http://dx.doi.org/10.18778/7969-032-9.06

Łukasz Wodarczyk
Lodz University of Technology, Lodz, Poland

A Quantitative Corpus-based Analysis of Linking Adverbials in Students' Academic Writing

1. Introduction

This study is an attempt to analyze the use of linking adverbials by Polish students in their academic writing. To the best of my knowledge, the use of linking adverbials by students with Polish L1 background has not been yet examined. Scheffler (2008) has analyzed learner language of Polish students of English in opposition to native language, however, his work has a broader scope focusing on word clusters.

The underlying assumption is that specific linking adverbials are overused by Polish students. This assumption was investigated in the past by Scheffler (2008) who showed that Polish students' essays are much more explicit than newspaper editorial texts in marking logical relations or meaning connections between clauses and sentences (2008: 174). This paper will, however, examine more heterogenic corpora of texts.

2. Procedures

A bottom-up procedure was applied in this study. The enquiry began with collecting a representative corpus of students' writing in order to compare its quality with similar texts produced by native speakers.

One of the requirements of a learner corpus is its representatives and balance. Not only does a corpus need to represent the same language but also

the same domain and genre. To meet this requirement the body of texts has to be similar. Different essays produced by non native English language students deal with different topics, therefore balance and representatives would be difficult to achieve. However, BA thesis written by third year students of English may be considered similar in many respects. The texts are of similar length and follow the same rules of composition writing. The students in their BA thesis must apply what is called 'academic language,' a variety of formal written English language that is different from register appropriate for conversation language or written fiction.

My work began with collecting computer readable texts produced in 2009 by third-year students of English at the University of Łódź. All files had to be converted into raw *.txt* files that do not contain pictures, figures, graphs, Polish summaries or lists of contents. Finally, 26 BA theses were selected: 4 literature texts, 7 dealing with phonetics, 7 on the subject of pragmatics, 8 focused on English language teaching methodology. The BA learner corpus (BALC) contains 310 531 word tokens and 17 983 word types. The type/token ratio (TTR) that measures lexical diversity is: 5,79%. The rather small number confirms that the corpus contains similar texts that represent a certain type of language i.e. academic texts.

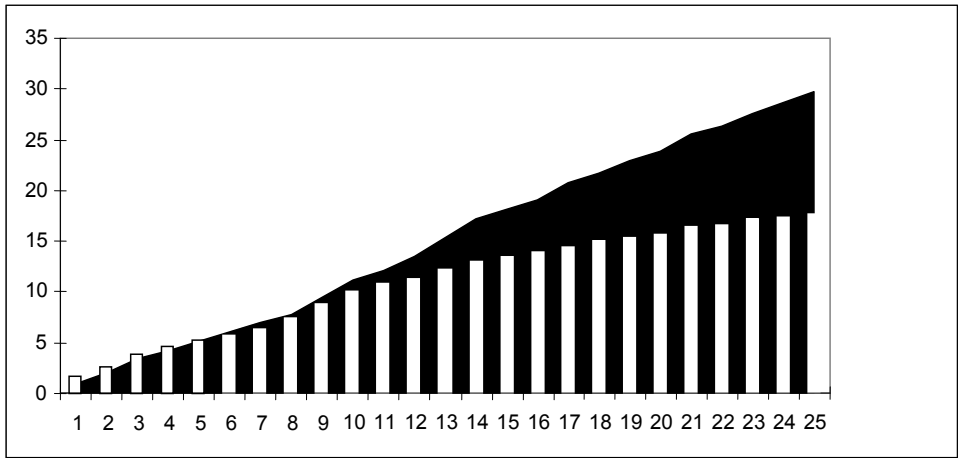

Figure 1. Type/token ratio (TTR)

Figure1 represents the TTR in BALC. Horizontal axis represents the number of texts in the corpus. The steady growing tokens (running words) are represented on the vertical axis in tens of thousands, while the word types (number of different words), which seem to have come to a standstill are represented in thousands. Looking at these figures we may assume that the corpus is balanced and representative.

Secondly, following the contrastive interlanguage analysis (CIA) principals the data from the learner corpus had to be compared with a referential corpus of academic texts. In order to create such a corpus individual files were selected from the British National Corpus.

2.1. British National Corpus

The British National Corpus (BNC) is a 100-million-word collection of samples of written language from a wide range of sources, designed to represent a wide cross-section of British English from the later part of the 20th century, both spoken and written.

The written part of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The spoken part (10%) consists of orthographic transcriptions of unscripted informal conversations (recorded by volunteers selected from different age, region and social classes in a demographically balanced way) and spoken language collected in different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

The compilation of the corpus began in 1991, and was completed in 1994. No new texts have been added after the completion of the project but the corpus was slightly revised prior to the release of the second edition *BNC World* (2001) and the third edition *BNC XML Edition* (2007).

