A Corpus of English Learners with Arabic and Hebrew Backgrounds

Omaima Abboud[1], Batia Laufer[2], Noam Ordan[3], Uliana Sentsova[4] and Shuly Wintner[4*]

[1]Dept. of English, Arab Academic College for Education, Israel.
[2]Dept. of English Language and Literature, University of Haifa, Israel.
[3]The Israeli Association of Human Language Technologies, Israel.
[4]Dept. of Computer Science, University of Haifa, Israel.

*Corresponding author(s). E-mail(s): shuy@cs.haifa.ac.il;
Contributing authors: omaima.abboud@gmail.com;
balaufer@univ.haifa.ac.il; noam.ordan@gmail.com;
uliana.sentsova@gmail.com;

Abstract

Learner corpora—datasets that reflect the language of non-native speakers—are instrumental for research of language learning and development, as well as for practical applications, mainly for teaching and education. Such corpora now exist for a plethora of native–foreign language pairs; but until recently, none of them reflected native Hebrew speakers, and very few reflected native Arabic speakers. We introduce a recently-released corpus of English essays authored by learners in Israel. The corpus consists of two sub-corpora, one of them of Arabic native speakers and the other consisting mainly of Hebrew native speakers. We report on the composition and curation of the datasets; specifically, we processed the data so that both sub-corpora are now uniformly represented, facilitating seamless research and computational processing of the data. We provide statistical information on the corpora and outline a few research projects that had already used them. This is the first and only learner corpus in Israel including two major native languages of people in the same educational system regarding the English syllabus. All the resources related to the corpus are freely available.

Keywords: Corpus linguistics, Learner corpora, ESL, Hebrew, Arabic

1 Introduction

Corpus data have been used as significant sources of evidence and insight in language-related research since the 1960s (Svartvik, 1991). In the 1980s, corpus-linguistic methods gained popularity in studying second language acquisition (SLA, an umbrella term for any number of non-native languages acquired either in the language speaking context, or as a school subject in the classroom), sparking the systematic collection of learner corpora in the early 1990s (De Knop & Meunier, 2015; Granger, 1988, 2002; Granger, Gilquin, & Meunier, 2015; Tono, 2003). Learner corpora opened the potential for quantitative analysis of learner data, providing a refined way of describing learner language and supporting countless studies in SLA and language training.

Numerous learner corpora have been curated since then,[1] such as the International Corpus of Learner English Granger (2003), the Longman Learners' Corpus[2] and the Cambridge Learners' Corpus.[3] Unfortunately, not all languages can benefit from such resources. The web site *Learner Corpora around the World*,[4] which lists over 200 datasets, includes none with Hebrew as the target language (L2) and only two with Hebrew as the native language (L1) of the learners. Very few datasets reflect the language of Arabic L1 speakers. The first learner corpus of Hebrew as the L2, with Arabic, French, and Russian as L1s, has very recently been published (Gafni, Prior, & Wintner, 2022; Nguyen & Wintner, 2022). We now complement this with a learner corpus that focuses on Arabic and Hebrew L1 speakers, with English as the L2.

We introduce a language resource consisting of around 3300 English essays written by non-native speakers of English in Israel. The corpus represents various education levels, spanning learners enrolled in primary, secondary, and tertiary educational institutes. The corpus also reflects the multilingual nature of the Israeli society: the essay authors collectively use twelve native languages, with Hebrew, Arabic, and Russian being the most common mother tongues. The new release unifies two subcorpora, resulting from two independent initiatives to collect learners' data in Israel. The subcorpora were compiled by separate research groups; we report on an effort to merge and consolidate them into one cohesive learner corpus. The result is the first, and so far the only, English learner corpus that includes speakers of the two major native languages spoken in Israel, reflecting people in the same educational system regarding the English syllabus.

[1] For an extensive bibliography covering learner corpus analyses, see the resources page of the *Learner Corpus Association*, http://www.learnercorpusassociation.org/resources/lcb/.
[2] http://www.pearsonlongman.com/dictionaries/corpus/learners.html
[3] http://www.cambridge.org/elt/corpus
[4] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

2 The Corpus

2.1 Compilation

The first initiative aimed at creating a corpus of texts written by Israeli English learners started in 2009 by a group of researchers at the University of Haifa, led by Batia Laufer. They collected essays from a variety of sources, authored by students with varying educational levels. Most texts were produced as part of the learners' coursework at Hebrew-speaking schools, grades five through twelve (in a twelve-year school system). Some of essays were written by college and university students majoring in English, during their first year of studies. Additionally, researchers collected several essays that were written as part of the entrance exam to the English Language and Literature Department at the University of Haifa. Although the corpus includes speakers of languages other than Hebrew, learners with Arabic as their native language are underrepresented in this resource.

