https://doi.org/10.18485/sfi.2026.2.ch21
378.147:811.21/.22

Saeed Safari[1]
University of Belgrade - Faculty of Philology
https://orcid.org/0000-0001-9559-4179
Mahbod Ghaffari[2]
University of Cambridge - Faculty of Asian and Middle Eastern Studies
https://orcid.org/0000-0002-4709-2462

Cambridge-Belgrade (CamBel) Persian Learner Corpus: Theoretical Frameworks and Methodological Insights

Learner corpora offer insights into learners' challenges and requirements and benefit teachers, materials producers, syllabus writers, and researchers by enabling them to meet the diverse needs of learners and improve learning experiences. Assessing the learning needs of students is essential for effectively teaching Persian to non-native speakers, and a valuable approach to identifying these needs is to construct and use specific learner corpora. This paper aims to present and introduce the design and methodology of the Cambridge-Belgrade (CamBel) Persian learner corpus. CamBel is an error-tagged (coded) learner corpus jointly compiled, recorded, and administered through academic research collaboration between Persian Studies at the Faculty of Asian and Middle Eastern Studies, University of Cambridge, and the Center for Persian Studies at the Faculty of Philology, University of Belgrade. The first version of the corpus (CamBel Version 0.1) consists of written texts produced by learners of Persian from around the world, with a particular focus on data collected from Serbian and English-speaking learners of Persian. CamBel adopts an "error-tagged" approach and applies a specific model for error annotations that combines linguistic categories, surface structure taxonomies, and error types. Errors in linguistic categories are categorized into six groups, namely, lexical, morphological, morphosyntactic, syntactic, orthographic, and style. Through the corpus's search and

[1] saeed.safari@fil.bg.ac.rs
[2] mg695@cam.ac.uk

query tools, various reports can be generated, allowing for in-depth examination of the frequency and types of errors made by learners. These analyses contribute to the theoretical understanding of language learning development, and linguistic error analysis through comparative studies, and identify the challenges learners face during their learning process. In addition, they provide objective suggestions for material writers and teachers, enabling them to create appropriate learning materials, learner dictionaries, and textbooks tailored to the needs of Persian learners. The corpus can also be used for language assessment and for the development of Persian language tests for non-native learners. The research findings and insights obtained can advance the field of Persian language teaching and provide teachers and learners with improved tools and resources to optimize language acquisition and proficiency.

Keywords: CamBel, learner corpus, error analysis, teaching Persian language

1. Introduction

Needs analysis of language learners based on their learning objectives is one of the most crucial aspects of designing language curriculum, selecting appropriate methodologies, and developing learning materials and textbooks. It involves analyzing the development of interlanguage and identifying the challenges that language learners face at different levels of language learning, as well as in various areas including phonetics, morphology, syntax, and semantics. While individual differences and personal learning strategies significantly impact language learning, the main questions are how to identify these challenges specifically and, subsequently, what solutions can be implemented to address them. Examining the linguistic productions of language learners (spoken, written, or mixed) and analyzing them, especially in terms of identifying language errors, is one of the most scientific and practical methods for determining the language needs of L2 learners. This paper introduces the theoretical framework and methodology for the preparation and production of the CamBel Persian Learner Corpus. CamBel is largely compiled from the linguistic productions of Persian learners whose first languages are English or Serbian. Following this introduction, the paper provides a review of key concepts and the background of learner corpus research. Subsequently, the design and compilation of the CamBel learner corpus are described in detail, and an error-analysis model is introduced. In the final section, the results of the error analysis and the broader implications of the corpus are discussed.

2. Review of Literature

Granger (2002) defines learner corpora (LC) as "electronic collections of authentic textual data collected according to explicit design criteria for a particular SLA/FLT purpose in a standardized format." Similarly, McEnery et al. (2006, as cited in Gilquin, 2015: 9) describe a corpus as a "collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety." Callies and Paquot (2015: 1) also emphasize the systematic collection of learner data and define learner corpora as "systematic collection of authentic, continuous and contextualized language use (spoken or written) by L2 learners stored in electronic format which is a special type of empirical data used by scholars in a variety of disciplines." On the other hand, although according to Gilquin (2020, 283) "Learner corpora are corpora representing written and/or spoken 'interlanguage', that is, language produced by learners of that language.", however, it is important to clarify that what is accessed through learner corpora does not represent the entirety of a learner's interlanguage, but rather specific and context-dependent aspects of it. Learner corpora are inevitably restricted to particular genres, such as written, spoken, or even multimedia productions, and therefore cannot fully capture the whole interlanguage system of a learner.

