Gyula Sankó: Scholarly texts flavoured with conversational features
Argumentum 10 (2014), 560-575
Debreceni Egyetemi Kiadó

Gyula Sankó

Scholarly texts flavoured with conversational features:

A corpus-based study of Hungarian EFL learners' academic writing

Abstract

This paper deals with the use of corpus linguistics to promote better academic writing. To explore Hungarian undergraduate EFL learners' awareness of academic vocabulary register conventions a learner corpus was compiled, and particular vocabulary items were then compared with the academic and spoken subcorpora of the BNC to establish similarities or differences with native academic writing or spoken language by counting frequency of use. The results largely confirm and coincide with previous (multicultural) research results, which demonstrate a considerable amount of stylistically inappropriate, colloquial lexical features in undergraduates' academic writing.
Keywords: learners' corpus, academic writing, conversational features

1 Introduction

Academic writing is relatively formal, using a highly conventionalized phraseology. This, in turn, means that in an essay colloquial words and expressions should be avoided. The rather fixed nature of academic writing is particularly challenging for non-native writers, who, mainly guided by the transfer from their mother tongue and their school language learning experience, are more familiar with casual style and use.

Corpus analysis of academic texts written by language learners compared with expert or native corpora can reveal potential problem areas as well as particular problems of this nature. Based on these findings suggestions can be made about incorporating some focal areas in academic writing classes. Some previous corpus-driven studies were dealing with learner academic writers' register awareness manifest in the underuse of the features required in academic writing and the overuse of certain colloquial lexical or grammatical characteristics (Gilquin et al., 2007, Gilquin & Paquot 2008, Lee & Chen 2009, Luzon 2009, Adel & Erman 2012, Lei 2012).

This research has taken major inspiration from Gilquin and Paquot's (2008) paper on learner academic writing and register variation. In their study Gilquin and Paquot use a 3.5 million-word learner corpus named ICLE, i.e. the International Corpus of Learner English based on academic writing data of a multilingual undergraduate community including Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, and Turkish students. The above study, however, does not break down its results to particular nationalities. Finding the most appropriate vocabulary with a suitable academic register, and a tendency to use colloquial items in academic essays, is an evergreen problem for Hungarian students of English, too. A pioneering corpus-based study of this nature, investigating the written production of Hungarian students of English, was conducted by Horváth (2001).

The primary objectives of this paper are to identify the problems of register which Hungarian undergraduate students of English experience in the course of their academic writing and to suggest some ideas as to how to improve students' awareness of the academic register in writing classes. The study limits itself to examining the academic writing of 445 different upper-intermediate to advanced level Hungarian undergraduate students studying English as a major or minor subject at the University of Debrecen.

The paper is organised as follows. Section 2 describes the corpora used and the methods applied to conduct the research. Section 3 provides a classification of potential vocabulary register problems and displays the findings. The article ends by discussing some pedagogical implications of the findings of the study, suggesting potential ideas to raise undergraduate students' academic register awareness.

2 Corpus and Method

2.1 The corpus

For the purposes of this study a learner academic corpus was compiled. The corpus, which includes 5,764,902 tokens and 78,114 types, contains the work of 445 undergraduate upper-intermediate to advanced level English majors and minors of a three-year period, with each text showing the output of a different student. The corpus is called University of Debrecen English Learner Corpus, henceforward UDELC. The UDELC is a raw corpus which is currently made up of four subcorpora: British literature and civilization, American literature and civilization, theoretical linguistics, and applied linguistics including ELT methodology. The British literature and civilization subcorpus has 2,175,519 tokens, the American literature and civilization 1,601,530 tokens, theoretical linguistics includes 930,743 tokens, and the size of the applied linguistics and ELT methodology part is 1,034,193 tokens. With all four subcorpora belonging to the humanities, and in some way related to learning English, UDELC represents a relatively homogeneous discipline rather than a chimera (cf. Hyland & Tse 2007). To get some more insight into the nature of the vocabulary found in the UDELC, it was compared with the General Service List (West 1953), the University Word List (Xue & Nation 1984), the Academic Word List (Coxhead 2000), and Paquot's (2010) Academic Keyword List using the text comparison tool Lextutor v.6.2 online software (Cobb 2009). As Table 1 illustrates, the UDELC, which includes 56,456 word families (wfs) is quite comprehensive by its nature. Although this comparison gives just a superficial picture about the words in UDELC providing limited details about the distribution of vocabulary, it can clearly be seen that the corpus includes the vast majority of vocabulary used in academic writing.

Word List	Word Families[1] shared with UDELC
GSL (977 wfs. in 2,000 words)	975
UWL (807 wfs.)	802
AWL (570 wfs.)	570
AKL (685 wfs.)	685

Table 1: The number of word families in GSL, UWL, AWL, and AK shared with UDELC

[1] Using the family as a unit of comparison means that if *cat* is in Text 1 and *cats* in Text 2 then this is considered a repetition of the cat family, i.e. these are equivalent tokens. Note also that in this routine, "unfamilied" items revert to classification by type (e.g., repeated proper nouns) (Cobb 2009).

2.2 Method

In addition to using some of the words studied by Gilquin and Paquot (2008), university and college writing labs and academic services (cf. e.g., Starkey 2004, Sherlock 2008) were consulted for selecting and gathering the target vocabulary for this study. These sites list the vocabulary which should be avoided in academic writing because the items are too colloquial, overcomplicated, and consequently they do not correspond to the conventions of academic writing style.

