LALEj

LANGUAGE AND LITERACY EDUCATION JOURNAL

1(2) (2025)

29 - 47

UNIVERSITI MALAYA
FAKULTI PENDIDIKAN
Faculty of Education

LEARNER CORPUS RESEARCH IN SLA FRAMEWORK: A SYSTEMATIC REVIEW OF EXPERIMENTAL DESIGNS

***Faizah Mohamad Nusri[1], Siti Zaidah Zainuddin[1], & Amir Rashad Mustaffa[1]**

[1]Universiti Malaya, Malaysia

****faizahnusri@gmail.com***

ABSTRACT

Learner Corpus Research (LCR) has become increasingly relevant to Second Language Acquisition (SLA), yet its systematic integration within experimental research designs remains relatively limited. This systematic review investigates how LCR has been incorporated into experimentally oriented SLA studies published between 2015 and March 2024. Guided by the PRISMA 2020 framework, ten studies were identified from the Scopus database that explicitly combined learner corpus data with experimental or quasi-experimental methods. The review examines these studies in terms of their targeted SLA constructs, types of learner data, methodological orientations, and analytical procedures. The findings show a strong preference for a quantitative approach and performance-based constructs, with particular emphasis on the complexity–accuracy–fluency (CAF) framework. The reviewed studies also show a clear preference for employing Natural Language Processing tools, particularly Coh-Metrix and syntactic complexity analysers such as SynLex. Across the reviewed studies, learner corpora consistently serve three key functions: accounting for experimentally observed effects, validating instructional outcomes through learners' extended performance, and facilitating triangulation across multiple data sources. Despite its clear methodological and theoretical value, the integration of LCR into experimental SLA research is still uneven.

Keywords: *Learner Corpus Research, second language acquisition, CAF, NLP.*

INTRODUCTION

Despite its youth, Learner Corpus Research (LCR) has experienced significant growth and evolution (Callies & Paquot, 2015; Granger et al., 2015). This progression is evident through various avenues, including an expanded corpus size, increased emphasis on longitudinal studies, and a heightened recognition within the learner corpus community regarding the importance of accounting for individual variability (Granger et al., 2015). LCR is known for its strong focus on practical applications (McEnery et al., 2019). While initially confined to the realm of foreign language instruction, it has now branched out into diverse fields, notably natural language processing (NLP) (Alexopoulou et al., 2017; Kyle, 2021). Additionally, the adoption of more advanced statistical methodologies in LCR has addressed previous limitations, enhancing its analytical capabilities (Gries, 2009). As a consequence, LCR has emerged as an interdisciplinary domain, intersecting with Corpus Linguistics (CL) and Second Language Acquisition (SLA) (Callies & Paquot, 2015). While this interdisciplinary nature offers fertile ground for exploration, it also necessitates the integration of diverse research strands that have yet to be fully synthesized, particularly regarding a key question that concerns how LCR can be meaningfully integrated with the experimental inclination of the SLA field.

LITERATURE REVIEW

Learner Corpus Research and Second Language Acquisition

Traditionally, learner corpus and experimentation have been linked to distinct paradigms: LCR and SLA research, respectively (Chambers, 2019; Granger et al., 2015; Myles, 2020). LCR primarily focuses on learners' performance, which refers to how they utilize the second language (L2), observable through their authentic production as documented in learner corpora. In contrast, SLA tends to emphasize learners' competence, which is more effectively investigated through experimentation (Bell & Payant, 2021; Gilquin, 2021; Iurato, 2022). Such experimentation allows researchers to manipulate various variables systematically aimed at uncovering the mental and psychological processes underlying language acquisition and discern their potential impact. Thus, experimentation aligns with one of the key objectives of SLA theory: to deepen understanding of learner language, its developmental trajectory, and the factors influencing both (Kersten & Greve, 2022).

Despite differing perspectives on the distinction between *competence* and *performance*, LCR, like conventional SLA research, shares the overarching goal of advancing our understanding of the complex processes involved in second and foreign language acquisition (Granger et al., 2015). Accordingly, the use of learner-generated written and spoken data has long been central to SLA inquiry (Fuchs & Werner, 2018; Volodina et al., 2019; Vyatkina, 2016), in a manner that enables researchers to capture the subtleties and dynamic nature of learners’ interlanguage development (Gass et al., 2011). By contrast, the experimental dimension of such research has only more recently begun to receive sustained attention.

Historically, the data used in SLA studies tended to be somewhat artificial, often stemming from highly controlled language tasks, typically of an experimental nature (Ellis, 2020; Gilquin, 2020; Lozano & Mendikoetxea, 2013; Meunier & Littre, 2013). Consequently, these datasets may not accurately reflect learners' behaviors in more natural communicative settings (Lozano & Mendikoetxea, 2013). Moreover, SLA investigations are often constrained by the inclusion of only a limited number of participants and the relatively modest size of the gathered data (Granger et al., 2015; Plonsky, 2013; Ziegler et al., 2017). Such constraints can restrict the generalizability of findings and limit researchers’ ability to observe low-frequency phenomena, developmental patterns, and individual variation over time.

These methodological limitations have prompted increasing interest in complementary data sources, such as learner corpora, which offer larger, more diverse, and more ecologically valid samples of learner language (Callies & Paquot, 2015). In applied linguistics research, ecological validity is commonly understood as the degree to which research tasks, data, or study conditions resemble real-world language use and learning contexts. A study is considered ecologically valid when it reflects how language is actually produced and used outside tightly controlled research settings, rather than relying solely on artificial or decontextualised tasks (Pusey & Butler, 2023; Verbeke, 2024).

Ecological validity has played a key role in establishing the relevance of LCR within the field of SLA (Gilquin & Gries, 2009; Granger, 2008; Vyatkina, 2016). LCR enables researchers to investigate factors influencing language acquisition using data that are typically high in ecological validity, such as individual differences (Larsson et al., 2022), learning strategies (Mukherjee & Götz, 2015; Winter & Le Foll, 2022; Vuuren et al., 2022), and sociocultural contexts (Czerwionka & Olson, 2020). It also endeavors to unravel how learners acquire vocabulary (Smith, 2020; Römer-Barron & Garner, 2019), grammatical structures (Newbery-Payton, 2022), and pronunciation patterns over time through exposure to the target language (Gut, 2007). By examining learner corpora, researchers can analyze patterns and trends in language use and shed light on the cognitive and socio-affective dimensions of language learning (Fernández-Mira et al., 2021; Gilquin, 2021; Pérez-Paredes & Díez-Bedmar, 2019)

Definitions of Learner Corpus

The term "learner corpus" is subject to various interpretations. As per McEnery and Gabrielatos (2006, p. 5), it denotes a compilation of 'machine-readable authentic texts, including transcripts of spoken data', chosen to 'accurately portray a specific language or language variation'. Nesselhauf (2005, p. 127) defines a learner corpus as a methodical 'computerized collection of texts produced by language learners.' Here, "methodical" implies that the texts are selected based on different external criteria, such as learner proficiency levels and native languages. This perspective diverges from earlier studies in SLA, as the objective is to ensure that the corpus represents the language variety under investigation.

