Diachronic investigation of learner language: Twenty years of the JPU corpus

József Horváth[1]

University of Pécs

DOI: 10.14232/edulingua.2020.1.3

Corpus linguistics studies have by now become a staple of linguists and teachers worldwide. Even practitioners who are not directly involved with corpus development or analysis are increasingly aware of this domain and its results. Thus, we can say that the time has come to investigate the long-term effects of the findings connected to corpus linguistics. This paper focuses on a specific sort of corpus: the learner corpus. It argues that what used to be a more traditional approach represented in the EFL (English as a foreign language) discipline has evolved into a perhaps more appropriate one represented in ELF (English as a lingua franca) partly because of the work of learner corpus research. To demonstrate any existing long-term effects of work with learner corpora on language education, an L2 corpus, the JPU Corpus, is presented. Five of the ten hypotheses originally set up in the early 2000s are revisited and critiqued by applying both quantitative and qualitative investigations. The results indicate that a diachronic learner corpus approach further establishes the shift from EFL to ELF approaches, a potentially useful and relevant change for students and their teachers across the world, especially within the framework of writing pedagogy.

Keywords: writing pedagogy, English as a lingua franca, learner corpora

Introduction

The educational and linguistic movements of the 1990s were truly momentous in Europe. It was the time, among others, that corpus linguistics experts began turning to learner language studies, building the first English as a foreign language (EFL) corpora as well as other L2 corpora. Today, looking back at those early beginnings, we can witness in those efforts the first winds of another change that would take place later: the emphasis on English as a lingua franca (ELF), a framework that offers equal status to all members of the English speaking community. In this sense, learner corpus (LC) researchers and practitioners can be said to have been the forerunners of an educationally, culturally, and socially significant change.

[1] Author's e-mail: horvath.jozsef@pte.hu

To probe a dimension of that change, the present article undertakes to demonstrate a unique case study, involving one of the earliest EFL corpora, developed by Horváth (2001).

Literature review

A corpus is a principled collection of written *or* spoken, or written *and* spoken texts that serves some descriptive purpose. Following the work of such classic corpus linguistic (CL) studies as Sinclair's (1991) and Kennedy's (1998) that identified and assessed the essential conceptual, theoretical, and empirical aspects of CL, the 1990 saw the first attempts to put this theory into practice as a driving force within learner language studies (see, for example, Granger, 1994, 1998). As described by Horváth (2001, p. 2), learner corpora can aid in the development of at least three types of pedagogically and linguistically relevant projects: those concerned directly with language use, those that yield valid research results, and those that may inspire pedagogical applications. It is worth considering how novel and unique the CL approach was at the time, when a large segment of the profession up to that point had been restrained to a prescriptive endeavor in many contexts, especially within the framework of error analysis.

With the advent of LC studies, however, a descriptive analytical focus started to gain ground, showing the first indications that a traditional EFL approach was slowly being replaced by an ELF principle, which represented a more modern, equitable, and possibly more sustainable educational cultural milieu. It would be hard to overestimate the significance of LC studies informing writing pedagogy, especially in the European educational context, but also in Japan and elsewhere (Mindt, 1997; Ooi, 2002). One example of this development was the increase in motivation at seeing one's work not as a repository of mistakes and errors but in its own right, as a text that is worth looking at from a more content-oriented and stylistic point of view. When all of this is done in order to collect evidence for valid descriptive, rather than prescriptive purposes, the result was a significant achievement, even if in some cases this may have been a mere side effect rather than a principled intention.

From these beginnings, we have witnessed a rapid spread of the approach, with the early 2000s and the past decade both yielding significant research informed by the LC stance. Authors such as Flowerdew (2014a, 2014b), Izumi and Isahara (2004) and Zhang (2014) have analyzed, respectively, notions pertaining to academic contexts of student writing in general, revisited error analytical tenets within the LC study framework, and compared and contrasted L2 and L1 data to capture the degree to which notions of overuse were justifiable.