The research followed the methodology applied by Gilquin and Paquot (2008) using the BNC as the corpus of comparison, and also the basic functional classification of the target vocabulary, although some new groups as well as several new vocabulary items were added to those suggested in the 2008 research. The target vocabulary in Hungarian EFL undergraduate EFL students' academic writing was classified into the following 17 features and functional categories: expressing cause and effect, adding information, expressing possibility, expressing certainty, listing items, using contracted verb forms, expressing repetition, run-on expressions, using the first person personal pronoun, expressing personal opinion, introducing topics and ideas, sentence starters, expressing concession, vague quantifiers, awkward expressions, speech crutches (fillers), and general, vague words.

The corpus was analysed using AntConc 3.3.5 text analysis concordance program (Anthony 2012). The most frequent output vocabulary items and collocations in the UDELC were also analysed using the academic section (15,331,668 words) and the spoken section (9,963,663 words) of the BNC (The British National Corpus 2007) for purposes of comparison and contrast.

The study was carried out with an unlemmatized UDELC corpus without changing the morphosyntactic features of the learner input in the academic papers, as the aim of the study was to gain insight into the learners' original written production. Using original, unmodified learner data facilitated the success of a contrastive analysis with certain characteristic features of native academic writing and native spoken language use as represented in the BNC.

3 Results

The results described below in detail indicate that Hungarian undergraduate EFL learners tend to overuse certain words and structures contraindicated in academic writing as too colloquial. Below you can read the detailed findings broken down into 17 functional/ conceptual categories.

3.1 Words and phrases expressing cause and effect

Besides sentence initial *So*, (139 occurrences) the most overused register violation was the use of *thanks to* (117 occurrences), probably due to the fact that secondary schools put more emphasis on teaching general English expressions which are used in everyday communication. Figure 1 illustrates how this phrase is represented in UDELC compared with BNC data.

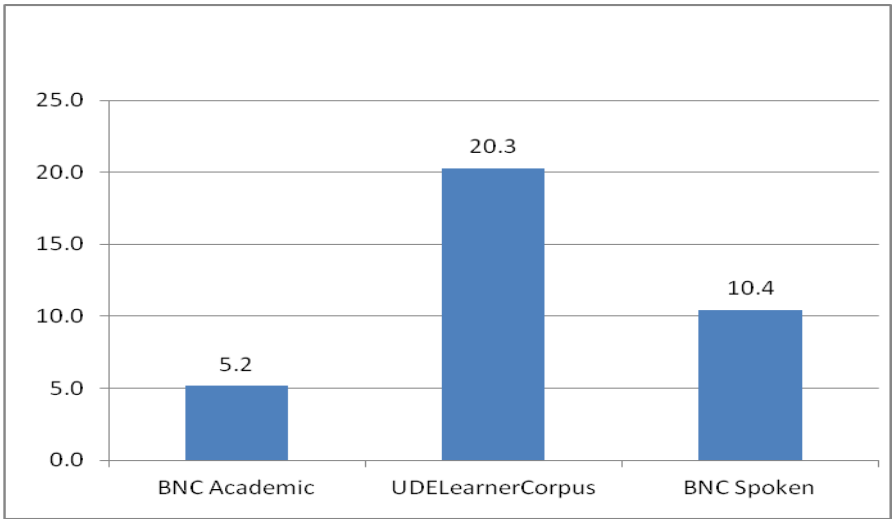

Figure 1: Relative frequency of thanks to *in native academic writing, learner academic writing and native speech (frequency per million words)*

3.2 Adding information

The two most outstanding features misused in this category were sentence-initial *And*, (1898 occurrences), together with the sentence-initial use of *Besides*, (239 occurrences). As shown in Figure 2, sentence initial *besides* is largely overrepresented when compared with data in the BNC academic subcorpus.

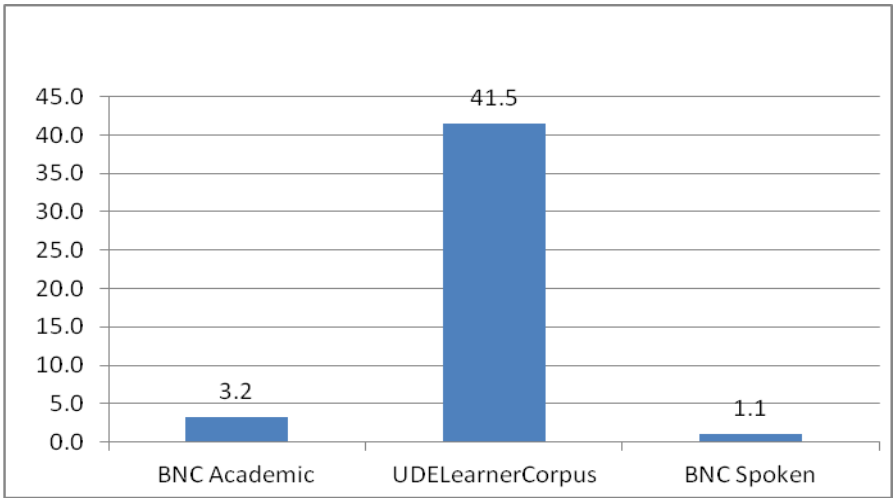

Figure 2: Relative frequency of sentence-initial Besides, *in native academic writing, learner academic writing and native speech. (frequency per million words)*

3.3 Expressing possibility

When expressing possibility, the items *perhaps* (804 occurrences), *it seems that* (267 occurrences), *it is possible that* (48 occurrences) were heavily used in UDELC. These phrases, however, proved to be relatively frequently used in the BNC academic section too. As can be seen from Figure 3, the item which stands out as inappropriately overused by non-native writers is *maybe* (409 occurrences).