This focus on representation and equilibrium in text selection, as highlighted by Lindquist and Levin (2018) and Webster et al. (2018), encompasses factors such as learners' native languages, gender, and proficiency, which are the key differences of learner corpora from other datasets commonly utilized in SLA research. Alongside considerations of representativeness and balance, another crucial aspect in defining learner corpora is the degree of naturalness they exhibit. Granger (2008, p. 338) defines learner corpora as 'electronic compilations of (near)natural texts' crafted by foreign or second language learners, structured according to 'explicit design criteria'. This emphasis implies that the texts within learner corpora may contain materials that, while not entirely natural, are curated in accordance with specific guidelines. Seen from this angle, LCR fits more comfortably within the broader SLA tradition, offering a methodologically sound way to examine learner language that bridges experimental control and authentic language use.

Methodological Approaches in LCR

LCR has traditionally been rooted in corpus linguistics, drawing on methodological principles that are closely aligned with theoretical linguistic inquiry (Gries, 2009). According to Paquot et al., (2017), LCR adheres to 4 key methodological principles of corpus linguistics; (1) it is grounded in empiricism, analysing written or spoken productions of learners; (2) LCR relies on the analysis of large and systematically collected collections of learner texts, chosen to represent a specific learner population; (3) utilisation of computer software tools is integral to LCR, facilitating tasks such as searching, organizing, and augmenting learner texts with diverse metadata and linguistic analyses; (4) LCR often employs a combination of quantitative and qualitative analytical methods.

Corpus research frequently employs a combination of quantitative and qualitative approaches. The methodological framework underpinning quantitative analyses in LCR typically follows a deductive approach, emphasizing outcomes and aimed at testing specific hypotheses. These hypotheses, as observed by Callies (2015a) and Lindquist and Levin (2018), can be confirmed, rejected, or refined and retested. Additionally, quantitative data in LCR are characterized as "hard" data, meaning they are identifiable, classifiable, and quantifiable, facilitating more precise statistical analysis. Such data possess greater generalizability and are more easily replicable compared to qualitative data (Callies, 2015a).

Conversely, researchers employing qualitative corpus analysis as the foundation of their studies embrace an exploratory, inductive methodology. They aim to empirically investigate the interplay between the meanings and functions of linguistic forms within the corpus and the various ecological factors influencing language communication (Hasko, 2012). These factors include the age and gender of speakers, their educational level and socioeconomic status, the context (time and place) of communicative events, the relationship between interlocutors, and the modality of speech, among others. The methodological structure guiding qualitative data is primarily heuristic and discoverydriven, emphasizing processes rather than results (Klag & Langley, 2013).

Mixed-method corpus analysis is increasingly employed in LCR as a way of balancing large-scale quantitative evidence with more fine-grained qualitative interpretation. In many studies, automated corpus tools are first employed to identify general patterns in learner language, such as frequency or complexity trends. These results are then examined more closely through qualitative analysis of concordance lines or learner texts to understand how linguistic features are realised in context. This combination helps reveal developmental variability and partial learning that may not be evident from statistical results alone (Granger, 2011; Paquot et al., 2017). In instructional and experimental research, mixed-method approaches also allow corpus findings to be interpreted alongside other forms of evidence, supporting more ecologically valid and well-grounded conclusions (Callies, 2015b; Vyatkina, 2016).

While LCR offers a multitude of methods and procedures, Callies (2015b) and Paquot et al. (2017) contend that the prevalent methodology and procedures utilized in LCR thus far have predominantly been ‘corpus-based, quantitative, cross-sectional and comparative’ (Ai & Lu, 2013; Estaji & Montazeri, 2022; Fujii & Kim, 2022). This preference can be ascribed to the prevalence of quantitative methodologies in corpus linguistics at large. In this domain, scholars often consider corpora as reflections of particular language variations and employ collective data to make broader generalizations beyond individual language users. Nonetheless, it's important to acknowledge that learner data often demonstrate considerable variability among different learners. As a result, researchers typically introduce control measures to manage this variability and make meaningful comparisons, effectively creating conditions that resemble experimental settings (Wulff & Gries, 2021)

Combining Learner Corpora and Experimental Design

In the evolving landscape of SLA and LCR, SLA researchers are increasingly incorporating LCR methodologies into their studies. This can largely be attributed to the fact that, as Ellis (2017) argues, investigating a complex phenomenon such as language acquisition requires drawing on insights, tools, and methodologies from multiple disciplines. These multiple techniques are particularly advantageous for researchers employing an analytic-deductive approach with a focused hypothesis to test. In such research designs, controlling numerous learner and contextual variables is often challenging, if not unfeasible, in non-experimental settings. Consequently, elicited data becomes a valuable resource, particularly in scenarios where certain constructs may be underrepresented. Therefore, it is not a surprise that Ellis (2017) highlights the scarcity of studies that effectively integrate multiple approaches.

Traditional SLA research has typically favoured (quasi-)experimental data obtained primarily through elicitation methods (Lozano & Mendikoetxea, 2013). Elicitation tasks are designed to encourage informants to generate particular linguistic elements without explicitly disclosing the research goals, thus preserving a naturalistic setting. When paired with corpus data, experimental data can supplement corpus findings, enhancing our overall comprehension of language acquisition processes (Rebuschat et al., 2017). This integration of experimental and corpus methodologies enhances the depth and reliability of SLA research findings, offering valuable insights into the complexities of language learning. References to corpora are prevalent in experimental linguistics and psycholinguistics, where researchers employ corpus information to different extents, ranging from complete exclusion to directly matching item types (Abbuhl et al., 2018; Börner et al., 2019).

There is no rigid dichotomy between corpora and experimental data. Instead, researchers propose a spectrum of naturalness, ranging from less controlled data elicitation methods to more controlled ones (Gilquin & Gries, 2009; Lozano & Mendikoetxea, 2013). Gilquin (2020) suggests that the integration of learner corpora and experimentation reflects the increasing interest in methodological pluralism, also known as multimethod or mixed-methods approaches. This approach advocates for triangulation, wherein multiple methodologies are employed to examine the same phenomenon. When these methodologies yield consistent findings, it indicates converging evidence (Gries et al., 2005; Römer et al., 2014). In SLA research, such convergence is especially valuable because it connects findings from controlled experimental settings with evidence drawn from more naturalistic language use, thereby strengthening both explanation and interpretation. When similar results emerge from multiple approaches, they are more plausibly interpreted as reflecting genuine characteristics of learner language rather than methodological artefacts.

A growing number of second language (L2) researchers are recognizing the immense potential of utilizing corpora. Meunier and Littre (2013) utilise a combination of corpus and experimental data to evaluate the use and comprehension of the simple and continuous present tense among French native speakers learning English as a foreign language (EFL). Initially, they present findings from an interpretation task, revealing that participants displayed a stronger preference for more typical uses of the tenses and aspects under investigation. Learners exhibited greater confidence in responding to these prompts. In the subsequent part of their research, they analyse written learner corpus data. The results indicate that students at the upper-intermediate to advanced levels continue to make errors related to fundamental functions of the simple and progressive tenses.