Furthering the scope of working with learner corpora are attempts that innovate the domain, and it is refreshing to see that such efforts are continually in the forefront of the applied linguistic arena – see, for example, the groundbreaking investigations of Bolton, Nelson and Hung (2004) and Doró (2015, 2016). Part of this trend in LC studies

is the work of Horváth (2001, 2013) which aims to bridge a gap between linguistic investigations with writing pedagogy.

There seems to be little doubt about the observation that CL and learner corpora are here to stay. The discipline has identified a valid and reliable methodology for corpus development and implemented appropriate ethical and legal procedures as it continues to encourage independent, autonomous student involvement (see, for example, the seminal work of Johns, 1991, who first started applying this technique with his students).

The JPU Corpus

This paper revisits a LC study from 2001, so it seems necessary to introduce some essential components of the corpus itself.

The name is JPU Corpus, one of the largest learner corpora twenty years ago. An English written corpus developed at a university in Hungary, it contains over 400,000 words in 332 scripts, each by a different student. (The corpus is available for concordancing applications on Cobb's website, 2020). According to Horváth (2001), it is made up of five sub-corpora, each representing a course where the texts (essays and research papers) were produced. The sub-corpora are:

- Early stages of corpus development (E), 30 scripts
- Language practice courses (LP), 74 scripts
- Writing and research skills courses (WRS), 130 scripts
- Russian retraining course (RR), 16 scripts
- Postgraduate courses (PG), 82 scripts

For more details on corpus development, contents, and rationale, see Horváth (2001). (See the Appendix from a representative script from the corpus.)

Ten hypotheses were set up after the development of the JPU Corpus, with the intention of identifying pedagogically and statistically significant facets of the contents of the scripts contained in the corpus. The ten hypotheses were as follows:

Hypothesis 1: The RR subcorpus will contain a number of inaccurate uses of the definite article. (One of the five hypotheses investigated in this study, this will be referred to as the *Definite article hypothesis* in the rest of the article.)

Hypothesis 2: The coordinating conjunction *but* will be most frequently used as a transitional phrase.

Hypothesis 3: There will be a relatively high frequency of *above* in anaphoric verbal phrases, with a significant verbal collocate being *mention.*

Hypothesis 4: Aims of writing will be primarily identified by the I *would like* + *to infinitive* structure. (The second hypothesis studied in this paper, this will be referred to as the *Aim hypothesis*.)

Hypothesis 5: There will be an overuse of the epistemic stem *I think.* (The third hypothesis studied in this paper, this will be referred to as the *I think hypothesis*.)

Hypothesis 6: The adverb *very* will have a high frequency in the JPU Corpus, but less significant in the PG and the WRS sub-corpora.

Hypothesis 7: The frequency of *case, thing, good, interesting,* and *etc*. will be lower in the PG and the WRS sub-corpora than in the rest of the JPU Corpus.

Hypothesis 8: There will be lower frequencies for *the fact that* and *in order to* in the PG and the WRS sub-corpora than in the rest of the JPU Corpus.

Hypothesis 9: There will be a correlation between the type of introductory sentence and the length and vocabulary of it. (The fourth hypothesis studied in this paper, this will be referred to as the *Introduction hypothesis*.)

Hypothesis 10: There will be a correlation between the type of concluding sentence and the length and vocabulary of it. (The fifth hypothesis studied in this paper, this will be referred to as the *Conclusion hypothesis*.)

Research question

Following these theoretical and corpus-development considerations, the question arises whether LC studies may hold long-term relevance for the teacher profession. Thus, the current paper undertakes to investigate five of the ten hypotheses first presented in the early 2000s to identify and evaluate the specific areas where relevance can be detected. The five selected for this study are the ones that seem to have most relevance after two decades.

Method

As we have seen, the JPU Corpus was developed in the late 1990s, with the first analyses published in the early 2000s (Horváth, 2001). For the current investigation, the method applied was as follows. Five of the ten hypotheses (*Definite article*, *Aim*, I *think*,

Introduction, and *Conclusion*) set up to investigate the JPU Corpus twenty years ago are briefly described with their original results, followed by an interpretation of the validity and reliability of those results as well as an interpretation of their potential validity in the current situation.

