INTRODUCTIONS AND CONCLUSIONS IN ADVANCED EFL STUDENTS' WRITING: EVIDENCE FROM THE CORPUS

József Horváth
University of Pécs
Department of English Applied Linguistics
joe@btk.pte.hu

A crucial issue in any analysis of language is the role of data. Evidence sought to support a theory of structure or language use provides the basis on which to evaluate the feasibility and applicability of a hypothesis. The role of linguistic evidence also has practical implications in language education, as it impacts on the manner in which a syllabus is presented (Seliger & Shohamy, 1989). As the examples may be either intuitive (coming from the linguist's own repertoire) or observed (recorded in some psycholinguistic elicitation or field work), the issues of competence and performance present the framework in which this question has been studied.

One of the leading figures in corpus linguistics applying machine-readable collections, Leech (1997), defined a corpus as "a body of language material which exists in electronic form, and which may be processed by computer for various purposes such as linguistic research and language engineering" (p. 1). Computer corpora of both spoken and written language, especially for English, have been used increasingly in linguistics (Sinclair, 1991; Kennedy, 1998; Renouf, 1997; Horváth, 1999). From the early 1990s, we have also seen a growing interest in collecting language samples from students of English -- these learner corpora represent an exciting new domain of linguistic and pedagogical pursuit (Granger, 1998; Kaszubski, 1998; Ringbom, 1998). In this paper, I will aim to demonstrate how a language teacher's pedagogical concerns may find a suitable framework in corpus linguistics (CL) studies. First, then, let's see why I care about CL.

Learner corpora

The rationale of corpus linguistics is to directly access, derive, and manipulate evidence from a collection of texts. Early CL achievements include the Brown and the LOB corpus. More recent, and more influential, large English corpora are the Bank of English and the British National Corpus, which contain millions of words in spoken and written samples, occurring in natural contexts.

Learner corpus projects can be seen as a natural extension of the interest in language sampling. They were launched in the early nineties, partly to satisfy a need to verify or refute claims about transfer from the mother tongue to the foreign language. Among these drives, the Louvain-based International Corpus of Learner English (ICLE) was the forerunner. Conceived by Granger, the ICLE collection of written texts by advanced students of EFL aims to be the

basis of lexical, grammatical and phraseological studies. The main objective is to gather objective data for the description of learner language. Besides, the ICLE's contribution has been in directing attention to the need for observation of this language so that the notion of L1 transfer may be analysed under stricter data control. The obvious potential outcome is for materials development projects, which will help specific classroom practices. Focusing on error analysis and interlanguage, the ICLE-based project enables researchers and educators to directly analyse and compare the written output of students from such countries as France, Germany, the Netherlands, Spain, Sweden, Finland, Poland, the Czech Republic, Bulgaria, Russia, Italy, Israel, Japan and China.

The JPU corpus

I have been teaching EFL at the University of Pécs (formerly Janus Pannonius University) for ten years, primarily Language Practice and Writing and Research Skills (WRS) courses. In 1992, I began collecting students' texts. Students were free to participate or not participate in the effort. For each text, two types of data were recorded: each script was saved to computer disk, and the information on the student and the course of origin for the script was also saved.

Over the years, the JPU Corpus has grown tremendously; by now it contains over 400,000 words. It is a semi-annotated collection: it has author, gender, year, course, and genre information tagged to it, but it does not take advantage of any of the robust tagging techniques available today (for a detailed discussion on the development of the corpus and other related issues, see Horváth, 1999).

Two major types of text are represented in the corpus, which also account for most of the assignments that students submit at the English Department of the university: essays and research papers. In this paper, I will focus on the latter component in one of the five sub-corpora, which is made up by scripts written by 130 students in my Writing and Research Skills courses (predominantly first-year students with little previous experience in reading and writing research papers even in their first language).

Two hypotheses: Introductions and conclusions in the JPU corpus

In terms of number of scripts and types of words, the Writing and Research Skills sub-corpus (WRSS) is the most representative, with its 130 texts (by 106 female and 24 male contributors). The text types represented by the WRSS are personal essays (23), with the rest of the collection (107 scripts) made up by research papers.