On the other hand, such words as *apparently* (209 occurrences), *presumably* (83 occurrences), *likely* (613 occurrences), and *assumption* (289 occurrences), characteristic of professional academic texts, were also frequently used, which is a heartwarming phenomenon.

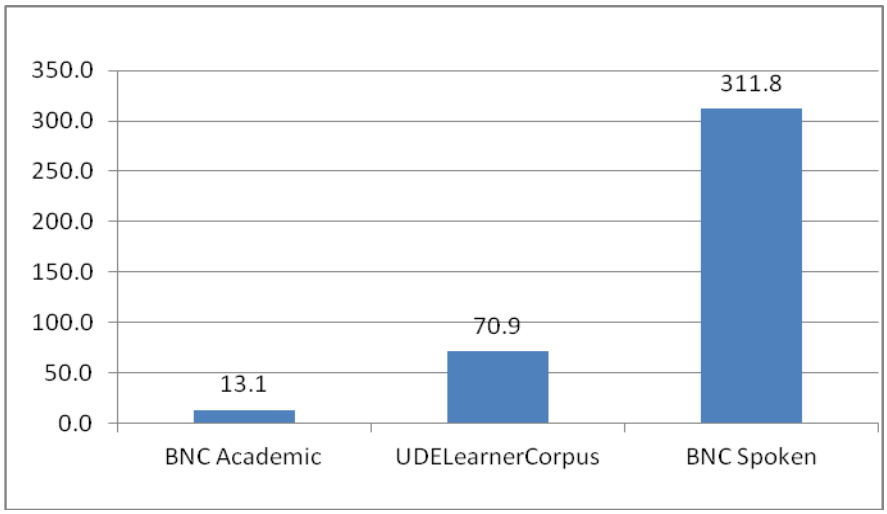

Figure 3: Relative frequency of maybe to *in native academic writing, learner academic writing and native speech (frequency per million words)*

3.4 Amplifying adverbs to express certainty

Among the adverbs amplifying the author's message, it was *really* (1,730 occurrences) and *totally* (546 occurrences) that violated academic norms. Other frequently used adverbs such as *of course*, *certainly*, *absolutely*, *definitely*, *quite* – although suspicious – proved to be used in academic language by native speakers, too.

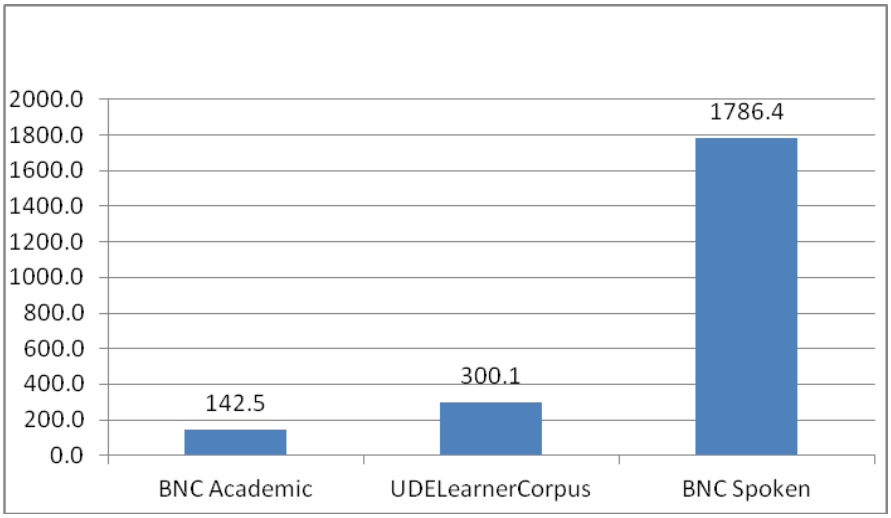

Figure 4: Relative frequency of really *in native academic writing, learner academic writing and native speech (frequency per million words)*

3.5 Listing items

Listing arguments or other items with adverbs such as *first, firstly, second, secondly, third, finally, in conclusion,* or *lastly* in a sentence-initial position was judged to be a critical issue for learners involved in the academic writing genre. The data in the BNC academic subcorpus, however, showed a relatively frequent occurrence of these words. As shown in Figure 5, the expression overused by learner academic writers was *first of all,* which was underrepresented in the academic corpus, occurring more frequently in spoken communication or narrative compositions.

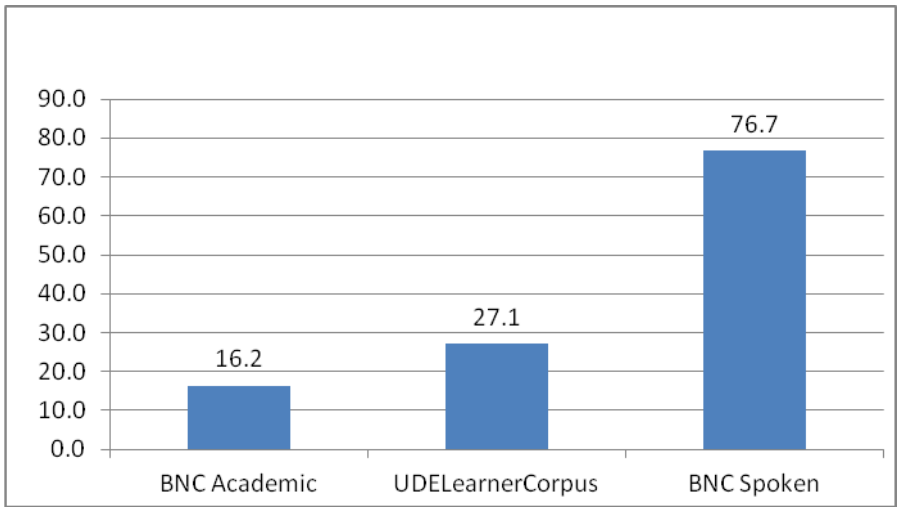

Figure 5: Relative frequency of first of all *in native academic writing, learner academic writing and native speech (frequency per million words)*

3.6 The use of contracted verb forms

Despite the fact that undergraduates are repeatedly warned against using contracted verb forms in academic writing tasks, they seem to forget about it and instinctively fall back on spoken language forms they are more used to in face-to-face communication. Table 2 shows the most frequently used contracted verb forms in the learner corpus.