In a different study, Callies (2009) explored the performance of German learners of EFL in terms of their production and comprehension of ways to highlight information through lexico-syntactic means. This research employed a triangulated approach, utilizing data from written learner corpora, experimental setups, and retrospective interviews to validate its findings. The results uncovered a notable prevalence of subject-prominent structures and a deficiency in certain lexico-grammatical devices aimed at focusing attention on learner-produced data. Furthermore, retrospective interviews indicated that advanced learners were not consciously aware of syntactic methods for highlighting information, unlike their awareness of lexical strategies such as intensifiers.

Research in LCR has also broadened to encompass spoken data. For example, Spina et al. (2025) examine patterns of adjective intensification in written L2 Italian produced by upper secondary school students in South Tyrol, comparing learner output with that of young native Italian speakers. Adopting a Diasystematic Construction Grammar framework, the study investigates how factors such as learners’ first language backgrounds, proficiency levels, and surrounding linguistic environments shape the selection and use of different intensifying constructions. Through quantitative corpus analysis complemented by qualitative examination of constructions, the study shows that learners’ choices of intensifiers are shaped not only by proficiency level but also by their L1 backgrounds and by the multilingual environment in which Italian is acquired and used.

Based on this literature, many LCR-based studies tend to foreground learner performance, particularly as observable output following an intervention or experimental condition. Development is therefore inferred from shifts in observable performance relative to earlier output, with competence remaining an implicit rather than explicitly measured construct. This change in measure makes progress, variability, and partial use of new forms observable without making strong claims about hidden competence. This focus also strengthens ecological validity, as it shows whether learning affects real language use, which fits well with usage-based and dynamic perspectives on SLA.

Another notable observation is that many of these studies adopt a mixed-methods approach. Learner corpus data, whether naturally occurring or elicited, are often analysed alongside other data sources to produce converging evidence. This practice reflects an emphasis on methodological pluralism and serves a triangulation function, allowing findings to be interpreted more robustly across multiple forms of evidence, as illustrated in the study by Callies (2009). In Spina et al. (2025), corpus data are used as a means of interpreting and accounting for patterns identified in the study, rather than serving a purely descriptive role. By comparing learner output with that of native speakers, the corpus evidence provides a concrete basis for interpreting why certain intensifying constructions are preferred or avoided. In this way, the learner corpus functions explanatorily in explaining variation observed across learner groups.

Research Question

Following the review above, the paper seeks to answer the following questions;

1. What are the types of data collected in combined experimental and corpus methodology?
2. How are the data collected in the combined experimental and corpus methodology analysed?

METHODOLOGY

Identification

In this review, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Page et al., 2021) methodology was employed, comprising three primary phases: identification, screening, and eligibility. These phases were utilized to select suitable publications for inclusion in the analysis. The first step of the systematic review process is to identify keywords and search for related, comparable terms using terms used in previous research. There are two search strings to retrieve the articles related to LCR in SLA research. It was searched separately as the search result differs for both strings. Table 1 shows the strings used to retrieve the articles, which resulted in 43 and 51 related articles from the Scopus database, respectively.

Table 1.
The Search Strings and Results

	SEARCH STRING	**RESULTS**
Scopus	**TITLE-ABS-KEY** ("corpus-based OR corpus-driven") AND (“SLA" OR "LCR”) AND (“ESL")	**101**
	TITLE-ABS-KEY ("Learner corpus" OR "LCR”) AND (“SLA" OR "Second Language Acquisition) AND (“ESL")	84
	Date: 06 March 2024	

By using auto tools available on the database, the review is limited to research articles published between 2014 and 2024. This period was chosen because LCR began to attract wider recognition during these years, particularly after the publication of *The Cambridge Handbook of Learner Corpus Research*, which marked an important point in the consolidation of the field. Research articles from a psycholinguistics journal are also excluded (n=76).

Screening

In the initial screening phase, 15 duplicate papers were excluded to ensure the integrity of the dataset. Recognizing the pivotal role of research articles in providing actionable insights, they were given precedence as the primary criterion for inclusion. Subsequently, during the second screening phase, articles under different categories, namely systematic reviews, general reviews, meta-analyses, and meta-syntheses, were omitted, causing the exclusion of 10 articles. This step aimed to maintain the focus on original research articles with empirical data. Following this refinement, an additional screening process was implemented to filter out articles that did not align with the corpus-based approach, relied on textbooks as primary sources of data, or were not specifically geared toward ESL/EFL contexts. This rigorous selection process guaranteed the inclusion of only pertinent and relevant articles in the final analysis, thereby bolstering the validity and credibility of the study's findings.

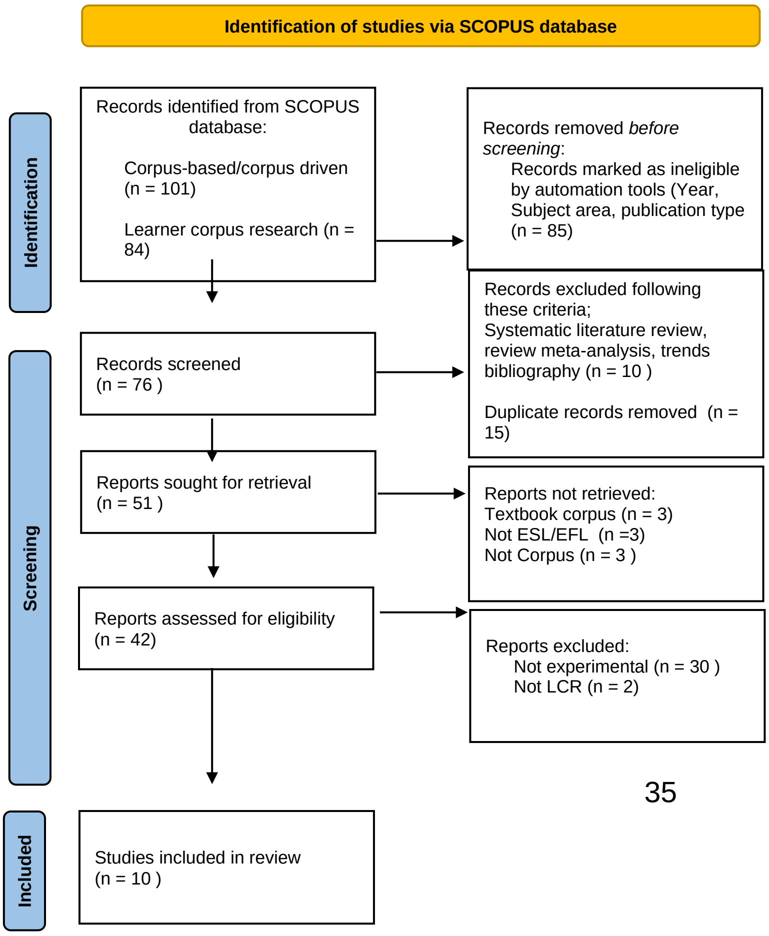
Eligibility

At this juncture, a total of 51 research articles have been identified for retrieval, marking the commencement of the eligibility phase. Each article title and its pertinent content on LCR in SLA were meticulously evaluated to ensure alignment with the study's inclusion criteria. Specifically, articles were deemed eligible if the research was framed within LCR and SLA theories and incorporated some form of experimental design. Consequently, 32 papers were excluded due to a lack of experimental design (n=30) or because they did not employ LCR methodology (n=2). Ultimately, 10 publications remained eligible for further review (refer to Table 3).

