József Horváth

What are BA and MA students proud of?

The case of the International Pride Corpus

Abstract

This paper presents the results of a new type of learner corpus analysis: one that is based on students' own choices. The study was part of an international questionnaire survey conducted in 2012, which aimed to collect valid data about attitudes toward and views of higher education of Hungarian and European students of English as a foreign language (EFL). After discussing the rationale for a new type of learner corpus, which in this case is the author's International Pride Corpus (IPC), the paper summarizes writing pedagogy studies dealing with notions of autonomy and originality, two values that have a central role in the current approach. The IPC is then submitted to thematic and lexical analyses, aiming to reveal differences between the BA and the MA subcorpora. Although the results appear inconclusive, the study is seen as potentially relevant for future corpus-building efforts as well as for the syllabus development of writing skills courses.
Keywords: writing pedagogy, corpus linguistics, vocabulary profile

The context for a new type of learner corpus

When working in courses devoted to the development of the writing skills of university students of English as a foreign language (EFL), we can implement a wide range of approaches and methods. Depending on individual and institutional considerations, we may choose to focus on form, rely on the achievements of the process approach of writing pedagogy, or establish a combination of grammatical, text-linguistic, pragmatic and, ever more widely, corpus linguistic tenets with which to construe and construct a syllabus (Grabe & Kaplan 1996; Horváth 2001; Hyland 2011). The role of students' input in this process cannot be over-emphasized, which is why there is an increased interest in contemporary writing pedagogy in questions that are posed (and often, answered) by the students themselves (Hyland 2002; Ivanič & Camps 2001; Zergollern-Miletić & Horváth 2009a, 2009b).

In this paper, I will present the results of an international survey conducted among BA, MA and PhD students of EFL. I was interested in what practices, behaviors and attitudes students would be willing to report about their studies in general and about writing skills in EFL in particular.

This paper (which extends one published in Hungarian: Horváth 2012) deals with the combination of EFL writing pedagogy (Grabe & Kaplan 1996; Pecorari 2003) and learner corpus research (Granger, 2003; Nesselhauf 2002). It is informed by the discovery that for pedagogical theory and practice to be meaningful in this domain, we need to collect valid data about both the processes and the products of writings skills development (Horváth 2001). In

addition, it acknowledges trends in current corpus linguistics that seek new ways of setting up corpora and implementing them in a variety of academic contexts (Andor 2010; Hollósy 1994; McEnery, Xiao & Tono 2006; Xiao 2008). In terms of the pedagogical aspect, according to Zinsser (1998), one cannot be sure that writing can be taught -- but he maintains, and provides valid examples to validate that claim, that one can learn it by following an appropriate, individualized approach.

The importance of the individual author's context, aims and abilities cannot be over-emphasized when dealing with two crucial values of writing pedagogy: autonomy and originality. It is via these two values that one can modulate a voice, an individual tone that one can regard as one's own (Matsuda 2001). Several practitioners and theoreticians of contemporary writing pedagogy point out the necessity of supporting the development of this voice via tasks and procedures that help novice writers position themselves in, rather than forcing them into, a writing context that fosters meaningful and relevant experimentation with topics, genres, and structures (Matsuda 2001; Zinsser 1998). The voice, or rather, voices thus developed are characterized and often determined by cultural and intercultural forces, besides being individual (Magnuczné Godó 2003). In the current study, the aim was to document a cross-section of mostly European students' views of where they saw themselves in this continuum by inviting them to lay the foundation of a new type of learner corpus: one that is pulled together by their own choices.

The notions of autonomy and choice are fundamental concepts of modern pedagogy (Géczi 2005). In terms of how they can and should manifest themselves in writing skills develop-ment, we have seen a number of attempts in the literature. For example, the need to modulate a personal voice in writing can be regarded as the appearance of both autonomy and choice. Closely connected to these notions, according to scholars such as Hirvela and Belcher (2002), Ivanič and Camps (2001) and Matsuda (2001) is the issue of originality and creativity (Zergollern-Miletić & Horváth 2009a, 2009b). Writing and any other form of communication is essentially a creative act: as several practitioners of writing pedagogy point out, one aim of this endeavor is the supportive development of personal voice. It is in this context that the aim to develop a new type of learner corpus was identified.

Research questions and procedures

The survey aimed to collect valid data about Hungarian and European university students' attitudes toward and development of their English writing skills. The items in the questionnaire invited participants, among others, to evaluate their current study programs, share their views on what they regarded as their best writing assets, and discuss plagiarism (Horváth 2012). The questionnaire was distributed electronically, using Google Form, a user-friendly tool that I had tested previously and found reliable.