Another important issue in learner corpora concerns the methods of annotation, as any language corpus must be refined and annotated according to specific research purposes. Like other types of language corpora, learner corpora differ in their annotation practices. As noted by Meunier (2020: 24), "initial productions by learners, often referred to as 'raw' texts, are frequently enriched through annotation to provide researchers with access to richly annotated data. Common linguistic annotations include automatic part-of-speech tagging, where each word in the corpus is assigned a part of speech (such as noun, verb, adjective) using automated software. Additionally, computer-aided error annotation (CEA) is used to identify areas of difficulty in learners' L2 production, aiding researchers in focusing on these aspects to enhance learners' proficiency." Among the different types of annotated learner corpora, error-tagged learner corpora, which focus specifically on identifying learners' errors—are considered particularly practical, as they provide deeper insights for language instructors, curriculum designers, and researchers into the nature and development of interlanguage, and they have led to the establishment of the methodology known as Computer-Aided Error Analysis (CEA). According to Díaz-Negrillo and Fernández-Domínguez (2006), and based on previous research (Granger 2003, 2004; Nesselhauf 2004; Fitzpatrick & Seegmiller 2004), learner corpora, by providing a comprehensive picture of learner interlanguage, being fully computerized and thus enabling large-scale data processing, offering systematically collected and representative datasets, and

containing both erroneous and error-free learner production, serve as highly reliable tools for error analysis and research in second language acquisition. These features give the Computer-Aided Error Annotation (CEA) approach a clear advantage over traditional methods of error analysis.

2.1 Learner Corpora Across Languages

Despite its relative youth compared with other areas of linguistic inquiry, Learner Corpus Research (LCR) has rapidly established itself as an influential domain situated at the intersection of Second Language Acquisition (SLA), Corpus Linguistics, and Foreign Language Teaching (FLT). Over the past three decades, a substantial number of learner-corpus projects have been developed for major world languages. English has the longest and most established tradition in learner corpus research, with major resources such as the Cambridge Learner Corpus (CLC)[3]. The CLC is a continually expanding corpus of more than 30 million words of learner-produced English from around the world. The International Corpus of Learner English (ICLE) project (Granger et al., 2020) is also regarded as a major milestone, representing the results of more than 30 years of collaborative work among a large number of universities worldwide. Some other sample projects on English learner corpora include British Academic Written English Corpus (Nesi et al., 2008), CORYL Corpus of Young Learner Language (Hasselgreen & Moe, 2006), ETS Corpus of Non-Native Written English (Blanchard et al., 2014), Römer and Swales (2010), and The Dresden Corpus (2004). Learner corpora have also been constructed for other widely taught languages, including Spanish (Lozano, 2009), German (Zinsmeister & Breckle, 2010), and French (Hilton, 2008). Moreover, significant initiatives exist for a broader range of languages, such as Chinese (Wang et al., 2015), Czech (Hana et al., 2010), Japanese (Izumi et al., 2004), Swedish (Hammarberg, 2010), Portuguese (del Río et al., 2016), Korean (Dickinson et al., 2010), Arabic (Alfaifi, 2015), and Russian (Rakhilina et al., 2016), illustrating the growing international engagement with corpus-based approaches to second-language learning.

2.2 Persian Learner Corpora

According to the Centre for English Corpus Linguistics at the Université Catholique de Louvain, which collects and indexes learner corpora as "Learner Corpora around the World"[4], most of the currently available corpora focus on English as the second language (L2). Among the 203 learner corpora listed,

[3] Cambridge University Press. (n.d.). Cambridge Learner Corpus (CLC). https://www.cambridge.org/elt/corpus/learner_corpus2.htm

[4] Centre for English Corpus Linguistics. *Learner Corpora around the World.* Louvain-la-Neuve: Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

most languages, excluding English, Spanish, French, and German, are deemed under-resourced in terms of available learner data.