Data Abstraction and Analysis

In this study, integrative analysis was employed as one of the assessment methods for examining different research designs (quantitative, qualitative, and mixed methods) and methodological approaches to LCR. The main objective is to identify the methodology in the experimental use of learner corpora. Themes were formulated according to these attributes. Figure 2 in the paper demonstrates the thorough examination of 10 publications for pertinent statements or content on the research inquiries. A log was kept during the data analysis phase to record relevant analyses, perspectives, and additional considerations. To address potential inconsistencies in the theme design process, the results were compared. In case of disagreements between concepts, discussions among the authors took place. Themes were refined to ensure consistency. For validation, examinations were conducted by two experts, one specialized in Applied Linguistics and the other in English as Second Language (ESL). Domain validity was established to ensure clarity, importance, and sufficiency of each theme. Adjustments based on the author's discretion were made in response to feedback and comments from experts.

Figure 1.
Identification of Studies via SCOPUS Database


Identification of studies via SCOPUS database
Identification
Screening
Included
Records identified from SCOPUS database:
Corpus-based/corpus driven (n = 101)
Learner corpus research (n = 84)
Records removed *before screening*:
Records marked as ineligible by automation tools (Year, Subject area, publication type (n = 85)
Records screened (n = 76)
Records excluded following these criteria;
Systematic literature review, review meta-analysis, trends bibliography (n = 10)
Duplicate records removed (n = 15)
Reports sought for retrieval (n = 51)
Reports not retrieved:
Textbook corpus (n = 3)
Not ESL/EFL (n =3)
Not Corpus (n = 3)
Reports assessed for eligibility (n = 42)
Reports excluded:
Not experimental (n = 30)
Not LCR (n = 2)
Studies included in review (n = 10)

RESULTS

As illustrated in Table 2, the integration of learner corpus and experimental methods remains limited, despite its acknowledged advantages. This pattern is evident in the modest number of publications that draw on both data sources over a 10-year period. Since the review period began in 2014, only a limited set of studies appeared in that initial year. It is also worth noting that 2014 coincides with the publication of *The Cambridge Handbook of Learner Corpus Research*, which may have raised awareness of LCR and encouraged its wider adoption, and established a trend that has continued in the years that followed.

Table 2.
Overview of Learner Corpus–Based SLA Studies Reviewed

No.	Study (Title & Year)	SLA Focus / Theory	Participants / Corpus	Method	Data Type	Analysis & Tools
1	Automated assessment of learner text complexity (2021)	NLP / Complexity	Russian learners of English (REALEC corpus)	Quantitative	CES	Frequency analysis; ANOVA
2	Learning to evaluate through that-clauses (2019)	Complexity	158 Chinese EFL learners; timed argumentative essays	Qualitative	CES	Descriptive statistics (frequency)
3	Syntactic and lexical development in an intensive EAP programme (2015)	Complexity	Chinese L1 undergraduates & postgraduates	Quantitative	CES	Wilcoxon signed-rank test; Coh-Metrix; Synlex analyzers
4	The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality (2016)	NLP / Cohesion	57 University EAP students	Quantitative	CES	ANOVA; Coh-Metrix; TAACO indices
5	The dynamics of monthly growth rates in the emergence of complexity, accuracy, and	CAF	The Written English Developmental Corpus of Polish Learners (WEDCPL)	Quantitative	CES	t-tests; Synlex analyzers

	fluency in L2 English writing at secondary school–a learner corpus analysis (2022)					
6	The role of asynchronous computer-mediated communication in the instruction and development of EFL learners' pragmatic competence (2015)	Pragmatics	Iranian EFL learners (n = 27)	Mixed-methods	NOS	DCT (Discourse Comprehension Task) Descriptive statistics; ANCOVA
7	The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study (2015)	Formulaicity	L2 English learners (multiple L1s)	Quantitative	CES	Paired t-tests; correlations; manual coding (SPSS)
8	Using corpus analysis to extend experimental research: Genre effects in L2 writing (2021)	CAF	Gachon Learner Corpus v2.1	Quantitative	CES	ANOVA; Coh-Metrix; lexical diversity measures (SPSS)
9	Learning to evaluate through that-clauses:	CAF	158 Chinese EFL learners; longitudinal essays	Mixed-methods	CES	Descriptive statistics (frequency)

	Evidence from a longitudinal learner corpus (2019)					
10	Can explicit instruction of formulaic sequences enhance L2 oral fluency?(2021)	Fluency	Freshmen spoken learner corpus	Quantitative	CES / EES	ANCOVA; Spearman correlation

Note. NOS = Naturally Occurring Data; CES = Clinically Elicited Data; EES = Experimentally Elicited Data.

RQ 1: What Are the Types of Data Collected in Combined Experimental and Corpus Methodology?

In the context of this review, these studies employ experimental designs that incorporate both learner corpus data and some measure of experimental condition. As a result, all data generated within these studies, regardless of their specific form or source, can be classified as experimental in nature. This fluidity in classification is acknowledged by scholars such as Gilquin and Gries (2009), who suggest that what may be considered a corpus by some researchers could be viewed as experimental data by others, and vice versa. This perspective underscores the complex interdisciplinary nature of research methodologies in the field of language acquisition.

In general, learner corpus data may be drawn from learners with either diverse first language (L1) backgrounds or a single shared L1. However, most LCR tends to focus on data collected from learners with a single L1 background, as evidenced in Table 3. This preference largely reflects the need to control for L1 influence as a key variable, particularly in studies with an experimental design.

For more detailed classification, the data reviewed are organised into three categories: naturally occurring data (NOS), clinically elicited data (CES), and experimentally elicited data (EES). As noted by Ellis (2008), SLA research frequently draws on multiple types of data, including learner language samples, clinically elicited tasks, and experimentally elicited measures, reflecting the inherently multifaceted nature of language acquisition research.

NOS are characterised by minimal researcher intervention and high ecological validity, as they capture learners' spontaneous language use in authentic communicative contexts; however, this authenticity comes at the cost of limited experimental control and an inability to support causal claims (Biber et al., 1998; Granger, 2008). CES occupy an intermediate position, as they are generated through guided tasks designed to prompt specific linguistic features while still allowing relatively free production, thereby striking a balance between naturalness and analytical focus (Ellis, 2003; Gass et al., 2011). In contrast, experimentally elicited data are produced under highly controlled conditions intended to isolate particular variables, making them well-suited for hypothesis testing and causal inference, although their artificiality may reduce their representativeness of real-world language use (Ellis, 2005; Gilquin & Gries, 2009). Within the LCR framework, learner corpus data typically fall into the first two categories, namely NOS and CES, as outlined by Granger (2011).

Table 2 shows that most of the datasets reviewed consist of CES, a pattern that reflects Granger's (2011) observation about the typical makeup of learner corpora. The CES aligns well with the experimental orientation of SLA research, and the majority of the studies reviewed seek to control variables such as task type, topic, or proficiency level while still collecting extended learner output. By contrast, only one study relies on genuinely naturally occurring data (NOS), drawing on email communication as its primary data source. This limited use can be attributed to the practical and ethical challenges associated with collecting and managing NOS. Concerns related to privacy and informed consent, uneven and unpredictable data distribution, and limited comparability across learners make such data less feasible for systematic analysis.

