RL Vol. 16.3 2018

Research in Language

WYDAWNICTWO
UNIWERSYTETU
ŁÓDZKIEGO

ŁÓDŹ 2018

EDITORS

Iwona Witczak-Plisiecka, Ewa Waniek-Klimczak, Jan Majer – University of Łódź
Faculty of Philology, Department of English Language and Applied Linguistics
90-236 Łódź, Pomorska 171/173

All papers published in "Research in Language" have been blind-reviewed by at least two independent readers

INITIATING EDITOR
Agnieszka Kałowska

EDITING AND TYPESETTING
Michał Kornacki

COVER DESIGN
Katarzyna Turkowska

Printed directly from camera-ready materials provided to the Łódź University Press

© Copyright by Authors, Łódź 2018
© Copyright for this edition by Uniwersytet Łódzki, Łódź 2018

Published by Łódź University Press

First Edition. W.09031.18.0.C

Printing sheets 7.375

ISSN 1731-7533
e-ISSN 2083-4616

Łódź University Press
90-131 Łódź, Lindleya 8
www.wydawnictwo.uni.lodz.pl
e-mail: ksiegarnia@uni.lodz.pl
phone (42) 665 58 63

CONTENTS

EXPERT-LIKE USE OF HEDGES AND BOOSTERS IN RESEARCH ARTICLES WRITTEN BY POLISH AND ENGLISH NATIVE-SPEAKER WRITERS

KATARZYNA HRYNIUK
University of Warsaw, Poland
k.hryniuk@uw.edu.pl

Abstract
The present study compares the use of main interpersonal metadiscourse markers - hedges and boosters - in a corpus of 40 research articles from the area of applied linguistics, written in English by native speakers and Polish writers. Used as communicative strategies, these words and expressions increase (boosters) or reduce (hedges) the force of arguments. In order to gain an in-depth insight and to achieve greater precision, in the analysis the author utilizes a concordance tool WordSmith 6.0 (Scott 2012). The results point to important discrepancies in the usage of these text features by authors representing different native languages and cultures. The study has important implications for developing competence in writing for publication in English as a Foreign Language.

Keywords: booster, corpus, hedge, metadiscourse, research article, writing for publication

1. Introduction

Gaining expertise in writing for publishing is presently increasingly important for academics in all disciplines, since publications in highly rated international journals have a great impact not only on knowledge construction through the process of writing, which is of primary importance, but also on their basis, universities are funded and scholars are evaluated. In Poland as well, the current academic evaluation system requires publication in English as a Foreign Language (EFL), also called English as an Additional Language (EAL) in this context, in prestigious international journals, where Anglo-American conventions prevail. At the same time, previous research shows that writers from other than Anglo-American cultural regions face many challenges when writing for publication in English because of distinct conventions that they follow, which are shaped by different literacy traditions.

Before discussing the issue of writing for publishing in EAL and the use of metadiscourse, first two concepts need to be distinguished, namely, second-language proficiency for general language use and academic writing expertise (Cumming 1989), because they are often understood as equivalent. Certainly, advanced foreign language proficiency is a prerequisite for successful writing in EAL. As Cumming (1989) claims, it has an additive value, because greater or

lesser level of it influences the quality of the text, but it is not sufficient. Expertise in academic writing requires also being able to engage in a highly complex composing process and this ability is attained with great effort in any language. As Weigle (2005) states, it is only rarely achieved even in the first language.

Academic writing, from socio-constructivist point of view, is a socially and culturally situated activity. Hence, expertise in writing for publishing develops through a writer's socialization into academic discourse community which shares a set of values and cultural preferences as to what 'good' writing should be like (e.g., Duff 2007, 2010; Flowerdew 2013; Hyland 2009). Expert writers, in order to complete the complex task of writing, must use a number of appropriate strategies and areas of knowledge, such as: topic and language knowledge, genre knowledge, audience knowledge, task schemas, and metacognition, which is "higher order thinking involving active management of the cognitive processes engaged in complex tasks" (Weigle 2005: 135).

Weigle (2005) makes a distinction between the engagement in cognitive activities by unskilled and skilled academic writers. As she claims, for experts writing is not less effortful than for novice writers. However, the main differences between them are that not only do skilled writers attend to conventions and orthography in writing, but also they make appropriate choices of syntactic structures and words to convey their messages, and they simultaneously monitor and evaluate their choices, bearing in mind a representation of a reader. They try to predict what will be persuasive for the audience, which rhetorical devices will be the most convincing, and how the readers will respond. They consider the readers' background and expectations. As Weigle (2005: 132) writes, "skilled writers are able to attend to a wider variety of considerations simultaneously, to use their resources flexibly in solving rhetorical and content problems, and to adjust their message to meet the needs of their audience." For novice EAL writers, on the other hand, the task of writing for publishing may be much more challenging, because they often lack appropriate knowledge of the conventions of writing or genre knowledge, and by imitating their native language ways of expression, they make inappropriate choices of metadiscourse. They frequently have worse awareness of the rhetorical effect that specific language resources can have on readers. Therefore, raising awareness of these language items use, especially among second language writers through corpus-based Data Driven Learning (DDL) (Johns 1991), as it is the case in the study carried out in this paper, can facilitate improvement in their writing.

Before outlining previous research results and describing the study carried out for the purposes of the present paper, the following terms need to be clarified: *metadiscourse*, and *hedges* and *boosters* (i.e., epistemic markers), belonging to the group of interpersonal metadiscourse markers, because they are the focus of the analysis. The most extensive work on *metadiscourse*, including a chapter discussing the definition of this term, is Hyland's (2005a) book published under this title. As the author writes, the term *metadiscourse* was coined by Zellig Harris

in 1959 and it was introduced into the applied linguistics vocabulary in the 1980s (Hyland 2005a). In essence, as Hyland (2005a: 16) writes, "'metadiscourse' is an umbrella term, used to include an apparently heterogenous array of cohesive and interpersonal features which help relate a text to its context." It can be defined simply as "discourse about discourse." It is considered as a fuzzy term which encompasses a wide collection of language items used to describe both the organization of discourse and the ways in which we relate to our listeners or readers. The adjective 'fuzzy' here means that the concept lacks clear-cut boundaries. In other words, sometimes it is hard to make a precise distinction between what is and what is not metadiscourse (Ädel 2006). Although this does not eliminate the fuzziness of the term, in a wider sense, as applied linguists, composition theorists, and rhetoricians agree, metadiscourse refers to "the various linguistic tokens employed to guide or direct a reader through a text so both the text and the writer's stance is understood" (Hyland 2005a: 18). Hyland (2005a: 37) finally arrives at the following explanation: "Metadiscourse is the cover term for the self-reflective expressions used to negotiate interactional meanings in a text, assisting the writer (or speaker) to express a viewpoint and engage with readers as members of a particular community." This definition introduces a typology of the lexical items.

The most general classification of metadiscourse markers into *interactive*, also called *textual* (i.e., guiding the reader through the text), and *interactional* (i.e., involving the reader in the text) was also made by Hyland (1998a, 2004a, 2004b). Each of these two categories includes five types of metadiscourse markers with *hedges* and *boosters* (in other words, *emphatics*) belonging to the category of the *interactional* ones. Definitions of hedges and boosters have been widely discussed, for example, by Crompton (1997) and Hyland (1998b). For the purposes of this paper, the following definition of *boosters*, referring to their function, will be used: "[boosters] express conviction and assert a proposition with confidence, representing a strong claim about a state of affairs … [they] mark involvement and solidarity with an audience, stressing shared information, group membership, and direct engagement with readers" (Hyland 1998b: 350). These are expressions, such as: *of course, clearly, obviously*, etc.

The definition of a *hedge*, which seems the most adequate, was formulated by Lyons (1977: 797; as cited by Crompton 1997: 281) in the following way: "an item of language which a speaker uses to explicitly qualify his/her lack of commitment to the truth of a proposition he/she utters." Examples of such expressions are: *possible, might, perhaps*, etc. Generally, hedges and boosters, which belong to the main interpersonal metadiscourse markers contribute to "the rhetorical expression of the relationship between writer and reader" (Hyland 2004b: 87). Used as communicative strategies, they increase (boosters) or reduce (hedges) the force of arguments. In other words, hedges are used to "move away from what can be safely assumed or experimentally demonstrated", and boosters

to express ”conviction or the significance of the work” (Hyland 2004b: 101). A convincing argument requires the use of both.

The occurrence of hedges and boosters in academic discourse written by Polish authors has not been explored extensively so far. Thus, the aim of this paper is to present a cross-linguistic and cross-cultural study comparing the use of these lexical items in the corpora of 20 research articles written by native English speakers, and in 20 articles written by Polish writers, specialists in the area of applied linguistics. It includes both quantitative and qualitative analyses of their appearance in the text. In the quantitative exploration, a chi-square test was conducted in order to compare the number of hedges and boosters in the corpora, and to establish whether the differences were statistically significant. It is argued here that the discrepancies between the usage of hedges and boosters by the two groups of writers may be linked with the national cultures that they represent. Before reporting on the results of the analyses carried out for the present paper, however, previous research results will be outlined below.

2. Previous research

The most extensive research on English metadiscourse, including hedges and boosters, has been carried out by the same author as cited above, i.e. Hyland (1998a, 1998b, 1999, 2004a, 2004b, 2005b, 2010), and by Hyland and Tse (2004). Most of the studies included analyses of disciplinary differences in the use of these lexical items in various genres. Many of them were also based on interviews with the writers of the texts (Hyland 1998b, 2004a, 2004b, 2005b, 2010). There are also cross-linguistic and cross-cultural studies by other authors outlined below. This review focuses on studies exploring genres such as research article, scientific letter[1], academic textbook and dissertation. What they have in common is that they are usually antecedents and serve as models to be followed by novice scholars writing their first research articles.

In one of the studies published in 1998, Hyland analyzed 28 research articles written in English as a mother tongue in the following disciplines: microbiology, marketing, astrophysics and applied linguistics (Hyland 1998a). He found that there were 20% more metadiscourse markers, hedges in particular, in marketing than in any other disciplines. Also, in applied linguistics there were more interactional metadiscourse markers found. It was concluded that hedges play an important role in research writing, especially in the humanities and social sciences. Similar results Hyland (1998b) obtained in another study which was an

[1] Scientific letter (also called ‘squib’ or ‘quick report’) is a very popular genre, especially in disciplines such as physics, chemistry or microbiology. It is usually less than four pages long, published monthly or weekly. Its aim is to announce new breakthroughs so it should be written in an understandable way to both researchers in the same and in other fields (Hyland 2004a).

analysis of 56 research articles in mechanical engineering, electrical engineering, marketing, philosophy, sociology, applied linguistics, physics and microbiology. The focus of this exploration were only hedges and boosters. Thus, there were four times more hedges and boosters found in philosophy than in physics. In the whole corpus there were three times more hedges than boosters. The most frequently occurring ones were: *may*, *would*, and *possible.* Over 70% of hedges occurred in the humanities and social sciences. Interestingly, the largest number of boosters was in philosophy and the least (less than 7%) in electrical engineering. In both studies, hedges were the most frequent metadiscourse markers in the whole corpus, which points to the writers' need to present claims with caution and deference to the views of the audience.

Hyland's 1999 study compared the use of metadiscourse markers in 21 textbook extracts and a similar corpus of research articles in the following disciplines: microbiology, marketing and applied linguistics. He found that in textbooks one-third of all metadiscourse markers were interactional, while in research articles, they constituted a half of them. As in the study mentioned above, their number was especially large in research articles in the area of marketing and applied linguistics. Moreover, three times more hedges were found in research articles than in textbooks. This seems logical as textbook writers most often present established knowledge to the readers, rather than cautiously introduce their new claims. Thus, the author points to the limitations of using only textbook extracts to teach research writing where metadiscourse knowledge is crucial.

In his book from 2004, Hyland also explored the use of metadiscourse markers in 56 textbook chapters in the following eight disciplines: philosophy, sociology, applied linguistics, marketing, electronic engineering, mechanical engineering, physics and biology (Hyland 2004a). Generally, epistemic markers (i.e., hedges and boosters) comprised half of all interactional discourse markers in the corpus, which points to the conclusion that the textbook content "is not simply an unreflecting repetition of uncontested disciplinary facts. Writers obviously have something to say on the epistemological status of what they report." (Hyland 2004a: 114). Hyland (2004a) found that texts from the 'soft' knowledge disciplines (the social sciences and humanities) included more interactional forms than texts from sciences. In the analyzed corpus, there was a smaller proportion of hedges in physics and engineering, and bigger proportion of boosters in engineering. In another study by Hyland (2005b), in which 240 research articles from the same eight disciplines were analyzed, similarly, more hedges were found in the 'soft' disciplines.

In the same publication, the use of hedges and boosters in 90 scientific letters from letters journals was explored by the author (Hyland 2004a). They were in the area of biology, chemistry and physics. Hyland (2004a) found that there was little difference between the disciplines in the occurrence of hedges and boosters in the letters. Hedges constituted about two-thirds of all such expressions in each field of science. However, there was around 50% more boosters in the scientific letters

than in the research articles analyzed earlier by the same author, especially in the introductions and conclusions, which points to the strong need to promote the findings published in these venues.

Due to the fact that hedges usually constitute the largest proportion of interactional metadiscourse markers, their use alone in texts from various disciplines was also extensively researched at the end of the 1990s (e.g., Hyland 1996, in molecular biology; Varttala 1999, in medicine). Hyland (1998c) again was the author of the most comprehensive, in-depth descriptions of their occurrences in academic discourse. However, a review of the main arguments and research results included in this book goes beyond the scope of this article. Other lines of research need to be outlined here, because of their immediate relevance to the study carried out for the purposes of this paper, i.e. the analyses focusing on the use of hedges and boosters by writers in English as a second language, cross-cultural comparisons of the use of these devices by native and non-native speakers of English, and by native speakers of different languages writing in their mother tongues.

Accordingly, in a series of studies, Hyland (2004b, 2010) and Hyland and Tse, (2004) have analyzed 240 doctoral and masters dissertations written by English as a second language (ESL) Chinese students from universities in Hong Kong. The dissertations were in the following six disciplines: electronic engineering, computer science, business studies, biology, applied linguistics and public administration. Generally, more metadiscourse markers were found in doctoral dissertations. The majority of devices found in the whole corpus were interactional. Hedges were the most common (41% of all interactional metadiscourse) with modal verbs such as *may, could,* and *would* appearing with the highest frequency. There were 60% more of interactional metadiscourse markers, hedges in particular, in the 'soft' disciplines, i.e. business studies, public administration, and applied linguistics. The results point to the fact that students seem to be aware of the need to present their claims to supervisors and examiners in persuasive and acceptable ways (see also: Hyland 2005b).

More recently, a study comparing the use of hedges and boosters in research articles written in English as a foreign language (EFL) by Iranian and by Anglo-American writers was carried out by Abdollahzadeh (2011). He focused on 60 conclusion sections of research articles in applied linguistics and found that the texts written by Anglo-American writers generally use more metadiscourse markers. Both groups of writers used an equal proportion of hedges but Anglo-American writers employed more boosters in their writing. These markers also performed slightly different functions in the texts of the two groups of writers. The author ascribed the differences to the writers' varied rhetorical sensitivity and awareness of the audience.

Finally, two other cross-cultural and cross-linguistic studies comparing the use of hedges and boosters in research articles written in English and in Chinese are worth mentioning. In the first one by Hu and Cao (2011), 195 research article

abstracts were analyzed. The corpus comprised three sub-corpora: Chinese abstracts published in Chinese-medium journals, abstracts in English published in Chinese-medium journals, and English abstracts in English-medium journals. The main results showed that more hedges appeared in the article abstracts in English-medium journals. The other two sub-corpora did not differ significantly in this respect. Also, there were more occurrences of boosters in the Chinese abstracts, published in Chinese-medium journals than in the other two sub-corpora. The latter did not differ in this respect. Thus, the English writers' arguments seemed more cautious, and the Chinese – more self-confident. The authors also compared the use of these devices in the abstracts of empirical and non-empirical academic articles (i.e., review, theoretical, methodological articles, etc.) and found that there were more occurrences of boosters in the former. The researchers claimed that the results can be ascribed to the cultural differences in the use of rhetorical strategies by the two groups of writers. In the second study, carried out by Mu, Zhang, Ehrich and Hong (2015), the authors compared the use of metadiscourse markers in 20 research articles in Chinese, and 20 in English. They were all from applied linguistics journals. The results showed that significantly more interactional metadiscourse markers appeared in the English sub-corpus than in the Chinese one. Hedges appeared most frequently in both sub-corpora. Similarly to the case in the previously described study, hedges occurred more often in English research articles. Chinese writers, on the other hand, employed more boosters. The two groups of writers also used hedges and boosters for slightly different purposes. As in the abovementioned study, the authors explained the results with the differences in cultural writing conventions followed by the writers.

From the above overview of the studies one can conclude that there are clear differences between texts in sciences, on the one hand, and the humanities and social sciences, on the other, in the use of hedges and boosters by native English speakers and ESL writers in many genres, except for scientific letters where the differences across disciplines were not significant. Moreover, in the studies comparing the use of hedges and boosters in research articles written by native English speakers and EFL writers, more differences were found when the scholars wrote in their mother tongues.

Few studies comparing Polish and Anglo-American research writing have been carried out, especially focusing on hedges and boosters. Apart from the research started in the 1990s by Duszak (1994) and Golebiowski (1998), the differences in cultural conventions followed by Polish and Anglo-American writers received scant attention so far (for an overview, see e.g., Hryniuk 2017). However, the abovementioned authors already noted that in the case of EFL writers, differences linked with the national cultures that they represent may play a crucial role. As Hyland and Milton (1997: 186) also observed, "students from different cultures may have preconceptions about the formal features of culturally and rhetorically appropriate writing which may differ from those which operate in English academic settings." Being educated in other cultures, novice writers may have a

different sense of the audience and the writer. The main factor having impact on the cultural differences explored in this paper may be the fact that Polish academic writing is reader-responsible as opposed to Anglo-American one, which takes into consideration the audience (Duszak 1994). Thus, a Polish writer may not feel responsible for guiding the reader through the text or engaging him in the argument by using hedges and boosters.

One of the few more recent comparative studies on Polish and Anglo-American research writing in medicine was Donesch-Jeżo's (2011) analysis of metadiscourse use. She analyzed 30 research articles and concluded that the factor which is expected to influence the use of boosters by Polish writers is academic modesty, highly valued in Polish culture. It does not allow them to describe their own work as interesting or useful. It is the reader who should make such judgments rather than the writer (Donesch-Jeżo 2011). Anglo-American writing, on the other hand, is characterized as more assertive and direct (Duszak 1994). More extensive research on epistemic modality markers used by the two groups of writers in 400 research articles in linguistics was also carried out by Warchał (2015). The results showed that Anglo-American writers used the markers of certainty and doubt twice more frequently, and boosters almost three times more frequently. Also, more of them could be found in the final sections of the articles. It must be noted, however, that both Donesch-Jeżo (2011) and Warchał (2015) compared texts written in English by Anglo-American writers with texts written in Polish by Polish writers. In the present study, all analyzed texts were written in English, so the outcome of the analyses described in the following section may be different.

3. The study

This study quantitatively and qualitatively investigates the differences and similarities in the use of hedges and boosters in the corpora of 40 research articles in the area of applied linguistics – 20 written by native English speakers, and 20 by Polish writers. All of the articles were in English. The aim of this study is to compare the frequency of their use and the location of hedges and boosters in particular sections of the IMRD structure (i.e., Introduction-Method-Results-Discussion) of research articles, and to explore if there are any significant differences in the types of hedges and boosters used in the two sub-corpora.

Thus, the study addresses the following research questions:

1. Are hedges and boosters used with the same frequency in the research articles written by Polish and Anglo-American writers?
2. Are there any significant differences in the distribution of hedges and boosters in particular sections of the IMRD(C) structure in the two sub-corpora?

3. Are there any significant differences in the types of hedges and boosters most frequently used by Polish and Anglo-American writers?

The study results will have implications for writing instruction in EFL aiming to develop sociolinguistic competence and expertise in writing. It involves raising awareness of different rhetorical conventions and the knowledge of the rhetorical effects that specific language resources can have on readers. Such instruction should assist novice writers in joining the target discourse community of experts.

3.1. Corpus and methodology

In order to address the research questions, the same corpus of 40 research articles was used as in my previous studies (see: Hryniuk 2015, 2016, 2017). It consisted of 20 research articles written by Polish writers, and 20 by native English speakers. The former set of articles was collected from two English-medium journals published in Poland, and the latter from two international journals published in the U.S. The main criteria for their selection were the following: they were all written in English; they were all from the area of applied linguistics; and they were all published in representative, peer-reviewed, highly reputable journals in the years 2009-2013. The equivalence of the sub-corpora content, also called *tertium comparationis* (Krzeszowski 1990), was achieved by using these criteria, in order to make meaningful comparisons, that is to compare the elements which can be compared and draw valid conclusions. One aspect in which the two sub-corpora differed considerably was that the Anglo-American sub-corpus consisted of 191,423 words and the Polish one of 135,358. Consequently, the articles in the Anglo-American sub-corpus were on average by 2,800 words longer. However, according to the accepted methodology, in the analyses of such small specialized corpora full texts should be used (Bowker and Pearson 2002; Flowerdew 2004). Therefore, the number of the hedges and boosters occurrence per 1,000 words (i.e., the frequency of use) was counted as well.

All articles in the corpus had the IMRD structure, typical of articles in the experimental sciences, with each section performing different communicative function. The conclusion section was added because it was present in 95% of the Polish sub-corpus. The IMRD structure is prescribed in the American Psychological Association style manual (APA 2010) and two journals from which the articles were selected referred to this style manual directly. However, it must be noted that, unlike in sciences, in applied linguistics this structure is not always strictly followed. As for example Abdollahzadeh (2011: 291) noticed in his corpus of articles in applied linguistics, "most of the articles have conclusions, some others discussion sections, yet some had results and discussion merged." Similarly, in the present study in 40% of the articles of the Polish sub-corpus the discussion sections were merged with the preceding ones or they were missing, and in 40% of the Anglo-American sub-corpus, the same happened with the conclusion sections.

In the present analyses, the taxonomy of metadiscourse markers, including hedges and boosters, introduced by Hyland (2004a, 2005a) was used. The list of hedges consisted of modal auxiliaries (e.g. *would, might, could*), epistemic adjectives and adverbs (e.g. *perhaps, mainly, likely*), epistemic lexical verbs (e.g. *seem, suggest, assume*), and other (e.g. *assumption (that), In general*). The list of boosters was comprised of modal auxiliaries (e.g. *must* to express possibility, *will*), epistemic lexical verbs (e.g. *demonstrate, find, show*), epistemic adjectives and adverbs (e.g. *actually, always, clearly*) and other (e.g. *it is well known (that), the fact that*) (Hu and Cao 2011: 2800). The complete list of 180 lexical expressions can be found in Hyland (2004a: 188-189). This compilation is based on much literature and research so it is the most reliable and appropriate for the purposes of this study.

In order to arrive at an in-depth insight into the use of hedges and boosters, and to achieve precision, in the present analysis a concordance program WordSmith Tools 6.0 (Scott 2012) was utilized. Concordance lines were generated and hedges and boosters were analyzed in context. First of all, the examples of hedges and boosters which were used in the utterances expressed by the writers' informants, rather than by the writers themselves, were excluded. The importance of context must be emphasized in this analysis because the distinction has to made between propositional and epistemic meaning of the expressions. For instance, in the sentence number (1) below, the word *about* functions as a hedge (i.e. approximation) indicating that the number is accurate enough in this text.

(1) There were about 300 punctuation marks in this text.
(2) Many authors write about the importance of the quality of the input.

In the example sentence number (2), the preposition *about* is only employed for signaling the proposition which follows; it is not a hedge (i.e., approximation).

Also, it must be noted that the same word can function both as a hedge and as a booster, depending on the context. For example, the word *quite* can perform a function of a hedge (e.g. quite good) or a booster (e.g. quite remarkable) (Hyland and Milton 1997). Thus, because of highly contextual nature of metadiscourse, all instances of hedges and boosters were individually, carefully analyzed in their sentential context first by the researcher to determine their actual functions. Then, the second rater analyzed the disputable cases till the satisfactory agreement of approximately 89% was reached. Finally, a chi-square analysis was undertaken to determine whether there are statistically significant differences between the number of hedges and boosters in the sub-corpora.

3.2. Results

Table 1 shows that even though the total number of boosters is larger in the Anglo-American sub-corpus than in the Polish one, the frequency of their use per 1,000 words is almost the same. When we take into consideration hedges, the number of

them in the Anglo-American sub-corpus is also larger. However, if we take into account the frequency, one can clearly see that it is higher in the Polish sub-corpus and the difference is statistically significant.

Table 1. The number and the frequency of hedges and boosters use per 1,000 words in each sub-corpus in the brackets. Last column – chi-square test results.

	Am.	**Polish**	**χ^2 (P)**
Boosters	1075 (5.6)	767 (5.7)	0.004 (0.843)
Hedges	2303 (12.0)	1980 (14.6)	39.48 (<0.001*)

*statistically significant difference between the Anglo-American and the Polish sub-corpora.

When we look at the numbers and frequency of hedges and boosters used in particular sections of the IMRD structure (Table 2), we can see that in most of the sections of the IMRD structure there are more both hedges and boosters in the Anglo-American corpus, but there are a few exceptions. As far as the number of boosters in the results sections is concerned, in the Anglo-American sub-corpus it is larger, but the frequency of their use per 1,000 words is higher in the Polish sub-corpus, and the difference is statistically significant. The frequency of boosters use in the discussion sections in the Polish sub-corpus is also slightly larger than in the Anglo-American one, but the difference is not statistically significant. In the conclusion sections, both the total number and the frequency of boosters use is much larger in the Polish sub-corpus and the difference is statistically significant.

As far as the use of hedges is concerned, both their total number and the frequency of use per 1,000 words is larger in the results sections of the Polish sub-corpus than in the Anglo-American one, and the difference is statistically significant. Although in the discussion sections the number of hedges is bigger in the Anglo-American sub-corpus, the frequency of their use is larger in the Polish sub-corpus. However, the difference is not statistically significant. Finally, in the conclusion sections the number and the frequency of hedges use is much larger in the Polish sub-corpus and the difference is statistically significant.

Table 2. The number and the frequency of hedges and boosters use in particular sections of the articles, and chi-square test results.

Markers	**Corpus**	**Article section**		
		Introduction	**Method**	**Results**
Boosters	Ang.-Am.	328 (5.5)	194 (4.6)	239 (4.9)
	Polish	257 (4.8)	101 (4.3)	215 (5.9)
	χ^2 (P)	2.94 (0.086)	0.36 (0.549)	4.45 (0.035*)
Hedges	Ang.-Am.	747 (12.5)	404 (9.6)	394 (8.0)
	Polish	723 (13.6)	235 (9.9)	517 (14.2)
	χ^2 (P)	1.77 (0.184)	0.21 (0.646)	76.8 (<0.001*)

Markers	Corpus	Article section		
		Discussion	**Conclusion**	
Boosters	Ang.-Am.	294 (8.2)	20 (4.3)	
	Polish	96 (8.6)	98 (9.5)	
	χ^2 (P)	0.15 (0.703)	11.24 (<0.001*)	
Hedges	Ang.-Am.	723 (20.2)	35 (7.5)	
	Polish	250 (22.4)	246 (23.8)	
	χ^2 (P)	1.97 (0.160)	46.84 (<0.001*)	

*statistically significant difference between the Anglo-American and the Polish sub-corpora.

Many differences between the sub-corpora can also be noticed when we look at the types of the metadiscourse markers used. In the tables below we can see what types of hedges and boosters appeared in the two sub-corpora in particular article sections in the highest number, i.e. on the first place, on the second and on the third place. The numbers of their occurrences are given next to the expressions. Those which were generally the most numerous in the whole sub-corpus, notwithstanding the section, are in bold type, and those which were on the second place are underlined.

Table 3. The types of boosters appearing in the largest numbers in the Polish sub-corpus.

	Introduction	Method	Results	Discussion	Conclusion
1.	particularly 23	at least 8	**the fact that** 27	must, **the fact that** 9	**the fact that** 12
2.	**the fact that** 22	establish 7	show 20	clearly 7	will 11
3.	indeed 18	show 6	particularly, will 16	indeed, show, will 5	must 7

Table 4. The types of boosters appearing in the largest numbers in the Anglo-American sub-corpus.

	Introduction	Method	Results	Discussion	Conclusion
1.	**will** 49	**will** 59	**will** 23	**will** 26	clearly 5
2.	determine 30	at least 22	given that, at least 18	evidence 22	at least 3
3.	evidence 23	must 21	show 17	determine 17	necessarily, quite 2

As far as boosters are concerned, Polish writers used *the fact that* most often in almost all article sections, except for the method section, and the second most frequently appearing word was the adverb *particularly* (see Table 3). Anglo-American writers used *will* most frequently in almost all article sections, except for the conclusion, and the second most often used item was the lexical verb *determine* (see Table 4).

Table 5. The types of hedges occurring in the largest numbers in the Polish sub-corpus.

	Introduction	**Method**	**Results**	**Discussion**	**Conclusion**
1.	**may** 169	**may** 22	**may** 65	**may** 52	should 38
2.	often 37	would 18	might 36	might 19	**may** 37
3.	should 32	could 16	rather 28	should 15	might 30

Table 6. The types of hedges occurring in the largest numbers in the Anglo-American sub-corpus.

	Introduction	**Method**	**Results**	**Discussion**	**Conclusion**
1.	**may** 138	would 60	would 42	**may** 115	would 4
2.	rather 55	could 32	**may** 34	would 70	could, **may**, suggest 3
3.	often 45	**may,** possible 25	rather 27	could 65	possible, quite, seem 2

With regard to hedges, there is no difference in the type of most frequently used word between the sub-corpora (see Table 5 and 6). The modal verb *may* was the most often occurring hedge in both the Polish and the Anglo-American sub-corpus. However, while in the articles written by Polish writers modal verbs mainly performed the function of hedges, in those written by Anglo-American writers more variety was noticed. There were also epistemic adjectives and adverbs and epistemic lexical verbs used more frequently.