WORD/ EXPRESSION	FREQUENCY
don't	109
it's	97
can't	36
doesn't	24
didn't	23
let's	13
wouldn't	12
won't	8

Table 2: The frequency of contracted verb forms in UDELC

3.7 Repetition: phrases referring to the paper or the reader

For some mysterious reason, learner academic writers like repeating their earlier utterances, and they also tend to explicitly inform their readers that a previous thought is going to be repeated (see Table 3). Moreover, they tend to do this using conversational word collocations or even ungrammatical phrases. For instance, the form **as it was mentioned* is used more than ten times more frequently than the correct *as was mentioned*, which was found in the corpus only 6 times.

WORD/ EXPRESSION	FREQUENCY
as I mentioned	102
as I have mentioned	97
as it was mentioned	72
as was mentioned	6

Table 3: The frequency of phrases indicating repetition in UDELC

As figure 6 illustrates, the phrase *as I mentioned* is much more typical of the spoken language than of academic texts, and it is heavily overused in UDELC even when compared with the spoken subcorpus of the BNC.

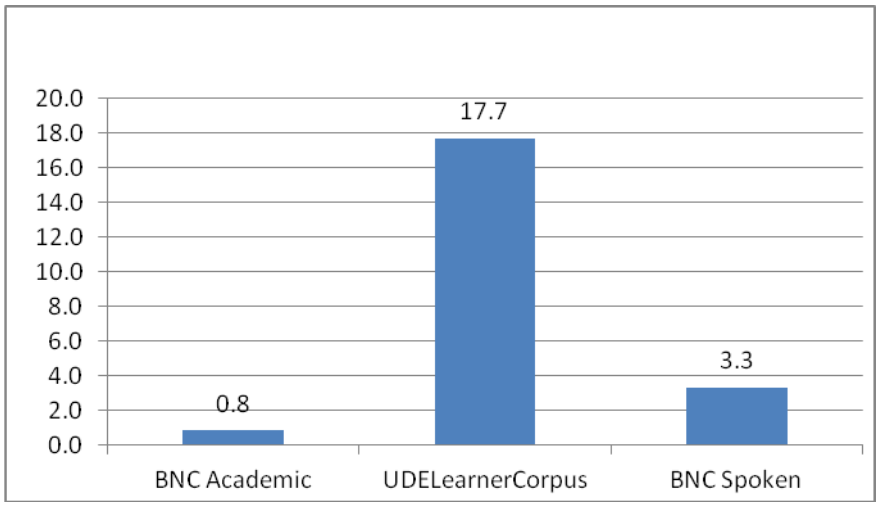

Figure 6. Relative frequency of as I mentioned *in native academic writing, learner academic writing and native speech (frequency per million words)*

3.8 Run-on expressions

Experts in academic writing do not like run-on expressions when listing examples because they give the utterance an air of vagueness, which should be avoided in scholarly papers. Instead, academic English prefers the use of *such as*, or *e.g.* Nevertheless, the BNC academic subcorpus demonstrated a relatively frequent use of expressions like *and so on*, or *and so forth* even *and so on, and so forth.* From Figure 7 it can be seen that the item which was incongruent with BNC academic use was the Latin run-on abbreviation *etc.*, which was considerably overused by Hungarian student writers.

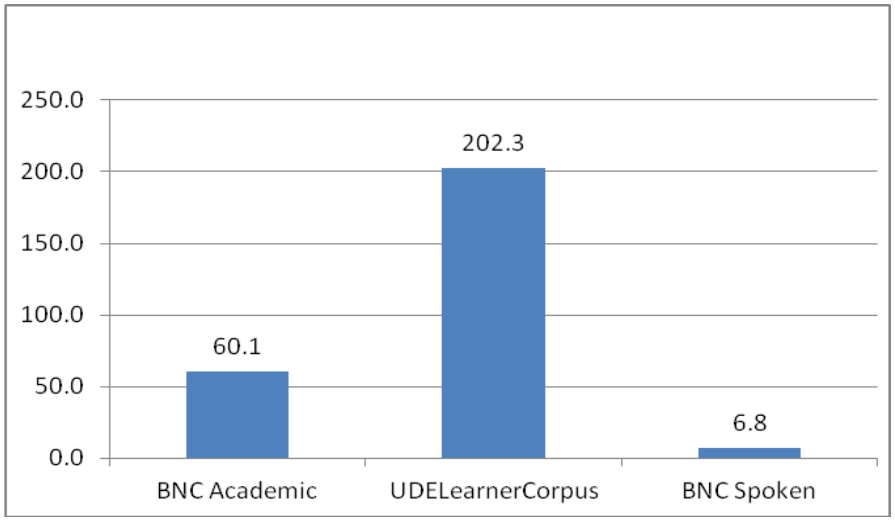

Figure 7: Relative frequency of etc. *in native academic writing, learner academic writing and native speech (frequency per million words)*

3.9 The use of personal pronouns referring to the author

Written language is objective rather than personal, it should have fewer words that refer to the writer. The main emphasis should be on the information intended to be conveyed and on the arguments to be made, rather than on the writer. Most university writing centres advocate that avoiding the use of emotive, personalized language presents objectivity to your work through the observations of a dispassionate, unbiased researcher.