To sum up, BNC is a monolingual corpus. It deals with modern British English, not other languages used in Britain. However non-British English and foreign language words do occur in the corpus. It is synchronic: it covers British English of the late twentieth century, rather than the historical development which produced it. Finally it is general: it includes many different styles and varieties, and is not limited to any particular subject field, genre or register. In particular, it contains examples of both spoken and written language[1].

Five hundred fifteen files were selected out of the BNC to create a referential academic corpus. The texts in the BNC are divided into genres. The following genres were selected: school essays, academic essays, written academic texts concerned with:

- humanities,
- arts,
- medicine,
- natural science,
- politics,

[1] http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=intro

- law,
- education,
- social science,
- technology,
- engineering.

The academic subcorpus of BNC has 10 711 531 word tokens and 201 924 word types which gives the TTR 1,9%. This rather small number confirms that the corpus contains similar texts that represent a certain type of language i.e. academic texts.

2.2. Academic language

Academic language is one of varieties of language that are referred to as language registers, used for a particular purpose or in a particular social setting. Language registers differ significantly from one register to another (Biber 2006: 6).

Academic language refers to the ways of thinking and using language which exists in the academy. Textbooks, essays, conference presentations, dissertations lectures and research articles are central to the academic enterprise and are written in the academic register. Individuals use language to write, frame problems and understand issues in ways specific to particular social groups and in doing these things they form social realities, personal identities and professional institutions (Hyland 2009: 1).

No new discovery, insight or invention has any significance until it is made available to others and no university or individual will receive credit for it until it has seen the light of day through publication (Hyland 2009: 2). Each academic text must comply to a certain set of rules. A view must be framed within a context of what is already accepted and using an argument carefully crafted for a particular audience. Ultimately a theory prevails because it is presented in a way which academics recognize as persuasive.

What is more, academic discourse treats events as existing in cause and effect networks, disguises the source of modality of statements, foregrounds events rather than actors, and engages with meanings defined by the text rather than in the physical context (Hyland 2009: 7).

To sum up, the complexity of academic discourse is not always recognized by tutors and administrators, which means that academic literacy tends to be misrepresented as a naturalized, self-evident and non-contestable way of participating in academic communities. The general assumption is that there is a single, overarching literacy which students have failed to master before they get to university, probably because of gaps in school curricula or faults in the learners themselves, and this deficit can be corrected by a few top-up English classes. More widely, the idea that university students cannot write is central to the official and public debate about literacy, and generic labels such as 'academic English' or

'scientific English' give the impression that literacy can be taught to students as a universal set of skills usable in any situation (Hyland 2009: 9).

Another point made by Furneaux (1995) is that the students who are to produce writing that will satisfy academic standards need to develop an understanding of what such writing involves. Many students have never been thought how to write; their schooling has never given them a lot of writing practice, but the focus is usually on the product and not the process—how you produce it. Writing in a foreign language requires the transfer of skills from the mother tongue, where they exist, or acquisition of them in the foreign language. Because of the possible L1 transfer we may assume that L2 students' academic writing may differ from the native writing.

2.3. Genre comparability

Having established that the variety of language examined is academic language, we have to consider the comparability of texts in the corpora. Not only do the examined corpora comprise of texts written in the same register but also of the same or similar genre.

The French word genre, meaning 'type' or 'kind', when applied to English literature, has been used to denote literary categories (such as types of novel, or short story) involving categorization of texts in terms of a range of structural and stylistic features (Bruce 2008: 6). Subsequently other non-print media, such as film, stage drama and graphic art have appropriated the term genre as a categorizer of creative outputs. In the last few decades, genre has also been applied to categories of non-literary written texts, sometimes for the purpose of characterizing the features of such texts for the teaching of writing. For example, newspaper editorials, letters, obituaries and different types of academic texts have also been identified as genres. These are often characterized in terms of similarities of content, the staging of the content, and the linguistic resources employed.

The genre of the texts in the BNC referential corpus is academic research paper. The examined BA theses resemble academic research papers in their construction and syntax. The following subsection will expand on the research article as an academic genre.

2.4. Research articles

Hyland (2009:10) notices that the research article (RA) remains the pre-eminent genre of the academy. Beginning life in the form of the letters published in The Philosophical Transactions of the Royal Society in the mid-seventeenth century, the RA is now the principal site of disciplinary knowledge-making. One reason for this pre-eminence is the value attached by the scholarly establishment to

the processes of peer review as a control mechanism for transforming beliefs into knowledge. Another is the prestige attached to a genre which restructures the processes of thought and research it describes to establish a discourse for scientific fact-creation. Language becomes a form of technology, which attempts to present interpretations and position participants in particular ways as a means of establishing knowledge. Consequently, the RA is a genre which has generated such a volume of research that it defies easy summary (Hyland 2009: 68-69).

The BA thesis follows similar principals as research articles. It may be longer, with more emphasis put on the literature review. Still their aim is the same as that of research articles produced by academics. They are to review existing knowledge, and through following the cause and effect relation indicate 'new knowledge' so that it may be accepted by the reader. In that respect choosing academic texts from the BNC and comparing them with non-native academic texts seems to be reasonable.