4. Discussion

The results of the study are quite unexpected. The quantitative analyses of the texts mainly show that, overall, Polish writers used more hedges than Anglo-American ones, which is contrary to what was found in the research by Warchał (2015). However, as it was mentioned in section 2. of this paper, Warchał (2015) compared the use of epistemic modality markers by Polish writers writing in Polish with Anglo-American writers writing in English. In this study, all articles were written in English, which may explain the differences in the outcomes. The larger number of hedges in the Polish sub-corpus can be the result of cultural differences in writing. It may be the consequence of following the accepted ways of expression in Polish writing, exhibiting academic modesty, but not as predicted by Donesch-Jeżo (2011) by avoiding the use of boosters, but by using larger amount of hedges. This makes their writing more tentative, less assertive and indirect as compared with Anglo-American style of writing (Duszak 1994).

Another outcome of the present study is that higher concentration of both hedges and boosters was found in the results and the conclusion sections of the

Polish sub-corpus. Warchał (2015) also showed in her research that more epistemic modality markers can be located in the final sections of articles, which seems logical as these are discussions and the conclusion sections where writers usually try to introduce their new claims and guide readers through the arguments based on results of their studies. Larger concentration of hedges and boosters in the results sections of the Polish sub-corpus may be explained with the fact that Polish writers tend to merge the discussion with the results sections, which was the case in 40% of articles from the Polish sub-corpus, and begin to shape their argument already in the results sections. Also, higher frequency of metadiscourse markers in the conclusion sections of the Polish sub-corpus may be due to the fact that this section was present in all but one article there, while in 40% of the Anglo-American sub-corpus it was not distinguished. It was merged with the discussion section or missing. Thus, because writers in applied linguistics do not always follow strictly the conventional IMRD(C) structure of research article, the content of particular sections of articles is often shifted both in writing by Polish writers (from the discussion to the results sections) and by the Anglo-American writers (from the conclusion to the discussion sections), and so is the concentration of hedges and boosters in particular article sections (see also: Hryniuk 2017).

The differences in the types of hedges and especially boosters used by Polish and Anglo-American writers were also noticed in the qualitative analyses in the present study. The most frequently appearing lexical items performing the function of boosters in the Polish sub-corpus were *the fact that* and *particularly*. This may be the result of transfer from the mother tongue, as the same words and phrases are also very frequently used in Polish. Moreover, they do not seem to express as much confidence as the words *will* and *determine*, used most frequently by the Anglo-American writers in the present study. Employing the former ones in combination with hedges in the co-text allows Polish writers to preserve the overall impression of being more tentative and modest, and to follow their culturally shaped ways of expression in this way, even though in Anglo-American culture it may be regarded as a sign of weakness. Finally, as far as hedges are concerned, Anglo-American writers, by definition more linguistically skilled, tended to use more variety of hedging devices than Polish writers. It seems that this characteristic feature of native-speaker competence should be more often focused on in EFL writing instruction (see also: Hryniuk 2015).

A few limitations of the study must be also acknowledged. First of all, the study results cannot be generalized to all research writing by Anglo-American and Polish writers, as the number of articles analyzed was not very large. Also, it must be admitted that some of the abovementioned explanations may be regarded as speculations. Further research would benefit from explorations of a larger corpora and more qualitative analyses of how writers use hedges and boosters in context, as well as from interviews with writers about the reasons behind the use of specific rhetorical devices and about their perceptions.

5. Conclusions

Summing up, it must be stated that the results of the present study do not completely support previous research. They indicate that the differences between the use of metadiscourse markers by writers representing different cultures exist, but more research is needed in this area in order to arrive at clearer explanations of the differences and accurate interpretations of research results. We need to acknowledge that native culture conventions are not the only factors influencing written communication. The use of hedges and boosters may also be impacted by individual factors, such as self-confidence and experience of writers. Their use depends on context as well. Moreover, it is often regarded as unreflective and automatic. Thus, more research findings would contribute both to the development of knowledge in this area and to the improvement in EFL writing instruction.

It seems that improvements in EFL writing instruction based on more research in this area are very much needed. Writers representing other than Anglo-American cultures, trying to publish in anglophone journals, would benefit from the instruction focused on specific metadiscourse markers, not only to increase or decrease their amount in their writing, but also to learn how to use them in context in order to achieve the desired rhetorical effects. Explorations of corpora, as it was done in the present study, by applying corpus linguistics tools and DDL on academic writing courses, would lead to raising EFL writers awareness of the effective rhetorical strategies which they can use depending on the audience. It would be conductive to developing expertise in writing by EFL scholars as well. Finally, publication gatekeepers – editors and reviewers – would benefit from more research in this area, and the recognition that the same texts can be perceived differently by culturally diverse audiences. But first and foremost, what we all need is better understanding of how complex combinations of cultural and individual factors influence writing.

References

Abdollahzadeh, Esmaeel. 2011. Poring over the Findings. Interpersonal Authorial Engagement in Applied Linguistics Papers. *Journal of Pragmatics* 43. 288-297.

American Psychological Association. 2010. *Publication Manual of the American Psychological Association* (6th ed.). Washington DC: Author.

Ädel, Annelie. 2006. *Metadiscourse in L1 and L2 English.* Amsterdam: John Benjamins.

Bowker, Lynne and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Crompton, Peter. 1997. Hedging in Academic Writing: Some Theoretical Problems. *English for Specific Purposes* 16(4). 271-287.

Cumming, Alister. 1989. Writing Expertise and Second-Language Proficiency. *Language Learning* 39(1). 81-141.

Donesch-Jeżo, Ewa. 2011. Comparison of Generic Organization of the Research Paper in English and Polish: Cross-Cultural Variation and Pedagogical Implications. *Journalism and Mass Communication* 1(3). 185-200.

Duff, Patricia A. 2007. Problematising Academic Discourse Socialisation. In: Helen Marriott, Tim Moore and Robyn Spence-Brown (eds.), *Learning Discourses and the Discourses of Learning*, 1-18. Melbourne: Monash University Press.

Duff, Patricia A. 2010. Language Socialization into Academic Discourse Communities. *Annual Review of Applied Linguistics* 30. 169-192.

Duszak, Anna. 1994. Academic Discourse and Intellectual Styles. *Journal of Pragmatics* 21. 291-313.

Flowerdew, John. 2013. English for Research Publication Purposes. In: Brian Paltridge and Sue Starfield (eds.), *The Handbook of English for Specific Purposes*, 301-321. Chichester: Wiley-Blackwell.

Flowerdew, Lynne. 2004. The Argument for Using English Specialized Corpora to Understand Academic and Professional Language. In: Ulla Connor and Thomas Upton (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*, 11-33. Amsterdam: John Benjamins.

Golebiowski, Zofia. 1998. Rhetorical Approaches to Scientific Writing: An English-Polish Contrastive Study. *Text* 18(1). 87-102.

Hryniuk, Katarzyna. 2015. Metalinguistic Expressions in Research Articles Written by Polish and English Native-Speakers: A Corpus-Based Study. In: Anna Turula and Maria Chojnacka (eds.), *CALL for Bridges in School and Academia*, 33-47. Berno: Peter Lang.

Hryniuk, Katarzyna. 2016. The Use of Citations in Research Articles Written by Polish and English Native-Speaker Writers. In: Halina Chodkiewicz, Piotr Steinbrich and Małgorzata Krzemińska-Adamek (eds.), *Working with Text and Around Text in Foreign Language Environments*, 143-157. Heidelberg: Springer.

Hryniuk, Katarzyna. 2017. Linguistics Research Articles Written in English: Comparing Native English Speakers and Polish Writers. *International Journal of Applied Linguistics* 27(1). 3-23.

Hu, Guangwei and Feng Cao. 2011. Hedging and Boosting in Abstracts of Applied Linguistics Articles: A Comparative Study of English- and Chinese-Medium Journals. *Journal of Pragmatics* 43. 2795-2809.

Hyland, Ken. 1996. Writing Without Conviction? Hedging in Science Research Articles. *Applied Linguistics* 17(4). 433-454.

Hyland, Ken. 1998a. Persuasion and Context: The Pragmatics of Academic Discourse. *Journal of Pragmatics* 30. 437-455.

Hyland, Ken. 1998b. Boosting, Hedging and the Negotiation of Academic Knowledge. *Text* 18(3). 349-382.

Hyland, Ken. 1998c. *Hedging in Scientific Research Articles.* Amsterdam: John Benjamins.

Hyland, Ken. 1999. Talking to Students: Metadiscourse in Introductory Coursebooks. *English for Specific Purposes* 18(1). 3-26.

Hyland, Ken. 2004a. *Disciplinary Discourses: Social Interactions in Academic Writing*. Ann Arbor, Mich.: University of Michigan Press.

Hyland, Ken. 2004b. Disciplinary Interactions: Metadiscourse in L2 Postgraduate Writing. *Journal of Second Language Writing* 13(2). 133-151.

Hyland, Ken. 2005a. *Metadiscourse: Exploring Interaction in Writing.* London: Continuum.

Hyland, Ken. 2005b. Stance and Engagement: A Model of Interaction in Academic Discourse. *Discourse Studies* 7(2). 173-192.

Hyland, Ken. 2009. English for Professional Academic Purposes: Writing for Scholarly Publication. In Diane Belcher (ed.), *English for Specific Purposes in Theory and Practice*, 83-105. Ann Arbor: The University of Michigan Press.

Hyland, Ken. 2010. Metadiscourse: Mapping Interactions in Academic Writing. *Nordic Journal of English Studies* 9(2). 125-143.

Hyland, Ken and John Milton. 1997. Qualification and Certainty in L1 and L2 Students' Writing. *Journal of Second Language Writing* 6(2). 183-205.

Hyland, Ken and Polly Tse. 2004. Metadiscourse in Academic Writing: A Reappraisal. *Applied Linguistics* 25(2). 156-177.

Johns, Tim. 1991. Should You Be Persuaded: Two Examples of Data-Driven Learning. *Classroom Concordancing. English Language Research Journal* 4. 1-16.

Krzeszowski, Tomasz P. 1990. *Contrasting Languages: The Scope of Contrastive Linguistics.* Berlin: Mouton de Gruyter.

Lyons, John. 1977. *Semantics.* Cambridge: Cambridge University Press.

Mu, Congjun, Zhang, Lawrence Jun, Ehrich, John and Huaqing Hong. 2015. The Use of Metadiscourse for Knowledge Construction in Chinese and English Research Articles. *Journal of English for Academic Purposes* 20. 135-148.

Scott, Mike. 2012. *WordSmith Tools* (version 6.0). Liverpool: Lexical Analysis Software.

Varttala, Teppo. 1999. Remarks on the Communicative Functions of Hedging in Popular Scientific and Specialist Research Articles on Medicine. *English for Specific Purposes* 18(2). 177-200.

Warchał, Krystyna. 2015. *Certainty and Doubt in Academic Discourse: Epistemic Modality Markers in English and Polish Linguistics Articles.* Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Weigle, Sara C. 2005. Second Language Writing Expertise. In: Keith Johnson (ed.), *Expertise in second language learning and teaching,* 128-149. New York: Palgrave Macmillan.

NOMINALIZATION IN APPLIED LINGUISTICS AND MEDICINE: THE CASE OF TEXTBOOK INTRODUCTIONS AND BOOK REVIEWS

ALIREZA JALILIFAR
Shahid Chamran University of Ahvaz, Iran
ar.jalilifar@gmail.com

SEYEDEH ELHAM ELHAMBAKHSH
Shahid Chamran University of Ahvaz, Iran
e.elhambakhsh@gmail.com

PETER R. WHITE
University of New South Wales, Australia
prr.white@unsw.edu.au

Abstract
Drawing on Systemic Functional Linguistics, this study explored variational use of nominalization in 600 textbook introductions and 200 book reviews in applied linguistics and medicine. The nominalized expressions were identified in the texts, the frequencies of the nominalization types were counted, and eventually a chi-square test was administered. Analysis of nominalization patterns across the different informational/promotional moves revealed divergent patterns in the two disciplines but insignificant differences across the genres in focus. The density of nominalizations was acknowledged in the applied linguistics introductions and book reviews. However, functional variations in the use of nominalizations were found only in the introductions. As for the proportion of nominalization to grammatical metaphor, results demonstrated a lower tendency towards nominalizing scientific information in the medicine corpus. Further research is needed to see how nominalization is exploited in other genres and other disciplines.

Keywords: book reviews, introduction, nominalization, systemic functional linguistics

1. Introduction

Over the last three decades, English for Specific Purposes (ESP) researchers have employed genre analysis extensively to examine a variety of academic genres including abstracts, presentations, lectures, theses, dissertations and textbooks and their related discourses (e.g., Bhatia, 1997, 2004; Bunton, 2002; Dudley-Evans, 1986; Hopkins and Dudley-Evans, 1988; Hyland, 2004; Hyon, 1996; Martin, Christie, and Rothery, 1987; Nwogu, 1997; Paltridge, 1997; Samraj, 2005; Swales, 2004; Thompson, 1994). While several studies have focused on

disciplinary writing (e.g., Brett, 1994; Holmes, 1997; Kuteeva, 2013; McCloskey, 1986; Peck MacDonald, 1990, 1992), other studies have explored particular sections of the research article (RA) or its overall structure cross-linguistically (see e.g., Coffin, Curry, Goodman, Hewings, Lillis and Swann, 2003; Hyland, 2009; Marefat and Mohammadzadeh, 2013; Martin, 2003). Many of the above academic genres begin by an introduction section wherein authors lay down their points of argument.

Introductory genres, which are conventionally used to introduce academic research articles and textbooks and their various manifestations, distinctly named as *introduction, foreword, preface, acknowledgement* and, occasionally, *trajectory, preamble* or *prologue*, have received prime attention in recent years (e.g., Abdollahzadeh and Salarvand, 2013; Bhatia, 1997; Jalilifar and Golkar Moosavi, 2016; Kuhi, 2008; Sorayyaei Azar, 2012; Zepetnek, 2010). Research into academic introduction sections of textbooks has not been prolific. Bhatia (1997) presented *book introductions* and *book prefaces* as different categories of *academic introductions*, arguing that the former fulfills an informative function while the latter fulfills both a promotional and an informative one. He finally acknowledged that one other purpose of all *academic introductions*, as an example of an interested genre, seems to be promoting the work, which even sometimes takes precedence over the original purpose (i.e., introducing the work).

Book reviews (as another focus of this study), on the contrary, are considered as a sub-genre (Bhatia, 1993: 21) or a member of the family of *review genres* (Giannoni, 2009). In terms of communicative purposes, review genres vary along a continuum extending from the most promotional (arguably blurbs (Bhatia, 2004; Cacchiani, 2007; Gesuato, 2004)) to the most critical (e.g., expert reviews). In book reviews, the purpose switches from endorsement to criticism, as the reviewer is a (supposedly neutral) third party acting as a gatekeeper on behalf of the academic community (Giannoni, 2009: 19). Book reviews, as examples of a disinterested genre, are defined as *promotional* (Bhatia, 1997, 2002; Lorés-Sanz, 2012) and *evaluative* (e.g., Gea Valor, 2000-2001; Groom, 2009; Hyland, 2004; Lorés-Sanz, 2012; Römer, 2005, 2008; Shaw, 2004, 2009; Tse and Hyland, 2009; Vassileva, 2010) and are meant to act as *critical windows* which open to the novelties and advances of a given discipline, and, in that sense, they may well contribute to the construction and development of disciplinary knowledge.

Valuable works on book reviews and introductory genres have brought an insight into their macro-structures and linguistic analyses. These studies, however, vary in their foci from disciplinary and cross-disciplinary variations to cross-linguistic differences of these genres and their micro-structure features. Given the variations of the introductory sections of textbooks and the importance attached to them as well as their seemingly close relationship with book reviews in presentation of an overview of the textbooks, the absence of more comparative research on micro-linguistic features in this regard is especially prominent. The fact that book reviews and most of the introductory sections of academic textbooks share at least one communicative purpose, that is to introduce the book in focus,

seemingly causes a considerable overlap, but some of them are sometimes appropriated by publishers to promote their product (Jalilifar and Golkar Moosavi, 2016).

It has been demonstrated that whereas textbook introductory sections seem to chiefly reinforce the positive aspects of the book, book reviews attend to both merits and demerits, looking at the book in focus with a critical mind from the outside (Alcaraz Ariza, 2010; Diani, 2007; Hyland, 2004; Hyland and Diani, 2009; Lindholm-Romantschuk, 1998; Motta-Roth, 1998; Salager-Meyer and Alcaraz Ariza, 2004). Given these functional differences, our assumption is that these aspects might partly transpire in the nominalizations used. Authors may experience confusion if they are not fully aware of genre tendencies and linguistic characteristics.

The inspiration for a comparative study of textbook introductions and book reviews comes from the need to determine how far the nominalization patterns are distinct in two disparate disciplines of applied linguistics and medicine, representing soft and hard sciences. There is, therefore, a pedagogical rationale for extending the analysis of the academic texts into a comparative study of nominalization use across two disciplines. The study hypothesizes that differences in nominalization use might become even more explicit when disciplinary tendencies also intervene, especially when the disciplines appear to be far from one another.

2. Theoretical framework of the study

This study is grounded in Halliday's (1994) systemic functional linguistics (SFL). SFL interprets language as interrelated sets of options for making meanings and seeks to provide a clear relationship between functions and grammatical systems (Halliday, 1994). Systemists focus on "how the grammar of a language serves as a resource for making and exchanging meanings" (Lock, 1996: 3). That is, SFL is concerned with the grammatical patterns and lexical items used in texts, as well as choices of those items. The grammatical domain of language is considered an important area of inquiry, an offshoot of which is studied under grammatical metaphor (Halliday, 1994). Grammatical metaphor is defined as "a substitution of one grammatical class, or one grammatical structure by another" (Halliday and Martin, 2005: 87). Specialized technical discourse cannot be created without deploying grammatical metaphor (Martin, 1990). In the area of grammatical metaphor, for any given semantic configuration, there will be some realization in the lexicogrammar—some wording—that can be considered congruent or unmarked; there may also be various others that are in some respect incongruent, "transferred" or "metaphorical" (Halliday,1994: 342).

In SFL, nominalization is connected to grammatical metaphor used to indicate a process or an attribute. Halliday and Matthiessen (1999) categorize grammatical metaphor into 13 types of which four types are classified as nominalizations, in

terms of semantic shifts involved in transforming the congruent into the incongruent form (i.e., adjective > noun, verb > noun, conjunction > noun, and preposition (al phrase) > noun).

As an aspect of complexity in written language (Halliday and Matthiessen, 2004; Heyvaert, 2003), nominalization is used for embedding as much information into a few words as possible. A nominalized structure like *I have found a lot of appreciation and greater acceptance abroad*, for instance, is thus viewed as the metaphorical counterpart of the clause *The scholars abroad have greatly appreciated and accepted the book*. In order to fully grasp the meaning of nominalization as an additional dimension of meaning, the identification and the analysis of both the metaphorical and the congruent realizations are essential (Halliday, 1994; Heyvaert, 2003).

The use of nominalization in scientific discourse has been the subject of a wide array of studies in recent years, for example, the historical origins of nominalization in scientific discourse (Banks, 2005), the realization of grammatical metaphor in modern prose fiction (Farahani and Hadidi, 2008), the contribution of verb-based nominalization to cohesion in 892 pages of history texts (Susinskiene, 2009), nominalization in the writing of undergraduate students (Baratta, 2010), and the role of nominalization in the English medical papers produced by native English speakers and Chinese writers (Wenyan, 2012). Other studies on nominalization in scientific discourse (e.g., Banks, 2003; Baratta, 2010; Halliday and Matthiessen, 1999, 2004; Ho, 2010; Jalilifar, Alipoor and Parsa, 2014; Martin, 1993; Sušinskienė, 2009, 2010; Wenyan, 2012) have also stressed the crucial role played by nominalization in the skillful orchestration of academic discourse. In fact, considering the frequency and usage of different types of nominalization, research on nominalization indicates variation in abstracts and in research articles (Holtz, 2009), in British newspaper editorials (Sušinskienė, 2010), in essay writings of undergraduate students (Baratta, 2010), in request e-mails (Ho, 2010), in business letters (Văn, 2011), in the discussion sections of medical research articles (Wenyan, 2012). Yet, we doubt how nominalization is realized in textbooks introductions and book reviews across disciplines. In other words, it is not clear how nominalization use is related to typological similarities and differences between medicine and applied linguistics as examples of hard and soft applied sciences. Nevertheless, the realization between discipline specificity, text scientificity, and nominalization has yet to be adequately examined. Furthermore, an understanding of the functional role and textual consequences of grammatical metaphor is essential for a full understanding of the meaning of any text.

Notwithstanding the aforesaid studies on nominalization from various angles, further research is required to find out disciplinary and genre specificity in the use of nominalization. Thus, this study seeks to investigate the variational use of nominalization in applied linguistics and medicine textbook introductions, prefaces, forewords and in book reviews. The analysis of these texts involves four steps: The first step of analysis identifies the frequency of nominalized

expressions and grammatical metaphors in each text. In the second step, different types of semantic shifts in the process of nominalization are determined. In the third step, the density of nominalization is examined. In th fourth step, the proportion of nominalization to grammatical metaphor in each genre is calculateded and the grammatical patterns of nominalization deployment are also illustrated in detail. Accordingly, the following questions stand out:

1. What are the grammatical functions of nominal expressions and their relative distributions in the sample English textbooks introduction sections and book reviews in applied linguistics and medicine, and how do the functions and their relative frequency of deployment compare?
2. What types of semantic shifts (i.e. quality, process, circumstance and relator) in the process of nominalization are frequently used in English applied linguistics and medicine textbooks introduction sections and book reviews, and how do the types and their relative frequencies compare?
3. Is there any significant difference in the density of nominalization use between English textbooks introduction sections and book reviews in applied linguistics and medicine?
4. What are the grammatical patterns of nominalization deployment and their relative distributions in the sample English textbooks introduction sections and book reviews in applied linguistics and medicine, and how do the patterns and their relative frequency of deployment compare?

3. Method

3.1. Selection of the disciplines

Following the experience of scientometricians and external experts, Glanzel and Schubert (2003) propose a two-level hierarchical classification scheme for three main discipline areas: *Sciences*, *Social Sciences*, and *Humanities*. Their two-level scheme includes 12 first-level fields and 60 second-level subfields of the Sciences, as well as three major fields and seven subfields for the Social Sciences and Humanities. Coffin, et al (2003) added one more major area—*applied versus pure disciplines*--and provided some representative examples for these four main discipline areas.

Acknowledging the complexity of demarcating disciplines, the present analysis rested on the most convenient way of grouping disciplines into four main areas: *Sciences*, *Social Sciences*, *Humanities/Arts*, and *Applied Disciplines* (Coffin, et al., 2003; Glanzel and Schubert, 2003). Figure 1 demonstrates a revision of Hyland's (2006) continuum, adding the hard applied seciences, which include disciplines such as medicine.

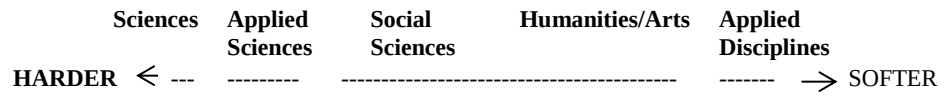

Figure 1. Continuum of disciplines (*Revised*)

Selection of the disciplines was motivated by the need to build a corpus representative of textbook introductions and book reviews in applied linguistics (closer to the soft end of the continuum) and medicine (closer to the hard end of the continuum). The motivation for selecting these disciplines as middle areas of science was to investigate representatives of two applied disciplines related to two major branches of science which can possess both similarities (due to the softer nature of applied disciplines) and differences (since each has a different tendency towards soft or hard sciences).

3.2. Selection of the textbook introductions

Three hundred English textbook introductions (100 samples from each variation of introduction, i.e., introduction, foreword, and preface) in each discipline were selected to allow comparisons across hard and soft applied sciences (a total of 600 samples). Textbook selection was to meet the following criteria:

i. The choice of textbooks was motivated by the need to control such variables as writer experience and expertise. The major criterion in selection was to include textbooks which were widely used in the syllabuses of applied linguistics and medicine courses in Iranian universities. Hence, a number of informants in each discipline were asked to recommend textbooks available in hard copies or those retrievable from downloadable databases that they considered as essential in their own field at two levels of BS/BA and PhD.

ii. To ensure the validity of analysis, textbooks written in English by English speaking authors were preferred.

iii. The selected corpus represented a span of 10 years (i.e., textbooks published in 2006-2016). The assumption was that a genre might change and evolve in response to changes in the communicative goals, as well as to "particular rhetorical needs" of the discourse community that regularly uses it (Abdollahzadeh, 2013: 424).

3.3. Selection of the book reviews

With regard to the selection of book reviews, initially, a list of applied linguistics and medicine journals publishing English language papers in the two disciplines was compiled. The major criteria guiding the identification of journals, from which book reviews in the corpus were taken, were reputation, accessibility, representativeness and dominance of the journals based on their impact factors, as

well as the period of publication of the book reviews. The criteria were shared with two independent applied linguistics experts and two independent medicine experts following panel discussions. The preliminary corpus for the pilot phase was drawn from the consented journals. The final corpus, consisting of 200 book reviews (100 from each discipline) was selected on the basis of stratified sampling procedure (see Table 1.). Similar to the introductions, selection of the reviews was restricted to a period of 10 years (2006-2016). Moreover, to qualify for the final corpus, all the book reviews had to be approximately 1000 words on average, to control length.

Table 1. Selected Journals in Medicine and Applied Linguistics

Applied Linguistics Journals	No. of BRs	Medicine Journals	No. of BRs
English for Academic Purposes	20	British Medical Journal	20
Second Language Writing	20	Annals of Otology, Rhinology and Laryngology	20
Language Teaching	20	Annals of Medicine and Surgery	20
Writing and Pedagogy	20	Annals of Emergency Medication	20
Studies in Second Language Acquisition	20	Asian Pacific Journal of Tropical Medicine	20
	167533		130335

3.4. Procedure

Prior to analyzing the data, the unit of analysis was assigned to be the clause complex. Clause complexes show "how the flow of events is construed in the development of text at the level of semantics" (Halliday and Matthiessen, 2004: 63). Eggins (2004) defines clause complex (i.e., parataxis and hypotaxis) as a "grammatical and semantic unit formed when two or more clauses are linked together in certain systematic and meaningful way" (p. 255). The clauses were coded in each text and the texts were coded in each genre, for instance, Bp. Med. #029 means text 029 which is a book preface in medicine. BI., BF., BR., and AL. stand for book introduction, book foreword, book review, and applied linguistics respectively.

One tricky and controversial category of nominals is gerunds. This study opts for consideration of gerunds denoting actions rather than situations as examples of verb > noun nominalization. Following Simon-Vandenbergen, Taverniers and Ravelli (2003: 82-83), this study assumed that as long as the gerund form can be

preceded by a premodifier, such as that of a possessive pronoun, it can be categorized as a nominal. However, in case the gerund only denotes modality, tense or process rather than action, it cannot be counted as nominalization.

In consideration for consistency in the analysis, those nouns which served as technical words in each discipline (e.g., *digestion* in medicine, and *competence* in applied linguistics) were excluded. As a further stage in the analysis, nominalization instances were tagged through querying for suffixes: nouns ending in the suffixes *-ity* and *-ness* were tagged as Type 1 (deriving from adjectives, originally realizing properties); nouns ending in the suffixes *-age* , *-al* , *-(e)ry*, *-sion / -tion*, *-ment*, *-sis*, *-ure*, -ing, and *-th* were tagged as Type 2 (deriving from verbs, originally realizing processes); and nouns not ending in suffixes were tagged through consulting dictionaries to find the related derivation from adjectives, verbs, prepositions, and conjunctions. Prepositional phrases metaphorically realized as nouns were tagged as type 3. Prepositional phrases often concern information about time and place; in other words, they deal with the circumstances of the events or states described in the text, hence called "circumstantial adjuncts" (Bloor and Bloor, 2004: 53). However, when they change into nouns metaphorically, they become the classifier of nominal groups. Consider the following nominalization instances derived from the corpus:

1. …*fourteenth-century* recognition of the connection between... (BR. Med. #014)
2. Teachers' supervision and assessment of *day-to-day performances* of students… *(BI. AL. #89)*
3. The fourth type, nominalization of conjunction, which is congruently presented by a conjunction, is metaphorically realized by a noun functioning as a participant in the clause. The only pattern manifesting this type of nominalization was as follows:
4. This *Handbook* is aimed at a diverse range of professionals and *for this reason,… (BI. AL. #051)*
5. …*For this reason,* color printing has been used to make… (BR. Med. #037)

In the above examples, the entity *reason* is transferred from the relator *because.* In 3, for example, the element *reason* is the metaphorical realization of the clause *because this Handbook is aimed at….*

Having identified the frequency, type, and density of nominalizations in the texts as well as the proportion of nominalizations to grammatical metaphors, in the next stage of analysis, we extracted the patterns of nominalizations. The basis for extracting these patterns was Halliday's (2004) suggestion that lexical expansion of nominal groups is attributed to pre/post-modification: a class of *things* is specified by nouns; and categorization within the class is typically expressed by one or more functional words organized around it. These functional elements – Deictic, Numerative, Epithet, Classifier, and Qualifier – serve to specify *things* within "different systems of the system network of the nominal

group" (Halliday, 2004: 312). The classes of the words which typically realize these functions are illustrated in Figure 2:

Deictic	Deictic 2	Numerative	Epithet	Classifier	**Thing**	Qualifier
determiner	adjective	numeral	adjective	noun/ adjective	**Noun**	Prepositional phrase/ (in)finite clause

Figure 2. Experiential functions and word classes

After about one month interval, the data were re-examined, and discrepancies on the method of analysis were resolved. Considering coding reliability, the data were cross-checked by a linguist to verify the accuracy of categorization of strategies. Then, to calculate the amount of inter-coder and intra-coder reliabilities, Phi correlation was employed twice. The indices obtained were 0.94 and 0.83, respectively. What follows provides quantitative and qualitative analyses of the materials.