Student writers, however, tend to refer to themselves as authors more overtly than expert academic writers would do. The data obtained from the analysis of the UDELC reveal that Hungarian EFL students also frequently use the first person singular "*I*" (and "*me*") and first person plural "*we*" (and "*us*") pronouns, as well as related possessive adjectives "*my*" and "*our*" in academic writing. As Luzon (2009) points out, the use of authorial *I* and *we* deserves more detailed attention from researchers because, contrary to popular belief, these pronouns are used by academic writers in a justified way for a range of purposes. What needs to be discovered is whether learner use is different from conventional use in expert native speakers' academic writing. In spite of being aware of the delicate nature of using the authorial first person, this study does not undertake the detailed examination whether these authorial pronouns were used in the correct way. The current study pursues a quantitative approach comparing native and learner academic corpora, and it limits itself to drawing en masse conclusions (cf. Table 4).

WORD/ EXPRESSION	FREQUENCY
I	21,750
we	14,802
my	6,292
our	4,449
us	3,709
me	2,712

Table 4: The frequency of using personal pronouns referring to the author in UDELC

Accordingly, Figure 8 shows that the relative frequency of using *I* was nearly three times as high in the UDELC learner corpus as in the native academic corpus.

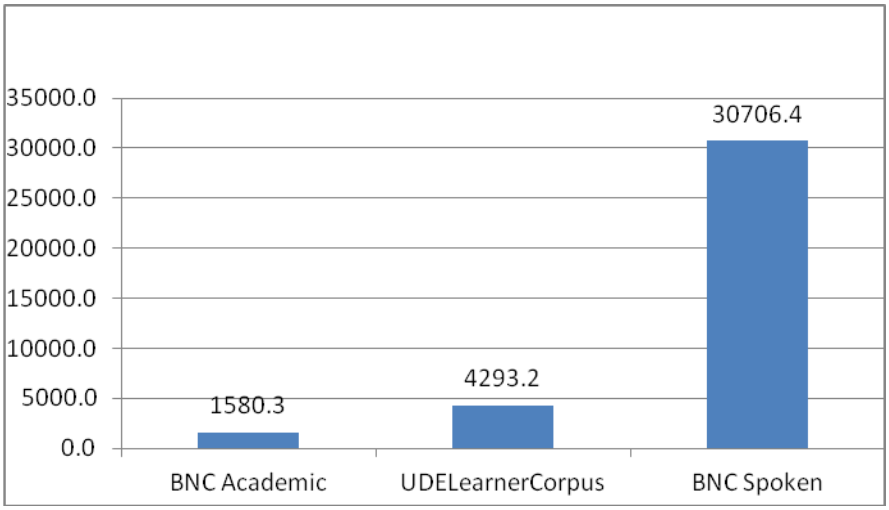

Figure 8: Relative frequency of I *in native academic writing, learner academic writing and native speech (frequency per million words)*

3.10 Expressing personal opinion (towards the writer's message)

Expressing the author's personal opinion in an explicit way is one of the fields which, unlike academic written language, is swarming with spoken-like expressions in undergraduate EFL learners' written works. As can be seen from Table 5, some of the most frequently used colloquialisms of this type were *I think, I believe, I like, I feel, I am sure, I must/ have to admit, from my point of view.* All of the phrases listed above are underrepresented in the BNC academic subcorpus and show a high frequency of occurrence in the spoken language subcorpus.

WORD/EXPRESSION	FREQUENCY
I think	973
in my opinion	351
I believe	282
I like	95
I feel (that)	91
in my view	80
I do not think	44
it seems to me	33
I am sure	26
I liked	24
to my mind	19
as for me	13
I must admit	12
I have to admit	10
from my point of view	6

Table 5: The frequency of using phrases expressing the author's personal opinion

As shown in Figure 9, the phrase *I think* is used three times more often in the UDELC than in the academic subcorpus of the BNC.

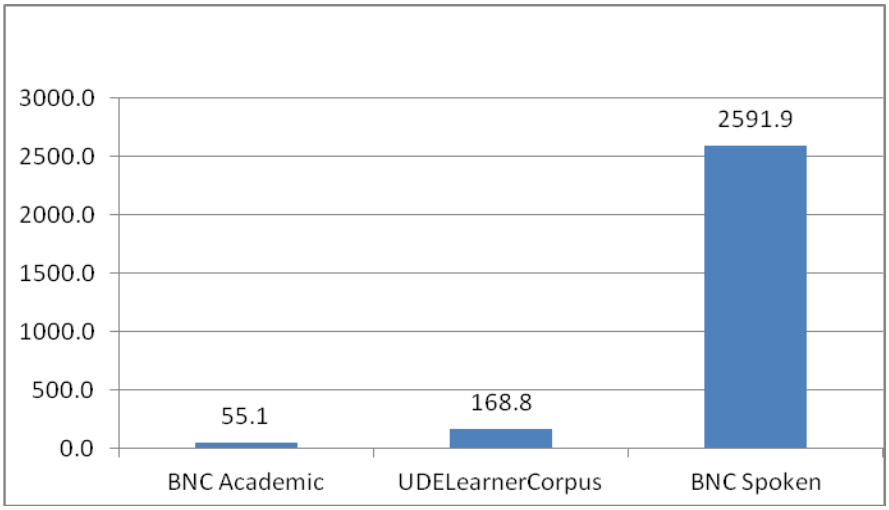

Figure 9: Relative frequency of I think. *in native academic writing, learner academic writing and native speech (frequency per million words)*

At the same time, it should also be noted that some more sophisticated academic expressions of this nature, such as *it seems* (636 occurrences), *it is worth noting that* (9 occurrences), *interestingly (*154 occurrences), *surprisingly* (131 occurrences*),* were also present in the learner corpus.