However, working with entire texts of students' BA theses may cause some valid objections. As it was mentioned, the BA theses include quite a lot of literature review, in other words, students apply paraphrase or direct quotations. Some paraphrases and all direct quotations cannot be treated as genuine students' production; therefore, measures have been applied to localize and exclude all quotations and a number of paraphrases from the examined texts in the BALC.

2.5. Software

As Anthony (2004) puts it, a corpus is virtually useless without some kind of computer software tool to process it and display results in an understandable way. Therefore, after gathering the text files and dealing with theoretical considerations concerned with the quality of the texts, time has come to choose a corpus software that will run specific queries. Despite the popularity of WordSmith Tools, I have decided to use more user-friendly, all-in-one, non-profit corpus analysis toolkit created by Laurence Anthony (2004).

AntConc hosts a comprehensive set of tools including a powerful concordancer, word and keyword frequency generators, tools for cluster and lexical bundle analysis, and a word distribution plot. For research purposes one may download the latest version of AntConc from the authors website http://www.antlab.sci.waseda.ac.jp/

Figure 2 shows a screenshot of AntConc while a user is operating the Concordancer Tool. Concordance lines are displayed in the middle of the screen. Each result, text line, is associated with the particular text (see: right-hand side of the screen). On the left-hand side there are all text files that constitute the corpus. At the bottom of the screen the user may key in the query. In the top-right corner the computer displays total number of hits i.e. raw frequency of a given item. All other options invisible at the screenshot are easily accessible and user-friendly.

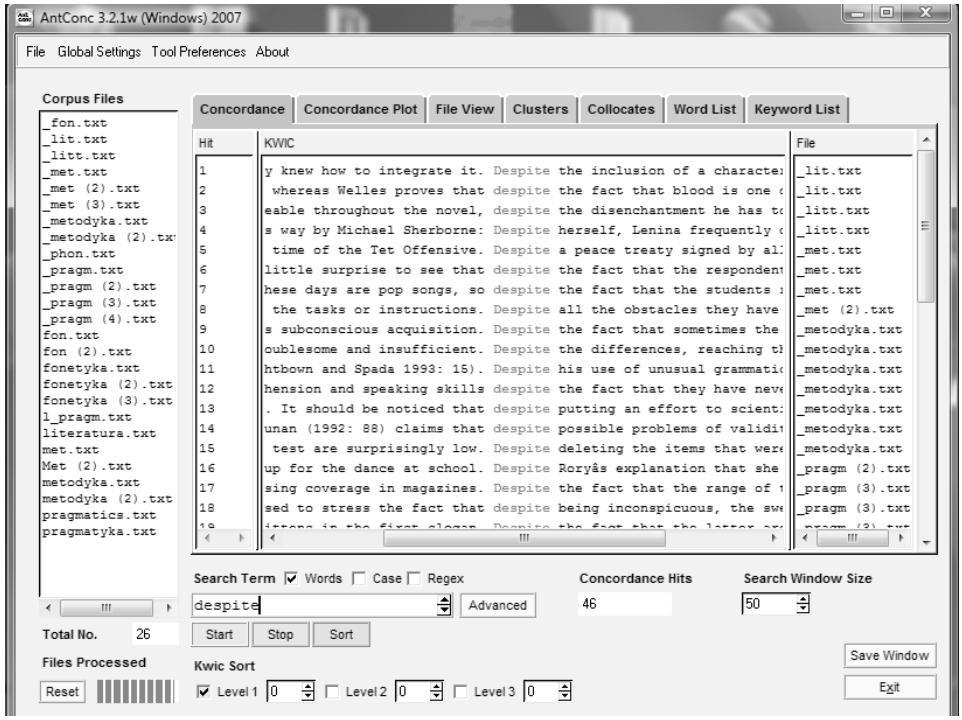

Figure 2. AntConc interface

3. Methods

On the basis of previous research on linking adverbials, their use by native speakers in academic writing, described in the second Chapter a number o linking adverbials were chosen.

As Shaw (2009) notices different text types have different text profiles of linking adverbial use. In academic prose the most frequent types are: *however, thus, therefore, for example,* and *then.* Unfortunately, *for example* has different orthographical realizations i.e. *for example* or *e.g.* While both realizations should be accounted for, the software can only search for one, or the other which will compromise the results.

Having chosen the types of linking adverbials, chi-square test was applied to determine whether word frequencies are significantly different from their general distribution reflected in the reference corpus (Pęzik 2010). The chi-square test is applied by preparing a contingency table (Table 1) where Corpus 1 and Corpus 2 are the working corpus and the reference corpus respectively.

Table 1.

	Corpus 1	Corpus 2
Frequency of word x	a	b
All other words	c	d

Then, the following formula is calculated for each item:

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)}$$

According to the chi-square results table each value grater that 1 is statistically significant and may be taken under consideration. The computation of each item is presented in Appendix 2 in an exel spreadsheet. The chi-square values for the examined linking adverbials were as following.