With regard to Persian specifically, only a small number of projects address learner Persian, and merely three learner corpora are indexed, as follows:

1. The Persian Learner Corpus (PLC), regarded as the first Persian learner corpus (Abdollahi et al., 2025), includes written narratives and essays aimed at intermediate and advanced learners, with restricted online access for academic use (Safari, 2012).
2. The Salam Farsi Learner Corpus (SFLC) was developed at the Faculty of Philology, University of Belgrade and focuses on written narratives and descriptive essays for beginner and upper-intermediate Persian learners, sourced from Serbian contexts (Safari, 2017).
3. The Learner of Persian Spoken Corpus (LoPSC), developed by Daghbandan, collects informal spoken conversations in Persian from English speakers at an upper-intermediate proficiency level, with an ongoing collection of approximately 30,000 recordings (Centre for English Corpus Linguistics).

It should be noted that there may be additional projects that are not listed in the "Learner Corpora around the World" database maintained by the Centre for English Corpus Linguistics at the Université Catholique de Louvain. For example, the Persian spoken learner corpus developed by Abdollahi et al. (2025), which is currently not listed in the database. However, to the best of the authors' knowledge, the projects mentioned above represent the currently known and published initiatives. Based on the available records, there is a noticeable lack of developed learner corpora for the Persian language.

Another significant contribution in this field is the Persian Learners Written Data (PLWD), a limited learner corpus developed at the Faculty of Asian and Middle Eastern Studies, University of Cambridge. The PLWD is a longitudinal learner corpus, where data are collected from the same subjects at different time intervals to reflect the development of their language skills over time. The corpus comprises 400 texts written by Persian learners, specifically students of Persian language at the faculty between 2015 and 2020, totaling more than 30,000 words. Both the Persian Error-tagged Learner Corpus (PLC) and the Salam Farsi Learner Corpus (SFLC) were developed at the University of Belgrade, Faculty of Philology. These corpora comprise written productions of Serbian students. The PLC began as a limited project with 100 written texts totalling about 10,000 words, where error annotations were manually conducted using Excel. In contrast, the SFLC evolved from the PLC into a separate and extensive research project, encompassing 500 texts and approximately 45,000 words. This corpus was annotated for errors using a specialized tagset capable

of identifying 22 types of errors across 5 domains: orthography, morphology, syntax, lexis, and style (Safari, 2018).

The Cambridge-Belgrade Persian Learner Corpus (CamBel) is formed and developed by merging and integrating the SFLC Error-tagged Learner Corpus, developed at the University of Belgrade, with the longitudinal learner corpus, Persian Learners Written Data (PLWD), from the University of Cambridge. This marks the ongoing development phase of CamBel (version 01), which was officially launched and introduced at the 4th International Conference on Teaching Persian Language and Literature, held at the Faculty of Asian and Middle Eastern Studies at the University of Cambridge in December 2023.

3. CamBel Design and Methodology

Regarding the major methodological approaches used in learner corpus research, Granger (2002) identifies two primary methods: Contrastive Interlanguage Analysis (CIA) and Computer-Aided Error Analysis (CEA). She described these methods as "linguistic exploitations of learner corpora". According to her, CIA involves both quantitative and qualitative comparisons between native language and learner language (L1 vs. L2), as well as between different interlanguage varieties (L2 vs. L2) (Granger, 2009).

CamBel, as a Persian error-tagged learner corpus, employs the Computer-Aided Error Analysis (CEA) approach within a mixed-methods framework that integrates both quantitative and qualitative analyses. According to Creswell (2014: 14), mixed-methods research is an approach to inquiry that involves collecting both quantitative and qualitative data, integrating these two forms of data, and using specific research designs that may incorporate philosophical assumptions and theoretical frameworks. The core assumption of this form of inquiry is that combining qualitative and quantitative approaches provides a more complete understanding of a research problem than either approach alone. Quantitatively, CamBel examines error counts based on learner metadata (such as gender, age, and proficiency level) and text metadata (including narrative, descriptive, essay, and exam paper), as well as error types across various domains like orthography, morphology, morphosyntax, and syntax. Qualitatively, it offers detailed descriptions and contrastive analyses of errors considering learners' first language and level of proficiency. Therefore, CamBel employs a comprehensive mixed-method approach, integrating both quantitative and qualitative analyses for a thorough examination of learner errors.