Results and discussion

*The **Definite article** hypothesis*

The original hypothesis twenty years ago claimed that the RR sub-corpus would contain a number of inaccurate uses of the definite article, the rationale being that students who contributed to this sub-corpus had had little exposure to English, having had to retrain from being teachers of Russian to teachers of English due to political changes. In addition, the lack of the definite article in Russian further exacerbated the issue. Upon investigating the scripts, no significant incorrect use of the article was found in the corpus, with the explanation that students in the RR course were exceptionally motivated for professional reasons.

Today, looking back at the essence of this investigation, we can identify in it an attempt to turn something negative (error) into something positive (solution). The LC served to test the hypothesis and revealed a significant result when the hypothesis was rejected.

Notwithstanding, the definite article continues to pose problems for many students around the world, and we can state with confidence that inquiring into how it is learned, taught, and used has not lost any of its relevance in CL as well as in writing pedagogy.

*The **Aim** hypothesis*

As Horváth wrote (2001, p. 126),

> the study of [the issue of various collocates of *I* in expressing the aim of texts] was necessitated by a potential pedagogical outcome: I wished to gather data on what the 332 writers of these texts identified as their aims and methods in their texts, either in explicit thesis sentences and statements of method or in topic sentences referring to a particular point made in the main body of the text. This information is necessary to form an overall view of the types of aims students identified for their scripts, and can serve as the basis of evaluating writing strategies in students' texts.

It is not unusual that students, in Hungary and elsewhere, have vague notions about the aim of their essays and research papers, which may have been one reason for the preponderance of the stem "I would like to" in this position. Other, seemingly more academic expressions were seldom employed by them.

What we can add today is that the perhaps naïve sounding expression of an aim does not necessarily have to be regarded as inadequate – rather, a stage of the development of the learners who are on their way of integrating with the academic discourse community specifically in terms of the written standards. How this written form of expression is related to spoken interactions with peers and professors can be an exciting direction for future research. It is hypothesized that in delivering lectures, professors, too, tend to overuse the "I would like to" phrase, which may then become an analyzed chunk relegated to the acceptable academic cluster.

*The **I think** hypothesis*

The third hypothesis revisited is concerned with the epistemic stem *I think*. Its frequency in the JPU Corpus (normalized for 200,000 words) was contrasted with a similarly normalized frequency in the International Learner English Corpus (ICLE) and a native English written corpus (L1). The JPU data (21) came in between the L1 corpus result (3, least frequent) and the ICLE result (72, most frequent).

It is in relation to such comparative and contrastive investigations that currently the views may be different from those twenty years ago. Without knowing more details about the contexts of the actual texts in which students were working in the three situations and a closer analysis of what the thoughts were that followed the epistemic stem it would be hard to argue that a simply quantitative analysis would be the right basis for any meaningful pedagogical action. Moreover, today it may sound not just a little odd to consider L1 student users of English the standard with whom, for example, non-British L2 users of English be compared. Changes in how we perceive writing and, in general, literacy, have meant that in the current ELF context a more equitable approach be, if not adopted, at least adapted for the specific circumstance in which one operates.

There is also the question, tightly connected to the previous argument that the time may have come for a re-interpretation of the notions of overuse and underuse. In fact, even in the past two decades, one may argue that, from a strictly descriptive CL standpoint, the relative value loading of frequencies as can be detected in L1 and various L2 corpora was fundamentally flawed, as it introduced a prescriptive element, which some may be inclined to classify as alien to the CL endeavor. Whether that is the case remains to be seen.

