Marine Makhatadze

LEARNER CORPUS PROFILES: LEXICAL PECULIARITIES OF GEORGIAN EFL STUDENTS

Abstract

This paper focuses on the usability of learner corpus in foreign language research. Teachers, lexicographers and researchers use learner corpus data to measure the most fundamental aspect of second/foreign language knowledge – lexical profiles of the learners. Learner corpora provide a mirror for the learners’ language competence.

The paper aims to analyze an ongoing project on the learner corpus of Georgian EFL (English as a foreign language) university students. The language productions that make up our yet ‘baby’ learner corpus include written texts. The reason for bringing learner productions into a corpus rather than examining them individually is the desire to arrive at generalizable findings about language acquisition, i.e. which words can the learners produce and with what degree of appropriateness. Our research seeks to answer the following questions: a) what kind of lexical behaviour is represented in the learner corpus data, and b) what is the potential of giving learners/teachers access to the learners' data?

The study and the quantitative information described in the work reflect the characteristics of learner English in terms of part-of-speech distribution and collocation usage. The usefulness and advantages of the corpus-based approach will be demonstrated by employing learner corpus-based activities that can be implemented in an educational environment.

Keywords: *learner corpus*, *lexis*, *language production*, *pedagogy*, *data-driven learning*.

Linguists love corpora; where two or three linguists are gathered, there shall you find heavy-breathing fetishism about the size, scope, all those possibilities, all that data.
Yet all the data in the world is useless unless you can find someone to parse and interpret it.
- *Kory Stamper*, *Word by Word: The Secret Life of Dictionaries*

1. Introduction

Learner corpus research emerged as an offshoot of corpus linguistics, which has shown great potential to explore native languages, although it neglected the non-native varieties. The area of learner corpus research has connected two previously dissimilar fields of corpus linguistics and foreign or second language research with each other. Learner corpora can be used for a wide range of objectives in language acquisition and production research. Compared to earlier foreign language acquisition studies, modern learner corpora's authenticity and representativeness of the language variety is sophisticated and, therefore, some pedagogical approaches in ELT can benefit from learner corpus research.

Learner corpora provide an enhanced description of learner language and improve the foreign language teaching process. These goals are achieved by using the main principles, tools and methods from corpus linguistics.

There are two significant advantages when it comes to collecting L2 data electronically. Firstly, as these collections are produced by a great number of learners, they are less prone to the representativeness problem.

Secondly, the learner texts can be analysed with a whole battery of software tools, for example, part-of-speech taggers, which assign a tag (in our case, a grammatical category) to each word in a learner corpus. The process facilitates investigations of learners' use of specific grammatical categories. On the other hand, concordance programs reveal the lexis and phraseology of foreign language learners. The concordances generate frequency lists of linguistic items, such as words and phrases and present them in an immediate linguistic context.

The idea of authenticity and genuineness is somewhat problematic in the case of learning English, as foreign language teaching context usually involves an unspontaneous nature. Due to that, several learner corpora involve control from the compilers. In narrative essays, for example, learners are free to write what they like rather than having to produce what research is interested in.

However, the issue is the task variables which give the learner corpus data some degree of artificiality, such as topic or time limit (Granger, S., Gilquin, G., & Meunier, 2015).

Still, as essay writing is an authentic classroom activity, learner corpora of essay writing can be considered valid written data. They form useful experimental data types which can give a distorted view of learners' language production reality (Selinker & Gass, 2008).

1.1 Learner corpus typology

Learner corpus typology is often described in terms of dichotomies, differing along some dimensions. While determining how the learners' data will be collected and turned into a corpus, we should disambiguate the learner corpora of written texts and transcriptions of spoken discourse. Today, written learner corpora are more common than spoken ones. Spoken corpora are more laborious to collect and involve extensive financial effort. Some learner corpora even include both written and spoken data, some of them are multimodal (or audio-visual) learner corpora (like MAELC, the Multimedia Adult ESL Learner Corpus; Reder et al. 2003), which include video recordings and give access to new domains of investigation like the analysis of learners' gazes or even gestures.

The second dimension is that of the genre. Most learner corpora to date are general as they correspond to language as used for general purposes, but recently language for specific purposes (LSP) learner corpora have made their appearance.