3.11 Introducing topics and ideas

Introducing topics and ideas is another critical area where learner writers are prone to borrow and use expressions from oral communication, rather than work with phrases whose register would be well suited for the academic writing style. Table 6 summarizes the most frequent examples of this kind found in the UDELC.

WORD/ EXPRESSION	FREQUENCY
I would like to	464
I am going to	348
Let me	125
By the way	48
I have to mention	30

Table 6: The frequency of using colloquial phrases when introducing topics or ideas

Figure 10 shows that the phrase *I would like to*, which is very rarely used by native academic writers, is used 21 times more frequently by Hungarian learner academic writers in the UDELC.

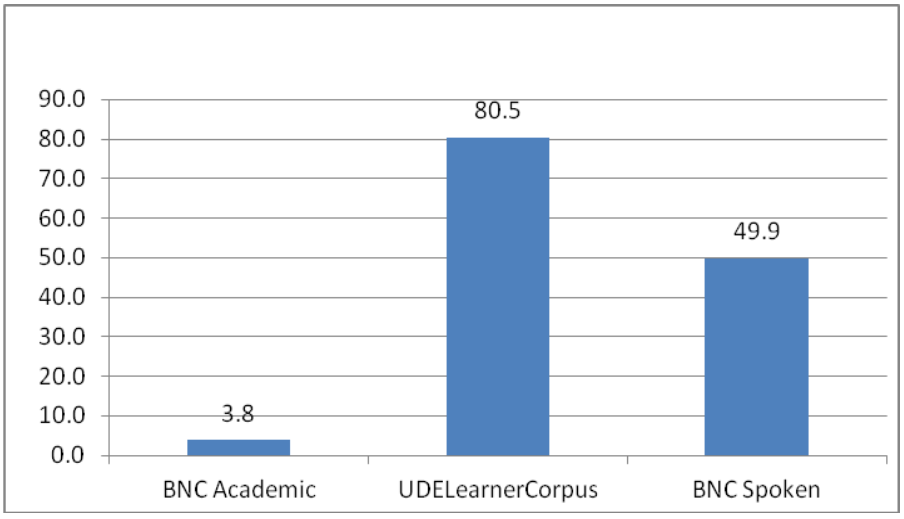

Figure 10: Relative frequency of I would like to. *in native academic writing, learner academic writing and native speech (frequency per million words)*

3.12 Expressing concession

Albeit not in large numbers, similar to Gilquin and Paquot's (2008) findings, the colloquial feature of placing the conjunction *though* at the end of a sentence to express concession could be spotted in Hungarian undergraduates written work, too. Figure 11 depicts the relative frequency of occurrence of this word in UDELC compared with two BNC subcorpora.

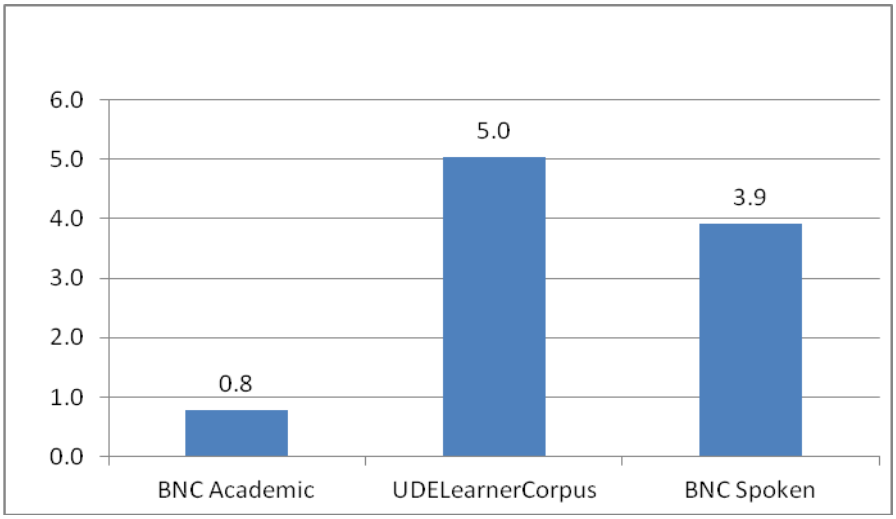

Figure 11: Relative frequency of the sentence-final linking word though. *in native academic writing, learner academic writing and native speech (frequency per million words)*

3.13 Vague quantifiers

In academic writing, facts and figures are given precisely, and expert writers advise against using vague adjectival quantifiers. Despite this fact, similarly to the UDELC learners' academic writing output, the academic section of the BNC displayed a relatively frequent use of expressions like *many* (5,933 occurrences), *several* (2,472 occurrences), *various* (1,459 occurrences), *a number of* (517 occurrences), *numerous* (290 occurrences), *a variety of* (269

occurrences), As Figure 12 shows, the item which was overused by Hungarian student writers, but underrepresented in BNC academic use, was *lots of* (185 occurrences). It can also be seen from the data representing spoken native use that this item is characteristic of oral, rather than written communication.