Enumeration and addition linking adverbials:

- *first* 6,19
- *firstly* 54,23
- *moreover* 424,75
- *furthermore* 115,41
- *then* 94,11
- *finally* 83,64
- *in addition* 0.55
- *secondly* 1,58
- *additionally* 816,13.

Summation:

- *in conclusion* 0.04
- *to sum up* 44,22.

Apposition linking adverbials:

- *in other words* 27,25

Result /inference:

- *therefore* 5,10
- *thus* 0,96
- *as a result* 7,31
- *so* 36,77.

Contrast/concession:

- *however* 95,72

- *in contrast* 5,19
- *in spite* 0,01
- *despite* 0,00
- *on the other hand* 161,75
- *nevertheless* 8,44
- *nonetheless* 1,16.

Unfortunately a few items had to be ignored because of the unsatisfactory chi-square value, among them *thus,* which is quite frequent in academic language.

In the following subsection I will present the overuse/underuse of linking adverbials which are statistically significant in the present study. Patterns of use will be provided, and a qualitative analysis conducted.

4. Results

Density of linking adverbials in BALC is 131 and in Academic BNC subcorpus 151, which means that every 131 word in BALC and every 151 word in Academic corpus is one of the linking adverbials listed in the previous subsection. Overall the difference is not impressive. It seems that the frequency of usage of linking adverbials does not differ a lot. Only when we look at individual types of linking adverbials can we notice some interesting results.

4.1. Enumeration and addition linking adverbials

Table 2. Enumeration and addition linking adverbials per million words

	BALC	Acad BNC
First	1307	1153
Then	493	1064
Moreover	451	85
Finally	319	127
Furthermore	203	54
Additionally	203	8
Firstly	97	26
Secondly	71	54

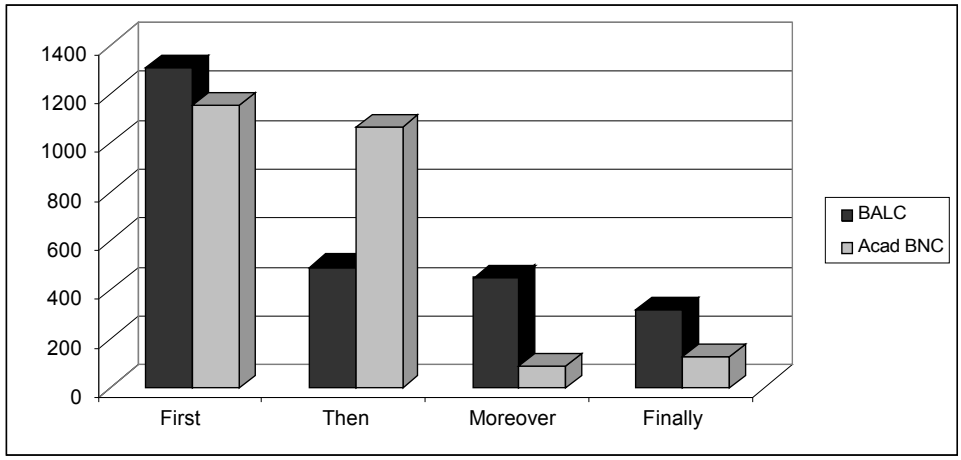

Figure 3. Enumeration and addition linking adverbials in BALC and Acad BNC subcorpus

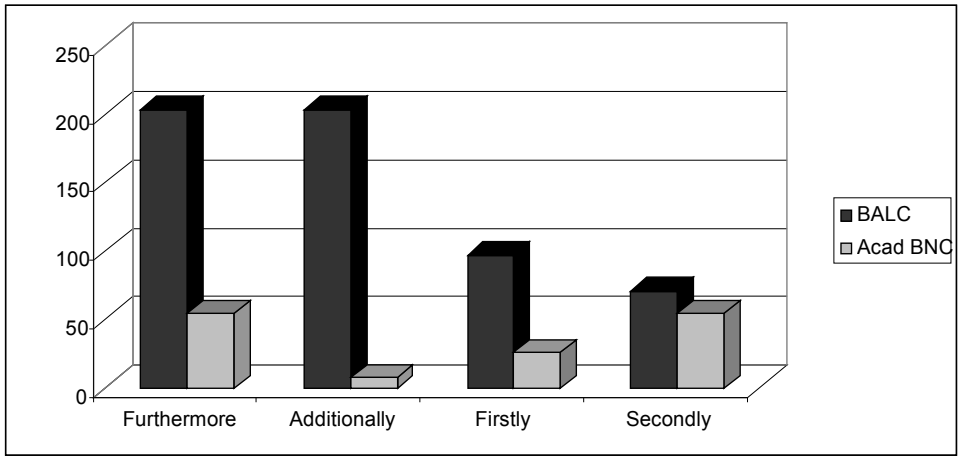

Figure 4. Enumeration and addition linking adverbials in BALC and Acad BNC subcorpus