This corpus, called the *Israeli Learner Corpus of Written English (ILCoWE)*, has never been released publicly, nor reported on as a stand-alone linguistic resource. It was first mentioned in Laufer and Waldman (2011); this study focused on the use of collocations in learners of English and included a comprehensive description of the corpus compilation. Later, the corpus was used by Levitzky-Aviad and Laufer (2013) for an analysis of the lexical proficiency of Hebrew speakers.

In 2019, researchers from the Arab Academic College for Education in Haifa, led by Omaima Abboud, started a separate initiative to collect English essays. The effort was solely focused on authors with Arabic as their native language. The data were collected from high school students as well as from college students who pursue the track of B.Ed. or M.Ed. in English at the College. As in the case of ILCoWE, the researchers included in the corpus essays written for the admittance exam to the college. This corpus, known as the *Arab College Corpus (ArabCC)*, was made available publicly via *GitHub*[5] in 2021, and has been available since 2022 via *SketchEngine*[6](Kilgarriff et al., 2014), but it has never been published.

We report on a public distribution of the two corpora (ILCoWE and ArabCC), organized and represented uniformly under one umbrella, and made available to the entire research community.

2.2 Composition

The unified corpus includes 3353 essays written by 1632 non-native learners of English between the years 2009 and 2020. The corpus represents speakers of twelve native languages in various educational levels – from early years of secondary education to graduate studies. The corpus consists of approximately 540k tokens with 162 tokens per essay on average. Table 1 lists some basic statistics of the two datasets.

[5] https://github.com/ulicoder/arcc/
[6] https://www.sketchengine.eu/arabcc-learner-corpus-of-english-essays/

	ArabCC	ILCoWE	Total
Learners	304	1328	1632
Essays	957	2396	3353
Sentences	9699	22303	32002
Tokens	204872	337795	542667
Types	6317	13635	16411
Tokens per essay	214.1 (127.5)	140.9 (169.2)	162 (161.8)
Sentences per essay	10.1 (5.7)	9.3 (8.5)	9.5 (7.8)

Table 1 Corpus statistics. Numbers in parentheses denote standard deviation.

The conditions under which the learners produced the texts were not reported. While some of the essays were timed, others were not; in several cases the learners were allowed to use dictionaries while writing. Most of the essays were written under supervision in the classroom, although some essays were part of homework.

Essay collection lasted several years, which led to several learners contributing more than one essay at different study levels, as they advanced in their education tracks. As a result, more than 500 essays produced by 70 learners can be viewed as a small yet non-negligible *longitudinal* data sample.

2.3 Metadata

To attain a cohesive umbrella corpus, the metadata underwent unification and normalization. However, some variables preserved distinctive features of the original metadata as a result of merging two differently structured subcorpora. In this section, we describe the metadata and indicate differences between subcorpora, wherever needed.

Author's L1.

The learners reported speaking one of the following native languages: Hebrew, Arabic, Russian, English, Amharic, Spanish, Portuguese, Romanian, German, Armenian, Polish, and Ukrainian (listed in the order of their usage frequency in the corpus). 86% of the essays include metadata about the native language of their authors, with Hebrew, Arabic,[7] and Russian accounting for approximately 95% of speakers' native languages. In a small subset of the data, the L1 was not documented, and is denoted as "Hebrew/Arabic". Table 2 summarizes statistics on the native languages of the learners.

Other L1.

While most students indicated Hebrew or Arabic as their L1, the corpus also includes texts written by learners who speak other languages at home as well (e.g., Amharic, Russian, or Ukrainian). For some essays, the researchers explicitly documented cases of multilingualism: out of 1632 authors, at least 203 (roughly 12%) reported native-level proficiency of another language, Hebrew

[7] More specifically, Palestinian Arabic.

L1	Authors	%	Essays	%
Hebrew	788	48.3	1710	50.1
Arabic	318	19.5	971	29.0
Hebrew/Arabic	439	26.9	439	13.1
Russian	63	3.9	173	5.2
Other	15	0.9	39	1.2

Table 2 Distribution of native languages in the corpus.