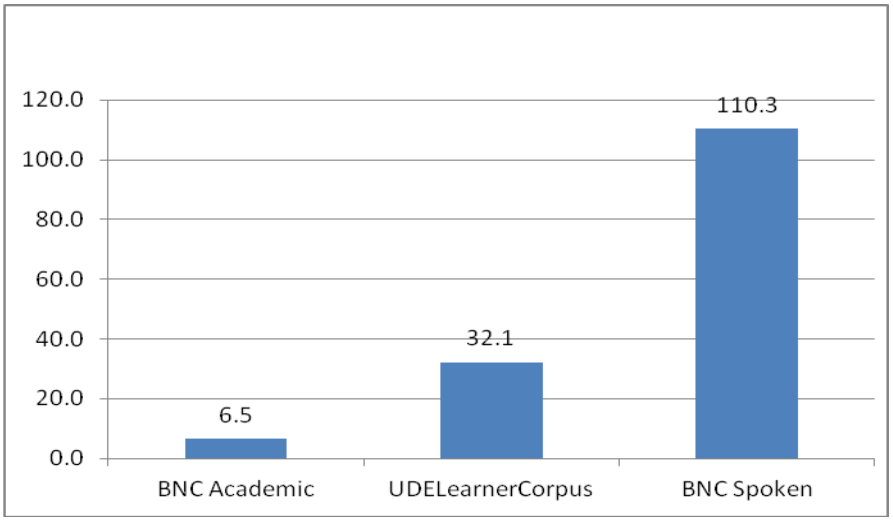

Figure 12: Relative frequency of the quantifier lots of *in native academic writing, learner academic writing and native speech (frequency per million words)*

3.14 Speech crutches or fillers

Fillers, these apparently meaningless words or phrases that mark a pause or hesitation in speech, clearly do not belong to the realm of academic writing. Nonetheless (as can be seen in Table 7), probably governed by speaking habits, student writers used some of these speech crutches in UDELC.

WORD/ EXPRESSION	FREQUENCY
basically	345
Well,	240
I mean	143
Better to say	15

Table 7: The frequency of using some speech fillers in the UDELC

Figure 13 shows that although *basically* is characteristic of spoken language, it is used nearly three times more often in the UDELC than by native academic writers.

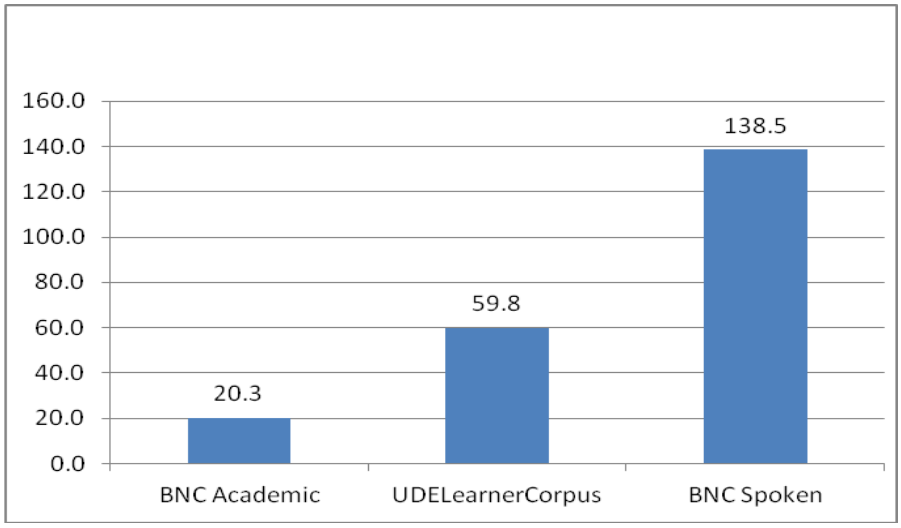

Figure 13: Relative frequency of basically *in native academic writing, learner academic writing and native speech. (frequency per million words)*

3.15 Awkward phrases

The phenomenon described in this part of the study is not related to using conversational vocabulary items in academic writing. Also it could be as characteristic of beginner native academic writers as of Hungarian English major undergraduates. As Oppenheimer (2005) notes, even though most experts on writing encourage authors to avoid overly-complex constructions or words, most undergraduates tend to deliberately increase the complexity of their vocabulary to appear smarter, give the impression of intelligence, and make the content of the paper look more valid. As shown in Table 8, Hungarian undergraduate writers also often opted for the more complicated solutions. They used *considered to be* instead of simply writing *considered; so as to* instead of simply writing *to; each and every* instead of writing either only *each* or *every* by itself; *due to the fact that* instead of writing *due to* or simply *because*; and wrote *as to whether* instead of writing *whether*. Surprisingly, despite advice from writing specialists to use uncomplicated structures, the data in the BNC academic subcorpus show that the above structures are frequently used by native academic writers as well.

WORD/ EXPRESSION	FREQUENCY
Considered to be	398
So as to	129
Each and every	110
Due to the fact that	99
As to whether	16

Table 8: The frequency of using some speech fillers in UDELC

3.16 General colloquial and vague words

As academic writing needs to be more precise than face-to-face, general oral communication, it is advisable to avoid common, but vague words and phrases, such as *get*, *nice*, or *thing*. As

Table 9 shows, undergraduates tend to overuse such conversational and vague words as *thing* or *things*, as well as the neutral, but nearly meaningless *interesting* or *nice.*

WORD/ EXPRESSION	FREQUENCY
things	2,067
thing	1,274
interesting	1,260
nice	332
has got	195
try and	33

Table 9: The frequency of using some colloquial and vague words in UDELC

Figure 14 illustrates that although the word *things* is underrepresented in native academic texts, being clearly typical of spoken language, it was overused in the UDELC.