The CamBel corpus design is organized into three primary stages: 1) collecting data and constructing the corpus, 2) proposing a system of error annotation, and 3) developing tools and software. These stages encompass practical phases such as the systematic collection of data and metadata, defining

the corpus design criteria, creating the error tagset, and developing the corpus interface, software, and specific tools. The CamBel software includes five main tools designed to function as an error-tagged learner corpus and generate statistical reports. The data collected in this corpus primarily consists of written works on a variety of topics by Persian learners at different proficiency levels (A1- C2) from around the world. Consequently, the corpus showcases the authentic written output of Persian learners with diverse first languages. The learners represent all genders, various ages, different educational backgrounds, and are situated in a range of educational or academic environments. Table 1 provides a brief overview of the design structure of the CamBel, which is organized according to three core components: corpus design criteria, the error annotation model, and the corpus tools.

Table 1: CamBel (version 0.1): Design, Annotation, and Tools

Corpus Design Criteria	Error Annotation Model	Corpus Tools
1. Profile Criteria 2. Data Criteria 3. Metadata Criteria 4. Learner Criteria 5. Annotation Criteria	1. Error Surface structure 2. Error Domain 3. Error Classes 4. Error Type 5. Error Discourse-Pragmatic Layer	1. DSMT[5] 2. EACT[6] 3. CQFT[7] 4. CSAT[8] 5. CPAT[9]

As outlined in Table 1, five categories of criteria were defined for the design of CamBel; each is introduced briefly as follows:

1. Profile Criteria, which include the mode of the corpus, its size, and its availability;
2. Data Criteria, covering factors such as text type, task type, genre, and the creation of metadata records for the corpus;
3. Metadata Criteria, relating to information about learners and the texts they produce;
4. Learner Criteria, which include learners' first and target languages, proficiency levels, and other relevant background data;
5. Annotation Criteria, involving the creation of an error taxonomy and the development of an error-tag set.

Table 2 presents detailed information on the application of these criteria in CamBel.

[5] Data Submission and Metadata Management Tool (DSMT)
[6] Error Annotation and Classification Tool (EACT)
[7] Corpus Query and Filtering Tool (CQFT)
[8] Corpus Statistics and Analytics Tool (CSAT)
[9] CEFR Profiling and Assessment Tool (CPAT)

Table 2: CamBel (Version 0.1): Design Criteria

<table>
<tr><th>Profile Criteria</th><th>Data Criteria</th><th>Metadata Criteria</th><th>Learner Criteria</th><th>Annotation Criteria</th></tr>
<tr><td>Mode of corpus: written</td><td>Type of data: written</td><td>Gender, Age, L2, L3</td><td>First language: Serbian, English</td><td>Error taxonomy</td></tr>
<tr><td>Size:
900 documents totaling 75,000 words</td><td>Type of tasks: composition, article, free, exam papers</td><td>Country
City
Institution
Grade</td><td>Target language: Persian</td><td rowspan="4">Error tagset</td></tr>
<tr><td>Purpose: research/ educational</td><td rowspan="3">Genre: descriptive, narrative, argumentative</td><td>Text ID
Text title
Year
Production</td><td rowspan="3">Proficiency framework tagset:
A1
A2
B1
B2
C1</td></tr>
<tr><td>Accessibility: limited</td><td rowspan="2">Timing
Reference</td></tr>
<tr><td>Users: researchers, language instructors, learners</td></tr>
</table>

4. CamBel (Version 01) Error Tagging and Reporting

The error-annotation model used in CamBel is fundamentally based on the error-detection framework proposed by Dulay, Burt, and Krashen (1982), as well as on the methodological experience gained from the SFLC error-annotation model (Safari, 2018). It has been carefully adapted and customized to meet the specific analytical and linguistic requirements of an error-tagged learner corpus for Persian. The model systematically categorizes and annotates learners' errors into five complementary layers, allowing for precise and multidimensional error detection. The model is organized as follows:

1. Errors in Surface Structure: This level identifies the underlying form of errors and includes Addition, Omission, Substitution, and Permutation.
2. Error Domain: This level identifies the linguistic subsystem in which the error occurs, including Orthography, Morphology, Morphosyntax, Syntax, Lexis, and Style.
3. Error Classes: This level includes 12 major error categories such as Noun, Adjective, Adverb, Verb, and others.