As a teacher of these courses, one of my concerns was how I could help students discover and experiment ways in which authors in introducing a text can arouse interest, both in personal writing and research reports. Another concern has been the various techniques authors use to conclude their texts. Thus, a number of classes, written assignments, readings, and office-hour meetings have been devoted to these two structural units (for a discussion of my writing pedagogy, see Horváth, 1999).

For the present investigations, I selected the research paper samples of the WRSS. The majority of scripts, 107, were submitted as the final research paper requirement of the course (the rest of the sub-corpus, the 23 essays, were excluded from this analysis). This collection represents a valid basis on which to test two hypotheses: one related to introductions, the other to conclusions.

The investigation of the types and composition of these introductions and conclusions was motivated by the linguistic and pedagogical concern with the importance of drafting and revising introductory and concluding matter. By looking closely at this sample, we can gather useful information on students' choices, using authentic data that can be exploited for future language education.

The introductory sentences

Of the 107 papers, 33 discuss aspects of Hungarian newspaper articles published on the day students were born. This option was designed to include a personal intrinsic motive for students to begin to want to do research. The high number of such papers seems to prove that the approach was successful. However, a large number of other content and method types are also represented in this sub-corpus – these themes are listed in Table 1.

Table 1: Themes of the 107 research papers in the WRSS

Type	Number
Newspaper articles from the day student was born	33
Analysis of students' writing	30
Survey among students	20
Word processing for writers	4
Types of revision	3
Analysis of WRS course tasks, readings, procedures	2
Analysis of Umberto Eco's writing	2
Survey among teachers	2
Analysis of teacher's comments on portfolios	1
Analysis of essay test markers' comments	1
University syllabus analysis	1
Analysis of writing textbooks	1
Introductions in an anthology of essays	1
Analysis of introductions in scholarly papers	1
Analysis of narrative essay	1
Analysis of Zinsser's notion of simplicity	1
Models of paragraph	1
Analysis of structure in research papers	1

Proficiency test for high-school students	1

The hypothesis claimed that the type of introductory sentence chosen by students would affect the length and vocabulary of the first sentence. Besides, I aimed to gather descriptive information on the frames of this language sample (Andor, 1985). To test the hypothesis, the first sentence of each introduction was saved as a separate document, which was then processed by the concordancer, also calculating tokens, types, and average sentence length in different groups. In short, the introductory sentences were treated as a mini corpus. Besides these measures, a table was also designed, listing the types of introductions observed.

The mini corpus of these sentences contained 1,946 words, of 579 types, a ratio of 3.36. The average length of a sentence was 18.18 words.

To test the validity of the hypothesis, I performed a content analysis of the sentences, using categories. Initially, I identified five categories to capture the types of frames of the introductions, representing different approaches I knew students employed in their texts. These included

- ∑ describing a *personal* incident related to the theme (e.g., “Having read the newspaper issue of Kisalföld [a Hungarian regional newspaper] of 14th September 1978, a whole new world opened to me.”)
- ∑ identifying a relevant *historical* detail (“In June 1979 Leonid Brezhnev paid a visit to Hungary.”)
- ∑ opening with a *narrative* (“The first thing that many people do in the morning is opening one of the daily newspapers and browsing among the articles.”
- ∑ giving a *definition* of a field, an issue or a problem (“Students’ opinion about syllabi can influence the popularity of courses.”)
- ∑ beginning the text with *five* semantically germane nouns, verbs or adjectives (“Clutch, weep, glare, jerk, loathe.”)

The last of these introductory frames was first employed and practiced, primarily for personal descriptive and narrative essays, in the WRS course in the Spring 1998 semester.

In categorizing the introductory sentences, I scanned them for traits of these frames. As some introductions did not fit into the original categories, new ones were set up:

- ∑ stating a matter clearly *obvious* for the intended reader, often containing determiners such as *every*, *each*, *all*, or adverbs like *always* (e.g., “Newspapers are used for informing the population about how the society works and what goes on all over the world.”)