4. Results

To address the first and second questions raised in this study, word count was run and the data were normalized afterwards in order to be consistent in our analysis because the number of clauses in the introductions and book reviews was different. The nominalized expressions were, then, counted. A glance at Table 2 reveals that nominalized expressions in applied linguistics outrun the corresponding expressions in medicine in the respective texts:

Table 2. Nominalized Expressions across Disciplines and Genres

	Applied Linguistics				**Medicine**			
	Tb. Intros.	Tb. Pres.	Tb. Fors.	Br. Arts.	Tb. Inros.	Tb. Pres.	Tb. Fors.	Br. Arts.
	F (%)	F (%)	F (%)	F (%)	F (%)	F (%)	F (%)	F (%)
Nominalized expressions	16008 (8.07)	10431 (8.37)	9367 (9.13)	15796 (9.42)	12783 (6.26)	8971 (8.61)	6891 (6.89)	11321 (8.68)
Grammatical metaphors	17941	12765	11651	20981	14368	11509	8593	18524
Clauses	18310	11769	10976	16867	18735	9873	8441	12958
Total words	198314	124611	102528	167533	203977	104122	99983	130335

Note. Tb. Intros: Textbook Introductions; Tb. Pres.: Textbook Prefaces; Tb. Fors.: Textbook Forewords; Br. Arts.: Book Review Articles.

Table 2 demonstrates the total number of the nominalized expressions in the analyzed texts. These results reveal the proportion of nominalization instances to grammatical metaphors (i.e. 51602 nominals vis-à-vis 63338 grammatical metaphors in applied linguistics, and 39966 nominals vis-à-vis 52994 grammatical metaphors in medicine). The dominance of nominalization in the categories of grammatical metaphor evinces the valuable role that this strategy plays in formulating scientific discourse. In order to compare the use of nominalization (i.e., adjective to noun (=Type 1), verb to noun (=Type 2), preposition to noun (=Type 3), and conjunction to noun (=Type 4), with their different types of semantic shifts, i.e. quality, process, circumstance, and relator respectively) in detail, the frequency of each nominalized phrase was counted and they were put in appropriate categories (see Table 3):

Table 3. Semantic Shifts in the Use of Nominalized Expressions across Disciplines and Genres

1.	**Tb. Intros. (Applied Linguistics)**	**Tb. Intros. (Medicine)**			
	F(%)	F(%)	df	X^2	P value
Type 1	1021(6.37)	772(6.04)	1	277.64	0.000
Type 2	14460(90.32)	11453(89.59)	1	581.89	0.000
Type 3	482(3.01)	507(3.96)	1	105.58	0.000
Type 4	45(0.28)	51(0.39)	1	7.92	0.005
2.	**Tb. Pres. (Applied Linguistics)**	**Tb. Pres. (Medicine)**			
	F(%)	F(%)	df	X^2	P value
Type 1	752(7.20)	571(6.36)	1	329.93	0.000
Type 2	9249(88.66)	8128(90.60)	1	308.35	0.000
Type 3	411(3.94)	250(2.79)	1	247.49	0.000
Type 4	19(0.18)	22(0.25)	1	3.12	0.077
3.	**Tb. Fors. (Applied Linguistics)**	**Tb. Fors. (Medicine)**			
	F(%)	F(%)	df	X^2	P value
Type 1	725(7.73)	406(5.89)	1	481.84	0.000
Type 2	8261(88.19)	6161(89.40)	1	372.15	0.000
Type 3	358(3.82)	304(4.41)	1	128.20	0.000
Type 4	23(0.25)	21(0.30)	1	7.10	0.008
4.	**Br. Arts. (Applied Linguistics)**	**Br. Arts. (Medicine)**			
	F(%)	F(%)	df	X^2	P value
Type 1	835(5.29)	756(6.68)	1	262.53	0.000
Type 2	14551(92.12)	10099(89.21)	1	857.77	0.000
Type 3	356(2.25)	408(3.60)	1	60.48	0.000
Type 4	52(0.33)	57(0.50)	1	10.13	0.001
5.	**All Textbook Introduction Genres (Applied Linguistics)**	**Br. Arts. (Applied Linguistics)**			
	F(%)	F(%)	df	X^2	P value
Type 1	2498(5.91)	835(5.29)	1	1299.68	0.000
Type 2	31970(75.69)	14551(92.12)	1	285.46	0.000

Type 3	1251(2.96)	356(2.25)	1	1412.98	0.000
Type 4	87(0.20)	52(0.33)	1	53.54	0.000
6.	**All Textbook Introduction Genres (Medicine)**	**Br. Arts. (Medicine)**			
	F(%)	F(%)	df	X^2	P value
Type 1	1749(6.10)	756(6.68)	1	143.31	0.000
Type 2	25842(90.21)	10099(89.21)	1	245.10	0.000
Type 3	1061(3.70)	408(3.60)	1	1361.76	0.000
Type 4	94(0.32)	57(0.50)	1	56.82	0.000

Note. Tb. Intros: Textbook Introductions; Tb. Pres.: Textbook Prefaces; Tb.Fors.: Textbook Forewords; Br. Arts.: Book Review Articles.

The results of chi-square analyses showed significant differences between the genres in focus in the two disciplines under study. Table 3 reveals the most and the least nominalized expressions used in the corpus. That is, verb to noun was extremely common and unmarked in the two disciplines. Adjective to noun ranked second in order of frequency in these academic texts. As shown by chi-square analysis, preposition to noun was used more frequently in applied linguistics than in medicine. Finally, though not significantly different, conjunction to noun was very scant in the focused texts for analysis and proved to be similarly employed in the two disciplines. The results marked verb to noun to be characteristic of the discourse of the two disciplines.

Table 4. Density of Nominalized Expressions in Textbook Introductions and Book Reviews

1.	**Tb. Intros. (Applied Linguistics)**	**Tb. Intros. (Medicine)**			
	F(%)	F(%)	Df	X^2	P value
Nominalized expressions	16380(87.42)	12783(68.23)	1	684.50	0.000
Number of clauses	18735	18735			
2.	**Tb. Pres.**	**Tb. Pres.**			
	F(%)	F(%)	Df	X^2	P value
Nominalized expressions	10431(88.63)	9710(82.50)	1	409.81	0.000
Number of clauses	11769	11769			
3.	**Tb. Fors.**	**Tb. Fors.**			
	F(%)	F(%)	Df	X^2	P value
Nominalized expressions	9367(85.34)	8961(81.63)	1	259.98	0.000
Number of clauses	10976	10976			
4.	**Br. Arts.**	**Br. Arts.**			
	F(%)	F(%)	Df	X^2	P value

Nominalized expressions	15796(93.65)	14736(87.36)	1	465.84	0.000
Number of clauses	16867	16867			

Note. Tb. Intros: Textbook Introductions; Tb. Pres.: Textbook Prefaces; Tb. Fors.: Textbook Forewords; Br. Arts.: Book Review Articles.

As for the third research question, Table 4 demonstrates the density of the nominalized expressions in the clauses in the four datasets. The chi-square revealed a statistically significant difference with regard to the density of the nominalized expressions in the focused genres. That is, the amount of the chi-square was higher than the critical value (3.84) at the level of $p < 0.05$. The density of the nominalized expressions in applied linguistics exceeded the corresponding expressions in medicine, showing that the writers in applied linguistics tend to condense and package a larger amount of information into single lexical items than in medicine.

Table 5. Nominalized Expressions and Grammatical Metaphors in Introductions and Book Reviews

1.	**Tb. Intros. (Applied Linguistics)**	**Tb. Intros. (Medicine)**			
	F(%)	F(%)	df	X^2	P value
Nominalized expressions	16008(89.22)	15962(88.96)	1	402.50	0.000
Grammatical metaphors	17941	17941			
2.	Tb. Pres. (Applied Linguistics)	Tb. Pres. (Medicine)			
	F(%)	F(%)	df	X^2	P value
Nominalized expressions	10431(81.71)	9950(77.94)	1	292.02	0.000
Grammatical metaphors	12765	12765			
3.	Tb. Fors. (Applied Linguistics)	Tb. Fors. (Medicine)			
	F(%)	F(%)	df	X^2	P value
Nominalized expressions	9367(80.39)	9343(80.19)	1	236.13	0.000
Grammatical metaphors	11651	11651			
4.	Br. Arts. (Applied Linguistics)	Br. Arts. (Medicine)			
	F(%)	F(%)	df	X^2	P value
Nominalized expressions	15796(75.28)	12823(61.11)	1	615.06	0.000

Grammatical metaphors	20981	20981

Note. Tb. Intros: Textbook Introductions; Tb. Pres.: Textbook Prefaces; Tb. Fors.: Textbook Forewords; Br. Arts.: Book Review Articles.

Table 5 reveals the final stage of the quantitative analyses which demonstrates the proportion of the normalized nominalized expressions to the total number of grammatical metaphors in each of the four genres under scrutiny in the two disciplines. The chi-square statistics revealed a statistically significant difference (at $p < 0.05$). The nominalized expressions in applied linguistics were more prevalent than in medicine. This shows that the presentation of information in applied linguistics is facilitated more by the use of nominalized expressions through the expansion and elaboration of nominal elements than in medicine.

With regard to the fourth research question, in all focused genres, Type 2 (i.e., conversion of verb to noun (process)) was reported to be more prevalent than other types of nominal expressions. There were different patterns in which Type 2 occurred. Table 6 summarizes the most frequent patterns with their related examples.

Table 6. Summary of Patterns and Related Examples

Pattern No.	Patterns and related examples	Frequency AL.	Frequency Med.
# 1	*Nominal + Qualifier*	9904	9277
	With careful cross-referencing and *provision of explanations and examples*, we have … (BI. AL. #004)	19.19%	23.21%
# 2	*Preposition + Nominal*	1423	1245
	In comparison, this manual is a collective effort to provide simple, practical solutions to… (BF. Med. #010)	2.76%	3.12%
#3	*a/an/the/- + nominal*	2153	1287
	…, and the reconstructive flap illustrations are well-done and reproducible for broad study and *recall*. (BR. Med. #081)	4.17%	3.22%
#4	*there/is/are/was/were + nominal*	2243	1066
	There are illustrations added in this edition wherever important points could be made more clear,…(BP. Med. #016)	4%	2.67%
# 5	*Nominal + Prepositional Phrase*	9147	7454
	…but *treatment of other contact phenomena* is less sure…(BR. AL. #085)	18%	18.65%
# 6	*Preposition + Nominal + Prepositional Phrase*	205	189
	We are pleased that Springer has taken this title under its direction and has helped to improve its quality *in preparation for international release (BP. Med. #089).*	0.39%	0.47%

# 7	*Classifier + Nominal* …provides examples specific to healthcare on how hospitals have greened their operations and facilities, ranging from healthy food procurement, to ***hospital waste***, to measuring and…(BR. Med. #014)	5043 10%	6359 15.91%
# 8	*Nominal as classifier + Nominal/ Noun* Chapter 1 gives an overview of the green ***healthcare movement***, …(BR. Med. #014)	5045 9.77%	2034 5.09%
# 9	*Classifier + Classifier + Nominal* In recognition of the growing excitement and potential of ES cells as models for both the advancement of ***future clinical applications*** and, …(BP. Med. #003)	937 2%	1536 3.84%
# 10	*Numerative + Nominal* ***One concern*** is to explore the nature of temporal frames of reference…(BP. AL. #098)	1821 4%	1342 3.36%
# 11	*Nominal + Participle* The ***information contained*** herein… (BF. Med. #011)	1952 4%	2374 5.94%
# 12	*Nominal + Relative clause* … A more reasonable ***expectation that*** interested readers will simply select the chapters that … (BI. AL. #001)	3575 7%	2242 5.61%
# 13	*Nominal + Gerund* However, in view of rapid ***changes occurring*** in medical science,… (BF. Med. #011)	971 2%	552 1.38%
# 14	*Nominal + Adjunct* This is addressed in greater ***depth in chapter 11…*** (BI. Med. #085)	729 1%	184 0.46%
# 15	*Nominal + Infinitive* We are most grateful to him for his ***permission to do*** this. (BP. Med. #036)	973 2%	367 0.92%
# 16	*Nominal + Adjective/Adverb as postmodifier* …about the accuracy of the scientific information communicated by many….. (BR. Med.#009) … to base their practice individually… (BR. Med. #008)	842 2%	362 0.91%
# 17	*Adverb as classifier + Nominal* No attempt was made to do experimental tests under ***carefully controlled plans***… (BF. AL. #001)	631 1%	345 0.86%
#18	*This/that/these/those+Nominalization* This reference is a not-so-quick one… (BR. Med. #070)	4011 7.77%	1751 4.38%
Σ		51602	39966

Note. Med: Medicine; AL.: Applied linguistics.

The most dominant pattern was nominalization + qualifier (#1). In the examples below, the congruent forms of Example 5 are the part of the section that remains addresses, and the fact that hypertext and modern media can influence comprehensibility…, and the congruent forms of Example 6 are the students' efforts have come out.., and They considerably experienced teaching biochemistry… The words remainder, influence, outcome, and experience

function as things in these nominal expressions, and the words section, hypertext, and modern media, efforts, and teaching which serve as qualifier in metaphoric forms, are, in fact, the head of material processes in their congruent realizations. Therefore, they belong to the ideational grammatical metaphor because their grammatical functions are transferred from Head to qualifier:

5. *The remainder of* the section addresses issues like *the influence of* hypertext and modern media on comprehensibilty and translating professional documents (BR. AL. #004)
6. The textbook of Medical Biochemistry for the medical students is *the outcome of* the joint efforts of a medical and a nonmedical biochemist, who possess considerable *experience in* teaching biochemistry to undergraduate and postgraduate medical students of Indian universities. (BP. Med. #33)

In some cases, from the grammatical point of view, nominalizing a process allows the addition of both modifiers and qualifiers packing the flow of information into fewer words. Note the following examples:

7. The last decade has witnessed *an explosive growth of molecular data* ... (BP. Med. #018)
8. …that should be taken into account to give the reader *a scientific understanding of the writing process* relative to planning … (BR. AL. #009)

(The typical form of Example 7 is molecular data grows explosively. Example 8 is represented congruently as writing process is understood scientifically.)

The rest of the patterns excluded from the analyses indicated that the dense clauses are usually formed by nouns with multiple premodifiers and postmodifiers in both disciplines. This, in effect, creates a text that is tightly packed with information in the form of nominal phrases rather than clauses to add information (Gray and Biber, 2010).

5. Discussion

The main findings of this study with respect to introductions and book reviews in medicine and applied linguistics are discussed below.

As revealed by the results of the four research questions, the similarities and differences in nominalization deployment in the four genres is likely to illustrate different tendencies for packaging the information in academic texts which involve fluctuation over the use of this strategy in the different types of texts. Although all texts were replete with instances of nominalizations, the introduction sections of textbooks had comparatively the most frequent distributions of nominals, whereas the book review articles had the least number of nominals. Prefaces and forewords were fairly similar in their frequencies of nominals. These results confirm that grammatical metaphor is a powerful language resource that

"simultaneously builds cohesion, foregrounds meanings in static nominal groups, and backgrounds personal and subjective voice".

Furthermore, information density is intimately tied to disciplinary characteristics. In this respect, grammatical metaphor is a resource that language uses to condense information by expressing concepts in an incongruent form which is very valued in scientific registers as a way of expressing "objectification" and "abstraction" (Halliday and Martin, 2005: 33). However, unlike other studies (Halliday, 1994; Halliday and Martin, 2005; Xue-feng, 2010), the writers in both disciplines put ideas into abstract forms variably and thus, at the level of lexicogrammar, the disciplinary distinction is manifested in the degree of the nominal phrases used.

Besides the density of nominal phrases that distinguishes the two disciplines, there were a few patterns that made the applied linguistics texts distinct from the texts in medicine. For instance, adjective-derived nominalization in applied linguistics mostly occurs in the clause initial position. In the following example, the writer explains why writers are required to act uniformly in emphasizing consistency in the next clause:

9. *Consistency* is a necessary characteristic of polished, highly readable prose. (BI. AL. #076)

Another recurring pattern characterizing applied linguistics is the nominalization of adjective and qualifier or nominalization of adjective with another adjective as illustrated below:

10. … the *importance of accessibility* of curriculum to the language teacher as a tool for increasing … (BR. AL. #019)

The pattern exclusive to medicine, which establishes the cause and effect relationship between the nominal groups, is of simple construction, with one nominal group clause initially, *the importance of genes*, one nominal group clause finally, *their ability*, and one verbal group, *lies in*, pushed in between indicating the logical relation between the two phenomena. Note the following example which is congruently taken to be *because genes are able to control the formation of cell, they are important*:

11. *The importance of genes lies in their ability* to determine key personality traits, as well as… (BI. Med. #31))

A noticeable difference in the use of prepositional nominalization in applied linguistics and medicine is revealed in the next two examples. Whereas, in medicine introductory genres, the nominalization of preposition occurs with nominalization of process and qualifier, in applied linguistics, nominalization of preposition often occurs before nominalization of process as shown below:

12. As *language learning* is a *cumulative effort*, it *must* be consolidated outside ... (BR. AL. #007))
13. *The juristic basis of the classification of disease* is concerned with the legal circumstances in which death occurs. (BI. Med. #026)

Therefore, even if there arguably are core features and characteristics in academic discourse, it is important to acknowledge the fact that many variations exist when it comes to how certain disciplines struggle with the challenges of conveying information and achieving academic writing. Various disciplines in the natural sciences, technology, social sciences, and humanities have their specific, conventionalized ways of describing ideas, knowledge, methods, results and interpretations (e.g., Basturkmen, 2011; Hawes and Thomas, 2012; McGrath and Kuteeva, 2013; Parodi, 2010). This *discipline specificity*, which stresses the distinctive ways of meaning making and constructing discourse (Hyland, 2009), attempted to highlight the necessity to go beyond the generalized view of academic writing and to pin down specific characteristics of the scientific discourse in each of these disciplines.

One other major finding drawn from our analysis was the greater density of nominalization in applied linguistics than medicine. This being said, in formal written language, there are fewer clauses, as the ideational information of two or more clauses may be realized as one. Thus, the possibility of two or more cases of grammatical metaphor being combined in the same nominal group would mean that two or more clauses are being expressed as a single participant. This feature prevails in applied linguistics because the writers tend to put the focus on objects, states, and process all encoded by nouns rather than human agents and their actions which are, in turn, encoded by verbs (Jalilifar, Alipour, and Parsa, 2014). Thus, it seems reasonable to assume that information density is closely related to disciplinary characteristics. Previous studies (e.g., Galve, 1998; Halliday, 1994) have also measured lexical density by dividing the number of lexical items to the number of ranking clauses. Galve (1998) argued that when a language is more planned and more formal, lexical density is higher (over 0.40 per clause). When processes are repacked as participants, academic texts become more abstract and complex, and much of the complexity is due to the nominal group structure which allows an extended explanation to be condensed into a complex phrase, as depicted in the following example:

14. The earliest activities in the documentation and description of language have been attributed to… (BI. AL. #093)

Therefore, writers and speakers make choices from the various options that language makes available, according to the social and cultural context in which meaning is exchanged. As an interlocking set of grammatical systems, language enables its users to make different kinds of meaning for different purposes and contexts. Schleppegrell (2001) argues that register differences manifest

themselves both in choice of words or phrases and also in the way that clauses are constructed and linked. Therefore, the higher proportion of lexical density in applied linguistics in comparison to medicine reveals that the language that constructs knowledge is subject to disciplinary specificities. The choice of different lexical and grammatical options is related to the functional purposes that are foregrounded by the writers of different disciplines. Lexical density is one way of qualifying the differences in lexical choices.

6. Conclusion

The research undertaken in this study can contribute to better understanding of nominalization in textbook introductions and book reviews. In this regard, it can help those who attempt to know the role and function of nominalization in scientific writing and as a writing style of academic discourse. Nominalization is closely linked to the principles of economy (Zhou, 2012) . Being a form of condensation of information, nominalization is a very efficient means of bundling information and consequently frequently used in formal writing. When compared with verbs, nominalizations can be more ambiguous due to valency reduction but they also provide valuable opportunities to organize discourse and express abstract relations among processes in a more efficient way. Hence, the realized differences in deployment of nominal groups in textbook introductions of harder and softer sciences can be pedagogically inspiring. Indeed, developing an awareness of the functions of nominalization—for example, enabling writers to pack more information in fewer clauses and increase information load of the text, expressing particularity by using classifiers in nominal phrases, elaborating and clarifying concepts by using relative clauses as postmodifiers for nominalizations—helps novice writers understand how this writing feature might help shape their writing in their specific discipline in a more compact and dense manner.

Furthermore, in the domain of pedagogy, teachers can make students aware of the complexity of language and how language works to compress various meanings in a sentence. Instruction of such rhetorical strategies can create an awareness of how by use of nominalization a single clause compacts several complex abstract ideas and makes language complex for the students. Thus, they need to learn a basic knowledge of grammatical metaphor and the different ways it is expressed in academic discourse.

The present study investigated the role of nominalization in applied linguistics and medicine textbook introduction genres and book reviews based on the model of grammatical metaphor proposed by Halliday and Mathiessen (1999). As the study was based on a limited data set, the results cannot be seen as conclusive. Further studies working on other disciplines can create opportunities for researchers to reflect on disciplinary characteristics. Nominalization can also be examined in other genres to determine the way nominal items are realized in

different contexts. Furthermore, our knowledge of nominalization in languages other than English is very sparse. To offset the balance, the nominal expression types used in English scientific discourse can be compared with those used in other languages to see how cross-cultural differences might play a role in using this feature of language which leads to concomitant decisions on the text texture. Given that the study design was text-based, this investigation can be extended by enquiring into academic writers' intentions and awareness about using nominal expressions in their writing. Interviews might be designed so as to gain insights into why the academic writers make use of particular patterns of nominalizations in developing their texts.

References

Abdollahzadeh, Esmaeel and Hossein Salarvand. 2013. Book prefaces in basic, applied, and social sciences: A genre-based study. *Journal of World Applied Sciences*, 28(11), 1618–1626.

Alcaraz-Ariza, María Ángeles. 2010. Complimenting others: The case of English-written medical book reviews. *Fachsprache*, 31(1– 2), 50–65.

Banks, David. 2003. The evolution of grammatical metaphor in scientific writing. In: Anne-Marie Simon-Vandenbergen, Miriam Taverniers and Louise J. Ravelli (eds.), *Grammatical Metaphor: Views from Systemic Functional Linguistics*, 127–147. Amsterdam: John Benjamins.

Banks, David. 2005.On the historical origins of nominalized process in scientific texts. *English for Specific Purposes*, 24(3), 347–357.

Baratta, Alexander M. 2010. Nominalization development across an undergraduate academic degree program. *Journal of Pragmatics*, 42(4), 1017–1036.

Bhatia, Vijay K. 1993. *Analyzing Genre: Language Use in Professional Settings.* London: Longman.

Bhatia, Vijay K. 1997. Genre mixing in academic introductions. *Journal of English for Specific Purposes*, 16(3), 181–195.

Bhatia, Vijay K. 2002. Applied genre analysis: A multi-perspective model. *Ibérica,* 4(1), 3–19.

Bhatia, Vijay K. 2004. *Worlds of Written Discourse: A Genre-based View*. London: Continuum.

Bloor, Tom and Meriel Bloor. 2004. *The Functional Analysis of English: A Hallidayan Approach*. London: Arnold.

Bunton, David. 2002. Generic moves in PhD thesis introductions. In: John Flowerdew (ed.), *Academic Discourse*, 57–75. London: Longman.

Coffin, Caroline, Curry, Mary J., Goodman, Sharon, Hewings, Ann, Lillis, Theresa M. and Joan Swann. 2003. *Teaching Academic Writing: A Toolkit for Higher Education*. London: Routledge.

Diani, Giuliana. 2007. Reporting and evaluation in English book review articles: a cross disciplinary study. In: Ken Hyland and Giuliana Diani (eds.), *Academic Evaluation: Review Genres in University Settings*. 87–105. New York: Palgrave Macmillan.

Dudley-Evans, Tony. 1986. Genre analysis: An investigation of the introduction and discussion sections of MSc dissertations. In Malcolm Coulthard (ed.), *Talking about Text: Studies Presented to David Brazil on his Retirement*, 128–145. Birmingham: ELR.

Eggins, Suzanne. 2004. *An Introduction to Systemic Functional Linguistics*. London: Continuum.

Farahani, Ali Akbar and Yaser Hadidi. 2008. Semogenesis under scrutiny: Grammatical metaphor in science and modern prose fiction. *Iranian Journal of Applied Linguistics*, 11(2), 51–82.

Galve, Guillén I. 1998. The textual interplay of grammatical metaphor on the nominalization occurring in written medical English. *Journal of Pragmatics*, 30(3), 363–385.

Gea Valor, Maria-Lluisa and María Mar del Saz Rubio. 2000-2001. The coding of linguistic politeness in the academia book review. *Pragmalingüística*, 8–9, 165–178.

Gesuato, Sara. 2004. Read Me First: Promotional strategies in back-cover blurbs. Paper presented at *the 2nd Inter-Varietal Applied Corpus Studies Conference*, Belfast, 25th-26th June. [Online] Available from: http://www.units.it/~didactas/pub/unipd/presIVACS 2004.doc. [Accessed: 14th March 2015].

Giannoni, Davide Simone. 2009. Negotiating research values across review genres: A case study in applied linguistics. In: Ken Hyland and Giuliana Diani (eds.), *Academic Evaluation: Review Genres in University Settings*, 17–33. Houndmills: Palgrave Macmillan.

Glanzel, Wolfgang and András Schubert, 2003. A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, 56 (3), 357–367.

Gray, Bethany and Douglas Biber. 2010. Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *English for Academic Purposes*, 9(1), 2–20.

Groom, Nicholas. 2009. Phraseology and epistemology in academic book reviews: A corpus-driven analysis of two humanities disciplines. In: Ken Hyland and Giuliana Diani (eds.), *Academic Evaluation: Review Genres in University Settings*, 122–39. Houndmills: Palgrave Macmillan.

Halliday, Michael A. K. 1994. *An Introduction to Functional Grammar* (2nd ed.). London: Edward Arnold.

Halliday, Michael A. K. 2004. *An Introduction to Functional Grammar* (3rd ed., revised by Christian M. I. M. Matthiessen). London: Arnold.

Halliday, Michael A. K. and James R. Martin. 2005. *Writing Science: Literacy and Discursive Power*. Pittsburgh: University of Pittsburgh Press.

Halliday, Michael A. K. and Christian M. I. M. Matthiessen. 1999. *Construing Experience through Meaning: A Language-based Approach to Cognition*. London: Cassell.

Halliday, Michael A. K. and Christian M. I. M. Matthiessen. 2004. *An Introduction to Functional Grammar. London: Arnold.*

Heyvaert, Liesbet. 2003. Nominalization as grammatical metaphor: On the need for a radically systemic and metafunctional approach. In: Anne-Marie Simon-Vandenbergen, Miriam Taverniers and Louise J. Ravelli (eds.), *Grammatical Metaphor: Views from Systemic Functional Linguistics*, 65–99. Amsterdam: John Benjamins.

Ho, Victor. 2010. Grammatical metaphor in request e-mail discourse. *Applied Language Studies*, 14(1), 1– 24.

Holmes, Richard. 1997. Genre analysis and the social sciences: An investigation of the structure of research article discussion sections in three disciplines. *English for Specific Purposes*, 16, 321–337.

Holtz, Mônica. 2009. Nominalization in scientific discourse: A corpus-based study of abstracts and research articles. In: Michaela Mahlberg, Victorina Gonzalez-Diaz and Catherine Smith (eds.), *Proceedings of the 5th Corpus Linguistics Conference Liverpool, UK*. September 25th. [Online] Available from: http://ucrel.lancs.ac.uk/publications/cl2009/. [Accessed: 4th May 2015].

Hopkins, Andy and Tony Dudley-Evans. 1988. A genre-based investigation of the discussion sections in articles and dissertations. *English for Specific Purposes*, 7, 113–122.

Hyland, Ken. 2004. *Disciplinary Discourses: Social Interactions in Academic Writing*. Ann Arbor: The University of Michigan Press.

Hyland, Ken. 2006. *English for Academic Purposes: An Advanced Resource Book*. London: Routledge.

Hyland, Ken. 2009. *Academic Discourse: English in a Global Context*. London: Continuum.

Hyland, Ken and Giuliana Diani. 2009. *Academic Evaluation: Review Genres in University Settings*. London: Palgrave-MacMillan.

Hyon, Sunny. 1996. Genre in three traditions: Implications for ESL. *TESOL Quarterly*, 30, 693–722.

Jalilifar, Alireza, Alipour, Mohammad and Sara Parsa. 2014. Comparative study of nominalization in applied linguistics and biology books. *RALs*, 5(1), 24–43.

Jalilifar, Alireza, and Zeinab Golkar Musavi. 2016. Genre analysis and genre-mixing across various realizations of academic book introductions in applied linguistics. *Journal of Teaching Language Skills*, 35(1), 111–138.

Kuhi, Davud. 2008. An analysis of move structure of textbook prefaces. *Asian ESP Journal*, 7, 63–78.

Kuteeva, Maria. 2013. English in academic and professional contexts. *Nordic Journal of English Studies*, 13(1), 1–6.

Lindholm-Romantschuk, Ylva. 1998. *Scholarly Book Reviewing in the Social Sciences and Humanities: The Flow of Ideas within and among Disciplines*. Westport: Greenwood Publishing Group.

Lock, Graham. 1996. *Functional English Grammar: An Introduction for Second Language Teachers*. Cambridge: Cambridge University Press.

Lorés Sanz, Rosa. 2012. (Non-)critical voices in the reviewing of history discourse: A cross-cultural study of evaluation. In: Ken Hyland and Giuliana Diani (eds.), *Academic Evaluation: Review Genres in University Settings*, 143–160. Houndmills: Palgrave Macmillan.

Marefat, Hamideh and Shirin Mohammadzadeh. 2013. Genre analysis of literature research article abstracts: A cross-linguistic, cross-cultural study. *Applied Research on English Language, 2(*2), 37–50.