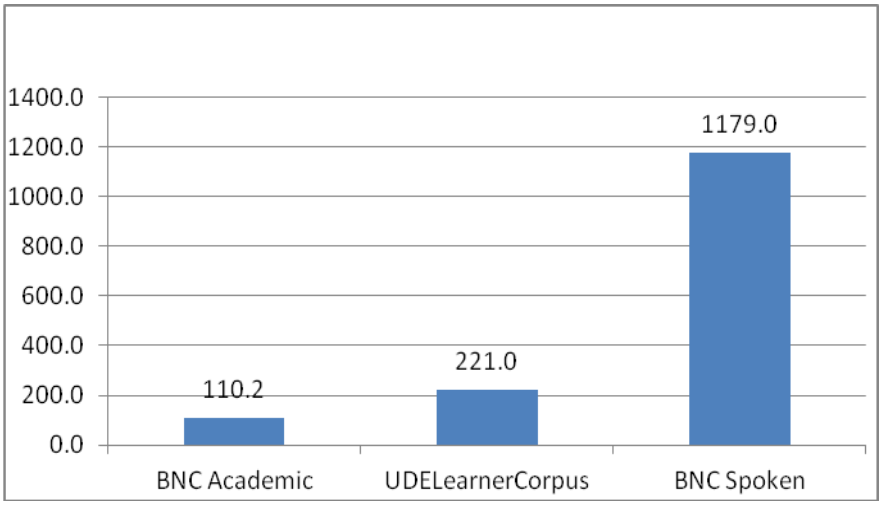

Figure 14: Relative frequency of things *in native academic writing, learner academic writing and native speech* (frequency per million words)

4 Conclusion and pedagogical implications

The present study aimed to analyse how learners use certain lexical items in academic writing to measure their awareness of the academic register. The results largely confirmed Gilquin and Paquot's (2008) findings about the academic writing style of a multicultural group of undergraduate students, but the study also highlighted some further error prone areas where Hungarian upper-intermediate to advanced level undergraduate students of English are likely to run into difficulties when writing an academic text. Student academic writers tended to overuse certain well-known and frequently used words and expressions, transferring them from the spoken language conventions they were more familiar with.

The findings of this study imply that teachers of academic writing should pay even more attention to sensitizing students to the differences between the spoken and academic genres (cf. Gilquin et al., 2007) by reading and analysing specialized texts, solving vocabulary exercises developed for this purpose, and doing many focused written tasks to raise their

consciousness of the style yet new to them. Luzon (2009) also suggests analyzing the co-text of the authorial first person pronoun, identifying phraseological patterns which help to achieve the desired rhetorical strategies typical of the academic genre. An additional tool for this could be introducing corpus-driven writing lessons, where learners can individually explore the features of academic texts by quantitatively and qualitatively analyzing both learner and native expert corpora (cf. Hyland 2002, Harwood 2005). By doing so, students can get closer to a well-developed academic literacy, and, in addition to being proficient in general spoken and written language skills, they can become more proficient academic writers as well.

References

Adel, A. & B. Erman (2012): Recurrent word combinations in academic writing by native, and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes* 31, 81-92.

Anthony, L. (2012): *AntConc* (Version 3.3.5) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.antlab.sci.waseda.ac.jp/

Cobb, T.M. (2009): *Compleat Lexical Tutor v.6.2.* [Online Computer Software]. Universite de Quebec: Toronto. Located at http://www.lextutor.ca/

Coxhead, A. (2000): A new Academic Word List. *TESOL Quarterly* 34.2, 213–238.

Gilquin, G., Granger, S. & Paquot, M. (2007): Learner corpora: the missing link in EAP pedagogy. In Thompson, P. (ed.): *Corpus-based EAP Pedagogy*. Special Issue of Journal of English for Academic Purposes. 6.4, 319-335.

Gilquin G. & Paquot, M. (2008): Too chatty: Learner academic writing and register variation. *English Text Construction* 1.1, 41-61.

Harwood, N. (2005): 'I hoped to counteract the memory problem, but it made no impact whatsoever': discussing methods in computing science using I. *English for Specific Purposes* 24.3, 243-67.

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

Hyland, K. (2002): Options of identity in academic writing. *ELT Journal* 56.4, 351-358.

Hyland, K. & Tse, P. (2007): Is there an "Academic Vocabulary"? *TESOL Quarterly* 41.2, 235-253.

Lee, D.Y.W. & Chen, S.X. (2009): Making a bigger deal of the smaller words: Function words and other key items in research writing by Chinese learners. *Journal of Second Language Writing* 18, 281-296.

Lei, L. (2012): Linking adverbials in academic writing on applied linguistics by Chinese doctoral students. *Journal of English for Academic Purposes* 11, 267-275.

Luzon, M.J. (2009): The use of we in a learner corpus of reports written by EFL Engineering students. *Journal of English for Academic Purposes* 8, 192-206.

Oppenheimer, D.M. (2005): Consequences of Erudite Vernacular Utilized Irrespective of Necessity: Problems with using long words needlessly. *Applied Cognitive Psychology* 20(2), 139-156.

Paquot, M. (2010): *Academic Vocabulary in Learner Writing: From Extraction to Analysis*. London & New-York: Continuum, 56-58.

Sherlock, K.J. (2008): *Advice on Academic Tone*. Retrieved from http://www.grossmont.edu/karl.sherlock/English098/Resources/Academic_Tone.pdf

Starkey, L.B. (2004): *How to write great essays.* New York: Learning Express LLC.

The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

West, M. (1953): *A General Service List of English Words*, Longman: London.

Xue G. & Nation, I.S.P. (1984): A University Word List. *Language Learning and Communication* 3.2, 215-229.

Gyula Sankó
University of Debrecen
Institute of English and American Studies
Pf. 73
H-4010 Debrecen
sanko.gyula@arts.unideb.hu