


LEARNER CORPORA AND SECOND LANGUAGE ACQUISITION RESEARCH

Mamatkulova Laziza Nodirbekovna

Mamatkulova Laziza Nodirbekovna, Jizzakh, Uzbekistan



Learner corpora have become one of the most powerful tools in contemporary Second Language Acquisition (SLA) research. As collections of authentic language produced by L2 learners, these corpora allow researchers to examine linguistic patterns, developmental sequences, learner errors, and cross-linguistic influence with a level of precision that traditional qualitative methods cannot achieve. The integration of learner corpora into SLA has resulted in new approaches to studying inter language, language pedagogy, and curriculum design. This research article examines the major contributions of learner corpora to SLA, the methodological approaches used in corpus-based studies, and the pedagogical implications of corpus findings. The article reviews key learner corpora such as ICLE, LOCNESS, and the Cambridge Learner Corpus, and highlights how they have shaped contemporary theories of acquisition, especially regarding error analysis, phraseology, and the development of writing proficiency. The findings demonstrate that learner corpora provide reliable evidence for identifying persistent learner difficulties, validating teaching materials, and creating data-driven learning tools. Ultimately, learner corpora support a more empirical and learner-centered approach to SLA, bridging the gap between linguistic theory and real classroom needs.

Keywords: learner corpora, SLA, inter language, corpus linguistics, error analysis

Purpose: The purpose of this article is to explain the concept of learner corpora and their role in SLA research and analyze how corpus-based methods contribute to understanding inter language development.

Methodology: This research article uses a descriptive and analytical approach based on:

A review of major learner corpus projects.

Comparative analysis of SLA theories supported by corpus findings.

Examination of how learner corpora are used for error analysis, phraseology studies, and pedagogical applications.

Synthesis of empirical results from existing corpus-based literature.

1. Introduction

In recent decades, the emergence of learner corpora has transformed SLA research by offering large, systematic, and searchable collections of learner language. This shift towards empirical data allows researchers to observe actual learner performance instead of relying on intuition-based descriptions. Learner corpora reveal developmental patterns, common errors, and cross-linguistic interference, thus supporting a more scientific understanding of inter language.¹

Historical development of learner corpus research

Text Learner corpus research began gaining prominence in the late 1980s and early 1990s, when researchers recognized the limitations of intuition-based descriptions of learner language². Early pioneers such as Granger argued³ that SLA research needed large, computerized collections of authentic learner data to examine real linguistic performance rather than idealized competence. This led to the creation of foundational projects such as the International Corpus of Learner English (ICLE) and the Cambridge Learner, which marked the beginning of systematic empirical investigation into learner production.

¹ (Granger, 1998; Ellis, 2008)

² (Granger, 1998)

³ (Granger, 2002)

During the 2000s and 2010s, the field expanded to include spoken learner corpora, longitudinal corpora, and multilingual datasets, reflecting a methodological shift toward evidence-based SLA. Today, learner corpora are central to research on inter language development, phraseology, error patterns, and cross-linguistic influence, supporting both theoretical and pedagogical advancements.

Importance of learner corpora in SLA research

Text Learner corpora are essential in SLA research because they provide authentic, quantifiable, and replicable evidence of learner language use⁴. Unlike small-scale elicited tasks, corpus data represent real learner performance across contexts and proficiency levels, making the findings more generalizable. These corpora reveal developmental sequences—such as article acquisition, verb morphology patterns, and lexical sophistication—that intuition and classroom observation cannot reliably capture.

They also allow for direct comparison between learner and native-speaker corpora, offering insights into inter language deviations, overuse, underuse, and transfer effects. Additionally, learner corpora support pedagogical applications, including the design of textbooks, assessment rubrics, automated writing evaluation systems, and data-driven learning tasks. Thus, learner corpora bridge theoretical SLA research with real-world teaching and learning practices.

Literature review

Definition and types of learner corpora

A learner corpus is a digital collection of texts produced by L2 learners. Examples include:

Written corpora (essays, exams)⁵

Spoken corpora (interviews, discussions)

⁴ (Ellis, 2008; Larsen-Freeman & Long, 2014)

⁵ (Granger, 2015; Nesselhauf, 2005)

Longitudinal corpora (tracking learner growth over time)

Major learner corpora in SLA

ICLE – International Corpus of Learner English: Advanced EFL argumentative essays.

LOCNESS: Native English corpus used for comparison.

Cambridge Learner Corpus (CLC): Contains learners' exam scripts from multiple proficiency levels.

ELLIPSE, SWEET, NICT JLE Corpus: Spoken learner corpora.⁶

Tab. 1: Learner errors identified by major learner corpora

Corpus name	Learner group	Most frequent errors
ICLE(International Corpus of learner English)	Advanced EFL learners	Articles misuse, lexical errors, collocation problems
Cambridge learner corpus(CLC)	Mixed global proficiency levels	Prepositions, verb tense choice, spelling, word order
Locness (Native control corpus)	Native speakers(control)	Minimal errors used as baseline for comparison
NICT JLE(Japanese learner English)	Japanese EFL learners	Overuse of simple verbs, avoidance of complex structure

Source: Granger et al. (2009); Nicholls (2003); Nesselhauf (2005); Dagneaux, Granger, & Meunier (1998).