Another aspect which serves to categorise the learner corpora is the time frame: data collected from one period in time is called synchronic corpus, which represents a snapshot of learners' language competence and from several periods – diachronic corpus, describing the evolution of language knowledge through time. For instance, The Longitudinal Database of Learner English (LONGDALE) is a project that aims to follow the same learners over at least three years and increases the number of collections per year to make the corpus denser. Belz and Vyatkina (2008: 33) use the term "developmental learner corpus" to refer to dense corpora.

Finally, from a pedagogic perspective, a distinction can also be drawn between global and local learner corpora. Global corpora are part of large-scale projects, while local learner corpora are typically collected by teachers among their students, who are both contributors and users of the corpus. The major aim of this approach is to identify learners' specific language needs through a corpus analysis and thus provide apt solutions to their problems.

2. Learner corpora and lexis

It is quite a journey for the learner of a foreign language to enrich the vocabulary. Understanding word meaning starts from its recognition in the context, and the next step involves the ability to provide a particular word in an appropriate context (production). In terms of the learning process, learner corpora shed light on word knowledge and reveal the items that have made or could not make it into productive use (Cobb, 2007). Learner corpus shows us whether a learner knows how the word collocates with other words and which multi-word units should be used in the context. For instance, full knowledge of a word like *wind* means knowing that it occurs numerously in sequences like *wind blowing* and, also, in less frequent idiomatic expressions like *gone with the wind*.

One of the core issues of learner corpus data is calculating the frequency of specific words. A learner corpus is also suitable to look for trends and patterns that are not readily evident to the naked eye. By way of illustration, Altenberg and Granger (2001) inspected whether French and Swedish learners of English over-or underused the verb *make* in their writing. The question was answered by the contrasting learner and native corpora. As a result, the Swedish learners were found to use *make* slightly more frequently than the native speakers, while the French speakers used it substantially less often.

Another study worth mentioning by Granger and Tyson (1996) found that French learners in their essays tended to overuse moreover and underuse however and therefore and generally overuse highly familiar all purpose-words, frequent nouns – Hasselgren (1994: 237) identifies them as “lexical teddy bears”.

Many other findings might be cited to highlight the fact that the powers of simple frequency counts throw light on learners’ lexical development.

3. Contribution of learner corpora to pedagogy

Learner corpora hold a tremendous potential for pedagogical studies. Despite the considerable number of studies about pedagogical learner corpora since the 1980s, it is still an emerging concept. Although the learner corpus is similar to a reference corpus, it is geared to the needs of learners. The corpora created for linguistic research can also be employed in language teaching.

Learner data can make the language acquisition process more focused by raising awareness of problematic areas or enabling the learners to consult native speakers’ corpora to correct errors that they or their teachers bring to attention. Moreover, learner corpora allow the learners to improve the accuracy of specific aspects of their writing (O'Sullivan and Chambers, 2006). For example, the Sketch Engine provides "word sketches" and summaries of a word's grammatical and collocational behaviour" (Kilgarriff et al. 2004). Frequency, in this sense, is a key factor, as corpus-based studies aim to give some descriptions of what is frequent and typical in the corpus under examination and are thus ideally suited for studying the linguistic features of academic discourse. It can highlight the words, phrases or structures most typical of the genre and how they are used.

Another pragmatic application of the learner corpus data is related to practical activities: learners can compare native speaker and local learner corpora to create exercises based on the non-native speaker data. This approach thus is seen as a development of John's (2002) concept of data-driven learning (DDL). For instance, Rankin and Schiftner's (2011) study of prepositions in the semantic field of aboutness can be mentioned in this respect. Interestingly, as the amount of local learner data was limited, they chose to add the L1 (mother tongue) German component of the International Corpus of Learner English (Granger et al. 2009). Having observed distinct patterns in the native speaker corpus, the authors asked the students to create vocabulary tasks based on the analyzed data.

There is an emerging trend in research in learner corpora and language learning when it comes to giving learners significant access to local learner corpus data (e.g. texts written by themselves or by their current and former classmates). Teacher mediation plays a particularly important role in this context. However, transferring this to the context of the everyday practice of teachers who are not researchers represents a challenge. Integrating the annotated learner data into the teaching environment can take time and effort. There is thus a substantial need for research regarding how to combine the learner corpus data in language learning and teaching in ways feasible for teachers who are not researchers in applied linguistics.