*The **Introduction** hypothesis*

The Introduction hypothesis and the Conclusion hypothesis followed the same pattern in their initial investigation twenty years ago, the only difference being which part of the scripts was involved. As far as the Introduction hypothesis is concerned, the

investigations revealed the stratification of introductory themes, presented in Table 1 (Horváth, 2001, pp. 132-133):

Table 1: Rank order of types and frequencies of introductions in the WRS sub-corpus of the JPU Corpus

Rank	Type	Frequency
1	definition	47
2	personal	15
3	obvious	12
4	historical	10
5	aim	7
6	method	4
7	five short terms	3
8	citation	2
9	reader ambiguous narrative question title	1

An ANOVA statistical test was run, which revealed no significant correlation among themes and sentence length. Although it was felt that ambiguous and obvious ways of introductions (that is, qualitative features) may correspond with token and type figures (that is, quantitative features), no such observation could be made.

Often, researchers prefer it when their analyses reveal significant results, as this is favored by the academic community at large. However, looking back, we can perhaps agree that it was just as well that no such correspondence could be quantified and verified, as this result lent further credit to the principle that in evaluating a student's approach to a theme, the relevant, inherently necessary, or even original and unexpected framing of a topic is what eventually matters. (The current paper, incidentally, uses a historical reference in its introduction.)

*The **Conclusion** hypothesis*

As we have seen, this hypothesis only differed from the previous one in that it looked at the conclusions of the scripts. The ANOVA test, however, did yield significant results for the Conclusion hypothesis. Table 2 shows the types and frequencies of conclusions in the sub-corpus.

Table 2: Rank order of types and frequencies of conclusions in the WRS sub-corpus of the JPU Corpus

Rank	Type	Frequency
1	qualitative	47
2	practical	26
3	obvious	9
4	unclear	7
5	quantitative	5
6	question	3
7	hypothesis limitation non-sequitur	2
8	citation reader	1

The statistical test revealed a significant relationship (p=0.02). As Horváth documented (2001, p. 138), the hypothesis "claiming that type of sentence affected length was verified." Further: whereas "the mean length of the qualitative and practical type of concluding sentences was almost identical (23.36 vs. 24.23 words), the length of the combined group of obvious and unclear type sentences was 15.62, for which the analysis confirmed significant variation.

That is, it was revealed twenty years ago that students who tended to have more vague notions of how to arrive at the end point of their papers showed that hesitation in cutting their ideas shot and tending to produce significantly shorter concluding sentences.

Today, we can add that even though this was a statistically verifiable result, its practical implications would need further substantiation. There is nothing wrong with a shorter than average sentence length. In fact, many such utterances may be effective and memorable, often requiring serious thought and continuous revisions to produce. (A preview: the concluding sentence of this article combines the limitation technique with that of the hypothesis.)

Conclusion

This paper has attempted to provide an example of how learner corpus studies can be conducted with a diachronic set of goals, by comparing initial investigations and their results with what we can observe by exploring the current educational landscape. Using the example of an early LC, the JPU Corpus, it has demonstrated what can be regarded as constant and what can be regarded as different when investigations are extended to the present and brought under critical scrutiny. It has also argued that notions of overuse and underuse may be regarded as potentially problematic modes of approach in LC research because of the disproportionate emphasis they often seem to lay on prescription on lieu of description.

Similar longitudinal and diachronic research can be implemented with other learner corpora as well as across different learner corpora. For this to be feasible, we need to establish factors, such as original aims of corpus development and content features, that would yield meaningful results as well as inspire application in relevant classroom contexts.

Inevitably, however, this paper has limitations, too. It had to remain, for the present purposes, within the confines of comparative and contrastive analyses, leaving questions such as data-driven learning applications and corpus data adaptations for such tasks unanswered. It is hoped that such studies will be conducted in the near future so that there may be a real continuity of efforts, some of which inspired by past and contemporaneous LC studies.