- ∑ stating the *aim* of the paper ("In this paper my aim is to compare two Hungarian daily newspaper issues...")
- ∑ defining the *method* of the investigation ("One possibility to gather information about a period of time is to read newspapers.")
- ∑ directly addressing the *reader* ("Reading old newspapers may make you realize what has and what has not changed during the years.")
- ∑ including a direct or indirect *citation* from a source ("According to Harris (1993, p. 81), a general point about writing is that it cannot be seen in isolation...")
- ∑ asking a *question* ("What is exactly a portfolio?")
- ∑ beginning with the *title* of a source ("Bits & Pieces.")

These labels were then assigned to the introductory sentences. To test the reliability of the categorization, the same procedure was conducted a second time. In only two instances was there a difference between the first and the second result, which were identified with a question mark, and the first and second label recorded. Altogether, I identified twelve types of introductions in the WRSS sample, with the 13th represented by the problematic examples. When these measures were taken, the frequency of types was rank-ordered. The results appear in Table 2. The table shows overwhelming preference for four types of introduction: those based on definition, personal incident, an obvious issue, and historical detail. Altogether, the four types account for the majority of the papers, 83 out of 107.

Table 2: The rank order of types of introductory sentences in the WRSS sample

Rank	Type	Frequency
1	definition	47
2	personal	15
3	obvious	12
4	historical	10
5	aim	7
6	method	4
7	five	3
8	citation, reader, ? (obvious- definition; obvious-historical)	2
9	narrative, question, title	1

To confirm or refute the hypothesis that the type of introduction affected the length of the first sentence, I devised the following procedure. Of the 107 sentences, I selected the 83 that

belonged to the most popular options. As the rest of the sentences were each represented by only seven or fewer examples, they were eliminated from the investigation, as their low frequency would not have given sufficient information on length distribution. After this, I calculated the length of the each of the 83 sentences in the four main groups. When these indices were obtained, I determined the effect of the type on length via one-way analysis of variance (ANOVA). Table 3 presents the statistics.

Table 3: Results of the analysis of variance on the data of length of first sentences

Source	df	SS	MS	F	Pr[X>F]
Between	3	199.14	66.38	1.20	0.31
Residual	80	4410.10	55.13		
Total	83	4609.24			

According to the figures in the table, the ANOVA findings are inconclusive: no significant differences were found ($F = 1.20$; $p = 0.31$). The type of sentence did not affect its length. This result points to the need to analyse the full introductory paragraphs, so as to reveal how type may affect its size and structure.

The concluding sentences

Similarly to the importance of how a research paper opens the theme for the reader, in writing the conclusion's last sentence, the author has an opportunity to make a last and maybe lasting impression. In this investigation, I analysed the final sentences of concluding sections of the 107 papers, looking for the same types of information as in the previous study. This hypothesis claimed that there would be a number of types of concluding sentences, which in turn would affect their length and vocabulary. The procedures for testing this last hypothesis were the same as for the previous one.

The mini corpus of the concluding sentences was made up by 105 sentences -- two fewer than in the introductory mini corpus, as two students did not include a conclusion in their submissions. The sample contained 2,389 words, representing 818 types, resulting in a ratio of 2.92. The rounded average length of sentences was 23 words. When compared with the same statistics for the introductory mini corpus, we can see that concluding sentences tended to be somewhat longer, using more types of words on average than the introductory ones. However, the differences cannot be regarded as marked, as shown in Table 4.

Table 4: Descriptive statistics of the two mini corpora

Index	Introductions	Conclusions
Tokens	1946	2389

Types	579	818
Ratio	3.36	2.92
Average length	18.18	22.75

As for the typology of the last sentences, the following eight categories were set up initially:

- ∑ summary of a *qualitative* result (e.g., "The more senses are involved in learning, the deeper the learning will be.")
- ∑ summary of a *quantitative* result ("From the foregoing it is clear that all of the analysed essays except for one or two are better than the average.")
- ∑ statement of practical implication ("I also learnt about the relationship between journalism and the political life.")
- ∑ identification of *limitation* of study ("As the other classes during the semester were more or less active than the one dealt with in this paper, this research paper and the results of it can be applied to this particular class."
- ∑ a direct or indirect *question* ("I wonder how many findings will apply to me and my peers in the future.")
- ∑ identification of *hypothesis* or problem for future study ("It could be used for finding out why some important information was left out from Hungarian papers, and what they were.")
- ∑ *non-sequitur* or irrelevant notion ("Only children were excited when they were waiting for Santa Claus to bring them presents.")
- ∑ stating the *obvious* ("Other sources can be used as well for doing similar research on this topic, which would certainly enrich knowledge about this field.")