There is also a single case in which CES is combined with experimentally elicited samples (EES). Although EES are not usually regarded as part of learner corpora, they represent a distinct form of data characterised by a high level of experimental control. In such settings, learners are guided to produce particular linguistic forms through tightly structured activities, including gap-filling or sentence translation tasks. Of the two subtypes of experimentally elicited samples, the study reviewed here uses L2 elicitation data (L2 ED), which are generated through experimental tasks designed to prompt the production of specific linguistic elements, such as words, phrases, constructions, or sentences. This pattern emphasises a methodological orientation in experimental LCR that favours ecological relevance alongside experimental control and a clear preference for production data that more closely resemble actual language use, possibly because they offer less insight into how learning is realised in learners' developing language.

RQ 2: How Are the Data Collected in the Combined Experimental and Corpus Methodology Analysed?

LCR can incorporate both quantitative and qualitative analytical approaches. However, one of its key strengths lies in the ability to quantify developmental measures, such as linguistic complexity, through a range of computerized tools. These automated measures offer clear advantages in terms of speed, consistency, and objectivity. At the same time, they have limitations, particularly in capturing broader proficiency gains, formulaic language use, or accuracy-related features. As a result, most studies rely on automated corpus analyses to measure select SLA constructs and adopt predominantly quantitative research designs.

The predominance of quantitative research designs has led many studies to rely on statistical techniques that test for significant differences across groups. These include analyses that compare the means of two or more groups, such as ANOVA, as well as analyses that examine adjusted means while statistically controlling for relevant covariates, such as ANCOVA. In addition, many studies report frequency patterns using descriptive statistics to provide an overview of learners' language use.

Given the tendency of these studies to examine the full complexity–accuracy–fluency (CAF) framework, or selected components of it, two of the most commonly used analytical tools are SynLex and Coh-Metrix. SynLex-type tools are commonly used to analyse syntactic features such as clause length, degrees of subordination, and phrasal expansion, as well as lexical features including density, sophistication, and diversity. These measures offer insight into how learners' grammatical and lexical resources develop over time. The SynLex tool is often combined with Coh-Metrix because they target different yet related dimensions of learner language. While SynLex focuses on syntactic and lexical complexity, Coh-Metrix examines cohesion and broader discourse features, and together they provide a more rounded account of L2 development. It combines information from various language resources, including LSA, the MRC Psycholinguistic Database, WordNet, and word frequency indices like CELEX, to extract linguistic, psychological, and semantic attributes from text (McNamara et al., 2014). Furthermore, Coh-Metrix integrates syntactic parsers into its functionality to enhance its analytical capabilities (Crossley et al., 2016; McNamara et al., 2017).

A key feature of LCR is its use of NLP techniques to identify and calculate linguistic features from learner texts. NLP allows language to be represented computationally and analysed automatically through procedures such as tokenisation, part-of-speech tagging, lemmatisation, and syntactic parsing. Once texts are processed, features relating to grammar, vocabulary, cohesion, and discourse can be measured systematically. This automation enables researchers to work with large datasets efficiently and consistently, making it possible to observe patterns in learner language that would be difficult to capture through manual analysis alone. In one of the reviewed studies, NLP-based tools were employed to extract cohesion- and discourse-related indices from learner texts using Coh-Metrix and TAACO. The resulting numerical measures were then subjected to ANOVA analysis to examine whether statistically significant differences existed between groups. This complex procedure enabled the researchers to assess group-level variation in discourse organisation and cohesion patterns in learner writing, providing quantitative evidence of how instructional or learner-related factors influenced text-level features.

Most of the studies reviewed rely primarily on quantitative methods, although a small number adopt a mixed-methods design. For instance, Man and Chau (2019) explored the use of *that*-clauses by examining concordance lines from a longitudinal learner corpus, using qualitative evidence to shed light on quantitative patterns. In a similar vein, Eslami et al. (2015) combined corpus-informed analysis with a discourse completion test (DCT) to examine the impact of asynchronous computer-mediated communication. Although such approaches require additional time and analytical effort, they allow researchers to draw on different types of evidence and arrive at more nuanced interpretations than would be possible with a single data source. This pattern echoes observations by Callies (2015b) and Paquot et al. (2017), who note that learner corpus research has so far been largely shaped by corpus-based, quantitative procedures.

DISCUSSION

Experimenting with SLA Constructs through Learner Corpus Research

The increasing theoretical convergence between LCR and SLA, as discussed earlier, has led to a reciprocal exchange of methodologies between the two fields. LCR now incorporates experimental designs and borrows theories or constructs from SLA. Among the underlying constructs and theories analysed in the articles, one that stands out is the complexity construct, often coupled with accuracy and fluency. This is because the performance, proficiency, and development of L2 learners are frequently evaluated based on these three fundamental dimensions, as proposed by Bui and Skehan (2018). While they can be considered distinct constructs, they are often viewed as components of the CAF triad (Craven, 2017; Housen & Kuiken, 2009; Thewissen, 2023). Accordingly, when learner performance and proficiency are the primary outcomes of interest, the CAF framework provides a suitable analytical lens.

As separate constructs, complexity tends to be favoured because it lends itself well to systematic and large-scale analysis. Many aspects of complexity can be captured using automated tools, such as the Synlex analysers, which makes it feasible to analyse sizeable datasets and to trace patterns across proficiency levels or over time. In comparison, examining accuracy usually involves detailed error coding and normative judgments that are both time-consuming and methodologically demanding. Fluency poses a different challenge, as it is difficult to operationalise in written corpora and depends largely on temporal information that is more readily available in spoken data. As a result, both accuracy and fluency are less frequently examined than complexity in corpus-based SLA research.

A review of these studies also reveals a recurring pattern in how the strengths of LCR are commonly conceptualised in SLA research. Focusing on performance-based corpus data allows researchers to see how intervention effects play out in learners' actual language use, rather than inferring change from abstract measures of competence. This perspective helps establish ecological validity by showing whether gains observed under controlled conditions carry over into more natural language production. In many studies, learner corpora are combined with other forms of data, which makes it possible to triangulate findings across methods. This is because even under controlled conditions, corpus data remain primarily outcome-oriented, capturing learners' linguistic production rather than the cognitive processes underpinning acquisition. As noted by Myles (2020), learner corpora, whether naturalistic or experimentally elicited, cannot independently provide direct evidence of mechanisms such as noticing, hypothesis testing, or restructuring. For this reason, experimentally generated learner corpora are best conceptualized as hybrid datasets that combine experimental rigor with ecological validity, functioning most effectively within plural-method research designs.

Viewed this way, LCR methodology offers a more cautious and well-grounded basis for interpretation. These observations are consistent with the patterns identified in the literature review, which indicate that LCR is frequently deployed in a functionally targeted manner aligned with specific research objectives. Across studies, learner corpus data may be used to account for patterns emerging from other data sources, to complement or extend existing datasets, or to validate findings obtained through experimental or assessment-based measures. In summary, combining experimental designs with learner corpora fulfils three principal roles: it helps explain observed effects, assesses the transfer of findings to authentic language use, and enables triangulation across complementary methods.