Enumeration and addition linking adverbials are used for structuring information. It seems that the enumeration and addition liking adverbials are the group of adverbials that are most frequently overused by Polish students. The most significant and surprising result is the overuse of *additionally*. It occurs 203 times per million words in Polish students' written work, while in the BNC academic subcorpus it is used only 8 times per million words. Sometimes, however, a frequency list based on simple frequencies may not reflect the typicality of word distribution in a corpus (Pęzik 2010). To verify the results dispersion plots were generated using the AntConc software—figure 5. Dispersion plot is the distribution pattern of certain items within the texts in the corpus. It seems that while *additionally* is used from one to fourteen times in most of BA theses,

in one thesis it occurs 23 times, which is significantly more that in other texts. This number must have had an influence on the general results. Still, even if this 'unrepresentative' text was to be excluded, the overuse would still be noticeable. Other addition linking adverbials like *furthermore* and *moreover* are overused as well. The results of *furthermore* are similar to *additionally* in this respect that one student overused this particular linking adverbial heavily. The other examined enumeration and addition linking adverbials are distributed equally among all texts. Thus the other results may be considered trustworthy.

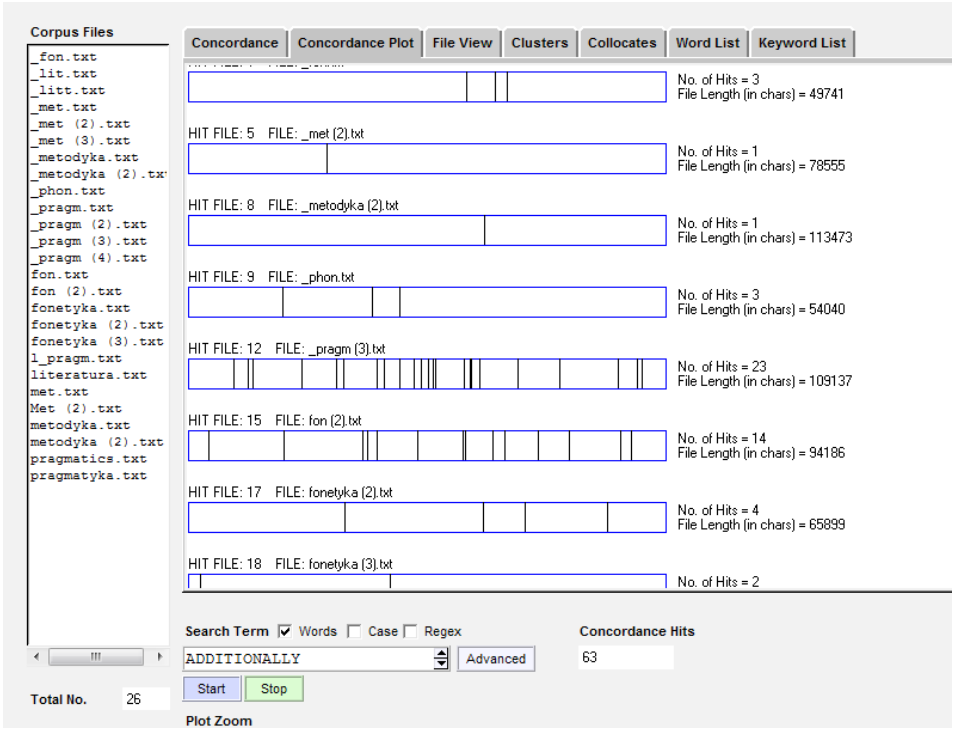

Figure 5. Dispersion plot of the item *additionally* in BALC

The pattern of overuse is broken only by *then.* It seems that Polish students avoid using it. It may be justified by the idea that *then* is regarded as an informal linking device. While, on the contrary it is one of the most frequently used additive and enumerative linking adverbials in academic writing according to the examined Academic subcorpus of BNC and Bibers's et al. (2007) data.

4.2. Summation linking adverbials

To sum up is used 23 times per million words by Polish students, while the frequency in academic language is only 2 times per million words. Summing up the most important ideas in the text is an essential part of any academic text. It seems that Polish students fail to notice the variety of options and choose the most common one, or, at least, the one that has been thought and applied by their English tutors.