Martin, James R. 1990. Literacy in science: Learning to handle text as technology. In: Frances Christie (ed.), *Literacy for a Changing World*. 79–117. Melbourne: Australian Council for Educational Research.

Martin, James R. 1993. Genre and literacy: Modelling context in educational linguistics. *Annual Review of Applied Linguistics: Issues in Teaching and Learning*, 13, 141–172.

Martin, James R. 2003. Beyond exchange: Appraisal systems in English. In: Susan Hunston and Geoff Thompson (eds.), *Evaluation in Text: Authorial Stance and the Construction of Discourse*, 142–177. Oxford: Oxford University Press.

Martin, James R., Christie, Frances and Joan Rothery. 1987. Social processes in education: A reply to Sawyer and Watson (and others). In: Ian Reid (ed.), *The Place of Genre in Learning: Current Debates*, 35–45. Geelong, Australia: Deakin University Press.

Motta-Roth, Désirée. 1998. Discourse analysis and academic book reviews: a study of text anddisciplinary cultures. In: Inmaculada Fortanet, Santiago Posteguillo, Juan C. Palmer and Juan. F. Coll (eds.), *Genre Studies in English for Academic Purposes*, 29–59. Castellón: Collecció Summa, Sèrie Filología/9.

Nwogu, Kevin N. 1997. The medical research paper: Structure and functions. *English for Specific Purposes*, 16(2), 119–38.

Paltridge, Brain. 1997. *Genre, Frames, and Writing in Research Setting*. Amsterdam: John Benjamins.

Römer, Ute. 2005. "This seems somewhat counterintuitive, though…": Negative evaluation in linguistic book reviews by male and female authors. In: Elena Tognini Bonelli and Gabriella Del Lungo Camiciotti (eds.), *Strategies in Academic Discourse*, 97–115. Amsterdam: John Benjamins.

Römer, Ute. 2008. Identification impossible? A corpus approach to realizations of evaluative meaning in academic writing. *Functions of Language*, 15(1), 115–130.

Salager-Meyer, Françoise and María Ángeles Alcaraz-Ariza. 2004. Las reseñas de libros en español: estudio retórico y diacrónico. *Spanish in Context*, 2(1), 29–49.

Samraj, Betty. 2005. An exploration of a genre set: Research article abstracts and introductions in two disciplines. *English for Specific Purposes*, 24(2), 141–156.

Schleppegrell, Mary J. 2001. Linguistic features of the language of schooling. *Linguistics and Education*, 12(4), 431–459.

Shaw, Philip. 2004. How do we recognize implicit evaluation in academic book reviews? In: Del Lungo Camiciotti and Elena Tognini-Bonelli (eds.), *Academic Discourse: New Insights into Evaluation*, 121–140. Bern: Peter Lang.

Shaw, Philip. 2009. The lexis and grammar of explicit evaluation in academic book reviews? In: Ken Hyland and Giuliana Diani (eds.), *Academic Evaluation: Review Genre in University Setting*, 217–235. Houndmills: Palgrave Macmillan.

Simon-Vandenbergen, Anne-Marie, Taverniers, Miriam and Louise, J. Ravelli. 2003. *Grammatical Metaphor: Views from Systemic Functional Linguistics.* Amsterdam: John Benjamins.

Sorayyaei Azar, Ali. 2012. The self-promotion of academic textbooks in the preface section: A genre analysis. *Journal of the Spanish Association of Anglo-American Studies*, 34(2), 147–165.

Sušinskienė, Solvegia. 2009. Textual functions of nominalizations in English scientific discourse. *Žmogus ir žodis*, 11(3), 58–64.

Sušinskienė, Solvegia. 2010. Nominalization as a cohesive device in British newspaper editorials. *Filologija*, 15, 142–150.

Swales, John. M. 2004. *Research Genres*. Cambridge: Cambridge University Press.

Thompson, Geoffrey. 1994. *Introducing functional grammar*. London: Arnold.

Tse, Polly and Hyland, Ken. 2009. Gender and discipline: Exploring metadiscourse variation in academic book reviews. In: Ken Hyland and Marina Bondi (eds.), *Academic Discourse across Disciplines*, 177–202. Bern: Peter Lang.

Văn, Luan. 2011. *A Study on Grammatical Metaphor in English Business Letters* (M.A. Thesis). Hanoi: Vietnam National University.

Vassileva, Irena. 2010. Critical book reviews in German. *International Journal of Applied Linguistics*, 20(3), 354–367.

Wenyan, Gao. 2012. Nominalization in medical papers: A comparative study. *Studies in Literature and Language*, 4(1), 86–93.

Xue-feng, Wang. 2010. Grammatical metaphor and its difficulties in application. *US-China Foreign Language*, 8(12), 29–37.

Zepetnek, Steven T. D. 2010. Towards a taxonomy of the preface in English, French, and German. *Neohelicon*, 37(1), 75–90.

COMPARING FORMULAICITY OF LEARNER WRITING THROUGH PHRASE-FRAMES: A CORPUS-DRIVEN STUDY OF LITHUANIAN AND POLISH EFL STUDENT WRITING

RITA JUKNEVIČIENĖ
Vilnius University, Lithuania
rita.jukneviciene@flf.vu.lt

ŁUKASZ GRABOWSKI
University of Opole, Poland
lukasz@uni.opole.pl

Abstract
Learner corpus research continues to provide evidence of how formulaic language is (mis)used by learners of English as a foreign language (EFL). This paper deals with less investigated multi-word units in EFL contexts, namely, *phrase-frames* (Fletcher 2002–2007), i.e. sets of n-grams identical except for one word (*it is * to*, *in the * of*). The study compares Lithuanian and Polish learner writing in English in terms of phrase-frames and contrasts them with native speakers. The analysis shows that certain differences between Lithuanian and Polish learners result from transfer from their native languages, yet both groups of learners share many common features. Most importantly, the phrase-frame approach highlights structural peculiarities of learner writing which are otherwise difficult to capture.

Keywords: EFL writing, learner corpus, Lithuanian EFL learners, phrase-frame, Polish EFL learners

1. Introduction

The rapid development of learner corpora continues to give impetus to lexical studies of learner language. Insights from lexical grammar on the one hand and the possibility of automated data extraction from corpora on the other have given rise to a number of studies of L2 learners' phraseological competence, which is broadly understood as their ability to use different formulaic sequences (Wray 2000: 465; Wray 2002: 9). Following the first publications of phraseological evidence in L2 language use (Pawley and Syder 1983; Kjellmer 1991), many studies have been undertaken to investigate the use of diverse multi-word combinations in learner corpora, for example, collocations (Nesselhauf 2005; Martelli 2006; Fan 2009), phrasal verbs (Waibel 2007), lexical bundles, also termed n-grams or recurrent word sequences (De Cock 2004; Chen and Baker

2010). This article deals with one of the least investigated multi-word unit in learner English so far, namely, a phrase-frame, first described by Fletcher (2002–2007). Identified using a bottom-up corpus-driven methodology, a phrase-frame is a set of variants of n-grams of any length identical except for one word, for example, *is the * of*, *it is * to*, *a part of **.[1] Phrase-frames (henceforth – PFs) constitute a theoretical concept designed to capture phraseological patterns in texts and in this respect they may be particularly interesting in learner language studies. Similarly to lexical bundles, PFs are automatically extracted from a corpus. Yet while lexical bundles offer a rather diverse lexical profile of recurrent word combinations which can be submitted to structural and functional analyses, the latter involving quite a few subjective and arguable choices for the researcher (cf. Ädel and Erman 2012: 89–90), PFs reveal a generalised picture of patterns in a corpus, which is especially valuable for a more holistic approach to the structural analysis of different language varieties, learner languages in particular.

In learner corpus research, the study of recurrent lexical combinations, PFs being one of them, usually follows one of the three research designs aimed at contrastive analyses of learner language varieties. First, such studies may be focused on one chosen EFL learner group vs. data from a comparable corpus of native speaker English (e.g. Ädel and Erman 2012; Baumgarten 2014; Chen and Baker 2010; De Cock 2004; Jalali 2013; Juknevičienė 2009; Kizil and Kilimci 2014). The second group of studies involves investigation of longitudinal or pseudo-longitudinal data representing learners at different proficiency levels (Hyland 2008a; Römer 2009; Vidakovic and Barker 2010; Juknevičienė 2013; Leńko-Szymańska 2014). Finally, the third research design is a contrastive analysis of data representing learners whose mother tongues are different (e.g. Paquot 2013; Paquot 2014; Wang 2016). Such studies usually offer an opportunity to highlight L1-specific features of the learner language varieties under study. In this respect, studies by Paquot (2013; 2014) present a significant contribution to the investigation of L1 transfer using learner corpora, most of all, owing to their methodology. It is this last research strand that the present study belongs to.

It has been only recently that PFs have become a unit of analysis in phraseological research. More specifically, PFs were explored in terms of their use and discourse functions in different registers and specialist domains (e.g. Stubbs 2007; Römer 2010; Gray and Biber 2013; Fuster-Marquez 2014; Grabowski 2015). These studies have shown that PFs may provide valuable insights into how fixed multi-word units are used in a given register and what degree of variation they exhibit (Römer 2009; 2010). Forsyth and Grabowski (2015) showed that PFs may be used not only for generalizing phraseologies in texts, but also for measuring the degree of formulaicity in language which allows

[1] On the surface, PFs bear resemblence to collocational frameworks described by Renouf and Sinclair (1991). However, the latter multi-word items are identified in a top-down corpus-based way, which means in practice that they are pre-selected by the researchers rather than automatically extracted from a corpus.

researchers to rank texts or corpora from the most to the least formulaic and, by implication, from the least to the most phraseologically varied.

PFs have been also explored in the context of English as a foreign language (EFL). For example, Römer (2009) found, first, that native and non-native students (whose L1 was German) of English often use the same PFs (with three or four words with a variable slot in the initial, medial and final position) yet with varying frequencies; second, that the students to a large extent share the slot-fillers used in the PFs; and, third, that much variation across PFs is content-related. Also, Römer (2009) found a number of PFs that occur in academic papers and yet they do not occur in native and non-native student writing, the finding that underscores the differences between expert and novice/learner language. In another study, using Michigan Corpus of Upper-level Student Papers (MICUSP), Römer and O'Donnell (2009) focused on positional variation of PFs (with 3–5 words with a variable slot in the medial position only) in native and non-native proficient academic writing, and they found that certain PFs have a strong preference for specific positions within sentences, paragraphs and texts as a whole (e.g. *it is * that* typically occurs in sentence-initial position as well as in text-final position); also, Römer and O'Donnell (2009) suggest that more research be conducted in the future on comparing student writing with expert academic writing (e.g. published research articles representing various disciplines). PFs have been also used as a unit of analysis in research on development of formulaic sequences in L1 and L2 student academic writing. For example, O'Donnell, Römer and Ellis (2013) compared the use of PFs (consisting of 3–5 words) in undergraduate native students essays collected in the LOCNESS corpus, undergraduate student writing produced by learners with eleven different L1s (sub-corpora extracted from the ICLE corpus), more advanced native and non-native student writing representing a variety of academic disciplines and collected in the MICUSP corpus as well as a corpus of expert academic writing (Hyland 1998). The said study revealed that although more advanced writers used more PFs than lower-proficiency writers (LOCNESS and ICLE), no significant effects were found of the level of language competence or native vs. non-native speaker status (O'Donnell, Römer and Ellis 2013). More importantly, the results of this study suggested that the variants of PFs should be analysed manually as otherwise no insights into their semantics or discourse functions are to be gained, and it is those functions which may help one distinguish between less and more advanced writers. In a more recent study (Garner 2016), focused on the exploration of PFs in learner language (L1 German learners of English) across five proficiency levels. The study revealed that PFs used by more proficient students exhibit a higher degree of variability and are more complex in terms of their discourse functions. An overview of PFs in EFL contexts shows that no research has been conducted so far on the comparison of the use of PFs by L2 learners with different L1 backgrounds, in particular with a focus on L1-transfer effects.

That is why the present research project was conceived as an exploratory corpus-driven analysis (Sinclair 2004; Tognini-Bonelli 2001: 65) to investigate

the phraseological competence of Lithuanian and Polish learners of English, speakers of two different first languages hardly ever contrasted before for their EFL competence. More specifically, we will try to answer the following research questions: (1) Are Lithuanian and Polish learners similar or rather different in terms of the use of frequent PFs and how similar are they to native speakers?; (2) What are the structural properties of PFs extracted from L2 English?

It should be noted that both Lithuanian and Polish are morphologically inflected languages with free word order in the sentence and hence typologically very different from English. As for the genetic typology, Lithuanian is a Baltic language and Polish a West-Slavic one, which makes the acquisition of L2 English, a West-Germanic language, obviously challenging. The two countries, Lithuania and Poland, exist in geographical proximity, and historically they have had periods of common history. The languages, however, apart from individual lexical cognates and rich albeit different morphology and syntax, have little in common and are mutually incomprehensible. Due to historical circumstances, the traditions of teaching EFL in both countries are fairly similar, which is another reason to compare learner English coming from the same geographical region. Since PFs have been rather underexplored in EFL research, our study is also an opportunity to test suitability of PFs as a unit of analysis in research on recurrent patterns in learner English, notably targeted at identification of L1-transfer effects.

2. Materials and methods

This study was designed as a contrastive interlanguage analysis (Granger 1996) aimed at highlighting L1 specific features characteristic of Lithuanian and Polish learners. To analyse their written English, we used two components of the ICLE corpus (International Corpus of Learner English): a subcorpus of Polish learner English (henceforth – PICLE) from the 2nd version of ICLE (Granger et al. 2009) and a corpus of Lithuanian learner English (henceforth – LICLE, Grigaliūnienė and Juknevičienė 2012), which is a new addition to the currently developed version of ICLE. LICLE and PICLE represent written English of advanced EFL learners who are senior undergraduate students at universities in Lithuania and Poland majoring in linguistics-based study programmes and whose first languages are Lithuanian and Polish, respectively. As a reference corpus, we used the Louvain Corpus of Native English Essays (LOCNESS, CECL 1998) consisting of argumentative and literary essays written by British and American students (excluding A-levels examination essays). Table 1 describes the corpora under scrutiny.

Table 1. Corpora used in the study

Corpus	Number of essays	Size (in words)
LICLE	335	191,570
PICLE	365	234,702
LOCNESS	298	262,339

The study was conducted in several stages. Firstly, frequency lists of PFs were generated using the kfNgram software (Fletcher 2002–2007) which retrieves four-word PFs with one variable slot, for example, *the * of the*. The variable slot may be realized in the corpora as *the beginning of the*, *the end of the*, *the importance of the* etc. To be included in the analysis, a PF had to have at least three realisations in the corpus, each with the minimum absolute frequency of 3. This decision was taken after we observed that the kfNgram software is not sensitive to capitalized letters and returned, for instance, * *of the most* (23 occurrences, 2 variants, LOCNESS) with the two realizations *one of the most* (12 occurrences) and *One of the most* (11 occurrences), which is of little value to the study. Furthermore, although the kfNgram program can generate PFs of varying lengths, in this study we decided to focus on four-word items which in the studies of recurrent sequences have been shown to be of the optimal length (cf. Hyland 2008b: 8; Chen and Baker 2010: 32) as they have a more readily recognizable range of structures and functions than the shorter sequences and are more frequent than the longer ones.

The next stage was related to the selection of PFs with respect to the position of the variable slot. In earlier studies, e.g. Römer and O'Donnell (2009) or Römer (2010), the decision was made to leave out PFs with variable slots in either the initial or final position (*BCD and ABC*) as they are often fragments of longer PFs and/or contain empty slots filled with function words. Function words, however, could be particularly important for this study because both Lithuanian and Polish learners have many difficulties in mastering the English articles and prepositions. Hence, we decided to include PFs with variable slots in any position.

Lastly, the frequency cut-off point was set at nine occurrences in LICLE and PICLE and ten in LOCNESS, which roughly corresponds to normalized frequency of 40–45 occurrences per million words. To avoid idiosyncratic effects, we also checked the dispersion of the least frequent PFs which was done by generating concordances of individual items using WordSmith Tools (Scott 2008, version 5). On average, a PF with the absolute frequency of nine occurrences has its textual variants in at least four different texts, which we considered to be an acceptable dispersion level. The statistical analysis program R was used to run statistical tests (R Core Team 2015).

The final stage of data selection involved manual revision of PFs in order to remove topic-specific items which, as demonstrated in earlier studies (e.g. Paquot 2013; 2014), can considerably distort the results, especially when the data is

retrieved from a specialised learner corpus. For this purpose, all PFs which could be linked to essay prompts were carefully checked. We considered a PF to be topic-specific if it included a lexical word from the essay prompt. Moreover, if a PF was realized as a sequence identical to a particular segment of the essay prompt on which the essay in question was written, it was excluded from further analysis. The resulting datasets are presented in Table 2 whereas a complete list of topic-specific PFs excluded from the analysis is given in Appendix 1.

Table 2. The number of topic-specific phrase-frames

	LICLE	PICLE	LOCNESS
Primary list of phrase-frames	149	163	98
Topic-specific PFs	27 (18%)	24 (15%)	26 (27%)
Topic-neutral PFs	122 (82%)	139 (85%)	72 (73%)

The relative frequency of topic-specific PFs in both learner corpora is considerably lower than in the native-speaker material. The smaller density of topic-specific PFs seems to indicate that argumentative texts written by non-native learners in comparison with those in LOCNESS lack at least one textual feature, namely, density of topical lexis, which is one of the lexical means to create cohesion (Halliday and Hasan 1976). This finding confirms observations reported in Juknevičienė (2009) which dealt with lexical bundles in learner English and found that less proficient learners underuse topic-related lexical bundles in comparison to more advanced EFL learners and native speakers. Similarly, Ädel and Erman (2012: 84) reported that "topic- and discipline-specific" lexical bundles were more numerous in their native-speaker material than in the non-native data. It is also interesting to observe that although EFL students usually make their best to exploit the lexis of essay prompts which naturally lend themselves to lifting, our findings confirm the results reported in Ädel and Erman (2012). Thus, it seems valid to assume that topic-specific lexis is indeed exploited the most in LOCNESS as compared with LICLE and PICLE. Since our study is targeted at learners' vocabulary rather than their discourse competence, topic-specific PFs were eliminated from the further analysis.

Hence, PFs will henceforth refer to four-word items with a gap in any position which meet the aforementioned frequency criterion, have at least three textual realisations and do not contain topic-specific lexical words. A full list of PFs is provided in Appendix 2.

3. Results and discussion

In the following, we report our findings starting with an overview of shared and corpus-specific PFs in LICLE and PICLE and the extent of overlap between each

of these two corpora with the LOCNESS data. This will enable us to check whether Lithuanian and Polish learners are similar or rather different in terms of the use of frequent PFs (research question 1). The second part of the analysis deals with the morphological properties of PFs. Firstly, we discuss PFs in terms of constituent words part-of-speech features. Secondly, we consider the gapped slots of PFs and their fillers in order to establish which lexical words are prone for frame-building in learner language. This stage of the study was undertaken to reveal the structural properties of PFs extracted from L2 English produced by Lithuanian and Polish students (research question 2).

3.1. Shared and corpus-specific phrase-frames

To answer the first research question we looked into shared and corpus-specific PFs. The analysis of the data showed that the three corpora have 33 identical PFs (Table 3). If the degree of overlap between each of the non-native speaker corpus and LOCNESS is considered, the results are similar for both groups of EFL learners: LICLE and LOCNESS share 20% of PFs whereas PICLE and LOCNESS have 19% identical PFs. In this respect, our results are similar to the ones obtained by Ädel and Erman (2012: 85) who explored lexical bundles and found that 22% of these multi-word units were shared by native speakers and advanced EFL learners (Swedish L1). The degree of overlap in our data, however, is slightly lower which might be related to the general lower level of proficiency in English of Lithuanian and Polish learners on the one hand and the type of items, viz. PFs rather than lexical bundles, on the other hand. Furthermore, the proportions of shared PFs reveal yet another interesting peculiarity of EFL learner writing. While in the case of LOCNESS the shared PFs account for the largest part of all PFs in this corpus (46 %), in the two learner corpora the 33 common PFs represent 27% of all PFs in LICLE and 24% in PICLE. Bearing in mind the fact that the primary data selection procedure involved a rather stringent removal of topic-specific items, it was not expected to find that the shared PFs account for less than one third of PFs in the learner corpora.

It was also found that LICLE and PICLE share between them quite many PFs; more specifically, 28% of PFs retrieved from LICLE and 24% from PICLE are identical. Moreover, if we add to this number PFs shared by all three corpora, the proportion of shared PFs between LICLE and PICLE is even greater and it certainly outnumbers those PFs that each of the learner corpora has in common with the native-speaker data. Bearing in mind the fact that LOCNESS represents a target language variety to advanced EFL learners, the picture is not very promising since both LICLE and PICLE seem to have less in common with LOCNESS than they have between themselves. This early observation was corroborated by further analysis.

As shown in Table 3, both LICLE and PICLE have a considerable number of corpus-specific PFs which only appear in one of the two corpora. While the greatest proportion (46%) of PFs in LOCNESS, as mentioned above, belongs to

the category of items established in all corpora, in the case of EFL learners, the greatest proportion is represented by the category 'corpus-specific.' It should be noted, however, that this data refers only to those items which meet the definition of PFs applied in this study; admittedly, some PFs were not included in our dataset even though they do appear in the corpora, albeit with lower frequencies. For instance, *as well as* * has only two realizations in LOCNESS (*as well as a* and *as well as the*) and was not included in the analysis.

Table 3. Proportions of shared and corpus-specific phrase-frames

Corpora	Number of shared / corpus-specific PFs	Percentages in respective corpora
LICLE, PICLE, LOCNESS	33	27% of all PFs in LICLE 24% of all PFs in PICLE 46% of all PFs in LOCNESS
LICLE and PICLE	34	28% of all PFs in LICLE 24% of all PFs in PICLE
LICLE and LOCNESS	5	4% of all PFs in LICLE 7% of all PFs in LOCNESS
PICLE and LOCNESS	8	6% of all PFs in PICLE 11% of all PFs in LOCNESS
LICLE	50	41% of all PFs in LICLE
PICLE	64	46% of all PFs in PICLE
LOCNESS	26	36% of all PFs in LOCNESS

One of the unexpected findings is the fact that the two foreign learners' corpora have more shared PFs between them than they have in common with the native-speaker data represented by LOCNESS. Both groups of EFL learners employ a number of PFs which are considerably less frequent or do not appear even once in LOCNESS. A closer examination of the data seems to suggest several explanations for the similarities between LICLE and PICLE. Firstly, owing to geographical proximity and a common cultural and historical past, the Lithuanian and Polish languages share a number of lexical similarities which apparently provide a common linguistic background to L1 Lithuanian and L1 Polish speakers. For example, both languages have equivalents for the English phrase *in this way* * which is a common lexical calque used in both Lithuanian and Polish: Lith. *tokiu būdu* and Pol. *w ten sposób*. The existence of a close equivalent in the learners' mother tongues most probably explains why *in this way* * is significantly overused by our learners in comparison to native speakers, in whose data set this PF does

not occur at all. Many more shared PFs between LICLE and PICLE, however, can be accounted for by the fact that our corpora represent inexperienced writers who are still learning to develop argumentative texts. Consequently, in comparison to native speakers they tend to overuse explicit markers of discourse organization (e.g. *as well as *, first of all *, in order to* *) and stance markers. For instance, both NNS learner corpora contain such lexical boosters as *more and more* *, * *more and more*, * *the most important, a great* * *of* etc., which add rather categorical undertones to the texts and which could be considered characteristic of novice writers (cf. Ädel 2006; Burneikaitė 2009). Another developmental feature of learner writing, which is common both to Lithuanian and Polish EFL writers, is a frequent use of gender-neutral references to people, namely, *he or she*, and *they do not* *. Obviously, the learners are demonstrating their awareness of sexist language; in addition, it is also evident that they have not yet internalized the general reference to people, i.e. *one*, which, as a matter of fact, does not exist either in Lithuanian or Polish. Lastly, some shared PFs could be linked to the common topics of the essays in LICLE and PICLE. As explained above, the data selection procedure allowed us to weed out many topic-specific PFs except for those which are not explicitly stated in any of the essay prompts. As a consequence, the topic effect could not be completely ruled out as evidenced by, for example, *the lack of* with a gap preceding or following the sequence. It often used in the essays where the questions of fortune making and (not) having money are dealt with.

The analysis of corpus-specific PFs in LICLE and PICLE was expected to shed more light on L1 transfer and L1-specific patterns. A close examination of corpus-specific PFs allowed us to identify items which could be categorized as specific features of learners sharing a mother tongue. As shown in Table 3, the largest number of items retrieved from both corpora appeared to be corpus-specific PFs, namely, 50 PFs (or 41%) in LICLE and 64 (46%) in PICLE were items not attested in the data set retrieved from the other corpora used in this study. To illustrate how corpus-specific PFs may serve as evidence of L1 influence, a more detailed discussion of two characteristic cases from each NNS corpus will be provided.

LICLE data include a number of PFs with the lexical verb *say*, namely, *say that * is*, * *be said that*, *it * be said.* All of them could be linked to the Lithuanian expression *sakoma, kad* 'it is said that' which is a passive form of *sakyti* 'to say' followed by a complement *that*-clause. This expression is typical of Lithuanian argumentative discourse where it usually introduces background information or common knowledge. While sequences with the verb *say* also appear in PICLE and LOCNESS, the only one that makes it into our data set is * *said to be* (PICLE, abs. freq. 16). While this frame does appear in LICLE (its absolute frequency of 7 is below our cut-off point), Lithuanian learners, in comparison to Polish, are significantly underusing it (Log Likelihood index 46.40, $p < 0.0001$). Instead, they are intensely exploiting such constructions which are verbatim renderings from their L1. Moreover, all PFs with *say*, with the only exception of the raising construction * *said to be*, are overused by Lithuanian learners in comparison both

to PICLE and LOCNESS data. Obviously, this finding points to the inter-L1-group heterogeneity (Jarvis 2000) and could be considered a candidate for the L1-induced constructions.

Similarly, corpus-specific PFs in PICLE also indicate such ways of expression which are overused by Polish learners. An interesting case is *in front of* * which has a few instances of specific use in PICLE. Consider the following examples:

(1) *But seven years ago the brand new world opened in front of Poles.*
(2) *In front of the unifying tendencies, in Europe at least, it would be tempting to think that the cultural boundaries (…)*

The reason for the overuse is clearly L1-induced. More specifically, the Polish preposition *przed* ('in front of', 'before', 'ahead of') can be used to indicate time, place or position with respect to someone or something else or in the presence of someone else, usually important. In the examples above, the intended meaning was to signal challenges facing Poles or Europe. That is why the use of the English preposition *in front of*, typically used to indicate place, shows that Polish learners of English tend to overgeneralize its use. Once again, we have a PF significantly overused in one of the corpora in relation to the other two (PICLE vs. LICLE Log Likelihood +4.21, $p<0.05$; PICLE vs. LOCNESS Log Likelihood +24.88, $p<0.0001$), and its idiosyncratic uses in PICLE point to possible transfer from the learners' L1.

Admittedly, not all corpus-specific PFs retrieved from LICLE and PICLE can be linked to a distinct feature of the learners' L1. While a full-scale study of transfer effects, following the methodology proposed by Jarvis (2000) and applied in Paquot's study of lexical bundles (2013; 2014), was beyond the scope of this research, the phrase-frame approach is undoubtedly a promising way forward to identify features of learner language which could be linked to L1 influence.

3.2. Structural analysis of phrase-frames

In the following stage of the study, a structural analysis of PFs was conducted to explore, first, which lexical or function words are prone to appear in PFs and, second, whether the tendencies are similar or different for both learner groups as compared with native speakers. The analysis was two-fold. Firstly, the morphological structure of PFs was taken into account, and they were grouped on the basis of constituent word classes. The second part of the structural analysis dealt with the words which appear in the variable slots of PFs, or, in other words, trigger clustering and, consequently, formation of PFs.

To analyse the morphological structure of PFs, we used the classification proposed by Gray and Biber (2013: 122) who distinguish three types of PFs, namely, (1) verb-based (V-based) PFs with one or more modal, auxiliary or lexical verbs; (2) PFs with content words other than verbs (C-based), and (3) PFs with

function words only (F-based). The results of the structural analysis are presented in Table 4.

Table 4. Distribution of phrase-frames across structural categories

Structural categories	LICLE		PICLE		LOCNESS	
	No	%	No	%	No	%
V-based	67	55%	73	52%	30	42%
C-based	37	30%	44	32%	19	26%
F-based	18	**15%**	22	**16%**	23	**32%**
Totals	*122*	*100*	*139*	*100*	*72*	*100*

The proportions of the structural categories in the three corpora are clearly different, although the effect size is small (Cramer's V 0.125). The χ^2 test of independence shows that differences in the frequencies of the structural categories in the three corpora are statistically significant (χ^2 10.3797, df = 4, p = 0.0345). To see which differences are the most important, we computed the residuals. It was found that it is the frequency of F-based PFs in LOCNESS that makes the statistically significant contribution to the χ^2 statistic value at the significance level of 0.05.

The underuse of F-based PFs in LICLE and PICLE in comparison to LOCNESS points to the fact that 'small' function words in the language of EFL learners do not build recurrent frames to the same extent as is the case in LOCNESS. Instead, in LICLE and PICLE patterns formed from lexical words are clearly dominating. One way of explaining it is the fact that non-native learners possess a rather limited repertoire of lexical words which inevitably leads to repetition of known words and familiar constructions and, as a result, yields a greater proportion of PFs incorporating a limited number of repeatedly used lexical words. This tendency was further confirmed by conducting a qualitative analysis of the data.

