Perspective*. Amsterdam & Atlanta: Rodopi, 123-145.
- 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*. Amsterdam: John Benjamins, 13-32.
- Granger, S. 2015. "Contrastive Interlanguage Analysis: a reappraisal", *International Journal of Learner Corpus Research* 1(1): 7-24.
- Granger, S. & Bestgen, Y. (forthcoming). "Using collgrams to assess L2 phraseological development: A replication study", in P. de Haan, R. de Vries & S. van Vuuren (Eds.), *Language, Learners and Levels: Progression and Variation*. Louvain-la-Neuve: Presses universitaires de Louvain.

- Granger, S., Dagneaux, E., Meunier, F. & Paquot, M. 2009. *International Corpus of Learner English v2 (Handbook + CD-Rom)*. Louvain-la-Neuve : Presses universitaires de Louvain.
- Granger, S., Gilquin, G. & Meunier, F. 2015. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press.
- Gries S. 2006a. "Some proposals towards more rigorous corpus linguistics", *ZAA* 54(2), 191-202.
- Gries S. 2006b. "Exploring variability within and between corpora: some methodological considerations", *Corpora* 1(2), 109-151.
- Gries, S. 2010. "Methodological skills in corpus linguistics: a polemic and some pointers towards quantitative methods", in T. Harris & M. Moreno Jaén (Eds.), *Corpus Linguistics in Language Teaching*. Frankfurt: Peter Lang, 121-146.
- Gries, S. 2013a. "Statistical tests for the analysis of learner corpus data", in A. Diaz-Negrillo, N. Ballier & P. Thompson (Eds.), *Automatic Treatment and Analysis of Learner Corpus Data*. Amsterdam & Philadelphia: Benjamins.
- Gries, S. 2013b. *Statistics for Linguistics with R. A Practical Introduction* (second edition). Berlin/Boston: De Gruyter Mouton.
- Gries, S. 2014. "Quantitative corpus approaches to linguistic analysis: seven or eight levels of resolution and the lessons they teach us", in I. Taavitsainen, M. Kytö, C. Claridge, & J. Smith (Eds.), *Developments in English: Expanding Electronic Evidence*. Cambridge: Cambridge University Press, 29-47.
- Gries, S. 2015a. "Statistics for learner corpus research", in S. Granger, G. Gilquin & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research*. Cambridge University Press, 159-181.

- Gries, S. (2015b). Quantitative designs and statistical techniques. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics* (pp. 50-71). Cambridge: Cambridge University Press.
- Gries, S. (2015c). Some current quantitative problems in corpus linguistics and a sketch of some solutions. *Language and Linguistics* 16(1): 93-117.
- Gries, S., & Deshors, S. (2014). Using regressions to explore deviations between corpus data and a standard/target: two suggestion. *Corpora*, 9(1), 109–136.
- Gries, S. & Wulff, S. 2013. “The genitive alternation in Chinese and German ESL learners. Towards a multifactorial notion of *context* in learner corpus research”, *International Journal of Corpus Linguistics* 18(3), 327-356.
- Hulstijn, J., Alderson, C. & Schroonen, R. 2010. “Developmental stages in second language acquisition and levels of second-language proficiency: are there links between them?”, in I. Bartning, M. Martin, & I. Vedder (Eds.), *Communicative Proficiency and Linguistic Development: Intersections between SLA and language Testing Research*. European Second Language Association: EUOSLA Series Monographs 1.
- Ioannidis J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. 2015. “Meta-research: Evaluation and improvement of research methods and practices”, *PLoS Biology* 13, e1002264. doi:10.1371/journal.pbio.1002264
- Kilgariff, A. 2001. “Comparing corpora”, *International Journal of Corpus Linguistics* 6(1), 97-133.
- Kilgariff, A. 2005. “Language is never, ever, ever random”, *Corpus Linguistics and Linguistic Theory*, 1-2, 263-275.
- Köhler, R. 2013. “Statistical Comparability: Methodological Caveats”, in S. Sharoff, R. Rapp, P. Zweigenbaum, & P. Fung (Eds.), *Building and Using Comparable Corpora*. Springer Berlin Heidelberg, 77-91.

- Larson-Hall, J. & Herrington, R. 2010. "Improving data analysis in second language acquisition by utilizing modern developments in applied statistics", *Applied Linguistics*, 31(3), 368-390.
- Larson-Hall, J. & Plonsky, L. 2015. "Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field", *Language Learning*, 65, S1, 127-1593.
- Levshina, N. 2015. *How to Do Linguistics with R? Data Exploration and Statistical Analysis*. Amsterdam/Philadelphia: Benjamins.
- Lipsey, M. W. 2009. "Identifying interesting variables and analysis opportunities", in H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of Research Synthesis* (2nd ed). New York: Russell Sage Foundation, 147-158.
- Liu, Q., & Brown, D. 2015. "Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing", *Journal of Second Language Writing*, 30, 66-81.
- Mahboob, A., Paltridge, B., Phakiti, A., Wagner, E., Starfield, S., Burns, A., Jones, R. H., & De Costa, P. I. 2016. "TESOL Quarterly Research Guidelines", *TESOL Quarterly*, 50, 42-65.
- Mackey & Marsden, E. (Eds.) 2016. *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages*. New York: Routledge.
- Marsden, E., Mackey, A., & Plonsky, L. 2016. "Breadth and depth: The IRIS repository", in A. Mackey & E. Marsden (Eds.), *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages*. New York: Routledge, 1-21.
- Meunier, F. 2010. "Learner corpora and English language teaching: checkup time", *Anglistik: International Journal of English Studies*, 21(1), 209–220.