4. Data and Methodology

The present study suggests a corpus-based approach to investigate the lexical behaviour of learners: (a) frequency of specific words, as well as (b) collocate analysis. To achieve these goals the learner corpus was annotated and tagged using the part-of-speech tagger.

Before the creation of the English learner corpus of Georgian students, it was clear that standard practice in corpus design had to be followed, as recommended by corpus designers

(McEnery et al. 2006), in particular, the design key principles and suggestions for basic considerations in the design of learner corpora. A carefully-constructed corpus must be guided by certain design criteria, such as representativeness, sampling and balance.

The annotated and analysed corpus consists of 40 essays, comprising 12,000 words. The essays were written by 40 students of the faculty of Humanities (concentration – English Philology), Ivane Javakhishvili Tbilisi State University. Following the principle of corpus design, before handing in their assignments, Georgian learners were asked for informed consent.

The assignments are anonymously written and include detailed information about their age and gender. These details are essential for fine-grained, quantitative analyses. As it is an ongoing project, data collection started in September 2021. Part of the assignments was written in an electronic format, and some of them were handwritten. In this case, they were digitized. This can be difficult as the texts have to be reproduced exactly in the same format, without any change, including the learners' errors but without introducing additional ones. Illegible handwriting can further complicate the task because converting the data through optical character recognition (another method of collection) has proved to be quite challenging.

The written production of the learner corpus consists of argumentative essays, narrative essays, as well as proposals.

Here are the titles of essays suggested to English learners:

- How important is family for you?
- Is the family relationship the most enduring of all?
- Importance of make-believe games for children's development.
- Recall one memorable day from your childhood.
- "Marriages, like chemical unions, release upon dissolution packets of the energy locked up in their bonding" - John Updike's "An orphaned swimming pool".
- Happiness is there, in front of our eyes, but we don't see it. Miracles do happen.
- There are plans to demolish an old and unused building in the town where you are a student. You feel that the building should be saved. You decide to write a proposal for the town council explaining why you think the building should be preserved, suggesting what could be done to modernise it and saying how the building could benefit the local people.

Once the raw texts were collected, some mark-up was added, such as a header containing a reference and details about the text or meta-textual information within the text, etc. The software tools used for the data collection and analysis were as follows: First of all, TagAnt (Anthony, 2010) was utilized for data annotation, which made it simple to tag the texts according to the parts of speech they represented. The tagged data were analysed in the Lancsbox tool (Brezina, McEnery, 2021). The advantages of tagging learner corpora include the following: (1) Lexical and grammatical patterns can be automatically extracted; and (2) Much more information could be readily extracted.

Antconc (Anthony, 2010) is the other piece of software used for the present study. This suite of software tools is powerful for lexical analysis, the most common tools being Concord, WordList, and KeyWords, GraphColl.

5. Results

Lancsbox tool (Brezina, McEnery, 2021) calculates the number of highly frequent lemmas in the learner corpus. The general results for each one of the subgroups are shown in the following table:

Words	Frequency	Dispersion
people	56	1.957246
students	55	2.516859
love	46	2.186112
make	44	1.777418
think	41	2.163837
believe	39	1.807033
children	39	2.076358
time	32	1.537274
relationship	32	2.335401

because	30	1.394043
also	28	1.594233
important	19	2.156316
therefore	18	2.660121

Table 1. List of lemmata according to the frequency

It is interesting to notice that the most frequent word (adjective) in the list is "*important*", which seems to indicate a tendency of students to overuse it. As for some linking device, the most frequent one is "*also*" and "*therefore*" (this shows some type of progression in their use of linking devices, but they still overuse some terms).

Several collocational patterns are suggested in our study for the qualitative analysis. (See table 2). The results show that Georgian EFL students excessively use adjective + noun collocational patterns. As for the adverb + adjective pattern, the learners feel comfortable using highly familiar word combinations like very good, and very special, which indicates their lack of lexical richness and sophistication.