4. Error Types: This level identifies 28 specific error types that correspond to the structural and usage patterns of Persian.
5. Discourse-Pragmatic Error Layer: This additional layer accounts for higher-level errors related to coherence and cohesion, register, pragmatic inappropriacy, and other phenomena that cannot be captured through surface-level, domain-based, or structural classifications alone.

The errors are categorized into five annotation levels based on the CamBel Tagset model, with each error identified by a unique six-letter tag. The first letter indicates the error in the surface structure, the second denotes the error domain, the next two letters represent the error class, and the final two letters specify the type of error. The fifth level, *the Error Discourse-Pragmatic Layer*, is applied separately and is used when an error involves coherence, cohesion, register, pragmatic appropriateness, or sociolinguistic factors. This layer does not alter the six-letter tag but can be added as an additional label when such higher-level features are present. This system provides flexibility, allowing for the selection and combination of errors across the three annotation levels. For example, in the error tag <O_M_VB_VA>, "O" signifies an "Omission" error in the surface structure, "M" indicates the error domain of "Morphology", "VB" represents the error class "Verbs", and "VA" specifies the error type "Verb Agreement". It should be noted that the fifth layer, the Error Discourse-Pragmatic Tag, is used less frequently in the annotation process and is applied only when necessary.

Table 3: An example of error annotation in CamBel

The sentence with an error	* دیروز <O_S_PR_PP > جلسه مهمی شرکت کردم.
ID for the text and for the learner	LIN126_T0389
Error tags	< O_S_PR_PP >
Error Surface Structure	Omission
Error Domain	Syntax
Error Class	Preposition
Error Type	Preposition

The reports generated by the corpus for analysis are highly diverse and comprehensive, allowing for the combination of multiple variables. For example, it is possible to compare the frequency of grammatical errors in the verb class and specific verb-agreement types among Persian learners at proficiency levels A and B. Additionally, comparative reporting makes it possible to determine error frequencies across different levels based on the learners' first language. In this phase, only errors from Serbian and English-speaking learners (English as their first language) have been included, enabling the corpus to produce a range of comparative reports between these two groups of Persian learners.

5. CamBel Applications and Implications

CamBel can provide a reliable resource for language teachers, textbook writers, and researchers in second language learning at both educational and research levels. Key uses and implications include:

1. Identifying Learners' Needs: Examining and analyzing language learning errors based on the learners' first language and proficiency level, the structural needs of each group in orthography, morphology, syntax, and style can be identified. For example, reports indicate that in the syntax domain, Serbian learners of Persian at the A1 level struggle more with sentence word order, while English-speaking learners at the same level have more challenges with verb omission. This highlights that learners at the same proficiency level but with different first languages have distinct needs.

2. Developing Syllabus Design and Teaching Methods: The analysis of corpus reports plays a crucial role in the selection of both the syllabus (what to teach) and teaching methods (how to teach). For example, addressing the specific needs of Serbian and English-speaking learners of Persian might involve tailoring the syllabus to their unique challenges. English-speaking learners often struggle with verb-related errors, while Serbian-speaking learners may face difficulties with sentence order. By understanding these common errors, teachers can design activities, exercises, and homework that specifically target these areas. This approach allows for more effective teaching and helps learners overcome their particular challenges more efficiently. CamBel analysis not only highlights frequent errors but also provides insights into the learners' interlanguage, enabling educators to create more relevant and impactful lesson plans. By focusing on these identified problem areas, teachers can improve their instructional methods and provide more personalized learning experiences.

3. Creating Educational Content and Textbooks: One of CamBel's most significant contributions is that it provides valuable resources for content creators and authors of Persian coursebooks. By analyzing error reports from different language groups and proficiency levels, educators can determine which aspects require the most emphasis. The corpus highlights the inadequacy of a one-size-fits-all textbook for learners with different first languages. For example, textbooks for Serbian-speaking learners at a particular level should differ from those for English-speaking learners at the same level, reflecting their unique learning challenges and needs. This ensures that educational materials are optimized for each group.