A closer examination of different structural types of PFs reveals that Lithuanian and Polish learners share a common feature which sets them apart from native speakers. As regards C-based PFs, the data sets from LICLE and PICLE include items which help express stance or act as boosters, for instance, * *more and more*, * *the most important*, *of the most **, *it is * difficult*, *in my opinion* etc. In contrast, the C-based PFs in LOCNESS are mostly referential expressions (* *the end of*, *the rest of **, * *the use of*, * *part of the*, *the use of* * etc.). So in their essays, Lithuanian and Polish learners resort to a more explicit marking of stance which, as our data shows, distinguishes them from native speakers and could

perhaps be viewed as a feature of less experienced writers. The other characteristic feature of non-native learner essays is discourse-organizing frames (e.g. *first of all *, as a result *, the same time *, as well as **). Interestingly, the only discourse-organizing phrase frame in LOCNESS, which is also attested in LICLE and PICLE, is *in order to **.

As to V-based PFs, the number of lexical words used in PFs is much larger in non-native English varieties than in LOCNESS. Only three forms of lexical verbs (*seen, say, continue*) appear in four PFs extracted from this corpus: *can be seen *, * seen to be, * to say that, * will continue to*. In contrast, the LICLE data set includes eleven PFs with the following forms of lexical verbs: *considered, think, say, said, sum up, want*; lexical verbs in PICLE, interestingly, are not so numerous (*want, said, take* and *realize*) yet in terms of frequencies both Lithuanian and Polish learners demonstrate a much more intense use of PFs with lexical verbs. Apparently, owing to limited vocabulary they inevitably rely on what could be seen as their 'lexical teddy bears' (Hasselgren 1994).

In an attempt to investigate which words in learner writing trigger the formation of a phrase frame, the second part of the structural analysis was focused on the variable slots. PFs retrieved from the three corpora were analysed in terms of the word class of the slot-fillers. For instance, *the * of the* may be realized by the nouns *end, beginning, majority* whereas *in order to ** is always realized by a verb. In other words, this analysis was undertaken to establish which words have the greatest potential for clustering and pattern building in the language of EFL learners and native speakers. Admittedly, some slots can be filled by different parts-of-speech, e.g. ** the fact that* is realised in LICLE by the verb *is*, preposition *to* and conjunction *and.* Such 'mixed' slots, with very few exceptions, usually occupy the initial or final position (*BCD and ABC*) of the phrase frame and they are often complete three-word formulaic sequences, e.g. ** the fact that, * in front of, * as a result *, in this way **. There are PFs, however, which are formed around one particular part-of-speech. Five morphological types of slot-fillers were identified, namely, nominal (nouns and pronouns), verbal, adjectival, adverbial and functional (conjunctions, determiners and prepositions). Table 5 below presents distribution of PFs on the basis of the morphological category of the slot-filler.

Table 5. Morphological types of slot-fillers in PFs

Morphological types of slot-fillers	LICLE	PICLE	LOCNESS
Nominal	40 (32%)	45 (32%)	36 (50%)
Verbal	19 (16%)	21 (15%)	10 (14%)
Adjectival	11 (9%)	11 (8%)	2 (3%)

Morphological types of slot-fillers	LICLE	PICLE	LOCNESS
Adverbial	2 (2%)	4 (3%)	-
Functional	9 (7%)	11 (8%)	7 (10%)
Mixed types	41 (34%)	47 (34%)	17 (24%)
TOTAL	122 (100%)	139 (100%)	72 (100%)

The majority of PFs in the three corpora have a variable slot for a noun or pronoun, namely, 32% in LICLE and PICLE and 50% in LOCNESS. Since the corpora under analysis represent written English, this finding is not unexpected as noun phrases are indeed a characteristic of the written discourse (Biber et al. 1999) and thus feature prominently in written learner language. Furthermore, our findings also confirm the results of earlier studies on learner writing which showed that the proportion of noun-based recurrent sequences is directly related to the proficiency of the learners (Juknevičienė 2009; Chen and Baker 2010). Hence, while a half of PFs in the most proficient variety of English in our data, i.e. LOCNESS, contain a variable slot for a noun or pronoun, the proportions in LICLE and PICLE (32% in both) are considerably smaller.

An interesting observation of structural peculiarities of PFs in the NNS data sets refers to the use of function words. A closer examination of PFs with a nominal/pronominal slot-filler offered an explanation why PFs incorporating functional words make a significant difference between native and non-native learners in this study. It turns out that Lithuanian and Polish learners are underusing phrases with the preposition *of* in comparison to native speakers. While PFs with *of* dominate in LOCNESS (26 out of 37, or 70%), their relative frequency is significantly lower in PICLE (22, or 50%) and even more so in LICLE (17, or 42%). Among *of*-frames, those that are formulaic expressions are particularly notable in LOCNESS, e.g. *the case of *, the rest of *, in favour of **. Although there are quite many shared PFs among the corpora, their frequencies significantly differ: the normalized frequency per 100,000 words of *the * of the* is 88 in LICLE, 77 in PICLE and 128 in LOCNESS, which shows a significant underuse of *of*-frames by EFL learners. This finding seems to be related to the fact that both Lithuanian and Polish are morphologically inflected languages, where prepositions occupy a very different place in the language system in comparison to English, while the Genitive is expressed by the case category rather than prepositional constructions equivalent to the English *of*-phrases. Undoubtedly, underuse of *of*-frames could be seen as an important feature of learner English produced by Lithuanian and Polish learners.

Another interesting finding is related to such PFs which contain a variable slot for adjective/adverb. As shown in Table 5, EFL learners significantly overuse such PFs in comparison with native speakers whereas the only ones which are

found in all three corpora are *it is * to* and *it is * that* yet even those are much more frequent in non-native English varieties. Consider their normalized frequencies per 100,000 words:

	LICLE	PICLE	LOCNESS
*it is * to*	52	48	15
*it is * that*	36	19	6

The frequent use of PFs with adjectival/adverbial slots is most probably related to the overall writing competence of EFL writers rather than any other peculiarities of learner writing. As argued above, expressions of evaluation and stance, in contrast to native speakers, are overused by EFL learners (cf. PFs with such lexical slot fillers as *better, easy, difficult, possible, impossible* etc.).

4. Conclusions

The analysis of PFs in advanced Lithuanian and Polish EFL learner writing was undertaken in order to investigate whether a structural approach involving the study of recurrent PFs in learner corpora might highlight differences between the two groups of EFL learners and, consequently, reveal L1-induced features of written learner English. The answer seems to be twofold. On the one hand, it was found that the largest proportion of PFs retrieved both from LICLE and PICLE are corpus-specific items, not attested in the remaining two corpora used in the study. Yet in order to measure to what extent corpus-specific PFs indeed indicate L1 influence, a more comprehensive study should be undertaken in the future following the framework proposed by Jarvis (2000) and focusing on measuring inter-L1-group heterogeneity in language learners' performance. Such a study may help verify statistically whether PFs explored in our study come from the same or different distribution. The qualitative analysis of selected individual PFs reported in the article seems to suggest that they could serve as a starting point for further investigation of L1 influence in learner English.

On the other hand, the study also revealed a number of shared PFs in Lithuanian and Polish learner writing that are not found in the LOCNESS corpus. These PFs often indicate developmental issues that the two learner groups are facing. Typically, the shared PFs are expressions of stance or text-organizing devices which are often favoured by less proficient learners. In this respect, it would be particularly interesting to consider PFs which are frequent in LOCNESS but underused by EFL learners. Possibly, they might represent a number of features that should be specifically targeted in EFL classrooms for at least two learner groups, i.e. Lithuanian and Polish.

A study like this one, i.e. conducted using basic quantitative methods and involving two corpora of learner language, can only be regarded as a preliminary one. There are many possible ways in which this research may be pursued further

in the future. One of the most obvious continuations would be application of the phrase-frame approach to corpora representing texts produced by learners of mother tongues other than Lithuanian and Polish. Next, if PFs indeed prove to be useful in EFL contexts, the natural line of research in the future would be to identify those PFs that carry the most salience to EFL learners of different L1 languages. In fact, similar studies have been already conducted using lexical bundles approach (e.g. Simpson-Vlach and Ellis 2010; Martinez and Schmitt 2012) even though L1 bias was beyond their focus. In this study, we focused on PFs based on contiguous sequences of four words and with a variable slot in any position. However, one may try employing longer or shorter phrase-frames in order to develop more comprehensive descriptions of phraseological patterns in learner language. Finally, bearing in mind specificity of the LOCNESS corpus, it would be possible to verify our findings by using other reference corpora representing more advanced argumentative essays, e.g. Michigan Corpus of Upper-Level Student Papers (MICUSP).

All in all, this descriptive and exploratory research may be useful for corpus linguists exploring phraseological patterns in learner language, notably when selecting phrase-frames as the unit of analysis, the concept that has been rather underexplored so far in ELF contexts.

References

Ädel, Annellie. 2006. *Metadiscourse in L1 and L2 English.* Amsterdam: John Benjamins.

Ädel, Annellie and Britt 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.

Baumgarten, Nicole. 2014. Recurrent Multiword Sequences in L2 English Spoken Academic Discourse: Developmental Perspectives on 1st and 3rd Year Undergraduate Presentational Speech. *Nordic Journal of English Studies* 13(3). 1–32.

Biber, Douglas et al. 1999. *The Longman Grammar of Spoken and Written English*. Harlow: Longman.

Burneikaitė, Nida. 2009. Metadiscoursal Connectors in Linguistics MA Theses in English. *Kalbotyra* 61(3). 36–50.

CECL (Centre for English Corpus Linguistics). 1998. LOCNESS. Louvain-la-Neuve: Universite catholique de Louvain. Available from https://www.uclouvain.be/en-cecl-locness.html. [Accessed: 12th September 2016].

Chen, Yu-Hua and Paul Baker. 2010. Lexical Bundles in L1 and L2 Academic Writing. *Language Learning and Technology* 14(2). 30–49.

De Cock, Sylvie. 2004. Preferred Sequences of Words in NS and NNS Speech. *BELL – Belgian Journal of English Language and Literature* 2. 225–246.

Fan, May. 2009. An Exploratory Study of Collocational Use by ESL Students. A Task Based Approach. *System.* [Online] ScienceDirect 37. 110–123. Available from: www.sciencedirect.com. [Accessed: 3rd January 2017].

Fletcher, William, H. 2002–2007. KfNgram. Annapolis: USNA. [Online] Available from: http://www.kwicfinder.com/kfNgram/kfNgramHelp.html. [Accessed: 20th November 2011].

Fletcher, William, H. 2010. Phrases in English. [Online] Available from: http://phrasesinenglish.org/. [Accessed: 20th September 2014].

Forsyth, Richard, S. and Łukasz Grabowski. 2015. Is There a Formula for Formulaic Language? *Poznań Studies in Contemporary Linguistics* 51(4). 511–549.

Fuster-Marquez, Miguel. 2014. Lexical Bundles and Phrase-frames in the Language of Hotel Websites. *English Text Construction* 7(1). 84–121.

Garner, James, R. 2016. A Phrase-frame Approach to Investigating Phraseology in Learner Writing Across Proficiency Levels. *International Journal of Learner Corpus Research* 2(1). 31–67.

Grabowski, Łukasz. 2015. Phrase-frames in English Pharmaceutical Discourse: a Corpus-Driven Study of Intra-disciplinary Register Variation. *Research in Language 13(*3). 266–291.

Granger, Sylviane. 1996. From CA to CIA and Back: An Integrated Contrastive Approach to Computerized Bilingual and Learner Corpora. In: Karin Aijmer, Bengt Altenberg & Stig Johansson (eds.), *Lund Studies in English 88: Languages in Contrast. Text-based cross-linguistic studies*, 37–51. Lund: Lund University Press.

Granger, Sylviane et al. 2009. *The International Corpus of Learner English*. *Handbook and CD-ROM. Version 2*. Louvain-la-Neuve: Presses universitaires de Louvain.

Gray, Bethany and Douglas Biber. 2013. Lexical Frames in Academic Prose and Conversation. *International Journal of Corpus Linguistics* 18(1). 109–135.

Grigaliūnienė, Jonė and Rita Juknevičienė. 2012. Corpus-based Learner Language Research: Contrasting Speech and Writing. *Darbai ir dienos* 58. 137–152.

Halliday, Michael, A.K & Ruqaiya Hasan. 1976. *Cohesion in English*. London: Pearson Education.

Hasselgren, Angela. 1994. Lexical Teddy Bears and Advanced Learners: A Study Into the Ways Norwegian Students Cope with English Vocabulary. *International Journal of Applied Linguistics* 4. 237–260.

Hyland, Ken. 1998. *Hedging in Scientific Research Articles*. Amsterdam: John Benjamins.

Hyland, Ken. 2008a. Academic Clusters: Text Patterning in Published and Postgraduate Writing. *International Journal of Applied Linguistics* 18(1). 41–62.

Hyland, Ken. 2008b. As Can Be Seen: Lexical Bundles and Disciplinary Variation. *English for Specific Purposes* 27. 4–21.

Jalali, Hassan. 2013. Lexical Bundles in Applied Linguistics: Variations Across Postgraduate Genres. *Journal of Foreign Language Teaching and Translation Studies*. [Online] Available from: http://efl.shbu.ac.ir/efl4/1.pdf. [Accessed: 3rd January 2017].

Jarvis, Scott. 2000. Methodological Rigor in the Study of Transfer: Identifying L1 Influence in the Interlanguage Lexicon. *Language Learning* 50(2). 245–309.

Juknevičienė, Rita. 2009. Lexical Bundles in Learner Language: Lithuanian Learners vs. Native Speakers. *Kalbotyra* 61(3). 61–72.

Juknevičienė, Rita. 2013. Recurrent Word Sequences in Written Learner English. In: Inesa Šeškauskienė and Jonė Grigaliūnienė (eds.), *Anglistics in Lithuania. Cross-Linguistic and Cross-Cultural Aspects of Study*, 178–197. Newcastle upon Tyne: Cambridge Scholars Publishing.

Kizil, Aysel S. and Abdurrahman Kilimci, A. 2014. Recurrent Phrases in Turkish EFL Learners' Spoken Interlanguage: A Corpus-driven Structural and Functional Analysis. *Journal of Language and Linguistic Studies* [Online] 10(1). 195–210. Available from: http://jlls.org/index.php/jlls/article/view/176/178. [Accessed: 3rd January 2017].

Kjellmer, Göran. 1991. A Mint of Phrases. In: Karin Aijmer and Bengt Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, 111–127. London: Longman.

Leńko-Szymańska, Agnieszka. 2014. The Acquisition of Formulaic Language by EFL Learners: A Cross-sectional and Cross-linguistic Perspective. *International Journal of Corpus Linguistics* 19(2). 225–251.

Martelli, Aurelia. 2006. A Corpus Based Description of English Lexical Collocations Used by Italian Advanced Learners. [Online] *Atti del XII Congresso Internazionale di Lessicografia: Torino, 6-*

9 settembre 2006, 1005–1011. Available from: https://dialnet.unirioja.es/servlet/articulo?codigo=4685334. [Accessed: 15th September 2016].

Martinez, Ron, and Norbert Schmitt. 2012. A Phrasal Expressions List. *Applied Linguistics* 33(3). 299–320.

MICUSP (Michigan Corpus of Upperlevel Student Papers). 2009. Ann Arbor, MI: The Regents of the University of Michigan.

Nesselhauf, Nadia. 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

O'Donnell, Matthew B., Römer, Ute and Nick C. Ellis. 2013. The Development of Formulaic Sequences in First and Second Language Writing. Investigating Effects of Frequency, Association, and Native Norm. *International Journal of Corpus Linguistics* 18(1). 83–108.

Paquot, Magali. 2013. Lexical Bundles and L1 Transfer Effects. *International Journal of Corpus Linguistics* 18 (3). 391–417.

Paquot, Magali. 2014. Cross-linguistic Influence and Formulaic Language: Recurrent Word Sequences in French Learner Writing. *EUROSLA Yearbook* 14. 240–261.

Pawley, Andrew and Francis H. Syder. 1983. Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In: Jack C. Richards and Richard W. Schmidt (eds.), *Language and Communication*, 191–225. London: Longman.

R Core Team. 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Viena, Austria. [Online] Available from: http://www.R-project.org. [Accessed 15th September 2016].

Renouf, Antoinette and John Sinclair. 1991. Collocational Frameworks in English. In: Karin Aijmer and Bengt Altenberg (eds.), *English Corpus Linguistics*, 128-143. New York: Longman.

Römer, Ute. 2009. English in Academia: Does Nativeness Matter? *Anglistik: International Journal of English Studies* 20 (2). 89–100.

Römer, Ute. 2010. Establishing the Phraseological Profile of a Text Type. The Construction of Meaning in Academic Book Reviews. *English Text Construction* 3(1). 95–119.

Römer, Ute and Matthew O'Donnell. 2009. Positional variation of phrase frames in a new corpus of proficient student writing. [Online] Paper presented at AACL conference. Edmonton, Canada, 9 Oct 2009.
Available from: http://www.ualberta.ca/~aacl2009/PDFs/RoemerODonnell2009AACL.pdf. [Accessed: 15th September 2016].

Scott, Mike. 2008. Wordsmith Tools. Version 5. Oxford: Oxford University Press.

Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Simpson-Vlach, Rita. and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics* 31(4). 487–512.

Stubbs, Michael. 2007. Quantitative Data on Multi-word Sequences in English: the Case of the Word 'World'. In Michael Hoey, Michaela Mahlberg, Michael Stubbs and Wolfgang Teubert (eds), *Text, Discourse and Corpora*, 163–190. London: Continuum.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Vidakovic, Ivana and Fiona Barker. 2010. Use of Words and Multi-word Units in Skills for Life Writing Examinations. *Research Notes* 41. 7–41.

Waibel, Birgit. 2007. *Phrasal Verbs in Learner English: A Corpus-based Study of German and Italian Learners.* [Online] Unpublished PhD dissertation. Freiburg: Albert-Ludwigs-Universität. Available from: https://freidok.uni-freiburg.de/dnb/download/3592. [Accessed: 15th September 2016].

Wang, Ying. 2016. *The Idiom Principle and L1 Influence*. Amsterdam: John Benjamins.

Wray, Alison. 2000. Formulaic Sequences in Second Language Teaching: Principle and Practice. *Applied Linguistics* 21(4). 463–489.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Appendices

Appendix 1

Topic-specific phrase-frames

LICLE	PICLE	LOCNESS
* as I lay	* are exposed to	* a loss of
* has done more	* have the right	* ethnic American literature
* in the society	* of mass media	* for the best
* men and women	* the development of	* invention of the
* money is the	* the European	* is for the
* of higher education	community	* Le Mythe de
* of the world	* the influence of	* loss of sovereignty
* same sex marriages	* the mass media	* of the absurd
* the English language	* the outside world	* of the play
* the European Union	* the right to	* the ##th century
* the higher education	* to adopt children	* the #th Republic
* the process of	approach to reality *	* the death penalty
* the quality of	have * right to	* the idea of
* the right to	in the * world	* the right to
language is a *	of mass media *	* the United States
of * in Lithuania	of the * world	for the best *
of higher education *	the * of mass	invention of the *
of the * language	the * of money	of the * Republic
quality of studies *	the development of *	the ##th century *
that writing is *	the influence of *	the * of optimism
the * of education	the opponents of *	the age of *
the * of money	the right to *	the death of *
the problems of *	the role of *	the death penalty *
the quality of *	to * a child	the people of *
the reform of *	to bring up *	the right to *
the right to *		the United States *
to study at *		

Appendix 2

Lists of PFs used for the analysis (in the frequency order). The first number indicates the absolute frequency of the phrase-frame and the second number shows the number of realizations it has in the corpus.

LICLE			PICLE			LOCNESS		
the * of the	169	42	the * of the	180	42	the * of the	335	50
one of the *	104	12	it is * to	113	12	in the * of	97	17
it is * to	100	13	in the * of	93	17	* the fact that	76	11
* a lot of	83	14	* the fact that	79	9	the fact that *	71	10
* one of the	82	8	it is not *	76	12	at the * of	61	7
it is not *	74	12	as a * of	71	7	* be able to	58	6
in order to *	69	13	one of the *	70	11	* the end of	58	6

LICLE			PICLE			LOCNESS		
it is * that	69	7	* one of the	66	7	the * of a	50	13
a lot of *	67	11	in order to *	64	13	to the * of	43	12
there is no *	58	9	there is no *	64	13	as a * of	43	7
is the * of	54	7	the fact that *	64	10	it is * to	38	6
in the * of	51	10	do not * to	55	4	end of the *	37	5
* there is no	51	9	* it is not	48	8	a * of the	35	5
that it is *	48	8	as well as *	47	7	in order to *	33	7
* the fact that	43	8	* more and more	45	6	* one of the	32	4
* be able to	42	5	* be able to	45	5	is * of the	32	4
do not * to	42	5	that it is *	44	8	that it is *	31	4
* it is not	40	6	it is * that	44	6	is * to be	30	5
first of all *	39	6	the other hand *	43	8	one of the *	30	5
the fact that *	38	6	they do not *	43	7	the rest of *	30	4
as well as *	37	8	in * of the	41	6	to the * that	28	3
* a part of	37	6	more and more *	41	6	for the * of	27	4
they do not *	36	7	as * as the	40	3	* that it is	24	6
* that it is	35	8	that is why *	39	7	that * is a	24	4
it is very *	33	4	* do not have	39	5	of the * of	23	6
* the most important	3	32	* not have to	39	3	is a * of	22	6
in the world *	31	8	do not have *	38	4	that * is not	20	3
* they do not	30	6	* that it is	36	8	in the * and	19	5
* in the world	29	6	* there is no	34	7	* it is not	19	3
the other hand *	29	4	they are * to	34	6	do not * to	19	3
is a * of	28	6	at the * of	34	5	the idea of *	19	3
is * to be	28	3	in front of *	34	5	the * of his	17	5
of the most *	28	3	as a result *	34	4	can be seen *	17	4
* more and more	27	6	* a lot of	33	7	it is * that	17	4
it is * a		27	for the * of	33	6	a part of *	17	3
the most important *	4	27	seems to be *	33	5	do not have *	17	3
* considered to be	27	3	* of the world	31	7	* that they are	16	4
* is one of	25	4	he or she *	31	7	* the use of	16	4
of the world *	24	5	of the * of	30	9	* they do not	16	4
he or she *	24	3	* aware of the	30	6	* have to be	16	3
* it is a	23	5	first of all *	30	5	* of the world	16	3
a part of *	23	5	a great * of	29	4	on the * of	15	4
of the * of	23	5	is not * to	28	5	* part of the	15	3
is very * to	23	3	* are able to	28	4	* seen to be	15	3
there is a *	22	7	to the * of	27	7	* not have to	14	4
that the * of	22	5	* at the same	27	3	the right to *	14	4
* it is the	22	4	is the * of	26	7	* the world and	14	3
as a * of	22	4	* they do not	26	5	* use to the	14	3
in my opinion *	22	4	to the * that	26	3	in favor of *	14	3
that * is a	22	4	* should not be	25	7	of the * and	14	3
the majority of *	22	4	is * to be	25	5	be able to *	13	4
as a * to	22	3	a lot of *	24	4	in * of the	13	4
do not think *	22	3	in such a *	24	3	* should not be	13	3
* be said that	21	3	it is very *	23	6	* there is no	13	3
* do not have	21	3	* to be a	23	3	a lot of *	13	3

LICLE			PICLE			LOCNESS		
the * that the	21	3	* out to be	22	3	the use of *	13	3
the * way to	21	3	there are * who	21	3	* to be a	12	4
* the lack of	20	5	they * not have	21	3	it is not *	12	4
* do not think	20	4	* have to be	20	4	* the rest of	12	3
do not have *	20	4	* seems to be	20	3	and the * of	12	3
it * be said	20	3	a * number of	20	3	I * that the	12	3
that * is not	20	3	does not * to	20	3	in a * of	12	3
* not have to	19	3	that there is *	20	3	is the * of	12	3
in the * world	19	3	to be the *	20	3	that this is *	12	3
the * of a	18	5	not * to be	19	4	the * of their	12	3
people do not *	18	4	* of the fact	19	3	they are * to	12	3
the * is that	18	4	we do not *	19	3	this is not *	12	3
to sum up *	18	4	the most important *18	18	5	* to say that	11	3
* at the same	18	3	the same time *	18	4	that the * of	11	3
* it does not	17	4	what is more *	18	4	the case of *	11	3
a * number of	17	4	* do not want	18	3	with the * of	11	3
in * of the	17	4	* they are not	18	3	* will continue to	10	3
of the * and	17	4	does not have *	18	3			
as * as the	17	3	the * of a	17	5			
because it is *	17	3	the majority of *	17	5			
more and more *	17	3	should not be *	17	4			
* are able to	16	4	* in front of	17	3			
a great * of	16	4	* the most important	17	3			
* the world and	16	3	in the world *	17	3			
considered to be *	16	3	is * it is	17	3			
is very important *	16	3	that * is a	17	3			
that * is the	16	3	* said to be	16	3			
is a * to	15	5	aware of the *	16	3			
it is a *	15	4	it is a *	15	4			
it is the *	15	4	* he or she	15	3			
* not able to	15	3	* it is a	15	3			
can be * that	15	3	it is * difficult	15	3			
say that * is	15	3	of the most *	15	3			
and it is *	14	4	the only * of	15	3			
in this way *	14	3	they are not *	15	3			
it can be *	14	3	we are * to	15	3			
most of the *	14	3	* the lack of	14	4			
they are * to	14	3	as long as *	14	4			
to * with the	14	3	that there are *	14	4			
would not be *	13	4	there are also *	14	4			
* do not need	13	3	* are not able	14	3			
most important thing *	3	13	* do not need	14	3			
* a number of	12	4	* people who are	14	3			
* in order to	12	4	* take into consideration	14	3			

LICLE			PICLE			LOCNESS
it is * for	12	4	are * to be	14	3	
the * and the	12	4	in this way *	14	3	
to the * of	12	4	* in the world	13	4	
* as a means	12	3	* the idea of	13	4	
* is the most	12	3	* most of the	13	3	
is * difficult to	12	3	is * difficult to	13	3	
that they * not	12	3	it is * a	13	3	
* amount of money	3	11	point of view *	13	3	
* should not be	11	3	they * to be	13	3	
* they want to	11	3	would not be *	13	3	
should not be *	11	3	* that they are	12	4	
the lack of *	11	3	fact that * are	12	4	
they want to *	11	3	* seem to be	12	3	
* is a very	10	3	a * variety of	12	3	
* that they are	10	3	from the * of	12	3	
i think that *	10	3	with the * of	12	3	
of a * language	10	3	* it possible to	11	3	
there * be no	10	3	* the number of	11	3	
there are many *	10	3	* the rest of	11	3	
* people do not	9	3	* the world and	11	3	
* to say that	9	3	* there is a	11	3	
is not * to	9	3	* we do not	11	3	
there are more *	9	3	are * likely to	11	3	
			by * of the	11	3	
			is * reason why	11	3	
			is the most *	11	3	
			that * is the	11	3	
			the idea of *	11	3	
			we must * that	11	3	
			* a kind of	10	3	
			* do not realize	10	3	
			it * not be	10	3	
			that the * of	10	3	
			the most * of	10	3	
			there are * many	10	3	
			to * about the	10	3	
			* would not be	9	3	
			in * to the	9	3	
			there are many *	9	3	
			there is * a	9	3	

DISCOURSE-PRAGMATIC AND PROCESSING-RELATED MOTIVATORS OF THE ORDERING OF REASON CLAUSES IN AN ACADEMIC CORPUS

ABBAS A. REZAEE
University of Tehran, Iran
aarezaee@ut.ac.ir

MAJID NEMATI
University of Tehran, Iran
nematim@ut.ac.ir

SEYYED EHSAN GOLPARVAR
University of Tehran, Iran
segolparvar@ub.ac.ir

Abstract
The present research is aimed at examining the relative importance of the competing motivators of the sequencing of reason clauses in a corpus of research articles of applied linguistics. All the finite reason clauses accompanied by their main clauses in this corpus were collected. Random forest of conditional inference trees is the statistical modelling in this study. The findings showed that sentence-final reason clauses outnumber sentence-initial ones. Moreover, subordinator choice and bridging, which are discourse-pragmatic constraints on clause positioning, emerged as the two more powerful predictors of the ordering of reason clauses in this corpus. Furthermore, the complexity of the clause turned out to be a stronger processing-related predictor than the length of the clause.

Keywords: positioning, reason clauses, subordinator, bridging, complexity

1. Introduction

Adverbial clauses are primarily positioned in initial and final slots (Aarts, 1988; Kirk, 1997; Diessel, 2001; Givón, 2011), each of which serve distinctively different functions in discourse, including academic discourse. Adverbial clauses that are sentence-final regularly have a local function: elucidating the situation of their matrix clause by specifying reasons, temporal circumstances, results, etc. Further, post-posed adverbial clauses are mainly unidirectional, i.e., they are associated with their main clauses that have been already mentioned. In addition, final adverbial clauses offer information that is more integrated with the main clause at the local level (Thompson, Longacre, & Hwang, 2007). Moreover, these adverbial clauses are often in the middle position of a paragraph, that is, final

adverbial clauses are consistently in the middle of a thematic chain which is tightly coherent (Givón, 2001). With respect to semantics, the information provided by sentence-final clauses is often in line with the information offered in clauses in coordination (Thompson, 1985; Ford, 1993; Givón, 2001).

Sentence-initial adverbial clauses, however, do not often have such a limited local function, but play a wider function in the organization of discourse, by introducing a new frame for the discourse that follows or connecting it back to the discourse that has come. Furthermore, the cohesive function of pre-posed adverbial clauses may realize at different levels, from the whole discourse to inter-paragraph and inter-sentential levels. The inter-sentential function can be considered as a local back-referencing function that establishes a close connection between two sentences, while the higher-level function marks the episode boundary or thematic discontinuity. It should be noticed that either local or global, the function of pre-posed adverbial clauses tends to be bidirectional. To put it differently, these clauses link what has been stated before to what is to be expressed. In addition, the semantic information that is offered by pre-posed clauses carries less significance since they regularly repeat or provide predictable information from what has already been mentioned (Thompson et al., 2007).

Therefore, the two sequencing patterns are not interchangeable in academic discourse and writers of research articles should know when to use each in their texts. Hence, exploring the factors that condition the ordering of subordinate adverbial clauses and the relative importance of these factors will provide us with fresh insight into the use of reason clauses in academic discourse.