In the present paper, after providing demographic details of the respondents, I will present and discuss the results of one of of the fourteen questionnaire items: Item 11, as follows:

> Please paste a paragraph of your writing in English that you are proud of. If there is an assignment that you have in electronic form, please copy and paste a paragraph from it here.

The rationale for this item was two-fold: more information is needed to evaluate students' own interpretation of their work if the educational goal is the maintenance and enhancement of learners' autonomy, and, from a corpus linguistic perspective, it seemed relevant that the design principle be based on students' own choices. In addition, knowing what students consider

relevant to themselves, that is, collecting information on intrinsically motivating work that they are proud of can be channeled back into the writing pedagogical process, enabling us to review our own practice. A further application of such an endeavor is the implementation of these texts as reading materials as well as the basis of revising and editing activities, a complementary approach to the management of feedback and peer feedback tasks (Zamel 1985).

As the majority of respondents were expected to be BA and MA students, I decided to investigate the research question of whether there were any differences across these two subgroups as far as lexical features such as range of vocabulary were concerned. I hypothesized, partly based on recent research into Hungarian higher-education issues (Doró 2011; Magnuczné-Godó 2003), that there would be a marked increase along these criteria from the BA to the MA level.

The corpus was created after ensuring the application of careful clean-text policy: the only changes I made to the texts shared by respondents were to correct obvious cases of typing errors. After that, Cobb's *Compleat Lexial Tutor* tools (2013) were used for the lexical analyses: vocabulary profiling.

Participants

The survey was available online for three weeks in March 2012. Students had total anonymity and freedom in answering questions: none of the items were mandatory. A total of 345 students of English as a foreign language responded to it, mostly from Europe, of whom 343 respondents' answers were finally included in the survey. Two responses had to be eliminated for technical reasons.

Results and discussion

Of the 337 students answering the question about where they were residing at the time of the survey, 145 reported Hungary, with the rest, the majority (192), reporting residence in a total of 32 other countries, mostly in Europe. The country list was headed by Austria and Poland (36 for each), followed by Croatia (29) Slovakia (22) and Germany (10). Countries represented by one single respondent were mostly non-European, such as Australia, Canada, India, and Mexico.

The gender ratio was 264 female and 73 male respondents (a total of 337 respondents answered this question).

See Table 1 for the age distribution of the 328 students who answered the question about age.

Age	Number of students
20 or younger	74
21-23	124
24-26	74
27-29	28
30-32	10
33-35	9
36-38	3
39 or older	6

Table 1: The age distribution of respondents (N=328)

As stated earlier, the focus in the current paper is on item 11 of the questionnaire, which invited each student to share a text that she or he was proud of. Even though the survey asked participants to submit paragraph-length texts, many chose to make available full essays and research papers. I decided to include all such submissions in the corpus. In this way, of the 334 respondents, a total of 178 students shared items via the questionnaire. This rate is just above 50%, which can be regarded as satisfactory, given the technical constraints of granting the request: many respondents said they would have liked to submit such works but were unable to, mostly because they were not on the computer where these scripts were stored.

The 178 scripts thus collected hold potential for an analysis of various thematic, textual and lexical purposes. Together, the scripts make up what I call the International Pride Corpus (IPC). As one aim of the investigation was to lay the groundwork for a more extended study, the current endeavor can be considered a pilot phase that is concerned with thematic and lexical features. It is hoped that the conclusions drawn from the study will inform future research in the area.

In terms of text types, besides creative and academic prose, some students shared their poetry. Table 2 presents the text type distribution of the IPC:

Letter of complaint	1
Poems	4
Short stories	5
Essays	37
Research papers	131

Table 2: Text types in the IPC (N=178)

Both the essays and the research papers in the IPC, that is, the academic content, show a wide thematic range. Essay samples cover topics as diverse as happiness, family, the economy, careers, arts, and sports. Full and partial research papers, the largest segment of the IPC, deal with linguistic and applied linguistic phenomena, sociology, media studies, as well as literature, marketing and history.

It is refreshing to see such a wide variety of themes and text types in the IPC. Obviously, one reason for this is that the student respondents represent a diverse pool of higher educational institutions and curricula, as well as because, as turned out from their answers to the other items in the questionnaire (Horváth 2012), many had a deep interest in writing in English.

One surprising content result is the inclusion of creative writing among the scripts: besides the four poems and the five short stories, many personal essays also featured highly creative text organization choices.