4. Conducting Pedagogical Research: At the research level, CamBel can be employed for a range of pedagogical and applied linguistic studies as follows: 1. Investigating the influence of learners' first languages on the acquisition of Persian in a comparative framework. 2. Investigating the development of interlanguage among learners of Persian language (Ghaffari, 2020). 3. Analyzing

errors of Persian learners by examining error fields, categories, types, and other metadata-related variables.

6. Conclusion

This paper provides a comprehensive overview of the theoretical background and literature on learner corpus research (LCR) and introduces the international project for the development and construction of CamBel (Cambridge-Belgrade Persian Learner Corpus). CamBel is an error-annotated learner corpus developed based on robust theoretical and methodological frameworks in corpus linguistics. It integrates the error-tagged learner corpus of SFLC, developed at the Faculty of Philology, University of Belgrade, with the longitudinal learner corpus of PLWD at the Faculty of Asian and Middle Eastern Studies, University of Cambridge. This study presents CamBel version 0.1, which represents the initial phase of an ongoing corpus-development project. The corpus remains under development, and subsequent versions will be released in the future. CamBel's methodology includes systematically collecting and annotating the written productions of Persian learners using an Error Tagset at five levels: Errors in surface structure: including addition, omission, substitution, and permutation; Error domain: encompassing orthography, morphology, morphosyntax, syntax, lexis, and style; Error Classes such as Noun, Adjective, Adverb, Verb, and others. Error types: consisting of 28 specific error types relevant to Persian construction and detecting Errors in Discourse-Pragmatic Layer. The corpus software provides various detailed reports, ranging from errors based on the first language, proficiency level, gender, age, and institution, to the frequency of errors or words throughout the corpus. These comprehensive reports are invaluable for error analysis, comparative linguistic studies, language teaching, and the development of textbooks and targeted educational materials. By providing detailed insights into the types and frequencies of errors made by Persian language learners, CamBel supports educators and researchers in enhancing instructional strategies and materials, thereby significantly contributing to the field of second language acquisition.

References

Abdollahi, E., Moloodi, A., Rahimi, M., & Kouhestani, M. (2025). Design, development, and annotation of the Persian spoken learner corpus. In H. Saeli (Ed.), *Handbook of teaching and learning Persian as a second language*. Springer Handbooks in Languages and Linguistics. Springer. https://doi.org/10.1007/978-981-97-1818-4_35-1

Alfaifi, A., & Atwell, E. (2015). *Arabic learner corpus* [Dataset]. Linguistic Data Consortium. https://doi.org/10.35111/5312-X803

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2014). *ETS Corpus of Non-Native Written English* [Dataset]. Linguistic Data Consortium. https://doi.org/10.35111/7EZ0-X912

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/10.1075/ijlcr.1.1.00ed

Centre for English Corpus Linguistics. (n.d.). *Learner corpora around the world.* Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. SAGE Publications.

del Río, I., Antunes, S., Mendes, A., & Janssen, M. (2016). Towards error annotation in a learner corpus of Portuguese. In *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* (pp. 8–17). LiU Electronic Press.

Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada, 19*, 83–102.

Dickinson, M., Israel, R., & Lee, S.-H. (2010). Building a Korean web corpus for analyzing learner language. In *Proceedings of the NAACL HLT 2010 Sixth Web as Corpus Workshop* (pp. 8–16). Association for Computational Linguistics.

Dulay, H. C., Burt, M. K., & Krashen, S. (1982). *Language two*. Oxford University Press.

Fitzpatrick, E., & Seegmiller, M. S. (2004). The Montclair electronic language database project. In U. Connor & T. A. Upton (Eds.), *Applied corpus linguistics: A multidimensional perspective* (pp. 223–237). Rodopi.

Ghaffari, M. (2020). Persian as an interlanguage. In P. Shabani Jadidi (Ed.), *The Routledge handbook of second language acquisition and pedagogy of Persian* (1st ed., pp. 546–566). Routledge.

Gilquin, G. (2015). From design to collection of learner corpora. In D. Gabrys-Barker & D. Gałajda (Eds.), *The Cambridge handbook of learner corpus research* (pp. 3–34). Cambridge University Press.

Gilquin, G. (2020). Learner corpora. In M. Paquot & S. Gries (Eds.), *A practical handbook of corpus linguistics* (pp. 283–303). Springer.