Table 3.
Functional Roles of LCR in SLA

Explanatory Function	Validation Function	Triangulation Function
Explains how and why experimental effects occur. • Reveals underlying linguistic mechanisms • Examines patterns, productivity, and development • Focuses on process-oriented interpretation	Examines whether experimental gains transfer to authentic language use. • Tests ecological validity • Observes spontaneous use in production data • Assesses durability and generalisation	Integrates corpus and experimental evidence. • Strengthens interpretive validity • Identifies converging or diverging findings • Supports methodological pluralism

Table 3 presents the three main ways in which LCR is typically used in SLA studies. In an explanatory capacity, corpus data help shed light on the linguistic processes that lie behind experimentally observed outcomes. When used for validation, learner corpora provide a means of checking whether gains identified through experimental measures are reflected in learners' more natural language use. Finally, in a triangulation role, corpus evidence is considered alongside experimental findings, allowing results to be compared and interpreted across methods and thereby strengthening the overall robustness of the analysis.

Theoretical Implications for Experimental Learner Corpus Research in SLA

The integration of LCR into experimental designs has important theoretical implications for SLA, particularly in how language development is conceptualised and evidenced. As discussed earlier, LCR and traditional SLA research have often prioritised different constructs, namely performance and competence. Bringing these approaches together challenges a rigid competence–performance dichotomy by highlighting the theoretical relevance of learner performance in understanding language development. Traditional experimental tasks tend to focus on abstract linguistic knowledge elicited in highly controlled settings, whereas learner corpus data draw attention to how language is actually used across extended stretches of discourse. Research that combines corpus evidence with experimental methods shows that learner performance should not simply be regarded as random fluctuation or error. Patterns in what learners actually produce, such as tentative use of new forms, variability across contexts, and gradual refinement over time, offer valuable insight into how linguistic knowledge develops. From this perspective, performance is not a distorted version of competence but a place where emerging competence can be observed in action. This view aligns with dynamic and usage-based approaches to SLA, which see competence as shaped through use and experience rather than as a fixed, underlying system (Ellis, 2008; Granger, 2008).

Secondly, experimental learner corpus research encourages a rethinking of what is validity in SLA research. When controlled experimental measures are considered alongside corpus evidence of how learners actually use language, validity is no longer confined to test outcomes alone but extends to whether learning is considered in communicative practice. In this way, experimental LCR supports a more multidimensional understanding of validity that links measured gains to authentic language use (Granger et al., 2015; Vyatkina, 2016).

Identified Gaps and Recommendations for Future Studies

The scope of this review is limited in two main respects. First, it focuses exclusively on studies of L2 English, rather than second language learning across a wider range of target languages. Second, the review is confined to a relatively narrow publication window, spanning from 2015 to the present. Although these boundaries ensure a focused and current overview, they also constrain the breadth of the findings. Future studies could address this gap by examining learner corpus research involving other languages and by extending the time frame to capture longer-term methodological and theoretical trends.

CONCLUSION

The current systematic review of literature has systematically tackled the primary research question regarding contemporary trends and practices in the analysis of LCR data within the SLA experimental paradigm. This was accomplished by examining ten articles published between 2015 and March 2024. These studies draw attention to the role played by learner characteristics, register, and task conditions, which has encouraged a more cautious interpretation of L2 complexity as an analytical construct. Through such work, researchers have been able to examine L2 complexity in greater detail, particularly in terms of how it evolves over time. The outcomes of this review offer an updated synthesis of LCR within the scope of SLA, encompassing methodologies and procedures for data analysis. With the convergence of LCR and SLA, more scholars will likely exceed methodological limitations and forge fresh connections between these fields as suggested by Gilquin (2021). Vyatkina and Housen (2021) also note that more studies are now using learner corpus tools to address questions that matter to SLA. Closer interaction between LCR and SLA has made this kind of methodological blending increasingly possible and has the potential to bring the two traditions into closer theoretical alignment. Even so, this integration is still far from complete, and there remains considerable room for LCR to be more fully embedded in SLA research.

REFERENCES

Abbuhl, R., Gass, S., Mackey, A., Podesva, R., & Sharma, D. (2018). Experimental research design. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 116–134). Cambridge University Press. https://doi.org/10.1017/CBO9781139013734.008

Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. *Automatic Treatment and Analysis of Learner Corpus Data*, 249–264. https://doi.org/10.1075/scl.59.15ai

Alexopoulou, D., Michel, M., Murakami, A., & Meurers, D. (2017). Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques [Research report]. https://doi.org/10.17863/CAM.7512

AlHassan, L., & Wood, D. (2015). The effectiveness of focused instruction of formulaic sequences in augmenting L2 learners' academic writing skills: A quantitative research study. *Journal of English for Academic Purposes*, *17*, 51–62. https://doi.org/10.1016/j.jeap.2015.02.001

Ali, R. (2023). *Implementation of blended learning in higher education: A case study of adoption and diffusion* (Doctoral thesis, University of Wollongong). University of Wollongong Research Online. https://hdl.handle.net/10779/uow.27666369.v1

Bell, P., & Payant, C. (2020). Designing learner corpora: Collection, transcription, and annotation. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 53–67). Routledge.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Bui, G., & Skehan, P. (2018). Complexity, accuracy, and fluency. *The TESOL Encyclopedia of*

English Language Teaching, *8*, 1-7.

Bybee, J. (2008). Usage-based grammar and second language acquisition. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 216–236). Routledge.

Callies, M. (2009). “What is even more alarming is …”: A contrastive learner-corpus study of what-clefts in advanced German and Polish L2 writing. In M. Wysocka (Ed.), *On language structure, acquisition and teaching: Studies in honour of Janusz Arabski on the occasion of his 70th birthday* (pp. 283–292). Wydawnictwo Uniwersytetu Śląskiego.

Callies, M. (2015a). Using learner corpora in language testing and assessment: Current practice and future challenges. In E. Castello, K. Ackerley, & F. Coccetta (Eds.), *Studies in learner corpus linguistics: Research and applications for foreign language teaching and assessment* (pp. 21–35). Peter Lang.

Callies, M. (2015b). Learner corpora in language testing and assessment: Prospects and challenges. In M. Callies & S. Götz (Eds.), *Learner corpora in language testing and assessment* (Studies in Corpus Linguistics, Vol. 70, pp. 1–9). John Benjamins Publishing Company. https://doi.org/10.1075/scl.70.01cal

Callies, M., & Paquot, M. (2015). Learner Corpus Research: An interdisciplinary field on the move. *International Journal of Learner Corpus Research*, *1*(1), 1–6. https://doi.org/https://doi.org/10.1075/ijlcr.1.1.00edi

Chambers, A. (2019). Towards the corpus revolution? Bridging the research–practice gap. *Language Teaching*, *52*, 1–16. https://doi.org/10.1017/S0261444819000089

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing*, *32*, 1–16. https://doi.org/10.1016/j.jslw.2016.01.003

Crosthwaite, P., & Baisa, V. (2024). A user-friendly corpus tool for disciplinary data-driven learning: Introducing CorpusMate. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.23056.cro

Czerwionka, L., & Olson, D. (2020). Pragmatic development during study abroad: L2 intensifiers in spoken Spanish. *International Journal of Learner Corpus Research*, *6*. https://doi.org/10.1075/ijlcr.19006.cze

De Knop, S., & Meunier, F. (2015). The learner corpus research, cognitive linguistics and second language acquisition nexus: A SWOT analysis. *Corpus Linguistics and Linguistic Theory, 11*(1), 1–18. https://doi.org/10.1515/cllt-2014-0012

Díez-Bedmar, M. B. (2020). Error analysis. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 90–104). Routledge.