The present study aims at investigating the constraints on the positioning of finite reason clauses in a corpus of research articles of applied linguistics. Further, this research purports to measure the weight of processing-related and discourse-pragmatic constraints on the ordering of finite clauses of reason by means of random forest modelling, which has been shown to be more efficient than ordinary regression models (Rezaee & Golparvar, 2017; Tagliamonte & Baayen, 2012; Wiechmann & Kerz, 2013).

2. Background

The sequencing tendency of adverbial clauses in English has been investigated by two approaches. The first approach postulates that the ordering of linguistic items, including finite adverbial clauses, is mainly influenced by information structure. Proponents of this line of research (Chafe, 1984; Birner & Ward, 1998; Greenbaum & Nelson, 1996) have put forward the argument that speakers and writers tend to produce new, inaccessible information reflected in the main clause after given, accessible information which is reflected in the dependent clause.

The users of a language usually put adverbial clauses in the initial slot in light of two factors, namely the ‘bridging’ function and the ‘setting the stage’ function. Adverbial clauses in the final slot play local functions, whereas sentence-initial

adverbial clauses have discourse-organizing functions. Establishing the link with the previous discourse or suggesting new frames for upcoming discourse are instances of discourse-organizing functions (Ford, 1993; Verstraete, 2004; Thompson et al., 2007; Givón, 2011).

In line with Wiechmann & Kerz (2013), in the present study, we solely examine one discourse-organizing function, which is bridging, referring to a context in which a sentence-initial adverbial clause serves a bridge-like function connecting the preceding and the upcoming discourse. The presence of an anaphoric item in a sentence-initial adverbial clause indicates the bridging function in that clause. In example (1), the underlined part is a sentence-initial reason clause and the anaphoric item **she** plays a bridging function, linking the sentence with the preceding discourse.

(1).

> "To find out why the teacher did the activities or made the choices recorded during the observations, a follow-up interview was held with the teacher. Since **she** was not aware of the specific research questions, in an unstructured interview she was asked to explain 'how' she taught the course and comment on course objectives, materials, in-class teaching and testing activities." (Saif, 2006, p. 20)

The other constraint explored in this line of research is the semantic nature of the subordinate clause. The semantic disparity detected among different types of adverbial clauses (i.e., adverbial clauses of time, condition, concession, and reason) leads them to assume different positions in a complex sentence. (Quirk et al., 1985; Biber et al., 1999; Diessel, 2005). Diessel (2001, 2005) found that conditional clauses are regularly pre-posed, clauses of cause are usually post-posed, and temporal clauses to be roughly equally divided between the two ordering patterns. In a similar vein, Diessel (2001) showed that adverbial clauses of reason and purpose largely follow their matrix clauses. Adverbial clauses of concession show a modest preference for the final slot (Biber et al., 1999; Diessel, 2001; Wiechmann & Kerz, 2013). Subtle meaning differences exist between clauses that are introduced by different subordinators. Thus, any subordinator selected for dependent clauses is viewed as a motivator of the sequencing of adverbial clause (Wiechmann & Kerz, 2013). For instance, concessive clauses introduced by ALTHOUGH are usually sentence-initial, while clauses headed by WHEREAS are mainly sentence-final (Wiechmann & Kerz, 2013).

The second approach trying to explain the sequencing of dependent clauses considers processing-related factors. These accounts investigate the ordering of an adverbial clause on the grounds of variables such as the relative length of the clausal string, its complexity, and its deranking status. The most famous supporter of this account is John Hawkins (Hawkins, 1994; Hawkins, 2004), who pointed out that the constituent order is mainly determined by processing difficulty. He has asserted that information structure matters only when two alternative orders are equally demanding with respect to processing.

The first factor conditioning the ordering of reason clauses is the length of the clause. Previous empirical research has vividly demonstrated that in languages like English longer constructions regularly come after shorter ones (Quirk et al., 1985). This tendency can be explained in light of the assumption that the online processing of the whole structure appears to be more efficient with this order (Hawkins, 1994; Hawkins, 2004; Gibson, 1998; Gibson, 2000). Based on Hawkins' performance-based theory of constituent ordering (Hawkins, 2004), constituents that are perceived as heavy tend to be placed in the final slot, because this pattern of sequencing is cognitively more efficient in languages that are head-initial, causing both production and comprehension to be easier.

In a similar vein, the dependency locality theory proposed by Gibson (1998, 2000) postulates that the processing complexity of a linguistic string is contingent upon the length of its syntactic dependencies. The ordering complexity effects are associated with the integration cost component which proposes that longer distance attachments are more expensive to make than shorter distance ones (Bever, 1970). Adverbial clauses of reason that are pre-posed introduce longer dependencies and are thus more demanding to process.

A pragmatic, information-structural account can also provide an explanation for the trend of 'lighter' constituents to precede 'heavier' ones based on the 'given-new' principle (Arnold et al., 2000), assuming that new information, in comparison with given information, requires more linguistic materials to be encoded. Discourse-pragmatic explanations have also demonstrated that for clauses and multi-clause constructions, the informativeness increases towards the end of each construction. Thus, length is a salient predictor of positioning of adverbial clauses of reason.

The other constraint on the ordering of reason clauses that is associated with processing difficulty is complexity. Several definitions have been proposed for complexity such as relative complexity (see Dahl, 2004; Vulanovic, 2007), absolute complexity (see (Miestamo, 2004), language complexity (Hawkins, 1994; Hawkins, 2004), and complexity with respect to informativeness (Li & Vitányi, 1997). Adverbial clauses can be complex in different degrees. It may be thought that sentence-initial adverbial clauses of reason are structurally less complex. Following Diessel (2008) and Wiechmann and Kerz (2013), in this study we regard as complex only those reason clauses that involve another subordinate clause of any kind. It should be noticed that there exists a close connection between linguistic complexity and the length of adverbial clause. Reason clauses that have another subordinate clause – complex reason clauses - tend to be longer and therefore are more burdensome to process. Consequently, we can make the assumption that complex adverbial clauses of reason are generally post-posed.

3. Method

3.1. Corpus

In order to conduct this study, a corpus of 100 research articles of applied linguistics were utilized. All the articles selected were written by native speakers of English, determined by the authors' affiliation. There are 801 tokens of reason clauses in this corpus. All the articles that are incorporated in this corpus deal with applied linguistics and language teaching and learning. The article length has not been considered as a variable. It should be noticed that the corpus of this research will include articles which are published from 2001 to 2014. All the journals used to collect the corpus are peer-reviewed both in terms of content and language. Ten articles were randomly selected from each journal. The title of these ten journals are as follows: Annual Review of Applied Linguistics, Applied Linguistics, ESP Journal, EAP Journal, Language Learning, Language Teaching Research, System, Second Language Research, language Testing, TESOL Quarterly.

3.2. Variables

The dependent variable in this study is the ordering of adverbial clauses of reason that is measured as a binary factor having two levels that are final (POS 1) and initial (POS 0). In addition, the predictive variables are subordinator, bridging, length, and complexity. Subordinator is a nominal variable with two levels, namely BECAUSE (SUB 0) and SINCE (SUB 1). According to Quirk et al. (1985), these two subordinators are the most frequent reasons subordinators in academic register.

Bridging is a categorical variable with two levels of having an anaphoric item suggesting a bridging context (BRG 0) and absence of such an item (BRG 1). Complexity is also a binary variable with two categories that are simple (COM 0) and complex (COM 1). Finally, length (LNG) is measured on a continuous scale which is defined as the proportion of the length of the reason clause to that of the whole complex sentence involving that clause. It should be noted that there were no instances of deranked reason clauses in this corpus; therefore, deranking, which is one of the processing-related constraints on clause positioning in Wiechmann and Kerz (2013), was excluded from this analysis.

3.3. Data Analysis

Conditional inference trees and the random forest developed from these trees is the modelling approach utilized in this research. Forests are a collection of multiple decision trees used for the purpose of variable selection. One single decision tree is simple and capable of coping with missing values; nevertheless, it might be unstable because minor changes in the input variables may cause huge

changes in the output. Consequently, a random forest of such trees is a more robust tool for selecting variables (Breiman, 2001).

The acceptance of random forest modelling lies in the fact that it is an unbiased method for selecting variables in the individual classification trees allowing us to reliably assess the relative weight of variables which are measured on different scales or that differ as regards to the number of their factor levels. This is the scenario where traditional tree-based models have trouble and the coefficients of logistic regression models are far more complex to interpret (Wiechmann & Kerz, 2013).

The purpose of classification trees in general is to predict a typically binary outcome on the basis of a number of predictors. The algorithms related to classification trees typically work through the data and determine a set of if-then logical (split) conditions producing accurate classification of cases. In other words, in the first step, the algorithm will split the data in accordance with the most salient predictor and will continue to split each resulting subset of the data until it can no longer find statistically meaningful associations between any of the predictors and the dependent variable (Breiman, 2001; Hothorn, Hornik, & Zeileis, 2006).

The random forest, nevertheless, is not prone to these kinds of problems, although the *cost of the computational complexity increases due to bootstrap resampling and permutation-based evaluation of variable importance. A researcher adopting random forest modelling will consider all variables in their own place, and determine which of these variables turn out to be more robust predictors. In a bid to specify how the variables operate together in the random forest, a conditional inference tree can be grown which will illustrate the way different predictors interact (Hothorn et al., 2006; Wiechmann & Kerz, 2013).

Random forests build a huge number of conditional inference trees (the random forest). Each tree in the forest is developed for a subset of the data that is produced by random sampling without drawing a replacement (subsampling) from observations and predictors. The statistical metaphor is to place part of the observed data into a bag. The data that is put in the bag is called the ‘in-bag’ observations, while the data points that are not included in the sample are referred to as the ‘out-of-bag’ observations. The result of this process is that for each tree a training set (the in-bag observations) is coupled with a test set (the out-of-bag observations). The accuracy of a tree’s predictions tends to be measured by drawing a comparison between its predictions for the out-of-bag observations and the actual values that are obtained for the out-of-bag observations (Hothorn et al., 2006). Figure 1 illustrates an instance of conditional inference tree modelling taken from Rezaee and Golparvar (2017).

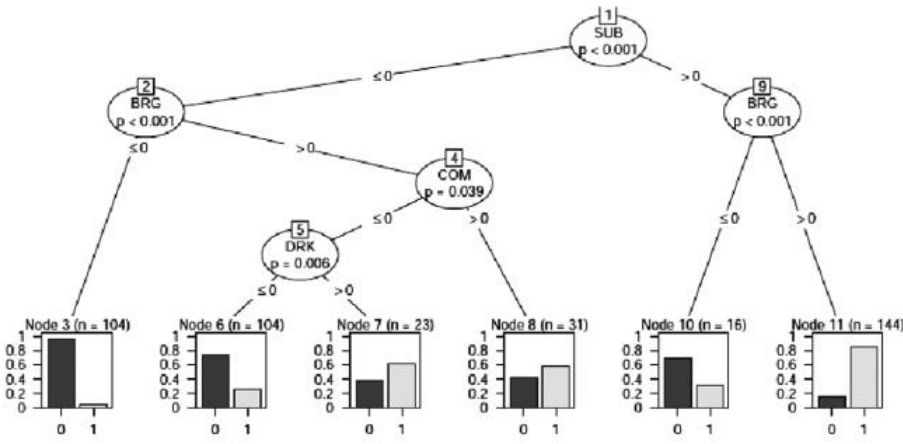

Figure 1. An Instance of Conditional Inference Tree

The analysis of this tree indicates that among the five predictors of the ordering of concessive clauses, four of them, i.e. subordinator, bridging, complexity, and deranking are significant predictors. The boxes at the bottom demonstrate the proportions of initial and final adverbial clauses in each subset, which are labeled as '0' and '1' respectively (0 represents initial clauses and 1 represents final clauses). In the white oval shapes in which the name of the variables is stated, the split variable and the *p* value indicating the significance level are observed. The numbers on the lines connecting the nodes of the tree suggest the particular categories of the nominal predictors or range of values of the numerical predictors (the only numerical predictor in this study is length).

4. Results

The results of this study demonstrated that a considerable proportion of adverbial clauses of reason (67.7%) are in final position and 32.3% of these clauses are sentence-initial. Moreover, the majority of them are simple (80%), have no anaphoric item suggesting a bridging context (88.2%), and are headed by *Since* (56.9%). Moreover, their average length relative to the size of the whole complex sentence is 0.45. Table 1 reports some descriptive statistics with regard to the sample.

Table 1. Descriptive Statistics for Reason Clauses

Dependent Variable	POS	Initial 32.3%	Final 67.7%
Predictors	BRG	Bridging 11.8%	Non-bridging 88.2%

	COM	Simple 80%	Complex 20%
	SUB	Because 43.1%	Since 56.9%
	LNG	Mean 0.45	Standard Deviation 0.25

The distribution of these five sequencing motivators across the two clause slots is demonstrated in figure 2. According to Figure 2, there exists a significant distribution difference between initial and final adverbial clauses with respect to subordinator and bridging. In addition, according to Figure 2, clauses having a bridging function are mostly in initial position, whereas those without a bridging context are mainly sentence-final. With regard to complexity, it is observed that in both simple and complex clauses, sentence-final clauses outnumber sentence-initial ones.

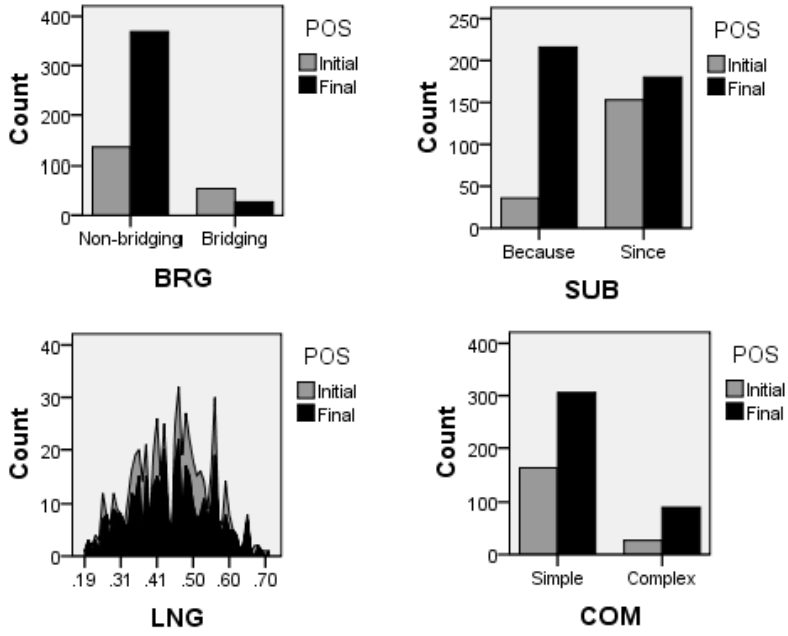

Figure 2. Distributions of the Positioning Motivators across the Two Positions of Reason Clauses

A total set of 500 trees were grown by means of bootstrapping technique, taking 500 different random subsamples from the original data The resulting model is statistically significant, indicating that three of the predictors exert a significant

effect on the positioning of reason clauses. The model shows good performance in predicting adverbial clause ordering. The index of concordance C (area under curve – ROC is 0.80) and the overall error rate of the model is 0.21. Figure 3 depicts the conditional inference tree for the positioning of reason clauses.

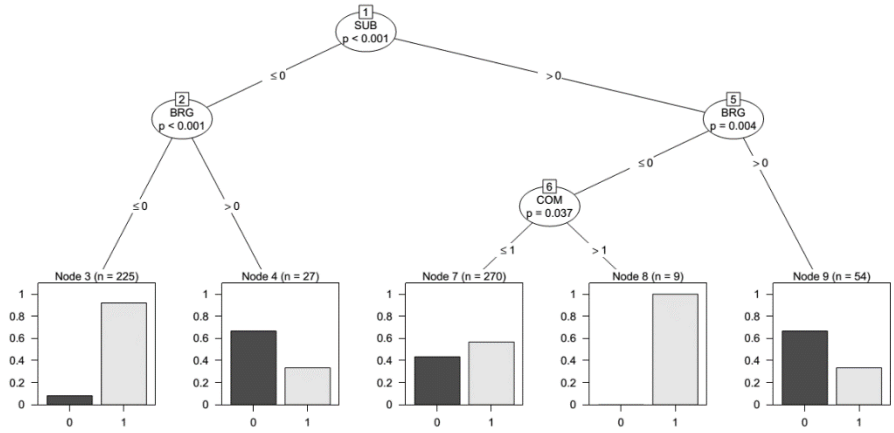

Figure 3. Conditional Inference Tree for the Positioning of Reason Clauses

Figure 3 illustrates the conditional inference tree. The analysis of this tree indicates that among the four predictors of the ordering of reason clauses, three of them, i.e. subordinator, bridging, and complexity are significant predictors. The boxes at the bottom show the proportions of initial and final adverbial clauses of reason in a given subset, which are labeled as '0' and '1' respectively (0 represents initial clauses and 1 represents final clauses). The rest of the symbols are the same as those in Figure 1.

In the first subset of the data, the first split is made based on subordinator (Node 1). The left split represents clauses of reason that are headed by *Because* (SUB ≤0) and the right one represents clauses of reason headed by *Since* (SUB > 0). Figure 3 illustrates that in both subsets of the data, a further split is made based on bridging (Node 2 and Node 5). Clauses that are headed by *Because* and do not have a bridging function (BRG ≤ 0, Node 3) are predominantly in final position. This is true for 225 cases, which is observed in Node 3. In contrast, *Because* clauses having an anaphoric item indicating a bridging context (BRG<0) are mostly sentence-initial (Node 4, 27 cases).

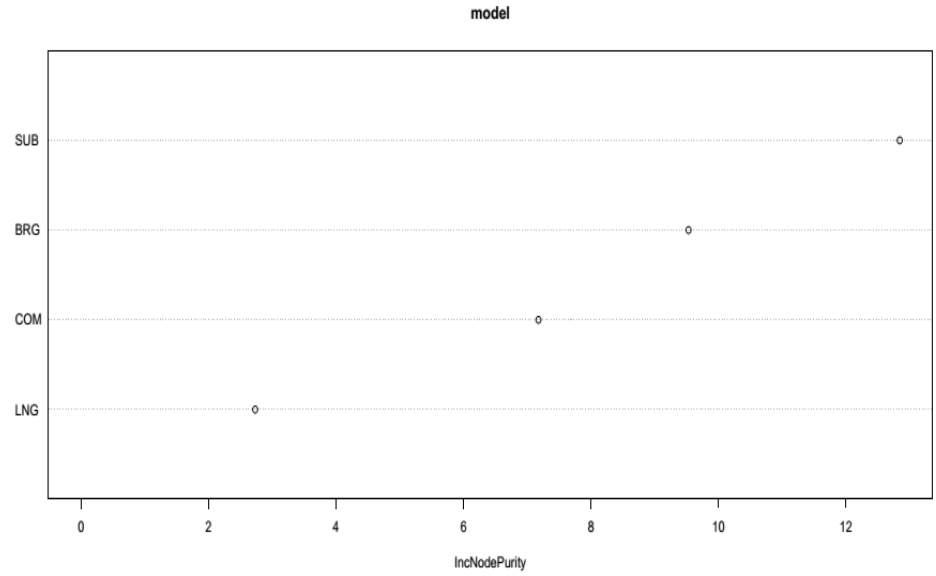

Figure 4. Variable Importance Plot for the Predictors of the Ordering of Reason Clauses

The right side of the tree depicts adverbial clauses of reason that are headed by *Since* (SUB>0). This subset of the data is further split based on bridging (Node 5). Adverbial clauses of reason beginning by *Since* with a bridging context (BRG>0) mostly precede their associate main clauses (Node 9, 54 cases), whereas those without such a function (BRG≤ 0) are further split based on their complexity (Node 6). Among clauses that are simple or complex (COM ≤ 1), sentence-final positions outnumber sentence-initial ones (Node 7, 270 cases).). Figure 4 depicts the variable importance plot for all predictors measured by the random forest model.

As shown in Figure 4, subordinator turns out to be the strongest predictor of adverbial clauses of reason, followed by bridging. Complexity turns out to be a stronger predictor than length. Finally, length has the lowest contributions to the prediction of clause ordering in this corpus of reason clauses.

5. Discussion

The analysis of a corpus of reason clauses produced by researchers of applied linguistics revealed that they tend to use these clauses in final position. This is in line with Quirk, et al. (1985), Biber et al. (1999), and Diessel (2001). Moreover, the majority of these clauses do not have an anaphoric item indicating a bridging context. In addition, only 20 percent of these reason clauses are complex,

containing another subordinate clause of any type. Moreover, clauses headed by *Because* outnumber those headed by *Since*.

The results of this study also demonstrated that *Because* clause are predominantly in final position. In addition, in reason clauses headed by *Since*, post-posed clause slightly outnumber pre-posed ones. Furthermore, in both *Because* and *Since* clauses, final clauses outnumber initial ones. It was shown that *Because* clauses are predominantly in final position, while sentence-final *Since* clauses outnumber sentence-initial ones. In other words, those reason clauses that are pre-posed are mostly headed by *Since*.

In addition, random forest modelling of conditional inference trees demonstrated that the ordering of reason clauses in a corpus of research articles of applied linguistics is firstly predicted by subordinator. To put it differently, whether the adverbial clause of reason is headed by *Because* or *Since* is the most important determinant of the sequencing of these clauses. Based on Wiechmann and Kerz (2013), the semantic disparity between reason subordinators is the most salient motivator of the positioning of reasons clauses. This finding is also in line with Rezaee and Golparvar (2017) who found that subordinator is the most powerful predictor of the sequencing of concessive clauses in a corpus of concessive clauses written by non-native speakers of English. Examples (2) and (3) illustrate this finding.

(2)

"The participants' OPI ratings were rather high even before studying abroad, most likely because they were highly motivated, enough to opt to study abroad." (Iwasaki, 2010, p. 50)

(3)

"*Since the test items were not discrete point but were nested within one of four tasks (each with their own theme),* by endorsing the interactionist view of construct definition, effects of these four themes (context) on individual items were also investigated." Vafaee, Basheer, & Heitner (2012: 1)

The random forest of conditional inference tree modelling revealed that having an anaphoric item indicating a bridging context is the second most powerful predictor of the sequencing of these clauses in research articles of applied linguistics. This is in line with Vandepitte (1993) mentioning that the information value of the reason clause impacts both its position in relation to matrix clause and the choice of its subordinator; therefore, reason clauses offering given, recoverable information usually occur in initial position, while reason clauses presenting new, unrecoverable information are placed in final position. This is also supported by the principle of end-focus and the principle of end-weight (Quirk et al., 1985; Mukherjee, 2001) asserting that the information in a message is often processed in a way to achieve a linear presentation from low to high information value.

This finding offers support for the fact that when the function of adverbial clauses of reason is to organize the flow of information in the discourse, and their

use is impacted by factors associated with information structuring and cohesion, they are mostly placed in the initial slot (Givón, 2001; Verstraete, 2004; Diessel, 2005; Diessel, 2008; Rezaee & Golparvar, 2017; Wiechmann & Kerz, 2013). To put it differently, the anaphoric relation in the discourse is the second most significant motivator of the positioning of adverbial clauses of reason in this academic corpus. Examples (4) and (5) illustrate this finding.

(4)

"This item required a gap in a dialogue to be filled with an utterance containing an expression which, in hindsight, even the highest ability candidates to the university were unlikely to have learned, hence possibly eliciting random guessing behaviors. Since **this** anomaly had to do with the correct option of this specific item, not with the number-of-options factor, it was decided to drop this item in subsequent analyses." (Shizuka et al., 2006, p.43)

(5)

"The English words were selected from the 5,000 most frequent words in Collins COBUILD corpus (Bank of English). Because **these** frequencies might not apply to FL/L2 learners, the selected words were checked against a word list based on EFL textbooks used in the Netherlands." (Schoonen et al., 2011, p. 45)

In (4) and (5), the underlined part is a reason adverbial clause in which 'this anomaly' and 'these frequencies' are anaphoric items indicating a bridging context. These anaphoric items and the reason clauses in which they are embedded create a link between the matrix clauses and the previous discourse. The results of this study showed that the majority of these bridging-functioning clauses are sentence-initial.

The variable that is most closely associated with processing-based explanations is complexity, which only emerged as the third predictor of ordering in reason clauses. To put it differently, adverbial clauses of reason that incorporate another subordinate clause tend to be put in sentence-final position; however, the impact of this constraint, i.e. complexity, is less than that of bridging and subordinator. This finding is in line with Wiechmann and Kerz 2013), demonstrating that processing-related factors are less powerful in predicting the positioning of adverbial clauses. This offers additional support for the assumption that the sequencing of adverbial clauses in general, and reason clauses in particular, is first and foremost determined by discourse-pragmatic motivators rather than processing-based constraints. Example (6) is an illustration of this point, in which the underlined part is a complex clause of reason and the bold part is a relative clause embedded in it.

(6)

"It is reasonable to expect working memory and short-term memory to be correlated because the tasks **that measure the two constructs** are very similar." (Trude & Tokowicz, 2011: 262)

6. Conclusion

This study investigated the positioning of adverbial clauses of reason in a corpus of 100 research articles published by writers of research articles of applied linguistics for whom English is considered as a native language. It was revealed that they tend to use these clauses in final position. Moreover, it was found that the ordering of reason clauses produced in this academic corpus is firstly predicted by subordinator type, and the presence of an anaphoric item indicating a bridging context is the second most powerful predictor of the sequencing of these clauses. In addition, this research lends further support for previous research on clause positioning (Diessel, 2005; Wasow, 2002; Diessel, 2008: Wiechmann & Kerz, 2013), indicating that the sequencing of adverbial clauses of reason is co-determined by principles of cognitive processing and discourse-pragmatics.

Further, motivators related to discourse-pragmatics (subordinator and bridging) are significantly more robust predictors of clause ordering than processing-related motivators (complexity and length). Moreover, the complexity of the dependent clause has a more significant contribution to the positioning of the reason clauses in comparison with other processing-related factors. Finally, random forest analysis proved to be a robust statistical means for predicting the relative weight of these constraints.

References

Aarts, Bass. 1988. Clauses of Concession in Written Present-day British English. *Journal of English Linguistics* 2. 39–85.

Arnold, Jennifer E., Losongco, Anthony, Thomas Wasow and Ryan Ginstrom. 2000. Heaviness vs. Newness: The Effects of Structural Complexity and Discourse Status on Constituent Ordering. *Language 76*(1). 28–55.

Bever, Thomas G. 1970. The Cognitive Basis for Linguistic Structures. In: John R. Hayes (ed), *Cognition and the Development of Language*, 279–362. Hoboken: Wiley.

Biber, Douglas, Johansson, Stig, Leech, Gwoffrey, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English.* London: Longman.

Birner, Betty J. and Gregory Ward. 1998. *Information Status and Noncanonical Word Order in English*. Amsterdam: John Benjamins Publishing.

Breiman, Leo. 2001. Random Forests. *Machine Learning* 45 (1). 5–32.

Chafe, Wallace. 1984. How People Use Adverbial Clauses. *Berkeley Linguistics Society* 10. 437–49.

Dahl, Östen. 2004. *The Growth and Maintenance of Linguistic Complexity*. Amsterdam: John Benjamins Publishing.

Diessel, Holger. 2001. The Ordering Distribution of Main and adverbial Clauses: A typological Stud. *Language* 77 (3). 433–455.

Diessel, Holger. 2005. Competing Motivations for the Ordering of Main and Adverbial Clauses. *Linguistics* 43 (3). 449–470.

Diessel, Holger. 2008. Iconicity of Sequence: A Corpus-based Analysis of the Positioning of Temporal Adverbial Clauses in English. *Cognitive Linguistics* 19 (3). 465–490.

Ford, Cecilia E. 1993. *Grammar in Interaction*: *Adverbial Clauses in American English Conversations*. Cambridge: Cambridge University Press.

Gibson, Edward. 1998. Linguistic Complexity: Locality of syntactic Dependencies. *Cognition* 68 (1). 1–76.

Gibson, Edward. 2000. The Dependency Locality Theory: A Distance-based Theory of Linguistic Complexity. In: Alec Marantz, Yasushi Miyashita and Wayne O'Neil (eds.), *Image, Language, Brain*, 95–126. Cambridge, MA: MIT Press.

Givón, Talmy. 2001. *Syntax: An Introduction.* vol. 1. Amsterdam & Philadelphia*:* John Benjamins.

Givón, Talmy. 2011. *Ute Reference Grammar*. Amsterdam: John Benjamins Publishing.

Greenbaum, Sidney and Gerald Nelson. 1996. Positions of Adverbial Clauses in British English. *World Englishes* 15 (1). 69–81.

Hawkins, John. A. 1994. *A Performance Theory of Order and Constituency*. Cambridge: Cambridge University Press.

Hawkins, John A. 2004. *Efficiency and Complexity in Grammar*. Oxford: Oxford University Press.

Hothorn, Torsten, Hornik, Kurt and Achim Zeileis. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15 (3). 651–674.

Iwasaki, Noriko. 2010. Style Shifts Among Japanese Learners Before and After Study Abroad in Japan: Becoming Active Social Agents in Japanese. *Applied Linguistics* 31 (1). 45–71.

Kirk, John M. 1997. Subordinate Clauses in English. *Journal of English Linguistics* 25 (4). 349–364.

Li, Ming and Vitányi, Paul. 1997. *An Introduction to Kolmogorov Complexity and Its Applications*. Heidelberg: Springer.

Miestamo, Matti. 2006. On the feasibility of complexity metrics. In FinEst linguistics, proceedings of the annual Finnish and Estonian conference of linguistics, Tallinn, 11-26.

Mukherjee, Joybrato. 2001. Principles of pattern selection. *Journal of English linguistics* 29 (4). 295-314.

Quirk, Randolph. et al. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Rezaee, Abbas Ali and Seyyed Ehsan Golparvar. 2016. The Sequencing of Adverbial Clauses of Time in Academic English: Random Forest Modelling. *Journal of Language Modelling 4*(2), 225-244.