(50%) and Russian (30%) being the most frequent cases. In reality, the actual number of multilingual authors is considerably higher: the majority of native speakers of Arabic in Israel are fluent in Hebrew, which is the official language of the country. During the collection of ArabCC the researchers did not obtain information about other native languages of the speakers; however, it is safe to assume that all essay authors with Arabic as a native language are also speakers of Hebrew. Based on these assumptions, the statistics on authors' other L1's are the following: 1502 essays (45% of all essays) were written by bilingual learners, and 507 authors are bilingual with Hebrew (80%) and (Russian 12%) as a second native language.

Educational level.

The essays were collected at school, college, and university. The educational level of the learners is documented on a fine-grained ordinal scale, with four categories: 1) lower secondary education, or middle school (grades five through nine); 2) upper secondary education, or high school (grades ten to twelve); 3) undergraduate studies towards a bachelor degree; and 4) graduate studies towards a master degree. Each category is further divided into subcategories that specify the grade or year of studies, e.g., grade nine, second year of undergraduate studies, first year of graduate studies, etc. Table 3 presents statistics on education levels, grouped by categories.

Education	Authors	%	Essays	%
Middle school	442	27.1	1264	37.7
High school	721	44.2	970	28.9
Undergrad. studies	434	26.6	821	24.5
Graduate studies	36	2.2	298	8.9

Table 3 Distribution of authors' educational level in the corpus.

Age.

Authors' precise age has been reported for 62% of the learners (mostly in the data collected in schools) and averages to 14.6 years. For another 19% of the learners, the age was reported as a general category of "over 21".

Gender.

71% of essays contain metadata about the gender of their authors. Only two values – female and male – were used, with a 1:0.7 female to male ratio. See Table 4.

Gender	Authors	%	Essays	%
Female	763	57.8	1290	54.3
Male	556	42.2	1085	45.7

Table 4 Distribution of authors' gender in the corpus.

Prompt.

The essays were written in response to a prompt, a book chapter or an article that students read as part of their coursework. In several cases, the essays were part of an admittance exam. The choice of the topics depended on the educational level of the students: while high school and university students responded to argumentative prompts, younger learners in grades five to eight were mainly asked to write passages of a narrative or descriptive nature in response to prompts such as "Describe a family event". Some topics were rather general such as "The most unexpected thing", whereas others were defined more strictly, e.g., "Many cities have begun banning cars from entering the city centre. Do you think this is a good idea? Write a passage for your school newspaper, stating your opinion and explaining the advantages and/or disadvantages of this policy". For school students, the prompts were written in Hebrew. Below are several examples of essay prompts:

- Describe a nice thing you did for someone or a nice thing someone did for you.
- Write a description of a school trip. Explain what you did or didn't enjoy.
- You learned about different theories of how human languages developed. Write 300 words on which theory, in your opinion, makes the most sense. Explain your answer.
- Advantages and disadvantages of day care.
- Clashes in the family.

2.4 Processing

The essays were collected in various formats; handwritten submissions were transcribed. Each essay has a unique ID and is stored as a separate text file, while the metadata are provided in an accompanying spreadsheet. To maintain anonymity, author names were replaced with a special omission token.

During the compilation of the ILCoWE subcorpus, middle-school students (grades five through eight) and some high-school students (only grades nine and ten) were allowed to use Hebrew words whenever they didn't know the corresponding word in English. These words were transcribed in English and

are preserved in the corpus in upper case, as in the following examples, where the English translation is added in parentheses:

- LIFAMIM (*sometimes*) I go to play tennis on 19:00 clock and I back to home in 20:00 clock.
- I told my frinds about the dream and they didnt HEEMINU (*believe*) me.
- In the center of the week I eat small ARUCHOT (*meals*).

For some research questions, spelling variations in learner language are of no interest and can even present an obstacle for automatic annotation tools such as part-of-speech taggers or dependency parsers. To facilitate research in such cases, a spell-checked version of the ArabCC subcorpus was produced; spell-checking was performed manually by the researchers. The ArabCC subcorpus was additionally annotated for erroneous use of the most frequent prepositions.

2.5 Availability

The corpus is freely available for research purposes from GitHub.[8] No license is required. The repository consists of the essays (as plain texts) and the metadata (in a comma-separated format). The prompts used to elicit the essays are also available, in JSON format.