SLA Theories supported by learner corpus research

Learner corpora provide evidence for:

⁶ (Granger et al., 2009; Nicholls, 2003)

- Inter language development
- Transfer and cross-linguistic influence
- Frequency effects (usage-based learning)
- Development of L2 phraseological competence⁷

Analysis and discussion

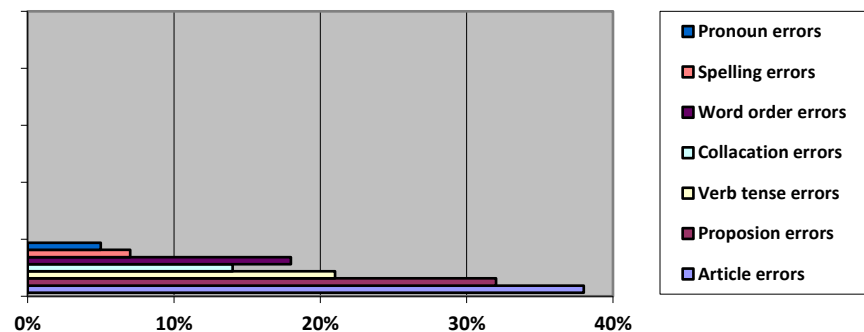
Error analysis and learner difficulties

Learner corpora help identify systematic errors such as:

- Article misuse
- Prepositional errors
- Verb tense choice
- Collocation problems (“strong rain”, “make a photo”)

Corpus evidence shows that these errors persist across proficiency levels, proving that they are not random but cognitively patterned.⁸

Fig. 1: Frequency of error types in learner corpora



Source: Dagneaux, Granger, & Meunier (1998); Nesselhauf (2005); Granger et al. (2009).

L2 Phraseology and formulaic language

Corpus studies demonstrate that learners struggle with:

⁷ (Ellis, 2008; Larsen-Freeman, 2011)

⁸ (Dagneaux, Granger, & Meunier, 1998)

Idiomatic expressions

Academic formulaic sequences

Multi-word verbs (phrasal verbs)

Collocations

These findings support the idea that phraseology is central to fluency and requires explicit teaching.⁹

Cross-linguistic influence

Learner corpora reveal how L1 affects L2 production. For example:

Uzbek learners may omit articles due to L1 article absence.

Japanese learners avoid certain verb tenses due to L1 structural differences.¹⁰

Pedagogical implications

Corpus-based findings contribute to:

Designing textbooks based on real learner errors

Creating data-driven learning (DDL) tasks

Improving assessment rubrics (writing tasks, speaking tasks)

Developing tools such as Grammarly, Write & Improve, and corpus-based vocabulary lists¹¹

Conclusion

Learner corpora have become indispensable in SLA research, offering a reliable empirical foundation for understanding how learners acquire and use language. They reveal systematic patterns of errors, highlight developmental sequences, and support theories of inter language and usage-based learning. Their pedagogical implications have reshaped modern teaching practices, encouraging data-driven learning and more accurate curriculum design. As learner corpora

⁹ (Nesselhauf, 2005; Wray, 2002)

¹⁰ (Jarvis & Pavlenko, 2008)

¹¹ (Boulton & Cobb, 2017)

continue to expand, they will further enhance SLA research and improve language education worldwide.

Acknowledgment

This paper is an output of the science project and reflects the author's effort to analyze the role of learner corpora in Second Language Acquisition research. The author acknowledges the valuable contributions of previous studies and corpus-based projects that provided the theoretical and methodological foundation for this work. Special appreciation is extended to the researchers and institutions responsible for the development of major learner corpora, whose resources and findings significantly supported the preparation of this article. Any errors or limitations remain the responsibility of the author.¹²

References

Granger, S. (1998). *The computer learner corpus: a new resource for foreign language learning and teaching*. *Applied Linguistics*, 19(4), 529–553.

International Corpus of Learner English (ICLE) — project page (UCLouvain / CECL).

Cambridge Learner Corpus (CLC) — Cambridge Research / demo & documentation.

Dagneaux, E., Granger, S., & Meunier, F. (1998). *A learner corpus approach to error analysis*. *International Journal of Corpus Linguistics*, 3(2), 161–182.

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

Boulton, A., & Cobb, T. (2017). “*Corpus use in language learning: A meta-analysis*.” *Language Learning*, 67(2), 348–393.

The Cambridge Handbook of Learner Corpus Research (eds. Granger, Gilquin, Meunier), Cambridge University Press (2015).

¹² (Granger, 2015)