In terms of the lexical features of the IPC, to be able to investigate differential features between scripts by BA and MA students, I established two subcorpora. The BA subcorpus consists of 84 scripts, whereas the MA subcorpus, 67 items (N=151). The rest of the submissions (27) were by PhD students or by respondents whose study programs belonged to pre-Bologna teacher training curricula or other systems. For the purposes of the current analysis, results from the BA and the MA subcorpora will be presented.

As Table 3 shows, the BA subcorpus consists of 12,200 words, that is, 145 words per script. The lexical profile of the subcorpus, based on Cobb's Vocabulary profiling tool (2013), shows a fairly advanced picture, with a type-token ratio of 0.24.

Tokens	12,200
Types	2,925
Type-token ratio	0.24
Tokens per type	4.17

Table 3: Vocabulary profile of the BA subcorpus of the IPC

Compare that with the data presented in Table 4, of the MA subcorpus. As can be seen, there are only minor differences. The average length of an MA script in the IPC is 152 words, just slightly higher than that of the BA subcorpus.

Tokens	10,169
Types	2,758
Type-token ratio	0.27
Tokens per type	3.69

Table 4: Vocabulary profile of the MA subcorpus of the IPC

However, when viewed with the tool allowing for more granular analysis, the lexical differences between the two subcorpora appear somewhat more clearly. Using the lexical profiling algorithm that compares text input with the K1 to K20 bands found in the lexis of the British National Corpus (Cobb, 2013), I have seen the following picture emerging. Table 5 presents this view of the lexical profile of the BA, whereas Table 6, that of the MA subcorpus.

Freq. Level	Families	Types	Tokens	Coverage (tokens)%
K1 Words :	754	1393	9824	80.52
K2 Words :	416	572	972	7.97
K3 Words :	208	254	341	2.80
K4 Words :	130	161	215	1.76
K5 Words :	81	88	109	0.89
K6 Words :	60	67	79	0.65
K7 Words :	43	45	65	0.53
K8 Words :	41	44	60	0.49
K9 Words :	31	32	34	0.28
K10 Words :	19	20	32	0.26
K11 Words :	16	16	17	0.14
K12 Words :	16	16	23	0.19
K13 Words :	11	11	12	0.10
K14 Words :	5	5	8	0.07
K15 Words :	3	3	3	0.02
K16 Words :	5	5	8	0.07
K17 Words :	4	4	4	0.03
K18 Words :	3	3	4	0.03
K19 Words :	2	2	2	0.02
K20 Words :				0.00
Off-List:	?	184	388	3.18
Total	1848+?	2925	12200	100%

Table 5: The BNC K1-K20 bands compared with the vocabulary of the BA subcorpus of the IPC

Freq. Level	Families	Types	Tokens	Coverage (tokens)%
K1 Words :	717	1306	7956	78.24
K2 Words :	373	515	844	8.30
K3 Words :	168	213	301	2.96
K4 Words :	135	168	237	2.33
K5 Words :	81	92	118	1.16
K6 Words :	48	51	65	0.64
K7 Words :	49	55	80	0.79
K8 Words :	41	47	64	0.63
K9 Words :	22	23	26	0.26
K10 Words :	17	17	20	0.20
K11 Words :	16	16	21	0.21
K12 Words :	13	14	18	0.18
K13 Words :	15	16	19	0.19
K14 Words :	7	7	10	0.10
K15 Words :	1	1	1	0.01
K16 Words :	3	3	5	0.05
K17 Words :	2	2	2	0.02
K18 Words :	1	1	1	0.01
K19 Words :	4	4	5	0.05
K20 Words :	1	1	1	0.01
Off-List:	?	206	375	3.69
Total	1714+?	2758	10169	100%

Table 6: The BNC K1-K20 bands compared with the vocabulary of the MA subcorpus of the IPC

As can be seen, there is a slight difference in terms of K1 and K2 lexis representation between the two subcorpora, with the MA scripts showing a slightly wider range. However, the differences are not major there, and the distribution of vocabulary across the other bands also appears fairly similar between the two subcorpora. The reason probably lies in the fact that both the BA and the MA students were among the highly motivated and proficient users of EFL, and thus the different educational levels did not result in more than a slight lexical profile difference between these two groups. Clearly, further studies involving concordances of lexical units and the analysis of the appropriateness of the lexis used are necessary.

Conclusion

This paper presented results of the analysis of a new type of EFL learner corpus: one that is based on students' own selection decisions. The International Pride Corpus has been established with the aim of relying on students' intrinsic motivation. This first such undertaking can be considered relatively successful: of the 343 respondents, the majority made their texts available for study purposes. The research question guiding the current analysis was directed at the viability of such an investigation, and we have some evidence that the approach may contribute to the validation of writing programs in higher education. In terms of the hypothesis concerning the differential lexical features of BA and MA scripts, it appears that the paper has yielded no conclusive evidence for improved lexical density from BA to MA levels. One reason for this can be seen in the limited scope of the corpus: with 178 students' scripts in it, it can hardly be regarded as a representative international corpus. However, as this has been a pilot project in this relatively new field, I hope that further corpus

building efforts may be able to reveal even more valid data, which can inform both writing pedagogy practice and learner corpus linguistic studies.