Lexical pattern	**Tags**	**Examples from the corpus**	**Frequency in texts**
Adjective + Noun	_JJ NN	Mental development; troubled relationship; major theme.	307 occasion; 37/40 texts

Noun + Noun	_NN NN	Out-of-box thinking; Self therapy;
		cocktail party;
		lower-class society.
			95 occasion;
			32/40 texts

Adverb + Adjective	_RB JJ	Always negative; really important;	94 occasion; 32/40 texts

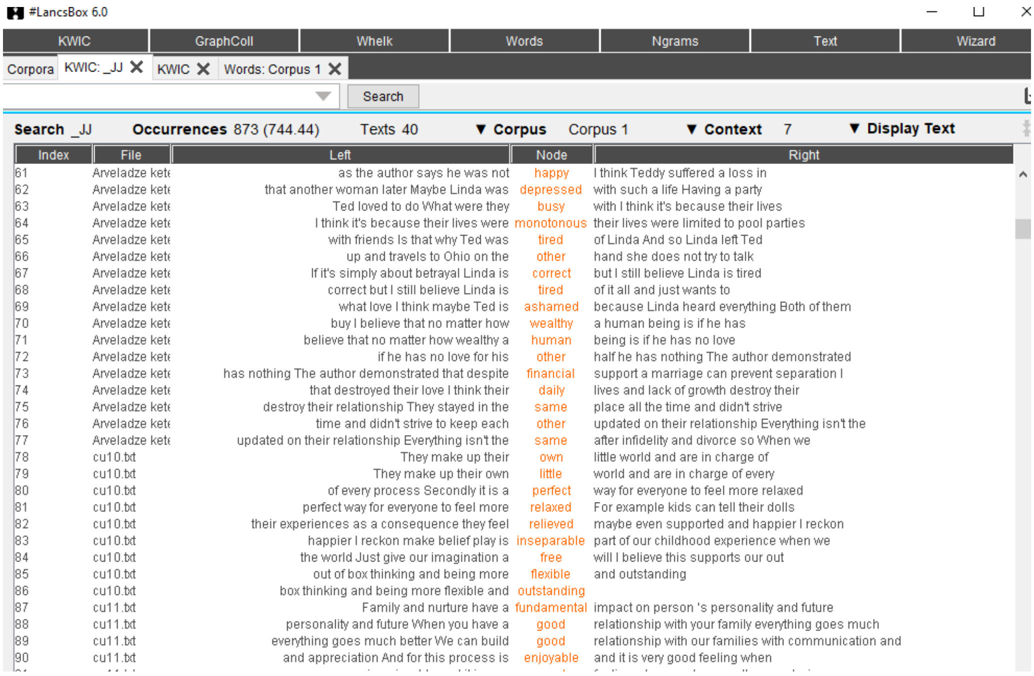
		very good; really happy; very special; particularly intriguing.	

Adverb + Verb	_RB V*	Actually be;
		wonderfully captures; frequently lack;
		finally left; really loved.
			237 occasion; 35/40 texts

Table 2. Collocate searches in Lancsbox Tool (Brezina, McEnery, 2021)

From a pedagogic perspective, some practical, learner corpus-based activities can be created. After analyzing adjectives in KWIC (key word in context) in Lancsbox tool, educators can motivate students by asking them to replace the key words (nodes coloured in red) either with synonyms or antonyms. To illustrate, see the contexts in Figure 1.

Figure 1. Key words in context in Lancsbox Tool (Brezina, McEnery, 2021)



Search _JJ Occurrences 873 (744.44) Texts 40 ▼ Corpus Corpus 1 ▼ Context 7 ▼ Display Text