References

Bolton, K., Nelson, G., & Hung, J. (2002). A corpus-based study of connectors in student writing: Research from the International Corpus of English in Hong Kong (ICE-HK). *International Journal of Corpus Linguistics*, *7*(2), 165-182. https://doi.org/10.1075/ijcl.7.2.02bol

Cobb, T. (2020). Concordance. *Compleat lexical tutor: For data-driven learning on the web*. Available at https://www.lextutor.ca/conc/eng/

Doró, K. (2015). Changes in the lexical measures of undergraduate EFL students' argumentative essays In P. Pietilä, K. Doró & R. Pípalová (Eds.), *Lexical issues in L2 writing* (pp. 57-76). Newcastle upon Tyne: Cambridge Scholars Publishing.

Doró, K. (2016). Linking adverbials in EFL undergraduate argumentative essays: A diachronic corpus study. In M. Lehmann, R. Lugossy & J. Horváth (Eds.), *UPRT 2015: Empirical studies in English applied linguistics* (pp. 152-165). Pécs: Lingua Franca Csoport.

Flowerdew, L. (2014a). Corpus-based analyses in EAP. In *Academic discourse* (pp. 105-124). Routledge.

Flowerdew, L. (2014b). Learner corpus research in EAP: Some key issues and future pathways. *English Language and Linguisics, 20* (2), 43-60. https://doi.org/10.17960/ell.2014.20.2.003

Granger, S. (1994). Learner corpus: A revolution in applied linguistics. English Today, 10 (3), 25-29. (ERIC Document Reproduction Service No. EJ 514 886)

Granger, S. (Ed.). (1998). *Learner English on computer. Studies in language and linguistics*. London: Longman.

Horváth, J. (2001). *Advanced writing in English as a foreign language: A corpus-based study of processes and products*. Pécs: Lingua Franca Csoport. Available at https://books.google.hu/books?id=XsmxiVsSPjAC

Horváth, J. (2013). What are BA and MA students proud of? *Argumentum*, *9*, 178-185. Available at http://argumentum.unideb.hu/2013-anyagok/kulonszam/08_horvathj.pdf

Izumi, E., & Isahara, H. (2004). Investigation into language learners' acquisition order based on an error analysis of a learner corpus. In *Proceedings of IWLeL 2004: An interactive workshop on language e-learning* (pp. 63-71). Available at https://pdfs.semanticscholar.org/1cb7/53bc79591978cc1df3a7cbd1baa3927b3e15.pdf

Johns, T. (1991). Should you be persuaded: Two examples of data-driven learning. *ELR Journal, 4,* 1-16.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Mindt, D. (1997). Corpora and the teaching of English in Germany. In A. Wichmann, S. Fligelstone, T. McEnery & G. Knowles (Eds.), *Teaching and language corpora* (pp. 40-50). London: Lomgman.

Ooi. V. (2002). From Shakespeare to Hungarian EFL writing: Using www corpora to motivate student learning. In M. Tan (Ed.), *Corpus studies in language education* (pp. 163-177). Thailand: Assumption University.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Zhang, R. (2014). Overuse and underuse of English concluding connectives: A corpus study. *Journal of Language Teaching & Research*, *5* (1), 121-126. https://doi.org/10.4304/jltr.5.1.121-126

Acknowledgment

This paper is dedicated to the memory and outstanding work of Tim Johns (1936-2009).

Appendix: A script from the JPU Corpus

W 057 F

INTRODUCTION

Students at Janus Pannonius University from Pécs have the opportunity to attend the Writing and Research Skills seminar where everyone can learn about and improve writing skills. For the sessions students have some writing tasks to do, for example: preparing essays on topics suggested by the teacher or chosen by the students; and accomplishing a small-scale research paper.

During the Fall 1998 semester we used for several times William Zinsser's book On Writing Well: An Informal Guide to Writing Non-Fiction, which was a great help for us in forming our writing style.

Zinsser is a free-lance writer, the author of humorous and non-fiction writings. He was a teacher as well: he taught even at Yale University. His main teaching and conception on writing well is to write as simply as possible and to avoid clutter. Cluttering is when a writer expresses ideas, thoughts with the help of more words than needed making confusing sentences. Simplicity is quite the contrary of clutter: it shows clear thinking. Clear and simple sentences are easy to understand but it is not always easy to produce them. Zinsser says: "Writing is hard work. A clear sentence is no accident. Very few sentences come out right the first time, or even the third time" (Zinsser 1998, 12).