Granger, S. (2002). A bird's-eye view of learner corpus research. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Language learning & language teaching* (Vol. 6, pp. 3–33). John Benjamins. https://doi.org/10.1075/lllt.6.04gra

Granger, S. (2003). Error-tagged learner corpora and CALL: A promising synergy. *CALICO Journal, 20*(3), 465–480.

Granger, S. (2004). Computer learner corpus research: Current status and future prospects. In U. Connor & T. A. Upton (Eds.), *Applied corpus linguistics: A multi-dimensional perspective* (pp. 123–145). Rodopi. https://doi.org/10.1163/9789004333772_008

Granger, S. (2008). Learner corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (pp. 259–275). De Gruyter.

Granger, S. (2009). The contribution of learner corpora to SLA and FLT: A critical evaluation. In K. Aijmer (Ed.), *Studies in corpus linguistics* (Vol. 33, pp. 13–32). John Benjamins. https://doi.org/10.1075/scl.33.04gra

Granger, S., Dupont, M., Meunier, F., Naets, H., & Paquot, M. (2020). *The International Corpus of Learner English* (Version 3). Presses universitaires de Louvain. https://dial.uclouvain.be/pr/boreal/object/boreal:229877

Hammarberg, B. (2010). *Introduction to the ASU corpus: A longitudinal oral and written text corpus of adult learner Swedish*. Department of Linguistics, Stockholm University.

Hana, J., Rosen, A., Škodová, S., & Štindlová, B. (2010). Error-tagged learner corpus of Czech. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 11–19). Association for Computational Linguistics.

Hasselgreen, A., & Moe, E. (2006, May 18–21). *Young learners' writing and the CEFR: Where practice tests theory* [Conference presentation]. 3rd EALTA Conference, Krakow, Poland.

Hilton, H. (2008). *Slabank French Parole Corpus* [Dataset]. TalkBank. https://doi.org/10.21415/T54C7R

Izumi, E., Uchimoto, K., & Isahara, H. (2004). The overview of the SST speech corpus of Japanese learner English. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. ELRA.

Kubanek-German, A. (2004). *Slabank German-English Dresden Corpus* [Dataset]. TalkBank. https://doi.org/10.21415/T5HS45

Lozano, C. (2009). CEDEL2: Corpus Escrito del Español como L2. In C. M. Bretones et al. (Eds.), *Applied linguistics now* (pp. 197–212). Universidad de Almería.

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

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

Meunier, F. (2020). Introduction to learner corpus research. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (Chap. 3). Routledge. https://doi.org/10.4324/9781351137904

Nesi, H., Gardner, S., Thompson, P., & Wickens, P. (2008). *British Academic Written English Corpus*. Literary and Linguistic Data Service. http://hdl.handle.net/20.500.14106/2539

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In J. McH. Sinclair (Ed.), *Studies in corpus linguistics* (Vol. 12, pp. 125–152). John Benjamins. https://doi.org/10.1075/scl.12.11nes

Rakhilina, E., Vyrenkova, A., Mustakimova, E., Ladygina, A., & Smirnov, I. (2016). Building a learner corpus for Russian. In *Proceedings of the Joint Workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition* (pp. 66–75). LiU Electronic Press.

Römer, U., & Swales, J. M. (2010). The Michigan corpus of upper-level student papers (MICUSP). *Journal of English for Academic Purposes, 9*(3), 249.

Safari, S. (2012). *Designing and developing a FFL learner corpus* (Master's thesis). Allameh Tabatabai University.

Safari, S. (2017). *Constructing and analysing an error-tagged learner corpus of Persian* (Doctoral dissertation). University of Belgrade.

Safari, S. (2018). The Salam Farsi Learner Corpus: Introducing the error tagging system. *Anali Filološkog Fakulteta, 30*(2), 249–263. https://doi.org/10.18485/analiff.2018.30.2.13

Wang, M., Malmasi, S., & Huang, M. (2015). The Jinan Chinese learner corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 118–123). https://doi.org/10.3115/v1/W15-0614

Zinsmeister, H., & Breckle, M. (2010). Starting a sentence in L2 German: Discourse annotation of a learner corpus. In P. Manfred et al. (Eds.), *Proceedings of the Conference on Natural Language Processing* (pp. 181–185). Saarland University.