Index	File	Left	Node	Right
61	Arveladze kete	as the author says he was not	happy	I think Teddy suffered a loss in
62	Arveladze kete	that another woman later Maybe Linda was	depressed	with such a life Having a party
63	Arveladze kete	Ted loved to do What were they	busy	with I think it's because their lives
64	Arveladze kete	I think it's because their lives were	monotonous	their lives were limited to pool parties
65	Arveladze kete	with friends Is that why Ted was	tired	of Linda And so Linda left Ted
66	Arveladze kete	up and travels to Ohio on the	other	hand she does not try to talk
67	Arveladze kete	If it's simply about betrayal Linda is	correct	but I still believe Linda is tired
68	Arveladze kete	correct but I still believe Linda is	tired	of it all and just wants to
69	Arveladze kete	what love I think maybe Ted is	ashamed	because Linda heard everything Both of them
70	Arveladze kete	buy I believe that no matter how	wealthy	a human being is if he has
71	Arveladze kete	believe that no matter how wealthy a	human	being is if he has no love
72	Arveladze kete	if he has no love for his	other	half he has nothing The author demonstrated
73	Arveladze kete	has nothing The author demonstrated that despite	financial	support a marriage can prevent separation I
74	Arveladze kete	that destroyed their love I think their	daily	lives and lack of growth destroy their
75	Arveladze kete	destroy their relationship They stayed in the	same	place all the time and didn't strive
76	Arveladze kete	time and didn't strive to keep each	other	updated on their relationship Everything isn't the
77	Arveladze kete	updated on their relationship Everything isn't the	same	after infidelity and divorce so When we
78	cu10.txt	They make up their	own	little world and are in charge of
79	cu10.txt	They make up their own	little	world and are in charge of every
80	cu10.txt	of every process Secondly it is a	perfect	way for everyone to feel more relaxed
81	cu10.txt	perfect way for everyone to feel more	relaxed	For example kids can tell their dolls
82	cu10.txt	their experiences as a consequence they feel	relieved	maybe even supported and happier I reckon
83	cu10.txt	happier I reckon make belief play is	inseparable	part of our childhood experience when we
84	cu10.txt	the world Just give our imagination a	free	will I believe this supports our out
85	cu10.txt	out of box thinking and being more	flexible	and outstanding
86	cu10.txt	box thinking and being more flexible and	outstanding	
87	cu11.txt	Family and nurture have a	fundamental	impact on person 's personality and future
88	cu11.txt	personality and future When you have a	good	relationship with your family everything goes much
89	cu11.txt	everything goes much better We can build	good	relationship with our families with communication and
90	cu11.txt	and appreciation And for this process is	enjoyable	and it is very good feeling when

For example, English learners can replace a node *good* by more complex equivalents, synonyms in order to enrich the vocabulary.

6. Conclusion

The close corpus-based analysis demonstrates that a crucial point about the exploration of concordances in Data-Driven Learning (DDL) activities is that students attempt to reach conclusions about usage through their own autonomous observation, enriching their lexis by searching for and replacing synonymous forms of the words. Moreover, learner corpora provide an enhanced description of learner language. According to our results, Georgian students tend to overuse highly frequent words and collocations.

Finally, there are outstanding possibilities of learner corpora in a process of language acquisition. Although, it is worth mentioning, that the use of learners' language for pedagogical treatment is something that teachers were doing long before learner corpora came onto the scene. The difference now is that this can be done with corpus linguistic techniques, such as using annotations, measuring, sorting, etc.). Consequently teachers can have more objective information about their students' difficulties, on the one hand, and more powerful tools with which to work on their students' data, on the other.

References

1. Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied linguistics*, *22*(2), 173-195.
2. Anthony, L. (2022). AntConc (Version 4.1.0) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from https://www.laurenceanthony.net/software
3. Belz, J. A., & Vyatkina, N. (2008). The pedagogical mediation of a developmental learner corpus for classroom-based language instruction.
4. Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). #LancsBox v. 6.x. [software package]
5. Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning & Technology*, *11*(3), 38-63.
6. Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, *15*(1), 17-27.
7. Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press.
8. Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics*, *4*(2), 237-258.
9. Johns, T. (2002). Data-driven learning: The perpetual challenge. In *Teaching and learning by doing corpus analysis* (pp. 105-117). Brill.
10. Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri-04-08 the sketch engine. *Information Technology*, *105*(116), 105-116.
11. McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Taylor & Francis.
12. O'Sullivan, Í., & Chambers, A. (2006). Learners' writing skills in French: Corpus consultation and learner evaluation. *Journal of second language writing*, *15*(1), 49-68.
13. Rankin, T., & Schiftner, B. (2011). Marginal prepositions in learner English: Applying local corpus data. *International Journal of Corpus Linguistics*, *16*(3), 412-434.
14. Reder, S., Harris, K., & Setzler, K. (2003). The multimedia adult ESL learner corpus. *TESOL Quarterly*, *37*(3), 546-557.
15. Selinker, L., & Gass, S. M. (2008). Second language acquisition. *Lawrence Erlhaum Ass*.

Author's email: *marine_makhatadze@yahoo.com*

Author's biographical data

Marine Makhatadze is a PhD student at Javakhishvili State University, currently working in Lexicography on the issue of learner corpora of English. Her interests include writing bilingual dictionary entries, translating both literary as well as scientific works. She has an experience of teaching English to multi-level classrooms. The author teaches at Tbilisi State University.