- Myles, F. 2008. "Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues", in L. Ortega & H. Byrnes (Eds.) *The Longitudinal Study of Advanced L2 Capacities*. New York & London: Routledge, 58-72.
- Norris, J. 2015. "Statistical significance testing in second language research: Basic problems and suggestions for reform", *Language Learning*, 65 (Supp. 1), 97-126.
- Norris, J., Plonsky, L., Ross, S. & R. Schoonen 2015. "Guidelines for reporting quantitative methods and results in primary research", *Language Learning* 65(2): 470-476.
- Ortega, L. 2009. *Understanding Second Language Acquisition*. London: Hodder Education.
- Ortega, L. 2013. "SLA for the 21st century: Disciplinary progress, transdisciplinary relevance, and the bi/multilingual turn", *Language Learning*, 63 (Supplement 1), 1-24.
- Ortega, L. & Iberri-Shea, G. 2005. "Longitudinal research in second language acquisition: recent trends and future directions", *Annual Review of Applied Linguistics*, 25, 26-45.
- Ortega, L. & Byrnes, H. 2008. "Theorizing advancedness, setting up the longitudinal research agenda", in L. Ortega & H. Byrnes (Eds.), *The Longitudinal Study of Advanced L2 Capacities*. New-York: Routledge, 281-300.
- Paquot, M., & Granger, S. (2012). "Formulaic Language in Learner Corpora", *Annual Review of Applied Linguistics*, 32, 130–149.
- Pendar, N. & Chapelle, C. (2008). "Investigating the Promise of Learner Corpora: Methodological Issues", *CALICO Journal* 25(2): 189-206.
- Plonsky, L. 2013. "Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research", *Studies in Second Language Acquisition*, 35, 655-687.
- Plonsky, L. 2014. "Study quality in quantitative L2 research (1990-2010): a methodological synthesis and call for reform", *The Modern Language Journal* 98(1), 450-470.

- Plonsky, L. 2015. "Statistical power, p values, descriptive statistics, and effect sizes: a 'back-to-basics' approach to advancing quantitative methods in L2 research", in L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research*. New York: Routledge, 23-45.
- Plonsky, L. (Ed.) 2015b. *Advancing Quantitative Methods in Second Language Research*. New York: Routledge.
- Plonsky, L. 2015c. "Quantitative considerations for improving replicability in CALL and applied linguistics", *CALICO Journal*, 32, 232-244.
- Plonsky, L. in press. "Quantitative research methods in instructed SLA", in S. Loewen & M. Sato (Eds.), *The Routledge Handbook of Instructed Second Language Acquisition*. New York: Routledge.
- Plonsky, L., & Derrick, D. J. 2016. "A meta-analysis of reliability coefficients in second language research", *Modern Language Journal*, 100, 538-558.
- Plonsky, L., & Gass, S. 2011. "Quantitative research methods, study quality, and outcomes: The case of interaction research", *Language Learning*, 61, 325-366.
- Plonsky, L., & Kim, Y. 2016. "Task-based learner production: A substantive and methodological review", *Annual Review of Applied Linguistics*, 36, 73-97.
- Plonsky, L., Egbert, J., & LaFlair, G. T. 2015. « Bootstrapping in applied linguistics: Assessing its potential using shared data", *Applied Linguistics*, 36, 591-610.
- Plonsky, L., & Oswald, F. L. 2015. "Meta-analyzing second language research", in L. Plonsky (Ed.), *Advancing Quantitative Methods in Second Language Research*. New York, NY: Routledge, 106-128.
- Plonsky, L., & Oswald, F. L. in press. Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*.

- Porte, G. (Ed.) 2012. *Replication Research in Applied Linguistics*. Cambridge: Cambridge University Press.
- Present-Thomas, R. L., Weltens, B., & de Jong, J. H. A. L. 2013. "Defining proficiency: A comparative analysis of CEF level classification methods in a written learner corpus", *Dutch Journal of Applied Linguistics*, 2(1), 57–76.
- Römer, U. 2009. "The inseparability of lexis and grammar. Corpus linguistic perspectives", *Annual Review of Cognitive Linguistics*, 7, 141-163.
- Ross, S. & Mackey, B. 2015. "Bayesian Approaches to Imputation, Hypothesis Testing, and Parameter Estimation", *Language Learning*, 65 (Supp. 1), 208-227.
- Selinker, L. 1972. "Interlanguage", *International Review of Applied Linguistics in Language Teaching*, 10, 209-231.
- Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Thomas, M. 2006. "Research synthesis and historiography: The case of assessment of second language proficiency", in J. M. Norris & L. Ortega (Eds.), *Synthesizing Research on Language Learning and Teaching*. Philadelphia, PA: John Benjamins, 279-298.
- Wells, K. & Littell, J. H. 2009. "Study quality assessment in systematic reviews of research on intervention effects", *Research on Social Work Practice*, 19, 52-62.
- Ziegler, N. 2016. "Methodological practices in interaction in synchronous computer mediated communication: a synthetic approach", in A. Mackey & E. Marsden (Eds.), *Instruments for Research into Second Languages: Empirical Studies Advancing Methodology*. New York: Routledge, 197-223.