3 Use cases

For several decades learner corpora have served as a resource for quantitative approaches to learner language, contributing greatly to SLA. ILCoWE served as such a resource for several studies in the field of EFL (English as a Foreign Language) research, facilitating research of various interlaguage-related phenomena. Levitzky-Aviad and Laufer (2013) used ILCoWE to examine the development of lexical proficiency by Hebrew-speaking English learners. Among other measurements of vocabulary acquisition, they examined the proportion of infrequent vocabulary, lexical diversity, and the frequency of collocation use across eight years of formal English learning at schools and universities, showing growth in lexical proficiency at specific learning stages.

Similarly, in a study by Laufer and Waldman (2011), ILCoWE data were used in a cross-corpora analysis: the study compared Hebrew-speaking English learners to native English speakers regarding the use of English verb-noun collocations. ILCoWE data allowed Laufer and Waldman (2011) to flesh out statistically significant differences in collocation use between native speakers and EFL learners as well as across three proficiency levels of the learners.

The corpus has also been used in a research focusing on English learners with Arabic as their mother tongue. A corpus-based study by Abboud (To Appear) analyzed the atypical uses of causative constructions with the verb *make* and established a connection between usage patterns in English and their

[8] https://github.com/HaifaCLG/ILCoWE

equivalent constructions in L1, making a case for a transfer between L1 and L2.

4 Conclusion

We reported on the recent release of a corpus of English learner essays, authored by young writers whose native languages are mainly Hebrew and Arabic. The essays were collected in two separate, independent efforts, and our distribution unifies their organization and representation. We trust that this resource will be invaluable for future study of learner language in general, and English learner language of native speakers of Arabic and Hebrew in particular.

Acknowledgements

We are grateful to Tami Levitzky-Aviad and Tina Waldman for collecting the ILCoWE corpus.

Author contributions

B.L. led the construction of ILCoWE whereas O.A. and N.O. led the construction of ArabCC. U.S.. compiled and reorganized the unified corpus. under the supervision of S.W. U.S. and S.W. wrote the article, which all authors reviewed.

Declarations

The authors have no competing interests to declare. No funding was received to assist with the preparation of this manuscript.

References

Abboud, O. (To Appear). *Good teachers make students love the subject: A corpus-based study of the atypical uses of causative* make *in Arab EFL learners.*

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.

Gafni, C., Prior, A., Wintner, S. (2022, June). The Hebrew Essay Corpus. *Proceedings of the language resources and evaluation conference* (pp. 5580–5586). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.598

Granger, S. (1988). The computer learner corpus: a versatile new source of data for SLA research. *Learner English on computer* (pp. 3–18). Routledge.

Granger, S. (2002). A bird's-eye view of learner corpus research. S. Granger (Ed.), *Computer learner corpora, second language acquisition, and foreign language teaching* (Vol. 6, pp. 3–33). Amsterdam: Benjamins. Retrieved from https://korpling.german.hu-berlin.de/svn/bibliographie/2002

Granger, S. (2003). The International Corpus of Learner English: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quarterly*, 538–546.

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

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... Suchomel, V. (2014). The sketch engine: ten years on. *Lexicography*, *1*(1), 7–36.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language learning*, *61*(2), 647–672.

Levitzky-Aviad, T., & Laufer, B. (2013). Lexical properties in the writing of foreign language learners over eight years of study: Single words and collocations. C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use. new perspectives on assessment and corpus analysis* (Vol. 2, pp. 127–148). European Association of second Language Acquisition.

Nguyen, I., & Wintner, S. (2022, June). Predicting the proficiency level of nonnative Hebrew authors. *Proceedings of the language resources and evaluation conference* (pp. 5356–5365). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.573

Svartvik, J. (1991, August). Corpus linguistics comes of age. J. Svartvik (Ed.), *Directions in corpus linguistics: Proceedings of nobel symposium 82 stockholm* (pp. 7–13). Berlin, New York: De Gruyter Mouton. Retrieved from https://doi.org/10.1515/9783110867275.7 doi:10.1515/9783110867275.7

Tono, Y. (2003). Learner corpora: Design, development and applications. D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), *Proceedings of the corpus linguistics 2003 conference* (pp. 800–809). Lancaster University: University Centre for Computer Corpus Research on Language. Retrieved from https://korpling.german.hu-berlin.de/svn/bibliographie/2003