In the chapter entitled "Simplicity" from the book On Writing Well, Zinsser presents two pages of an earlier version of the same book and I found the way he simplified his sentences very exciting.

I thought it would be interesting and useful for me as well as for other students to compare the two versions and to see how a professional writer gets his final edition.

METHOD

As its title itself shows, the chapter "Simplicity" deals with the importance of simple and clear writing and explains that "the secret of good writing is to strip every sentence to its cleanest components" (Zinsser 1998, 7).

As an example for how to simplify sentences, the author shows the reader on pages ten and eleven a piece of draft, an earlier version of the final, edited work. On the two pages of the draft we can see a text and a lot of corrections on it: words and sentences crossed out, some new words written between the lines replacing something dropped out.

I wanted to know how many words were dropped out from the text and how many were substituted with new words and expressions, so I counted every item from both versions. I was also curious to find out if the content of the final version changed as compared to the content of the draft.

RESULTS AND DISCUSSION

According to Zinsser, the draft shown on pages ten and eleven was already the fourth or fifth version and he could still find a great number of unnecessary words in it.
There were 538 words in the text shown on the two pages of the draft, and after the author revised it once more, he left out 123 elements, which means that almost one fourth of the text (22.8 % exactly) was left out finally. In spite of the large number of omitted words from the draft the final text does not change in meaning, it has the same content and meaning but it is shorter and more simple, thus it is easier to follow and to comprehend the ideas the author shares with his readers.

There were only seventeen additions or substitutions. Zinsser replaced several phrases or even sentences with shorter terms, for example: constructions of a definite article and a noun were replaced by personal pronouns; sentences were substituted with verbs; adjectives were left out where the noun carried the meaning of the adjective; noun phrases were replaced by nouns and long verb phrases with short verbs with the same meaning. In three cases he left out complete sentences without any replacement.

A very common problem of both professional and amateur writers is the use of redundant adjectives, I mean the use of adjectives that are not necessary for the understanding of the noun they belong to. That is why I expected adjectives to be the most numerous among those items that were crossed out but I was wrong. They were on the third place on the list. The group of verbs was the leading one of the list containing different parts of speech. The most frequent types of omission were: the verb ”to be” and verbs in the infinitive form. See Table 1 to find out about the number of words belonging to different parts of speech.

Part of speech	**The number of left out items**
Verbs	28
Adjectives	15
Adverbs	14
Nouns	13
Articles	8
Consjunkctions	3
Others	22

Table 1: The number of omitted words belonging to certain parts of speech in decreasing order

Zinsser considers revising to be very important. Revising our writings we can realize how many words and phrases can still be omitted, changed or replaced by shorter ones. He says: “Be grateful for everything you can throw away” (Zinsser 1998, 18).

Simplicity assumes brevity and clarity of thoughts and expressions, clutter is everything that can be left out without altering the meaning of what we want to express.

CONCLUSION

Simplicity makes a writing valuable. Sentences with many unnecessary elements in them, very elaborate and confusing sentences, or simply: cluttered sentences, will make the reading difficult.

Zinsser in his work gives several writing tips for those who want to improve their writing style, for those who would like to learn how to be simple in our writings. He convinces us to revise all the time what we write and to drop out as many elements as needed. The most important thing according o him is to be as simple and clear as possible.

Everyone who wants to be read must think first of all of the reader, “this elusive creature” (Zinsser 1998, 9), whose attention must be captured. If the reader gets lost among the confusing ideas of a writing, he will stop reading.

Cluttered sentences make the reading difficult. Simple sentences are easy to understand and we must always try to simplify for the sake of good writing and for the sake of better understanding.

We must try to substitute long subordinate clauses with some verbs that contain the same meaning, to use adverbs and adjectives where they are required, to be attentive at nouns that already carry the meaning that can be expressed also with an adjective.