Rezaee, Abbas Ali and Seyyed Ehsan Golparvar. 2017. Conditional Inference Tree Modelling of Competing Motivators of the Positioning of Concessive Clauses: The Case of a Non-native Corpus. *Journal of Quantitative Linguistics 24*(2-3), 89-106.

Saif, Shahrzad. 2006. Aiming for Positive Washback: A Case Study of International Teaching Assistants. *Language Testing* 23 (1). 1-34.

Schoonen, Rob. et al. 2011. Modelling the Development of L1 and EFL Writing Proficiency of Secondary School Students. *Language learning* 61(1). 31-79.

Shizuka, Tetsuhito, Takeuchi, Osamu, Tomoko Yashima and Kiyomi Yoshizawa. 2006. A Comparison of Three-and Four-Option English Tests for university Entrance Selection Purposes in Japan. *Language Testing* 23 (1). 35-57.

Tagliamonte, Sali A. and Harald R. Baayen. 2012. Models, Forests, and Trees of York English: Was/were Variation as a Case Study for Statistical Practice. *Language Variation and Change* 24 (2). 135–178.

Thompson, Sandra A., Rober A. Longacre, and Shin Ja J. Hwang. 2007. Adverbial Clauses. In: Timothy Shopen (ed.), *Language Typology and Syntactic Description*, 237–300. Cambridge: Cambridge University Press.

Thompson, Sandra A. 1985. Grammar and Written Discourse. Initial and Final Purpose Clauses in English. In: Talmy Givón (ed.), *Quantified Studies in Discourse*. *Special Issue of Text*, 5, 55–84.

Trude, Alison M. and Natasha Tokowicz. 2011. Negative Transfer from Spanish and English to Portuguese Pronunciation: The Roles of Inhibition and Working Memory. *Language Learning* 61(1). 259-280.

Vandepitte, Sonia. 1993. *A Pragmatic Study of the Expression and the Interpretation of Causality: Aonjuncts and Conjunctions in Modern Spoken British English*. Brussel: Paleis der Academiën.

Vafaee, Payman, Basheer, Nesrine and Reese Heitner. 2012. Application of Confirmatory Factor Analysis in Construct Validity Investigation: The Case of the Grammar Sub-Test of the CEP Placement Exam. *Iranian Journal of Language Testing* 2 (1). 1-19.

Verstraete, Jean-Christophe. 2004. Initial and Final Position for Adverbial Clauses in English: The Constructional Basis of the Discursive and Syntactic Differences. *Linguistics* 42 (4). 819–853.

Vulanovic, Relja. 2007. On Measuring Language Complexity as Relative to the Conveyed Linguistic Information. *SKY Journal of Linguistics* 20. 399–427.

Wasow, Thompson. 2002. *Postverbal Behavior*. Stanford: CSLI Publications.

Wiechmann, Daniel and Kerz, Elma. 2013. The Positioning of Concessive Adverbial Clauses in English. *English Language and Linguistics* 17. 1–22.

LEXICOGRAMMATICAL FEATURES IN JAPANESE ENGLISH: A STUDY OF FIVE SPEAKERS*

TOSHIKO YAMAGUCHI
University of Malaya, Malaysia
tyamag@um.edu.my

Abstract
Japanese English (JE) refers to the English spoken by Japanese citizens. This paper characterizes JE by examining its lexicogrammatical features produced by five speakers participating in experimental recordings. Drawing on the initiatives taken by Cogo and Dewey's seminal work (2012), this study presents nine lexicogrammatical features which are taken to be typical of JE. It is shown that one decisive factor in creating a new variant is the formation of an alternative form to its native counterpart and this mechanism is sourced from the speaker's multiple knowledge about two languages.

Keywords: creativity, Japanese English, lexicogrammatical features, multiple knowledge about two languages

1. Introduction

In this rapidly globalizing world, people from different countries and cultures communicate using English, a language which is not the mother tongue of the majority of those who speak it. Already in the 1980s, scholars witnessed a surge of non-native speakers of English which led the latter to outnumber the population of native speakers (Swan 1985; Strevens 1992: 27). For example, Swan (1985: 7) predicted the rise of "the new international English" which may, viewed from his EFL (English as a foreign language) perspective, shed many of the complexities of present-day native Englishes (e.g. British English, American English), such as in the tense system. In Japan, one learns English as a foreign language at school. Within the Japanese education system, English is a compulsory subject from the first year of junior high school (at which point pupils are 12–13 years old), but English has never become integral to the daily communication of Japanese nationals. The average Japanese citizen living in Japan with no outside contact has no need to communicate in English; Japanese is the language used to express oneself in all situations of everyday life (e.g. Browne and Wada 1998; Seargeant 2011; Abe 2013; Tsuneyoshi 2013; D'Angelo 2018). The need for communicative English is therefore restricted to specific domains, such as international business

* I am obliged to the second reviewer whose constructive comments were useful in revising an earlier version of this paper.

and academia or tourism. Despite its limited use, English is becoming increasingly important in Japan, particularly in universities (McKenzie 2008). The number of non-Japanese students enrolled at Japanese universities has increased and they are not all fluent in Japanese. Lectures are now conducted in English at some institutions, and there are many student exchange programs. These changes are reflected in recent publications, but scholars have often discussed the ambivalent status of English as a global language, whether it is a foreign language, a *lingua franca*, an international language, or a language operating as a touchstone for social and cultural issues in contemporary Japan (on the last point, see Seargeant 2011). Crystal (2010: 17) remarked that the notion of English as a global language may not only refer to common features found across the globe but also to regional features specific to individual languages. For example, *Yeah right* is an expression of suspicion about the content preceding it (e.g. *Of course I remember your name. Yeah right*). To understand an example such as *Let Paul fly us there. Yeah right*, however, Crystal asserted that we need cultural knowledge about Paul (a radio personality who owned two private planes and crashed and survived twice), a piece of knowledge shared collectively by local people (New Zealanders in this case). Scholars and journalists generally express pessimism about the teaching of English in Japan. Friedman (2016) has expressed concern about the future of English in Japan since the teaching of the language has not undergone a paradigm shift, especially in terms of methodology and textbooks, which are still rooted in Meiji-era practice (1868–1912). In a similar vein, Tsuboya-Newell (2017) has reported on teachers' lack of communication skills in English in *The Japan Times*. What is interesting about Tsuboya-Newell's article is her point that environmental factors appear to be decisive in the acquisition of a foreign language, that is, the amount of exposure to an English-speaking environment rather than simply contact with a teacher. It is notable that the literature, as reviewed above, has a tendency to rely upon attitudinal, educational, or sociolinguistic standpoints to describe English in Japan. More than ten years ago, in a review of Stanlaw's (2004) monograph on Japanese English, which itself focuses on English loans that have entered the Japanese lexicon, Smith (2004) noted the lack of a linguistic study of the English used by native Japanese speakers. More recently, McKenzie reiterated the same point: there are "no detailed descriptions of … linguistic features" (2013: 228).

This paper is an interim report on an ongoing project that currently has a sample of 25 Japanese speakers of English. We call the English produced by native Japanese speakers in spoken discourse "Japanese English" (JE). The main objective of this paper is to describe the linguistic features produced by five native Japanese speakers (four female, one male; J2, J3, J7, J8, and J12) talking about the topic of "weather" in an experimental setting.[1] Section 2 explains how recordings were conducted and outlines the participants' linguistic and social

[1] The letter "J" stands for a Japanese speaker who participated in the experimental recordings in 2016 and 2017.

backgrounds. Based on the non-native features collected, Section 3 presents a preliminary analysis of lexicogrammatical features produced by these Japanese speakers. It is important to note at the outset that non-native features often, if not always, co-occurred with their native counterparts, forming two alternatives. JE speakers used these alternatives effectively to construct new meaning. Section 4 closes the paper by highlighting that a new variant can be identified when an alternative to an existing native form is created and this is sourced from the speaker's multiple knowledge about two languages – in this case, Japanese and English.

2. Recordings: structure and participants

The recordings were conducted in 2016 and 2017. Each recording comprised three components: (i) reading a short text, (ii) reading words, and (iii) speaking about given topics in English and Japanese. The first two components have already been the subject of an acoustic phonetic analysis (Yamaguchi and Pétursson 2018). The present study focuses on the third component, the free talks. There were three topics – (i) "my current situation and future plans," (ii) "weather," and (iii) "an event/person I can't forget" – and participants were informed of them in advance (2–3 weeks prior to the recording) and allowed to bring keywords with them to help organize their talks. While participants spoke on all three topics in English, they also chose one to speak on in Japanese, thus producing Japanese and English spoken texts on one subject that were conceptually the same. The talks were "free" in that participants created a spoken text constrained solely by their linguistic capacity. Each free talk lasted about two minutes (120 seconds) and the recording was made in a professional studio where participants spoke alone into a microphone in a sound-insulated recording room.[2]

At the time of the recordings, all 25 project participants (J1–J25) of the larger project lived in Kuala Lumpur, the capital of Malaysia.[3] Five speakers were selected from among them for this analysis, for two reasons. One is that they chose "weather" as the topic of their free talk in Japanese, which enabled the author to compare and contrast English and Japanese texts that were conceptually similar. The other is that the spoken texts on weather were the most homogenous: speakers customarily began with a description of the hot weather in Malaysia, including comparison with the weather in Japan or another country, referring to personal experiences or sharing thoughts about hot or cold weather. Table 1 presents the beginning of each speaker's talk; all of them are concerned with either the heat or the rain considered to characterize the weather in Malaysia. I judged that such

[2] They were facing the control room and could make eye contact with people (among whom the author) in the control room.

[3] The choice of Malaysia is due to the author's affiliation with a university in Kuala Lumpur.

homogenous texts would ease analysis and offer quick insights into the general picture of JE.

Table 1. The beginning of the talk produced by five Japanese speakers

Speaker	The beginning of the talk
J2	The weather in Malaysia is very hot and I know some people don't like the weather, but for me (0.44)[4] it's very (0.5) comfortable to stay here.
J3	I (0.17) I like summer. (1.26) I like Malaysian weather (0.38) such as (0.46) hot and (0.51) rain. (0.96)
J5	So weather, uh (0.59), so about weather in Malaysia. I (0.43), yeah, it's really hot. It's just hot. (0.71)
J8	About weather, uh (0.67), about Malaysian weather I have three impression. First is hot, and second is humid and third is, uh (0.51), heavy rain and thunder clap.
J12	I'm from Sapporo, Hokkaido, where is north part of Japan. (0.91) So, the weather here is totally different from weather here. The big difference is rain and thunder.

In terms of demographics, the five speakers were either Japanese language teachers working for a university or foundation course (J2, J3, and J12) or undergraduate exchange students from Japanese universities (J5 and J8). The participants in the first group rarely had the opportunity to speak in English due to the nature of their job, although they spoke it occasionally (at meetings in the workplace, in conversations outside work). Due to their study program, the participants in the second group used English actively every day, mostly with classmates and roommates. All the participants had started to learn English substantively from the first year of junior high school (Section 1).[5] These speakers were not stereotypical Japanese citizens, routinely speaking Japanese with little or no exposure to English (Section 1). They had contact with the outside world and made good use of English as a means of communication; interestingly, most of their English communication occurred without the presence of native English speakers. Table 2 summarizes the five speakers' linguistic and social backgrounds as of December 2016 (J2, J3), May 2017 (J5, J8), and June 2017 (J12).

[4] The parentheses present the duration of a pause in seconds, which was measured by Praat (Boersma and Weenink 2017). The pauses are given only when they are prominent.

[5] A brief note on J5 might be useful since he stayed on Fiji for a year when he was a high school student and went to a local high school. He described this as an unforgettable experience in his third free talk.

Table 2. Japanese English speakers' linguistic and social backgrounds

Speaker, Gender, Age	Exposure to native English	Length of stay in Malaysia	Language in everyday life	Language at home (with the family	When do I use English?
J2 Female 37	Recently completed 2 years of English classes	6 years	● Japanese ● Malay	● Japanese ● English	● Workplace ● Conversation with Malaysians
J3 Female 51	NA	5 years	● Japanese ● Occasional use of English in public places	● Japanese	● Limited use in the workplace
J5 Male 21	Sharing a flat with an American	9 months	● Japanese ● English	● Japanese	● At school ● In the flat
J8 Female 21	Had English teacher at school in Japan	9 months	● Japanese ● English	● Japanese	● At school ● In the dormitory
J12 Female 38	Stayed in Wales for 10 months in 2004–2005	2 years	● Japanese ● Occasional use of English in public places	● Japanese	● Limited use in the workplace

3. Analyzing free talks

In analyzing the free talks, lexicogrammatical features typically departing from native English norms were examined carefully. This study follows the principle of Cogo and Dewey (2012) (hereafter C&D) by regarding non-native features as "innovative language forms" (C&D: 13) integral to the English produced by JE speakers. In using the term "innovative forms" rather than "learner errors", as did classical scholars for virtually the same type of data (Corder 1967; Selinker 1972), C&D regard new forms as exhibiting systematic occurrence and organized patterns within the "localized repertoire" (C&D: 21),[6] and hence, native Englishes are no longer viewed as the goal of learning and/or a language in international

[6] There are concerns about the notion of systematicness. Swan (2012: 386) observed that C&D described non-native features (e.g. definite article use) but questioned how systematic the occurrences are.

settings.[7] Where the present study differs from C&D is in its consideration of native-norm equivalents of non-native features and their Japanese equivalents. The reason for the first is that, as noted briefly in the Introduction, both non-native and native features co-occur often, if not always, in JE talks. The reason for the second consideration is that the analyzer can better grasp what the speaker has in mind when the same topic is provided in two languages. This process of comparison between Japanese and English talks boosted my understanding of how the participants managed the English language to verbalize their thoughts and ideas.[8]

3.1. Articles

The majority of non-native features occurred in the use of articles. Since Japanese has no articles (C&D: 64), one might argue that errors concerning articles originate from their absence in L1, but this rule of thumb does not generalize to all cases. Rather than choosing articles different from those found in English as a native language (ENL),[9] participants produced nouns with no article. The noun "weather" frequently occurred with no definite article when it was introduced as the topic of a talk (J5: So, (the) *weather, so* (the) *Malaysian weather*; J8: *About* (the) *weather*).[10] Articles were also used inconsistently. J12 produced a sentence consisting of two clauses comparing the weather in Japan and Malaysia: *So, the weather there is totally different from* (the) *weather here*. The word *weather* is accompanied by *the* in the first clause and appears without it in the second, meaning that the speaker might have known that weather needs the definite article but forgot to include it the second time. There is another problem with the definite article when it comes to general reference. In stating his opinion about hot weather in general, J5 first talked about a memorable stay in Fiji and said that he had chosen that destination because he likes hot weather: *And I chose Fiji because Fiji* (it) *has hot weather*. However, he indicated that his opinion about hot weather was changing in Malaysia as he had to walk a long way every morning to get to the bus stop and arrived at the classroom dripping with sweat. He said: *But now, after living in Malaysia for like nine month* (months), *I feel like maybe I don't really like* ***the*** (Ø) *hot weather*. Twenty seconds later, he rephrased his opinion: *So, maybe I like cold weather now*. That *cold weather* is not preceded by the article and has no reference in the given discourse indicates that he was making a generalization about hot weather by mentioning its opposite. These three

[7] Ranta (2018: 251–252), discussing the grammar of ELF, urges the explanation of the exact meaning of innovations.

[8] The author consulted a native speaker of English for analysis of non-native features.

[9] The term "English as a native language" is used interchangeably with "native form" and "native norm" in this paper.

[10] The parentheses give either alternative words replacing non-native features or new words added to the original utterance. These possibilities are not absolute but regarded as alternatives that fit into the uttered expressions by JE speakers.

examples allow us to infer that he might have known the rule that the article is left out for general reference but didn't follow it consistently, similar to J12. A variable usage of *weather* with and without the definite article is characteristic of J5, and is found in the following two examples. J5 said: *One thing I find interesting about* (the) *weather, like, something related to* (the) *weather*. Forty-eight seconds later, in order to conclude his talk, he said the same phrase with *the*: *Something related to* ***the*** *weather*. J8 tended to use nouns with the zero article where *the* is expected: … *I have to stay in* (the) *library or my class until the rain stop* (stops). But J2 placed *the* as expected: … *you don't want to go out and try to stay in* ***the*** *house where it's very warm*.

Indefinite articles were produced much less frequently than definite articles, in line with C&D (2012: 98, Table 4.5). The indefinite article was replaced either by the zero article or the definite article. To illustrate, three participants, J2, J5, and J12, did not use *a* for T-shirt (J2: *I can just wear* (a) *simple T-shirt*; J5: *I need to go to school with all sweat* (in a really sweaty) *T-shirt*; J12: *Usually, I* (only) *wear only* (Ø) (a) *T-shirt at home even in winter*). There was a single case in the data set in which first mention of the noun was preceded by the definite article (J8: *And actually my college's power supply was cut off by* ***the*** (a) *thunderstorm*.

Because of the complexity of the article usage and its variations (cf. C&D 2012: 62), the above discussion is summarized in Table 3. Note that five types (A–E) in the table are selective with special focus on cases where non-native and native features are put to alternative uses, often by the same speaker. While type E only occurred in J8's talk, it is included as it may be a typical feature of JE. The heading for each type in small capitals is the function of an article in the ENL system.

Table 3. Distributions of articles produced by five JE speakers

Type	alternative usage	Examples	Speakers				
			J2	J3	J5	J8	J12
			Total uses of articles				
			11	7	14	12	11
A	SPECIFIC REFERENCE						
	the unused	Ø weather in Malaysia	1	4	5	5	5
	expected	the weather in Malaysia	6		2	1	4
B	REFERENCE IN THE PHYSICAL ENVIRONMENT						
	the unused	in Ø library				3	
	expected	in the library	1				
C	GENERAL REFERENCE						
	the used	I prefer the hot weather		1	1		
	expected	I prefer Ø hot weather	1	2	3	1	
D	INDEFINITE REFERENCE						
	a unused	I wear Ø T-shirt	1		2	1	2
	the used	go to the country	1				
	expected	I wear a T-shirt; go to a country			1		

	FIRST MENTION						
E	*the* used	X was cut off by the thunderstorm				1	
	expected	X was cut off by a thunderstorm					

3.2. Plural formation

Plural formation is another problem for JE speakers. Like articles, Japanese does not possess the equivalent of the plural marker *-s* which is added to nouns productively.[11] Plural forms were correctly applied when they were clearly countable (e.g. *four seasons; maple leaves*). However, J8 used *impression* to mean "a point" or "an aspect," but did not pluralize it (J8: *I have three* ***impression*** (points), see Table 1). A similar example comes from J12, who was talking about Malaysians, shopping malls, and restaurants in general terms but did not pluralize them (J12: *Then, inside like in* ***shopping mall***, ***restaurant***, ***office***, *or* ***cinema***, *freezing*). These examples might indicate that the JE speakers distinguished between abstract and concrete nouns in their mind. While *points* are definitely abstract as they are intangible, *shopping malls*, *restaurants*, and *offices* are tangible. These tangible nouns might have been used without *-s* because J12 was listing them as general concepts: she did not consider them to be concrete and therefore countable. A similar case is found in J3's talk. When she was referring to events in general, she said: *In Malaysia, I can't remember when* ***the event*** (events) *was held* (happened). The fact that she did not pluralize *event* indicates that it was conceptualized as general. The addition of *the* reinforces the pattern of JE: general reference is signaled by the definite article (see C, Table 3). Back to J12: in her talk about "weather" she used a plural form once, for the noun *students*. She was talking about her students who visited her in her home town. That she met more than one student is an important piece of information, for it was real event and concrete and, above all, we recognize that *students* is clearly countable, like *seasons*. She said: *And in March, actually, I went* (0.25) *back to* (0.22) *I went* (0.31) *back to Sapporo and met my* ***students***. It follows that when it comes to plural formation in JE talks, the important consideration is apparently whether nouns are conceived of as abstract/general or concrete: the suffix *-s* tends to be attached to the latter but not to the former.

3.3. Possessive pronouns

The frequent absence of the possessive pronoun in JE talks is another feature that deserves attention. When J8 was talking about heavy rain in Malaysia, she said, *if*

[11] Japanese marks some restricted nouns, such as the first person singular pronoun *watashi* or common nouns such as *kodomo* "child" and *gakusei* 'student,' by adding *-tachi* (*watashi-tachi* "we"; *kodomo-tachi* "children"; *gakusei-tachi* "students"). However, this bound morpheme is not productive.

I forget ***the*** (my) *umbrella*, referring to her own. When J2 was talking about why she found the weather in Malaysia comfortable, she said, *I don't worry about* ***the*** (my) *clothes and* (Ø) *shoes and* (or) *anything* (anything else). As discussed in Section 3.2, J12 used *my* when referring to students she had taught: *And in March, actually, I went* (0.25) *back to* (0.22) *I went* (0.31) *back to Sapporo and met* ***my*** *students*. It is not clear whether the usage of the possessive pronoun *my* is explicable by reference to L1, but one aspect that is illuminating is that nouns in J2's Japanese spoken text[12] were zero nouns (i.e. *fuku toka kutsu toka* "clothes and shoes"), whereas in J12's Japanese text, *gakusei* "student" was accompanied by the reflexive pronoun *jibun* "self," clearly expressing the idea of possession. In Japanese, possession is often implicit and hence does not surface, as in J2's case. However, when the speaker underlines the aspect of "ownership," like J12, the possessive pronoun surfaces. That is, the native form *my* is realized when the speaker feels strongly that someone or something belongs to or is connected with him or her. Clothes, shoes, or umbrella do not impart the same level of ownership.

3.4. Assigning different lexical meaning

An innovative usage of word meaning is to change the original meaning slightly to make it suitable for a specific context or situation. For example, J3 used the verb *recognize* to express the meaning *understand* when she wanted to say that her destination can be deduced based on the clue of a specific season related to a specific event (*nobody can* ***recognize*** (work out) *when I went there*). J8 used *acceptable* to mean that she can tolerate or get used to the heat and humidity in Malaysia, and *impression* to mean "point" or "aspect" (*Ah, for the first and second* ***impression*** (points/aspects), *hot and humid, it was* ***acceptable*** (okay) *for me*) (see Section 3.2). An interesting contrast was found in J5's usage of *remember* and *recall*. When J5 moved to the second subtopic of his talk (i.e. the cultural relationship between seasons and events), he started to use *recall* to mean *remember*. For an ENL speaker, *remember* and *recall* are synonymous but differentiated through formality. In J5's talk, the two verbs were also synonymous but differentiated contextually. *Remember* was used as a neutral verb to describe what he considered to be a general situation (*Then* (Ø) *I will* (Ø) ***remember***, *like, ah, it was cold, and I was wearing a jacket and I was eating this kind of food*) and *recall* was apparently used to emphasize difficulty: when there is only one season and there is no link between seasons and events, one cannot remember as easily (*So, it's really hard to* ***recall*** (remember) *the past*).

[12] J8 was excluded here because she did not talk about forgetting of her umbrella in her Japanese talk.

3.5. Using adjectives in place of nouns

This section discusses the use of *hot* as a noun in place of *heat*. Violation of the system of lexical categories is a feature which Suenobu (1990: 261) listed as a typical L1 transfer. Note, however, that although J12 used the adjectives *hot* and *cold* as nouns in English (J12: *Malaysian is* (Malaysians are) *strong to both* ***hot*** *and* ***cold*** (tolerate the heat and the cold)), in her Japanese talk, she used derived nouns, namely, *atsusa* "the heat" and *samusa* "the cold," meaning that the nominal use of *hot* and *cold* may not, strictly speaking, be L1 transfer. Something similar can be seen in J8's talk: she had in mind a nominal expression in her Japanese text when she said *atsui to iu koto* "the fact that it is hot," referring to one of three features of Malaysian weather, but in her English talk, she simply said: *First is* ***hot*** (the heat). Comparison of their texts in English and Japanese suggests that, conceptually, these two speakers had a nominal expression at their disposal but did not end up with *the heat* when they spoke in English. Although the idea of L1 transfer may not be excluded completely, the use of adjectives as nouns may signal the presence of more than one language in the mind of JE speakers, or put differently, the speaker's knowledge about the target language might influence its production. It is possible that J8 and J12 simply did not know the nominal form of *hot*. J24 used *hotness* preceded by the definite article to express to the same meaning. He knew the nominal form of *hot*.

3.6. Personal pronouns

One example that might have been influenced by L1 is the use of *human being*, as in J3's talk (*but I think the* (Ø) ***we***, *human being* (human beings), *have always overcome, uh, winter*). An ENL speaker might simply have used *we* without *human beings*; the addition could be considered redundant or excessively formal. When J3 spoke in Japanese, she used the equivalent expression *watashi-tachi ningen* "we, human beings," an addition which native Japanese speakers would not consider excessively formal. Use of *ningen* "human being" underlines that J3 was talking generally, while the pronoun *watashi-tachi* "we," when it stands alone, does not directly impart collective meaning. J3's use of *human being* may be categorized, at first glance, as a prime example of L1 transfer, but it can also demonstrate, similar to Section 3.5, the presence of more than one language in J3's mind. That is to say, *we*, as used in her English talk, may not be, strictly speaking, a marker of collectivity, but simply an expression of the first person plural marker. This usage may also have derived from J3's knowledge of the target language. J13 and J14 used *me* as a collective marker (J13: *So, the story* (stories) *he gave* (told) *me* (us) *sounded really interesting*; J14: *He gave* (taught) *me* (us)

some kind of lecture (various courses)).[13] This fact indicates that *we* is interchangeable with *me* in JE spoken discourse and its collectivity may be much weaker than the native-English counterpart.

3.7. Relative pronouns

Talking about "weather," with the exception of J12, who used *where* once in place of *which*, the five JE speakers did not use any relative pronouns. J12's use appeared in the first sentence of her talk (see Table 1): *I'm from Sapporo, Hokkaido, **where*** (which) *is* (in the) *north part of Japan.* While J12 used *which* to modify "course" and "organization" in her talk about "my current situation and future plans,"[14] she apparently knew how to use this pronoun following the native norm, and her use of *where* represents an alternative usage. In J12's mind, places might be treated differently from objects such as "course" or "organization," since Sapporo is clearly the name of a place and the choice of *where* would make sense, while it does not for the other two nouns. In my entire data set, this usage of *where* occurs only once but it indicates that *where* is emerging as another relative pronoun besides *which* in JE's grammar.

3.8. Coreference

JE speakers tend to repeat the same noun within the same sentence, while the noun is coreferenced by the pronoun in a separate sentence. A neat pair of examples come from J5. Talking about his one-year stay in Fiji, he said: *And I chose Fiji because **Fiji*** (it) *has hot weather*. An ENL speaker would use the pronoun *it* to avoid repetition of *Fiji*. This non-native usage does not mean that J5 did not know the rule of coreference, as he made good use of it in the next sentence: *So, I used to really love **it***. The pronoun *it* refers to *hot weather*, the theme of his free talk. The same kind of repetition was produced by J2 when she said at the beginning of her talk (see Table 1): ***The weather** in Malaysia is very hot and I know some people don't like **the weather*** (it). There was no coreferential usage of *it* in her talk comparable to J5's.

3.9. Overdoing explicitness

This is a notion originally introduced by Seidlhofer (2004), quoted by C&D (2012: 48), to categorize examples such as *black color* instead of *black*. In my data set, one example that fits this category is *winter season*. J2 and J3 both produced this expression in their talks: the season "winter" was characterized as having a low

[13] J24 (Section 3.5) and J13 and J14 (Section 3.6) were the participants not included in the paper. Additional examples produced by these speakers were offered here to strengthen the arguments in both sections.

[14] I am referring to another talk by J12 to offer examples.

temperature causing coldness. Neither speaker liked the cold, especially because they had been in Malaysia for some time and had become comfortable with the hot weather. As shown in the following examples, J2 and J3 used *winter season* where *winter* was expected.

> J2: So since, since I, since I, since I (have) got used to Malaysia with the hot weather, I don't want to go to the (a) country, uh, which has, uh uh **winter season** (winter).
> J3: I don't like (the) winter season but I think we need **winter season** (winter) (1.83) to (2.0) appreciate (2.19) spring or summer.

In her Japanese equivalent of the same utterance, J2 used *fuyu no kisetsu* ("winter season") but J3 did not. Moreover, as shown by the example below, which was said after the above example, J3 used *winter* alone in her talk in English. Why didn't she use *winter season*?

> J3: I think it is too exaggerate (much of an exaggeration), but I think the (Ø) we, human being (human beings), have always overcome, uh, **winter**.

On closer look, we note that the meaning of *winter* is reduced in this utterance; that is, *winter* may not be understood as representing the cold but a season. Recall Crystal's (2010) earlier assertion (Section 1). I interpret *winter season*, as used here, as representing cultural knowledge about winter shared collectively by Japanese people, which is that it is typically cold and severe in contrast to the other three seasons. To express this cultural knowledge, speakers created a new variant, *winter season*, without losing the ENL form *winter*. This usage shows how culture is embodied in non-native spoken discourse, what C&D dubbed "localized repertoire." The choice of *season* might have derived from L1, since Japanese allows the same expression but its usage in L2 may not be a direct reflection of L1.

4. Conclusion

This paper has presented a preliminary analysis of lexicogrammatical features extracted from recorded free talks on the topic of "weather" in English and Japanese produced by five native Japanese speakers aged 21 to 51. All these speakers had a similar educational background in Japan, beginning to learn English at school. At the time of the recording, they had opportunities to use the language in Malaysia. Against this sociolinguistic background, the analysis of the data (Section 3), the heart of this paper, has demonstrated how some new variants, or what C&D called "innovative language forms," came into existence. This study has identified coexisting alternatives, so to speak, to native forms which, in some cases, carried specific meanings/functions which were largely individual-based and arose on the fly in spoken discourse.