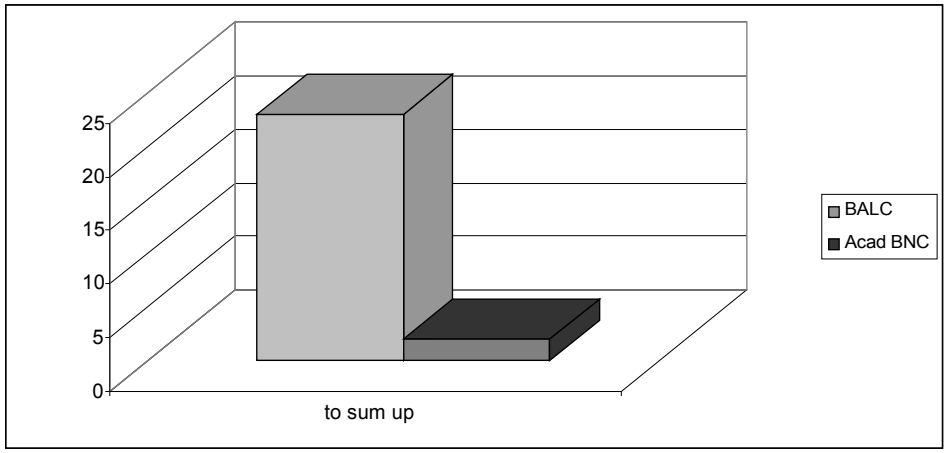

Figure 6. Summation linking adverbial

4.3. Apposition linking adverbials

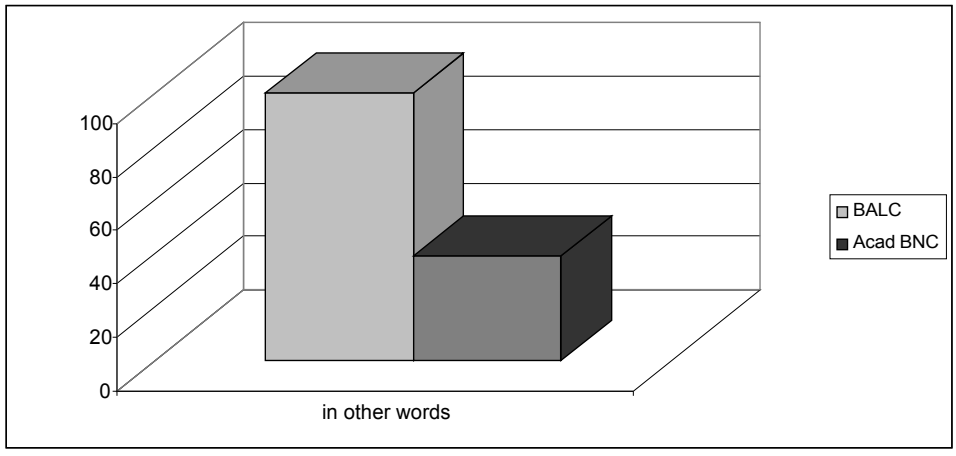

Figure 7. Apposition linking adverbial

In other words—an apposition linking adverbial is overused as well. One hundred times PMW compared to 39 times PMW in Academic subcorpus of BNC. Rephrasing a quotation in order to prove ones knowledge of the subject, or presenting the idea in a more comprehensive manner is a quite commonly used technique by students. This need may not exist on such a scale in native academic texts. What is more, native writers may use a wider range of apposition linking adverbials instead of *in other words*.

4.4. Result/inference linking adverbials

Table 3. Result/inference linking adverbials—use per million words

	BALC	Acad BNC
Therefore	303	383
As a result	148	99
So	1224	1673

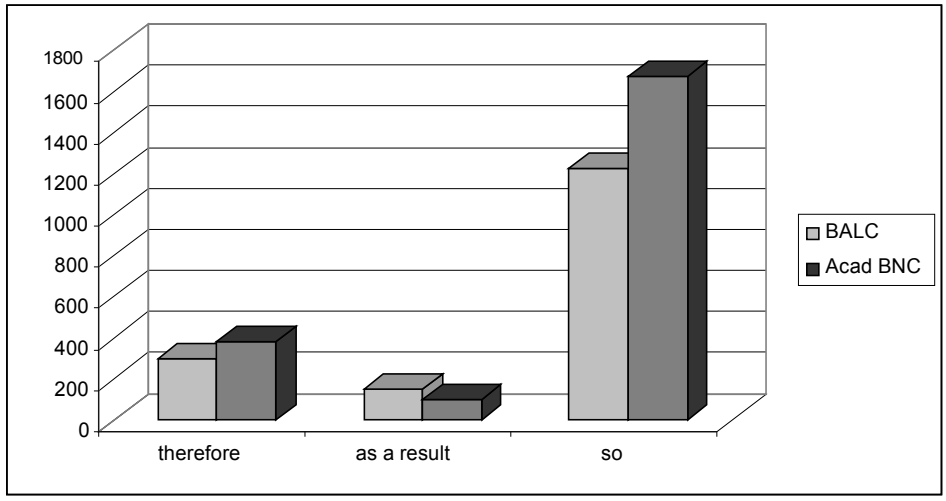

Figure 8. Result/inference linking adverbials

Among the examined result/inference linking adverbials the pattern of overuse is not recognizable; on the contrary, Polish students underestimate *therefore* and *so.*

The difference in the frequency of usage of the latter in noticeable. Polish students may have 'branded' *so* as informal and thus avoid using it as long as it is possible. *As a result* is slightly overused but the difference is relatively small. For some reasons students prefer to use *as a result* instead of *therefore.* The difference may result from individual preferences and is not crucial for this study.