Ellis, N. C. (2017). Salience in usage-based SLA. In B. MacWhinney & W. O’Grady (Eds.), *The handbook of language emergence* (pp. 21–40). Routledge. https://doi.org/10.4324/9781315399027-2

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: A psychometric study. *Studies in Second Language Acquisition, 27*(2), 141–172.

Ellis, R. (2008). Investigating grammatical difficulty in second language learning: Implications for second language acquisition research and language testing. *International Journal of Applied Linguistics*, *18*(1), 4–22. https://doi.org/10.1111/j.1473-4192.2008.00184.x

Ellis, R. (2020). A short history of SLA: Where have we come from and where are we going? *Language Teaching*, *54*, 1–16. https://doi.org/10.1017/S0261444820000038

Eslami, Z. R., Mirzaei, A., & Dini, S. (2015). The role of asynchronous computer-mediated communication in the instruction and development of EFL learners’ pragmatic competence. *System*, *48*, 99–111. https://doi.org/10.1016/j.system.2014.09.008

Estaji, M., & Montazeri, M. R. (2022). Native English and non-native authors’ utilisation of lexical bundles: A corpus-based study of scholarly public health papers. *Southern African Linguistics and Applied Language Studies*, *40*(2), 177–199. https://doi.org/10.2989/16073614.2022.2043169

Fernández-Mira, P., Morgan, E., Davidson, S., Yamada, A., Carando, A., Sagae, K., & Sánchez-

Gutiérrez, C. (2021). Lexical diversity in an L2 Spanish learner corpus: The effect of topic-related variables. *International Journal of Learner Corpus Research*, *7*, 230–258. https://doi.org/10.1075/ijlcr.20017.fer

Fuchs, R., & Werner, V. (2018). Tense and aspect in Second Language Acquisition and learner corpus research: Introduction to the special issue. *International Journal of Learner Corpus Research*, *4*, 143–163. https://doi.org/10.1075/ijlcr.00004.int

Fujii, Y., & Kim, M. (2022). A cross-sectional corpus-based study on the acquisition of the English definite article: A preliminary study of L1 Korean/Chinese learners of English. *Data Science in Collaboration, 5*, 112–119.

Gass, S., Mackey, A., & Pica, T. (1998). The role of input and interaction in second language acquisition. *The Modern Language Journal,* *82*(3), 299–307. https://doi.org/10.1111/j.1540-4781.1998.tb01206.x

Gilquin, G. (2020). Learner corpora. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 283–303). Springer. https://doi.org/10.1007/978-3-030-46216-1_1

Gilquin, G. (2021). Using corpora to foster L2 construction learning: A data-driven learning experiment. *International Journal of Applied Linguistics, 31*(2), 229–247. https://doi.org/10.1111/ijal.12317

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

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* (pp. 37–51). Lund University Press.

Granger, S. (2008). Learner corpora in foreign language education. In N. H. Hornberger (Ed.), *Encyclopedia of language and education* (2nd ed., Vol. 6, pp. 1427–1441). Springer. https://doi.org/10.1007/978-0-387-30424-3_109

Granger, S. (2011). How to use Foreign and Second Language Learner Corpora. In *Research Methods in Second Language Acquisition* (pp. 5–29). Wiley. https://doi.org/10.1002/9781444347340.ch2

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Gries, S., Hampe, B. & Schönefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, *16*(4), 635-676. https://doi.org/10.1515/cogl.2005.16.4.635

Gries, S. Th. (2009). What is Corpus Linguistics? *Language and Linguistics Compass*, *3*(5), 1225–1241. https://doi.org/10.1111/j.1749-818X.2009.00149.x

Gut, U. (2007). Learner corpora in second language prosody research and teaching. In J. Trouvain & U. Gut (Eds.), *Phonetic description and teaching practice* (pp. 145–170). De Gruyter Mouton. https://doi.org/doi:10.1515/9783110198751.1.145

Hasko, V. (2012). Qualitative corpus analysis. In *The Encyclopedia of Applied Linguistics* (pp. 1-7). Wiley-Blackwell.

Hasselgård, H., & Johansson, S. (2011). Learner corpora and contrastive interlanguage analysis. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *Learner corpora in contrastive linguistics* (Studies in Corpus Linguistics, Vol. 45, pp. 33–62). John Benjamins Publishing Company. https://doi.org/10.1075/scl.45

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in Second Language Acquisition. *Applied Linguistics*, *30*(4), 461–473. https://doi.org/10.1093/applin/amp048

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world* (Vol. 1, pp. 91–118). Kobe University.

Iurato, A. (2022). Learner Corpus Research meets Chinese as a Second Language Acquisition: Achievements and challenges. *Annali Di Ca' Foscari. Serie Orientale*. https://doi.org/10.30687/AnnOr/2385-3042/2022/01/024

Kersten, K., & Greve, W. (2022). Investigating cognitive-linguistic development in SLA:

Theoretical and methodological challenges for empirical research. In *Understanding variability in second language acquisition, bilingualism, and cognition* (pp. 3-38). Routledge.

Klag, M., & Langley, A. (2013). Approaching the conceptual leap in qualitative research. *International Journal of Management Reviews*, *15*(2), 149–166. https://doi.org/10.1111/j.1468-2370.2012.00349.x

Kyle, K. (2021). Natural language processing for learner corpus research. *International Journal of Learner Corpus Research, 7*(1), 1–16. https://doi.org/10.1075/ijlcr.00019.int

Larsson, T., Plonsky, L., & Hancock, G. (2022). On learner characteristics and why we should model them as latent variables. *International Journal of Learner Corpus Research*, *8*, 237–260. https://doi.org/10.1075/ijlcr.21007.lar

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated grammatical error detection for language learners*. Morgan & Claypool Publishers.

Lee, J. (2021). Using corpus analysis to extend experimental research: Genre effects in L2 writing. *System*, *100*. https://doi.org/10.1016/j.system.2021.102563

Lindquist, H., & Levin, M. (2018). *Corpus linguistics and the description of English*. Edinburgh University Press. https://doi.org/10.1515/9781474421713

Lozano, C., & Mendikoetxea, A. (2013). Learner corpora and second language acquisition: The design and collection of CEDEL2. In A. Díaz-Negrillo, N. Ballier, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (Studies in Corpus Linguistics, Vol. 59, pp. 65–100). John Benjamins Publishing Company. https://doi.org/10.1075/scl.59.06loz

Lyashevskaya, O., Panteleeva, I., & Vinogradova, O. (2021). Automated assessment of learner text complexity. *Assessing Writing*, *49*. https://doi.org/10.1016/j.asw.2021.100529

Man, D., & Chau, M. H. (2019). Learning to evaluate through that-clauses: Evidence from a longitudinal learner corpus. *Journal of English for Academic Purposes*, *37*, 22–33. https://doi.org/10.1016/j.jeap.2018.11.007

Mazgutova, D., & Kormos, J. (2015). Syntactic and lexical development in an intensive English for Academic Purposes programme. *Journal of Second Language Writing*, *29*, 3–15. https://doi.org/10.1016/j.jslw.2015.06.004

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus Linguistics, Learner Corpora, and SLA: Employing technology to analyze language use. *Annual Review of Applied Linguistics*, *39*, 74–92. https://doi.org/10.1017/S0267190519000096

McEnery, T., & Gabrielatos, C. (2006). English corpus linguistics. In B. Aarts & A. McMahon (Eds.), *The handbook of English linguistics* (pp. 33–71). Blackwell Publishing. https://doi.org/10.1002/9780470753002.ch3

McNamara, D. S., Allen, L. K., Crossley, S. A., Dascalu, M., & Perret, C. A. (2017). Natural Language Processing and Learning Analytics. In *Handbook of Learning Analytics* (pp. 93–104). Society for Learning Analytics Research (SoLAR). https://doi.org/10.18608/hla17.008

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge University Press.