References

Andor, J. (2010): What corpora can do, what you can do with corpora: An interview with Douglas Biber. In: Hegedűs, I. & Martsa, S. (eds.): *CrosSections: Selected papers in linguistics from the 9th HUSSE conference*. Pécs: Institute of English Studies, University of Pécs, 307-317.

Cobb, T. (2013): *Compleat lexical tutor*. lextutor.ca

Doró, K. (2011): Students' perceptions about their preparedness for undergraduate studies in English. In: Szabó, G., Horváth J. & Nikolov M. (eds.): *UPRT 2009: Empirical studies in English applied linguistics*. Pécs: Lingua Franca Csoport, 81-92.

Géczi, J. (2005): *Pedagógiai tudásátadás*. Pécs: Iskolakultúra.

Grabe, W. & Kaplan, R.B. (1996): *Theory and practice of writing*. London: Longman.

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* 37, 538-546.

Hirvela, A. & Belcher, D. (2001): Coming back to voice: The multiple voices and identities of mature multilingual writers. *Journal of Second Language Writing* 10, 83–106.

Hollósy, B. (1994): Corpora, concordancing and database management for the lexicographer. *Studies in Linguistics* III, 111-123.

Horváth J. (2001): Advanced writing in English as a foreign language: A corpus-based study of processes and products. Pécs: Lingua Franca Csoport.

Horváth, J. (2012): "Szeretem, ha egyénivé tehetem, amit írok: Egyetemisták körében végzett nemzetközi felmérés tapasztalatai." In: *Íráspedagógiai tanulmányok*. Budapest, Veszprém: Gondolat, 68-87.

Hyland, K. (1999): Disciplinary discourses: Writer stance in research articles. In: Candlin, C.N. & Hyland, K. (eds.): *Writing: Texts, processes and practices*. London: Longman, 99-121.

Hyland, K. (2002): *Teaching and researching writing*. Harlow: Pearson Education Limited.

Hyland, K. (2011): Corpora and EAP: Specificity is disciplinary discourses. In: Goźdź-Roszkowski, S. (ed.): *Explorations across languages and corpora*. Frankfurt am Main: Peter Lang, 317-334.

Ivanič, R. & Camps, D. (2001): I am how I sound: Voice as self-representation in L2 writing. *Journal of Second Language Writing* 10, 3–33.

Magnuczné Godó, Á. (2003): *Cross-cultural perspectives in academic writing*. PhD dissertation. Budapest: ELTE.

Matsuda, P. K. (2001): Voice in Japanese written discourse: Implications for second language writing. *Journal of Second Language Writing* 10, 35– 53.

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

Nesselhauf, N. (2002): Learner corpora and their potential for language teaching. In: Sinclair J. & Tognini-Bonelli E. (eds.): *Teaching and language corpora*. Amsterdam: John Benjamins, 125-152.

Pecorari, D. (2003): Good and original: Plagiarism and patchwriting in academic second-language writing. *Journal of Second Language Writing* 12, 317–345.

Xiao, Z. (2008): Well-known and influential corpora. In: Lüdeling A. & Kyto M. (eds.): *Corpus linguistics: An international handbook*. Vol. 1. Berlin: Mouton de Gruyter, 383-457.

Zergollern-Miletić, L. & Horváth, J. (2009a): Coherence and originality in university students' writing in EFL: The Zagreb – Pécs project. In: Lugossy, R., Horváth, J. & Nikolov M. (eds.): *UPRT 2008: Empirical studies in English applied linguistics*. Pécs: Lingua Franca Csoport, 145-151.

Zergollern-Miletić, L. & Horváth, J. (2009b): Eredetiség zágrábi és pécsi egyetemisták angol nyelvű írásaiban. *Iskolakultúra* 19 (9-10), 50-57.

Zamel, V. (1983): The composing processes of advanced ESL students. *TESOL Quarterly* 17, 165-187.

Zamel, V. (1985): Responding to student writing. *TESOL Quarterly* 19, 79-101.

Zinsser, W. (1998): *On writing well*. New York: Harper.

József Horváth
University of Pécs
Institute of English Studies
H-7624 Pécs
Ifjúság útja 6.
jozsefhor@gmail.com