4.5. Contrast and concession linking adverbials

Table 4. Contrast/concession linking adverbials

	BALC	Acad BNC
However	1336	822
In contrast	61	36
On the other hand	290	78
Nevertheless	167	111
Nonetheless	48	15

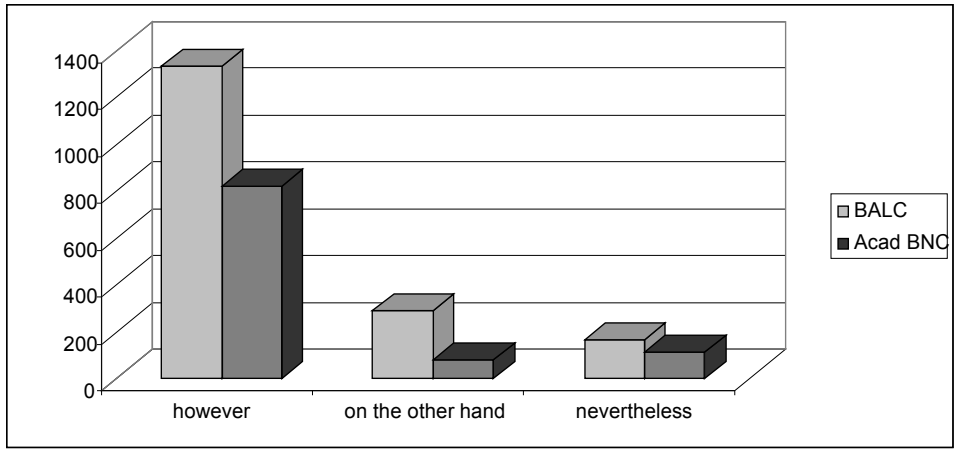

Figure 9. Contrast/concession linking adverbials

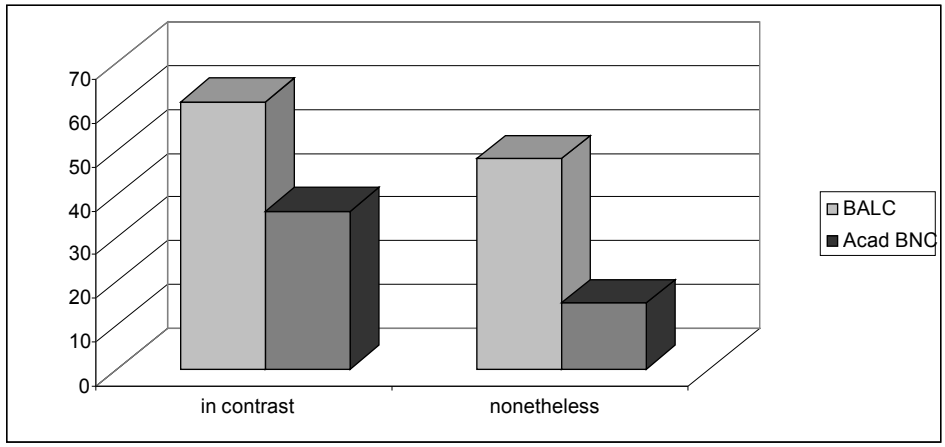

Figure 10. Contrast/concession linking adverbials

A noticeable pattern of overuse is visible in this category. *However* seems to be one of the most frequently used linking adverbials by Polish students. It is worth to notice that *but* and *however* are close in meaning to each other (Shaw 2009). It may be the case that *but* is underused, instead the more formal *however* is preferred. The fact that *but* occurs only 2302 times per million words in BALC, while in academic subcorpus of BNC 3848 times per million words may support this hypothesis. Other contrast/concession linking adverbials like *in contrast* and *nonetheless* follow the pattern of overuse. Contrast/concession linking adverbials are responsible for connection opposing ideas. Students who want to explicitly mark the logical relations between sentences, thoroughly present their topic and review existing literature will approach their subject from different angles and juxtapose different views. It may seem that native writers are more self-confident and they 'get straight to the point'. Experienced native research writers can assume that readers will fill in the links.

5. Discussion and limitations

Despite the best efforts the present paper did not exhaust the topic of non-native usage of linking adverbials. The presented results consistently follow the applied methodology, however the size of the learner corpus, despite its representativeness, is relatively small. A bigger corpus might provide more reliable results. What is more, one might argue that the texts in the learner corpus are characterized by a high degree of paraphrase. That is to say the learner language is not genuine students' production as it is based on other academic works concerned with the given topic. Moreover, one has to agree that the tutors involvement in the process of writing in each text could have had some influence on the final result. However,

while the role of the tutor is invaluable, the final choice of words is made by the student.

The present work limits itself to a quantitative study of linking adverbials' usage in academic texts produced by L2 students of English. However, it may be developed into qualitative study, where issues such as position of linking adverbials in a sentence or the concept of 'dummy adverbials' (Pęzik: unpublished), which might be connected with L1 transfer could be examined.