Again, not all concluding sentences could be grouped under these headings. The three new categories added were

- ∑ *citation* (e.g., "Such an essay test might be a torture for those students who dislike essay writing, but it 'continues to serve as a challenge for a number of students who have shown excellence in writing.' -- reports Horváth József….")
- ∑ addressing the *reader* ("Thank you for not leaving and reading the Research Paper.")
- ∑ *unclear* content or ambiguous ("With this paper I got the information, what I wanted to know.")

Each of the 105 sentences was coded, and the grouping double-checked. In the second analysis, the original division was found to be reliable. See types of concluding sentences in Table 5.

Table 5: The rank order of types of concluding sentences in the WRSS sample

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 two most popular last statements in the mini corpus were represented by the qualitative and the practical outcome types. This result is in line with previous pedagogical experience suggesting that student writers favoured these options. They also appear to be relevant for the types of research design the scripts were based on. However, the high ranking of the obvious type of sentence and of the unclear category calls attention to the need for more practice in the area of writing conclusions. This can be facilitated by channelling back the information on students' scripts to the writing course, using authentic student texts.

Finally, to test the relationship between type of concluding sentence and length, I employed a one-way analysis of variance test for types. I used the sentence-length data for the qualitative and practical groups, and the combined length for the obvious and unclear types. The results appear in Table 6.

Table 6: Results of the analysis of variance on the data of length of last sentences

Source	df	SS	MS	F	Pr[X>F]
Between	2	862.29	431.14	4.34	0.02
Residual	86	8539.22	99.29		

Total 88 9401.51

Qualitative Mean: 23.36

Practical Mean: 24.23

Obvious + Unclear Mean: 15.62

The table shows that the analysis revealed a significant effect of type of concluding sentence on length: $F = 4.34$; $p = 0.02$. 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. Thus, the hypothesis claiming that type of sentence affected length was verified.

This finding may imply that students who wrote the type of concluding sentences that were categorized as either unclear or obvious had difficulty ending their papers, and thus they opted to write much shorter sentences than others. However, factors such as grammatical accuracy of the sentences, the type of concluding sentence and the full concluding paragraph, and the appropriateness of the type of conclusion in relation to the body text of the research paper are to be investigated in the future.

Conclusion

In this article, I have attempted to illustrate how a language teacher may benefit from employing a CL method to test hypotheses about learner writing. By developing learner corpora, other teachers may conduct their own studies, relevant to their teaching aims and contexts. With the increasing interest in CL, and especially CL investigating learner English, there are also potentials for national and international cooperation, as shown by the example of the International Corpus of Learner English.

Students may not learn to write more effective introductions and conclusions as an automatic result of such analyses, but they, and their teachers, will have a clearer concept of what it is they write, which could have pedagogical significance.

References

Andor, J. (1985). On the psychological relevance of frames. *Quadreni di Semantica, 6* (2), 212-221.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). London: Longman.

Horváth, J. (1999). Advanced writing in English as a foreign language: A corpus-based study of processes and products. Ph. D. diss., Janus Pannonius University, Pécs [Online]. Available http://www.geocities.com/writing_site/thesis

Kaszubski, P. (1998). Enhancing a writing textbook: A national perspective. In S. Granger (Ed.), *Learner English on computer* (pp. 172-185). London: Longman.

Leech, G. (1997). Introducing corpus annotation. In R. Garside, G. Leech, & T.

McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 1-18). London: Longman.

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

Renouf, A. (1997). Teaching corpus linguistics to teachers of English. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 255-266). Applied linguistics and language study. London: Longman.

Ringbom, H. (1998). Vocabulary frequencies in advanced learner English: A cross-linguistic approach. In S. Granger (Ed.), *Learner English on computer* (pp. 41-52). London: Longman.

Seliger, H. W. & Shohamy, E. (1989). *Second language research methods*. Oxford: Oxford University Press.

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