Meunier, F., & Littre, D. (2013). Tracking learners' progress: Adopting a dual 'corpus cum experimental data' approach. *The Modern Language Journal*, *97*(S1), 61–76. https://doi.org/10.1111/j.1540-4781.2012.01424.x

Mukherjee, J., & Götz, S. (2015). Learner corpora and learning context. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 423–442). Cambridge University Press. https://doi.org/10.1017/CBO9781139649414.019

Myles, F. (2020). Commentary: An SLA perspective on learner corpus research. In M. Callies & S. Götz (Eds.), *Learner corpora and language teaching* (pp. 258–273). Cambridge University Press. https://doi.org/10.1017/9781108674577.013

Nergis, A. (2021). Can explicit instruction of formulaic sequences enhance L2 oral fluency? *Lingua*, *255*. https://doi.org/10.1016/j.lingua.2021.103072

Nesselhauf, N. (2005). *Collocations in a Learner Corpus* (Vol. 14). John Benjamins Publishing

Company. https://doi.org/10.1075/scl.14

Newbery-Payton, L. (2022). A corpus-based analysis of crosslinguistic influence on the acquisition of concessive conditionals in L2 English. *Asian-Pacific Journal of Corpus Research, 3*(1), 35–49. https://doi.org/10.22925/apjcr.2022.3.1.35

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). *The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. PLOS Medicine, 18*(3), e1003583. https://doi.org/10.1371/journal.pmed.1003583

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research, 3*(1). http://hdl.handle.net/2078.1/185993

Pérez-Paredes, P., & Díez-Bedmar, M. B. (2019). Certainty adverbs in spoken learner language. *International Journal of Learner Corpus Research*. https://doi.org/10.1075/ijlcr.17019.per

Plonsky, L. (2013). Study quality in SLA. *Studies in Second Language Acquisition*, *35*. https://doi.org/10.1017/S0272263113000399

Pusey, K., & Butler, Y. G. (2023). Investigating the ecological validity of second language writing assessment tasks. *System, 119*, 103174. https://doi.org/10.1016/j.system.2023.103174

Rebuschat, P., Meurers, D., & McEnery, T. (2017). Language learning research at the intersection of experimental, computational, and Corpus-Based Approaches. In *Language learning* (Vol. 67, pp. 6–13). Blackwell Publishing. https://doi.org/10.1111/lang.12243

Römer, U., Roberson, A., O'Donnell, M. B., & Ellis, N. C. (2014). Linking learner corpus and experimental data in studying second language learners' knowledge of verb-argument constructions. *ICAME Journal*, *38*(1), 115–135. https://doi.org/10.2478/icame-2014-0006

Römer-Barron, U., & Garner, J. (2019). The development of verb constructions in spoken learner English: Tracing effects of usage and proficiency. *International Journal of Learner Corpus Research*, *5*, 207–230. https://doi.org/10.1075/ijlcr.17015.rom

Rokoszewska, K. J. (2022). The dynamics of monthly growth rates in the emergence of complexity, accuracy, and fluency in L2 English writing at secondary school–a learner corpus analysis. *System*, *106*. https://doi.org/10.1016/j.system.2022.102775

Shimada, K. (2014). Contrastive interlanguage analysis of discourse markers used by nonnative and native English speakers. *JALT Journal*, *36*, 47. https://doi.org/10.37546/JALTJJ36.1-3

Smith, S. (2020). DIY corpora for Accounting & Finance vocabulary learning. *English for Specific Purposes*, *57*, 1–12. https://doi.org/https://doi.org/10.1016/j.esp.2019.08.002

Spina, S., Glaznieks, A., & Abel, A. (2025). Intensification in written L2 Italian: Insights from the multilingual region of South Tyrol. *International Journal of Learner Corpus Research, 11*(2), 276–308. https://doi.org/10.1075/ijlcr.23041.spi

Thewissen, J. (2021). Accuracy. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 305–317). Routledge.

Vandeweerd, N., Housen, A., & Paquot, M. (2021). Applying phraseological complexity measures to L2 French: A partial replication study*. *International Journal of Learner Corpus Research*, *7*, 197–229. https://doi.org/10.1075/ijlcr.20015.van

Verbeke, G. (2024). *On the role of ecological validity in language and speech research.* In J. Buysschaert & A. Lefèvre (Eds.), *Taalkunde nu* (pp. 69–95). Skribis.

Volodina, E., Granstedt, L., Matsson, A., Megyesi, B., Pilán, I., Prentice [Grosse], J., Rosén, D., Rudebeck, L., Schenström, C.-J., Sundberg, G., & Wirén, M. (2019). The SweLL Language Learner Corpus: From design to annotation. *The Northern European Journal of Language Technology*, *6*, 67–104. https://doi.org/10.3384/nejlt.2000-1533.19667

Vuuren, S. van, Berns, J., & Bank, M. (2022). Strategies of clausal postmodification in learner English. *International Journal of Learner Corpus Research*, *8*(2), 157–189. https://doi.org/10.1075/ijlcr.21013.vuu

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

Webster, K., Recasens, M., Axelrod, V., & Baldridge, J. (2018). Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, *6*, 605–617. https://doi.org/10.1162/tacl_a_00240

Winter, T., & Le Foll, E. (2022). Testing the pedagogical norm: Comparing If-conditionals in EFL textbooks, learner writing and English outside the classroom. *International Journal of Learner Corpus Research*. https://doi.org/10.1075/ijlcr.20021.win

Woodfield, H. (2008). Interlanguage requests: A contrastive study. In M. Pütz & J. Neff-van Aertselaer (Eds.), *Developing contrastive pragmatics* (pp. 231–264). Mouton de Gruyter. https://doi.org/10.1515/9783110207217.3.231

Wulff, S., & Gries, S. T. (2021). Exploring individual variation in learner corpus research. In S. Gries, & others (Eds.), *Cambridge handbook of learner corpus research* (pp. 401–422). Cambridge University Press.

Ziegler, N., Meurers, D., Rebuschat, P., Ruiz, S., Moreno-Vega, J. L., Chinkina, M., Li, W., & Grey, S. (2017). Interdisciplinary research at the intersection of CALL, NLP, and SLA: Methodological implications from an input enhancement project. *Language Learning*, *67*(S1), 209–231. https://doi.org/https://doi.org/10.1111/lang.12227