6. Conclusions

The results of the study show that most of the examined linking adverbials are overused by Polish students of English in academic texts. The results correlate with other studies. For example Shaw (2009) points out that a number of studies have shown a higher density of linking adverbials for L2 students compared to L1 students and/or professional L1 writers (Bolton et al. 2002; Green et al. 2000; Milton & Tsang 1993). However, 'informal' linking adverbials: *so* and *then* are underused. Students who have been thought to use only formal language avoid using the result/inference linking adverbial *so* and enumeration/addition linking adverbial *then* for fear of breaking the 'formal' register rules. On the other hand, other 'more sophisticated' linking adverbials were overused in comparison to native writers.

Overuse of linking adverbials may by justified by various explanations. It is possible that a high density of linking adverbials in learner genres is a consequence of text purpose: the writers have to show the readers/graders explicitly that they understand relationships between arguments correctly. What is more, Gardezi and Nesi (2009) observe, that linking adverbials are simple concrete items, easily taught and useful in drawing learners' attention to the importance of logical coherence. As a result, writers of learner genres, and particularly non-natives, are often encouraged to use linking adverbials to articulate the structure of their argument. Shaw (2009) claims that there is a sort of paradox here, in that within one learner genre, like the test essay, higher-rated products may have more linking devices, and development is associated with using more linking adverbials, but comparing learner genres with professional ones, fewer linking devices appear in the work of the presumably more skilled group. It, therefore, seems that the abilities acquired in the earlier stage of language teaching are used in the same way later in English education.

To tackle this problem, Tankó (2004) suggests some teaching implications. It seems that the teacher should supply a reliable and thorough introduction to linking devices. Information on the variety of linking adverbials and their frequency in various spoken and written text types can be given on the basis of such sources (e.g. Biber et al. 2007) that rely on corpus evidence. The teacher can furthermore

give valuable feedback concerning the number of linking devices used in students' texts as well as make explicit, relevant and therefore effective comments based on particular instances taken from students' texts concerning the question of when to use and when not to use linking adverbials. Their work may be more efficient if students hand in their written assignments in a computer readable format, which will allow for the use of corpus tools.

One should also bear in mind that, however, the presence, the frequency, and the distribution of linking devices in a particular text cannot be considered the ultimate indicator of text quality. A text that contains an acceptable number of stylistically appropriate linking adverbials applied in the right positions can still be devoid of either logic or content (Tankó 2004).

Although English is an international language, I believe that we should pursue the native norms, ways and forms in which native users of English express their ideas so that we may benefit more from our communication.

References

Anthony, L. 2004. AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit In: *IWLeL 2004: An Interactive Workshop on Language e-Learning*, pp. 7-13.

Biber, D. 2006. *University Language. A corpus based study of spoken and written registers.* Amsterdam/Philadelphia: John Benjamins Publishing company.

Biber, D. et al. 2007. *Longman Grammar of Written and Spoken English.* Edinburgh Gale: Pearson Education Limited.

Bolton, K., Nelson, G. and 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.

British National Corpus, http://www.natcorp.ox.ac.uk/corpus/index.xml?ID=intro, Accessed 3 January 2011.

Bruce, I. 2008. *Academic Writing and Genre : A Systematic Analysis.* London: Continuum International Publishing.

Furneaux, C. 1995. The Challenges of Teaching Academic Writing. In *BBC English: Teachers Suplement.*

Gardezi, A. and Nesi, H. 2009. "Variation in the Writing of Economics Students in Britain and Pakistan: The Case of Conjunctive Ties" In: Charles, M., Hunston, S., Pecorari, D. (Eds.), *Academic Writing: At the Interface of Corpus and Discourse*. London: Continuum International Publishing.

Hyland, K. 2009. *Academic Discourse.* London: Continuum International Publishing Laurence Anthony's Website, http://www.antlab.sci.waseda.ac.jp/, 3 January 2011.

Milton, J. & Tsang, E.S.C. 1993. A corpus-based study of logical connectors in: EFL students' writing: Directions for future research. In R. Perbertom & E.S.C. Tsang (Eds.), Lexis in Studies. Hong Kong: Hong Kong University, pp. 215-246.

Pęzik P. 2010. Computational and Corpus Linguistics. In: *New Ways to Language*. Łódź: Wydawnictwo Uniwersytetu Łódzkiego.

Polish and English Language Corpora for Research and Applications, www.pelcra.pl, Accessed 12 January 2011.

Scheffler P. 2008. *Native English and learner corpora: linguistic comparison and pedagogical applications.* Poznań: Wydawnictwo Naukowe.

Shaw, P. 2009. Linking Adverbials in Student and Professional Writing in Literary Studies: What Makes Writing Mature In: Charles, M., Hunston, S., Pecorari, D. (Eds.), *Academic Writing: At the Interface of Corpus and Discourse*. London: Continuum International Publishing.

Tankó, G. 2004. The use of adverbial connectors in Hungarian university students' argumentative essays. In Sinclair, J.McH (Ed.) *How to Use Corpora in Language Teaching* Amsterdam/ Philadelphia: John Benjamins Publishing Company.