Since Honna and Takeshita's (1998) paper on JE in the first volume of *Asian Englishes*, there have been calls for a paradigm shift in foreign language education in Japan, a shift that would promote an indigenous Japanese usage of the English in place of native English norms in the spirit of Kachru (2017: 148). Concomitantly, the focus of this research direction has been on criticism of "native-speakerism" (Houghton and Rivers 2013), or "Japan's propensity for native speaker English," to borrow Honna and Takeshita's (1998) expression. The stance taken by the present paper is essentially the same as Honna and Takeshita's in that I seek to define JE as a variety owned by its users and existing independent of native English – and this conception is accurate given the fact that JE speakers have few opportunities for contact or communication with ENL speakers. Besides this, the description of features, as undertaken above, should also contribute to defining an emerging variety and it is hoped that this paper is a step in this direction. Considering the fact that JE is essentially learned – it originates from formal education in a context in which English is not a language of daily communication – it can be concluded that innovative linguistic forms in JE develop from the speaker's L1 and his or her knowledge, albeit basic, of the native norm, the norm learners are taught in the classroom. This paper has illustrated the robustness of multiple knowledge possessed by JE speakers as L2 users. On a broader scope, this is what Cook (2016: 3) has defined as "multi-competence," "the overall system of a mind" which L2 users have at their disposal in handling two languages, and Mackenzie (2016: 494) sees ELF (English as a *lingua franca*) as an instantiation of multi-competence. Turning to the teaching of English (Section 1), I take the position, for now, that standard English – or, better, pedagogical core English – is the form of the language which Japanese people should learn as an input language, supplemented by non-native features, as discussed above, that draw on the "real-world phenomenon" (Jenkins 2018: 599) we experience in our lives – and from here we may be able to scrutinize the processes or features involved in shaping an output language.

One final point concerns Swan's (1985) prediction, quoted at the outset, of the rise of a new international language whose lexicogrammatical structure may be simplified. What we have discussed in this study turns this proposition on its head. The creation of alternative forms, if it continues, will definitely add to the structure of this new language.

References

Abe, Emiko. 2013. Communicative language teaching in Japan: current practices and future prospects. *English Today* 29 (2). 46–53.

Boersma, Paul and David Weenink. 2017. Praat: doing phonetics by computer (Version 6.0.40) [computer software]. Available from: http://www.praat.org/

Browne, Charles M. and Minoru Wada. 1998. Current issues in high school English teaching in Japan: an exploratory survey. Language, *Culture and Curriculum* 11 (1). 97–112.

Cogo, Alessia and Martin Dewey. 2012. *Analysing English as a Lingua Franca: A Corpus-driven Investigation*. London/New York: Continuum.

Cook, Vivian. 2007. The goals of ELT: Reproducing native-speakers or promoting multicompetence among second language users? In

Corder, Stephen Pit. 1967. The significance of learners' errors. *International Review of Applied Linguistics* 5. 161–170.

Crystal, David. 2010. The future of Englishes: going local. In: Roberta Facchinetti, David Crystal and Barbara Seidlhofer (eds.), *From International to Local English – And Back Again*, 17–25. Bern: Peter Lang.

D'Angelo, James F. 2018. The status of ELF in Japan. In: Jennifer Jenkins, Will Baker and Martin Dewey (eds.), *The Routledge Handbook of English as a Lingua Franca*, 165–175. London: Routledge.

Friedman, Jeffrey. 2016. English education in the era of Meiji Japan. *World Englishes* 35 (1). 3–17.

Honna, Nobuyuki and Yuko Takeshita. 1998. On Japan's propensity for native speaker English: a change in sight. *Asian Englishes* 1 (1). 117–134.

Houghton, Stephanie Ann and Damian J. Rivers. 2013. *Native-speakerism in Japan: Intergroup Dynamics in Foreign Language Education*. Bristol: Multilingual Matters.

Jenkins, Jennifer. 2018. The future of English as a lingua franca? In: Jennifer Jenkins, Will Baker and Martin Dewey (eds.), *The Routledge Handbook of English as a Lingua Franca*, 594–605. London: Routledge.

Kachru, Braj. 2017. *World Englishes and Culture Wars*. Cambridge: Cambridge University Press.

Mackenzie, Ian. 2016. Multi-competence and English as a lingua franca. In: Vivian Cook and Li Wei (eds.), *The Cambridge Handbook of Linguistic Multi-competence*, 478–501. Cambridge: Cambridge University Press.

McKenzie, Robert M. 2008. The complex and rapidly changing sociolinguistic position of the English language in Japan: a summary of English language contact and use. *Japan Forum* 20 (2). 267–286.

McKenzie, Robert M. 2013. Changing perceptions? A variationist sociolinguistic perspective on native speaker ideologies and Standard English in Japan. In: Stephanie Ann Houghton and Damian J. Rivers (eds.), Na*tive-speakerism in Japan: Intergroup Dynamics in Foreign Language Education*, 219–230. Bristol: Multilingual Matters.

Ranta, Elina 2018. Grammar in ELF. In: Jennifer Jenkins, Will Baker, and Martin Dewey (eds.), *The Routledge Handbook of English as a Lingua Franca*, 244–254. London: Routledge.

Seargeant, Philip (ed.). 2011. *English in Japan in the Era of Globalization*. Basingstoke: Palgrave Macmillan.

Seidlhofer, Barbara. 2004. Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics* 24. 209–239.

Selinker, Larry. 1972. Interlanguage. *International Review of Applied Linguistics* 10. 209–241.

Smith, Donald L. 2004. Review of *Japanese English: Language and Culture Contact* (James Stanlaw). *Asian Englishes* 7 (2). 126–132.

Stanlaw, James. 2014. *Japanese English: Language and Culture Contact*. Hong Kong: Hong Kong University Press.

Strevens, Peter. 1992. English as an international language: directions in the 1990s. In: Braj B. Kachru (ed.), *The Other Tongue*, 27–47. Urbana/Chicago: University of Illinois Press.

Suenobu, Mineo. 1990. Nihon Eigo [Japanese English]. In: Nobuyuki Honna (ed.), *Ajia no Eigo* [*Varieties of English in Asia*], 257–286. Tokyo: Kuroshio.

Swan, Michael. 1985. Where is the language going? *English Today* 3. 6–8.

Swan, Michael. 2012. ELF and EFL: are they really different? *Journal of English as a Lingua Franca* 1–2. 379–389.

Tsuboya-Newell, Ikuko. 2017. Why do Japanese have trouble learning English? *The Japan Times*. [Online] 29th October. Available from: https://www.japantimes.co.jp/opinion/2017/10/29/

commentary/japan-commentary/japanese-trouble-learning-english/.
[Accessed: 17th April 2018].

Tsuneyoshi, Ryoko. 2013. Communicative English in Japan and "Native speakers of English." In: Stephanie Ann Houghton and Damian J. Rivers (eds.), *Native-speakerism in Japan: Intergroup Dynamisms in Foreign Language Education*, 119–131. Bristol: Multilingual Matters.

Yamaguchi, Toshiko and Magnús Pétursson. 2018. Japanese English: Norm-dependency and emerging strategies. *English Today* 34 (2). 17–24.

GOOGLE BOOKS NGRAM VIEWER IN SOCIO-CULTURAL RESEARCH

ANNA ZIĘBA
Adam Mickiewicz University, Poland
azieba@amu.edu.pl

Abstract
The objective of this paper is to verify if Google Books Ngram Viewer, a new tool working on a database of 361 billion words in English, and enabling quick recovery of data on word frequency in a diachronic perspective, is indeed valuable to socio-cultural research as suggested by its creators (Michel et al. 2010), i.e. the Cultural Observatory, Harvard University, Encyclopaedia Britannica, the American Heritage Dictionary, and Google. In the paper we introduce a study performed by Greenfield (2013), who applies the program to her *Ecological Analysis*, and contrast the findings with a study based on similar premises, in which we follow the trends in changes in word frequency throughout the 19th and 20th centuries to observe if these changes correspond to one of the major socio-cultural transformations that took place in the studied period, i.e. mediatization. The results of this study open a discussion on the usefulness of the program in socio-cultural research.

Keywords: Google Books Ngram Viewer, word frequency, socio-cultural transformations, mediatization, news values

1. Introduction

It is tempting to believe that the arrival of a new tool giving access to a massive database, a corpus of 5,195,769 books scanned and digitized with the use of optical character recognition (OCR) will open new possibilities in many fields of science. As performing a study on such vast material has not been achievable before, a research on the corpus provided by Google Books Ngram Viewer seems a state-of-the-art endeavour that should provide reliable data.

It could also be valuable to the socio-cultural research that is based on linguistic material as such research is usually very time-consuming. Therefore, one of the merits of this tool is that it allows the researcher to spend more time on the analysis of data than on their collection.

Moreover, it might appear that since the lexical changes are gradual and relatively stable, the fluctuations in word frequency, upon which Google Books Ngram Viewer provides extensive data, are relevant and their study will improve our comprehension of the social changes and their consequences.

The paper first presents the tool and gives examples of its possible application to research in various fields as suggested by its creators. Then, a recent study in

human ecology proposed by Greenfield (2013) is introduced, a study which inspires us to perform a similar one on the relationship between one of the biggest socio-cultural transformations in the period under study, i.e. mediatization, and changes in frequency of words relevant to the subject. Next, we present the methodology of the study and the obtained data. In the later parts of the paper the results of our study are discussed and conclusions on the usefulness of Google Books Ngram Viewer in socio-cultural research are drawn.

2. Theory and background

2.1 Google Books Ngram Viewer

Linguists have hitherto worked with word frequency dictionaries or lists such as 450 million word Corpus of Contemporary American English (Davies 2010), 5 million word The American Heritage Word Frequency Book (Carroll, Davies & Richman 1971), or databases such as 1.7 billion word Dante (Atkins 2010) and WordNet (Fellbaum 2005) with little over 155,000 words. These tools seem modest in comparison with Google Books Ngram Viewer, a new tool introduced in 2010 by the Cultural Observatory, Harvard University, Encyclopaedia Britannica, the American Heritage Dictionary, and Google, as its creators constructed a corpus of 5,195,769 digitized books (4 percent of all books that have ever been published) from over 40 university libraries and individual publishers.

The texts were scanned and digitized with the use of optical character recognition (OCR). Having taken into account the quality of the texts' OCR and metadata, the team selected a group of over 5 million books for analysis to develop the corpus including 361 billion words in English, 45 billion in French, 45 billion in Spanish, 37 billion in German, 35 billion in Russian, 13 billion in Chinese, and 2 billion in Hebrew. The study was limited to the analysis of frequency of a given 1-gram, which might be understood as a single lexical unit, or an n-gram (a series of lexical units) over time, but occurring at least 40 times in the corpus. Michel et al. (2010) define the 1-gram as "a string of characters uninterrupted by a space" and an n-gram as "a sequence of 1-grams, such as the phrases 'stock market' (a 2-gram) and 'the United States of America' (a 5-gram)". The frequency was "computed by dividing the number of instances of the n-gram in a given year by the total number of words in the corpus in that year" (Michel et al. 2010).

In the article published in *Science* Michel et al. (2010) maintain that the corpus enables investigators to study cultural trends quantitatively, and that it has opened up a new field of research, namely *culturomics*, a field drawing a connection between changes in word frequency and linguistic and cultural shifts. The researchers give examples of such undertakings. They observe changes in the English lexicon studying the overall number of words to discover that the size of this language increased by over 70% in the past 50 years. They also follow the

changes in frequency of the 2077 headwords that entered the American Heritage Dictionary of the English Language in 2000 and notice that part of the words, still found in the dictionaries, were no longer used. These investigations lead them to the following conclusion: "Our results suggest that culturomic tools will aid lexicographers in at least two ways: (i) finding low-frequency words that they do not list, and (ii) providing accurate estimates of current frequency trends to reduce the lag between changes in the lexicon and changes in the dictionary" (Michel et al. 2010).

The team investigates grammatical changes and finds that frequency is an important factor in the shifts between regular and irregular forms of past verbs. The researchers also make an inquiry into collective memory, developing plots concerning the interest in various events between 1875 and 1975, in order to compare the rise and fall of fame of the most well-known people and uncover censorship in Nazi Germany.

At a later stage the researchers add a system that enables identification of parts of speech, searching for inflections or for multiple capitalization styles simultaneously and a feature called 'wildcards': for retrieving the ten most popular collocations.

Since its introduction in 2010 Google Books Ngram Viewer has been widely described and employed both in social and natural sciences. Berry (2012) describes it as an example of "the way in which code and software become the conditions of possibility for human knowledge, crucially becoming computational epistemes" (Berry 2012: 1), Rutten et al. treats it as a tool to overcome a "chronological distance, or time lag, between books and their subject matter in studies of memory" (Rutten 2013: 40) and Michalski et al. (2012) suggests the Ngram Viewer could be used "as a fast prototyping method for examining time-based properties over a rich sample of literary prose" (Michalski 2012: 1).

Google Books Ngram Viewer has been applied in various studies. Linguists used it to investigate biomedical domain literature in respect of terminology changes (Grigonyte et al. 2012), to follow word usage and cultural transformations in contemporary West Bengal (Phani 2012) and to illustrate diachronic variation of preferred adjective ordering (Hill 2012). It was also employed in social studies: Kesebir and Kesebir (2012) used it to prove that moral ideals and virtues decreased significantly in the American public conversation, Oishi et al. (2013) to analyze the concepts of happiness across time and cultures, Cockerill (2013) to trace the roots of industrial ecology education to the 1960s and 1970s, Lucier (2012) to study the relations of science and capitalism, Kumar and Sahu (2010) to trace the history of marketing, and Johnson (2011) to introduce the concept of information overload, not to mention Greenfield (2013) who applied it to a research into human ecology. It has also been used by Alcock (2012) to assess trends in the use of evolutionary concepts in non-technical literature and Crasto (2011) to study the use of the term 'bioinformatics' in literature.

Google Books Ngram Viewer also received criticism, which came mostly from Mark Davis (2014), who recognised the dataset as remarkable but perceived the

interface as too simplistic. He claimed it did not allow for the use of collocates in searches, searching by wildcards and a meaningful use of parts of speech. As the datasets had been made available online by their collectors, Davis incorporated them into his work and proposed an alternative architecture and interface that enabled more complex searches (e.g. with variables for a given part of speech or providing data on complicated grammatical constructions). However, his criticism towards Google Books Ngram Viewer does not seem fully grounded as GBNV does in fact allow for searches based on speech tags or wildcard searches (though these features were not possible at the original stage).

There also appeared questions concerning the accuracy of data acquired through the use of Google Books Ngram Viewer (mostly on blogs and forums): the long 's' mistaken by OCR for 'f' (which fell out of the English typeface in early 19th century), as well as semantic and spelling changes. However, these can influence a study of word frequency in the 19th and 20th century only marginally. Therefore, even if we take into consideration the imperfections of OCR, Google Books Ngram Viewer still seems to put socio-cultural research in a context whose significance is hard to question, especially if carried out cautiously and conscientiously.

2.2. Greenfield's ecological analysis

Our study, whose objective was to inspect whether the tool is indeed suitable for investigating socio-cultural changes, was inspired by the work of Greenfield published in Psychological Science in 2013. The researcher uses Google Books Ngram Viewer to study human ecology and finds confirmation for her hypothesis concerning a shift in this ecology from rural to urban. She also maintains that cultural features indexed by word frequencies reflect what is preferred by a population.

She generates the hypotheses on a theory of social change from *gemeinschaft* into *gesellschaft* environments. She focuses on individualistic values and behaviours such as: personal choice, materialism, significance of personal property, independence and assertiveness, becoming dominant in the modern world. Greenfield assumes in her study that the gesellschaft-adapted cultural traits, indicated by relevant words in the American English corpus should grow in number, and that the gemeinschaft-adapted features studied within the same corpus should decline.

In her study Greenfield uses high-frequency words, as advised by Michel et al. (2010), with a narrow range of semantic interpretations, relevant to the theory and their synonyms. She studies the changes in frequency of the following word pairs:

- 'oblige' (characteristic of the gemeinschaft environment) and 'choose' (characteristic of the gesellschaft environment), and their noun synonyms: 'duty' and 'decision',

- ‘give’ and ‘get’ (representative of gemeinschaft and gesellschaft environments respectively) and their noun synonyms ‘benevolence’ and ‘acquisition’,
- ‘act’ (exhibiting gemeinschaft comprehension of the social world in terms of action or behaviour) and ‘feel’ (representing inner psychological processes typical of the gesellschaft domain), and their noun synonyms ‘deed’ and ‘emotion’,
- and additional concepts to illustrate the historical pattern of shifts in values: ‘obedience,’ ‘authority,’ ‘belong’ and ‘pray’ (and their synonyms: ‘conformity,’ ‘power,’ ‘join’ and ‘worship’) as depicting gemeinschaft values, and ‘child,’ ‘unique,’ ‘individual,’ and ‘self’ (and their synonyms: ‘baby,’ ‘special,’ ‘personal’ and ‘ego’) as exemplary of the gesellschaft scene.

The results of the analysis confirmed Greenfield’s stance. The relative frequency of all words characteristic of the gemeinschaft environment decreased and the words characteristic of the gesellschaft environment increased. Putting the results in the context of other studies relevant to the field and replicating the analysis for each word in the corpus of British books validated the assumptions of the researcher. Therefore, Greenfield maintains that the transformation of the American culture from rural to urban is reflected in the American cultural products, i.e. books.

3. The analysis

The research conducted by Greenfield encouraged us to perform a similar study with the use of Google Books Ngram Viewer. Our analysis covered one of the major socio-cultural changes, which occurred in the last two centuries, namely mediatization (Hjarvard 2008, Lilleker 2008). With the development of media, a change in communication has taken place. As a consequence, entire societies are strongly influenced or even formed by mass media (Mazzoleni and Schulz 1999, Hjarvard 2013). Following Greenfield’s example it can be assumed that this in turn should lead to changes in the frequency of words relevant to the socio-cultural phenomenon in question.

3.1. Methodology

The semantic key upon which we performed the analysis was based on a set of features deciding on the newsworthiness of information, i.e. *news values*. The set, originally defined by Gatlung and Ruge in 1965 and by now well established in media studies, comprised 18 qualities: *negativity*, *recency*, *proximity*, *consonance*, *unambiguity*, *unexpectedness*, *superlativeness*, *relevance*, *personalization*, *eliteness*, *quality of attribution*, *facticity*, *continuity*, *competition*, *co-option*, *composition*, *predictability*, and *prefabrication*. We assumed that if Greenfield is

unmistaken, the analysis of the frequency of words relevant to the news values should show an increase in the studied period. Therefore, we selected ten features to be included in the analysis and prepared 5-word semantic keys presented in Table 1. We chose to work on an English (both British and American) corpus, as it is the largest database available so far.

Table 1. The selected news values and their representative lexical items

no.	News value	5-word semantic key
1.	negativity	awful, bad, dreadful, poor, unacceptable
2.	recency	currently, lately, presently, recently, today
3.	proximity	close, dear, familiar, nearby, neighbouring
4.	consonance	average, common, normal, standard, usual
5.	unambiguity	apparent, clear, distinct, evident, obvious
6.	unexpectedness	abrupt, rapid, sudden, surprising, unexpected
7.	superlativeness	best, first, last, least, most
8.	relevance	essential, great, important, significant, substantial
9.	continuity	constantly, continuously, regularly, steadily, still
10.	predictability	anticipated, expected, predictable, probable, supposed

Including all 18 values would not have been possible, as some of the features cannot be represented accurately. In the case of *personalization*, *facticity*, *co-option*, *composition*, and *prefabrication* the feasibility of the analysis was found to be very limited. It would be problematic to find words relating to a personal portrayal of information, intertextual references (in case of *co-option*) and *eliteness* as the objects of these strategies would be different in each text, and *facticity* would be manifested by an infinite number of names, dates and statistics. Likewise, it was difficult to identify words characteristic for a prefabricated message. Additionally, we assumed that *competition* (in the media occurring between agencies, editorial teams or journalists) and *composition* (maintaining a balance of different types of coverage) are irrelevant to our study since they affect language rather at the textual than lexical level.

Each semantic key includes 5 lexical items, listed in alphabetical order in Table 1. The items were selected as the most relevant to the given value, i.e. occurring in contexts representing the value. The keys include both adjectives and adverbs, as part of the news values are described predominantly by adjectives (*negativity*, *proximity*, *consonance*, *unambiguity*, *unexpectedness*, *superlativeness*, *relevance*, *predictability*) and part by adverbs (*recency*, *continuity*). The main criterion for selection was high-frequency (mean frequencies for all relevant items were measured).

As Greenfield points out, since Google Books Ngram Viewer works on a corpus of 361 billion words in English, the absolute percentage of any single word is naturally small. However, the focus of the study is on the change in frequency, not its height. For example, the word ‘great’ in Figure 8 starts in the year 1800 with a frequency of about 130 occurrences per 100,000 words but decreases to

about 30 occurrences per 100,000 words by the year 2000. It seems that this change is meaningful. Moreover, we took interest only in general trends in changes in the frequency of selected items, as focusing on each rise and fall in frequency of subsequent items in a 200-year period would not increase the value of the study, on the contrary it could blur the results.

We also considered using the advanced interface proposed by Davis (2014), but since the object of the study was the standard version and since in the study we included only basic searches available in both interfaces, such endeavour seemed pointless.

3.2. Findings of the study

The scores for each news value, presented separately for the purpose of clarity, are illustrated in the figures below. Comparing the results we included values accurate to four decimal places. The objective was to calculate the ratio of increasing to decreasing trends in the changes in frequencies of the selected words representing each of the 10 news values. In cases when the relative change between the values for 1800 and 2000 was less then (+/-) 30% the change was deemed insignificant.

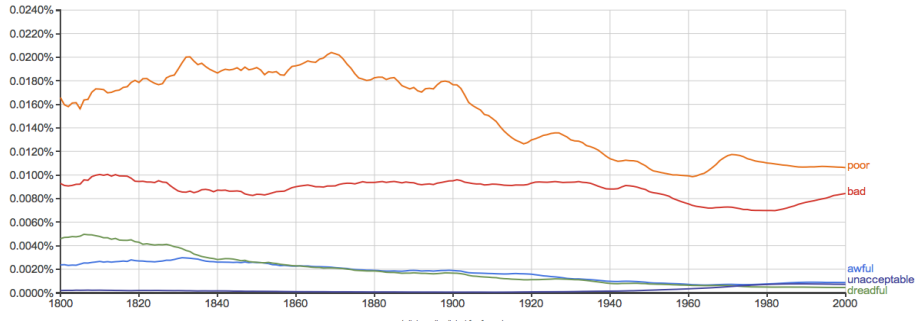

Figure 1. The frequency of the five words representative of *negativity* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

The first news value studied was *negativity*, which refers to higher rating of bad news than good news by the media. It was represented by the following words: *awful*, *bad*, *dreadful*, *poor* and *unacceptable*. Out of these five items just one (*unacceptable*) confirms the assumption that the values typical for mass media do influence the changes in frequency of words representing these values, as the word's frequency rises from 0.0002% to 0.0007%. The frequency of the other four items either decreases (*poor*, *awful*, *dreadful*) or does not change significantly (*bad*), i.e. the relative change is little over 10%.

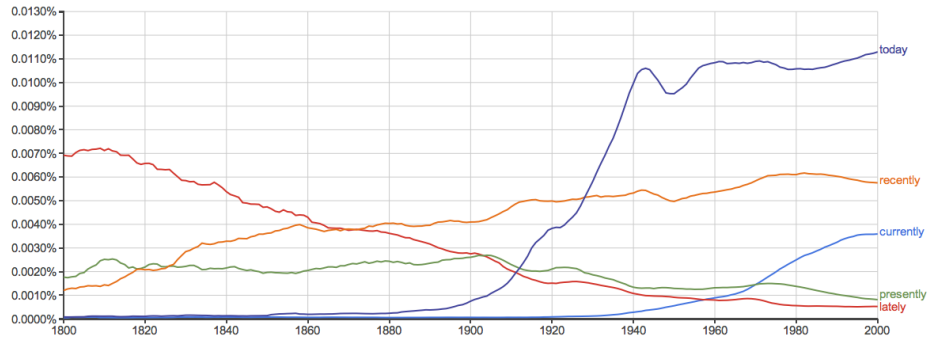

Figure 2. The frequency of the five words representative of *recency* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

The scores for the items representing the second value, i.e. *recency*, treating of the media's preference for breaking news, are more diverse. The frequency of three words increases quite rapidly since 1800: in case of *today* it rises over 130 times, in case of *recently* – almost five times, and in case of *currently* – over 70 times. On the other hand, the frequency of the other two items falls from 0.0069% to 0.0005% in case of *lately* and from 0.0018% to 0.0008% in case of *presently*.

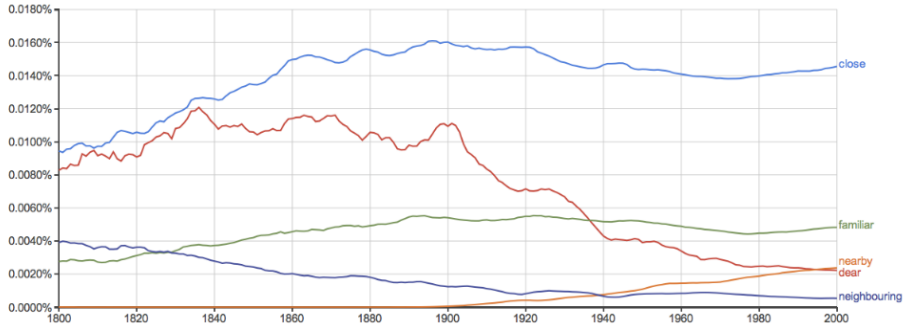

Figure 3. The frequency of the five words representative of *proximity* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

Likewise, the results for *proximity*, which relates to the closeness (either geographical or in terms of values) of the occurrence to the readers, are diverse. Three values rise and two fall. The most rapid change concerns the word *nearby* whose frequency increases over 200 times. The relative change in case of *close* is 35% and in case of *familiar* – 71%. The frequency of *dear* decreases from 0.0083% in 1800 to 0.0022% in 2000, exhibiting relative change of -73% and the frequency of *neighbouring* changes by -87%.

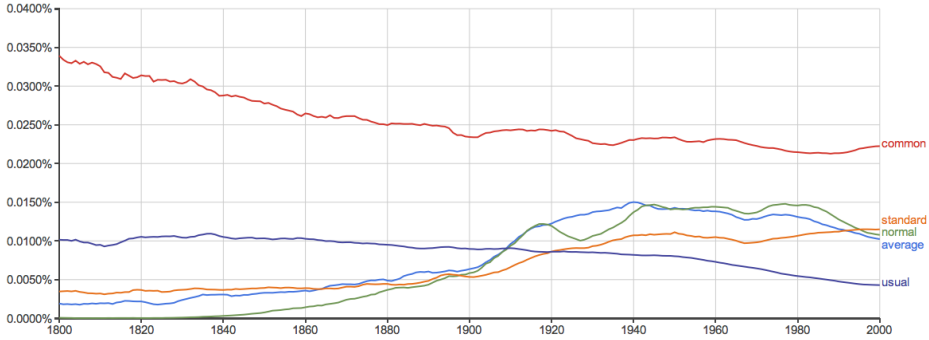

Figure 4. The frequency of the five words representative of *consonance* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

The words relevant to *consonance*, a value referring to high newsworthiness of occurrences following regular, familiar patterns, provide a similar model: the frequency of three items rises, and the frequency of two words decreases (by over one third in case of *common* and by half in case of *usual*). The most rapid change can be noted with *normal*, whose frequency increases over 100 times. The relative changes of frequency of *standard* and *average* are also substantial: 229% and 437% respectively.

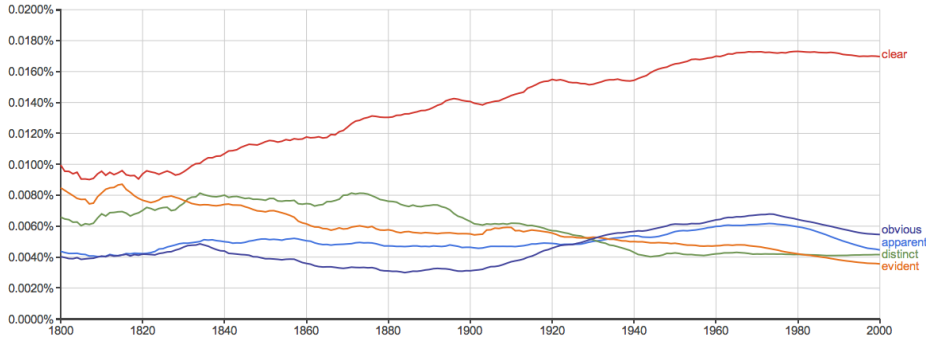

Figure 5. The frequency of the five words representative of *unambiguity* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

The next news value studied was *unambiguity*, which refers to the media's preference of clarity of the information and interpretation (preferably limited to one) of events. The biggest change was noted for the word *clear*. Its frequency rises as many as 17 times in the studied period. The frequency of *obvious* increases from 0.0040% to 0.0054% (relative change 35%) and the frequency of both

distinct and *evident* decreases. The relative change in the first case is -38% and in the second -59%. The frequency of *apparent* does not change significantly.

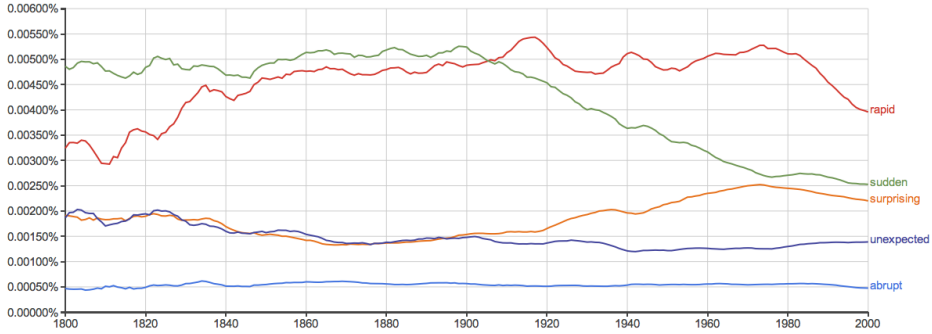

Figure 6. The frequency of the five words representative of *unexpectedness* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

As far as *unexpectedness* is concerned, none of the items exhibit a dramatic change in frequency. As shown in Figure 6 *abrupt*, *unexpected*, *surprising* and *rapid* note similar values in 1800 and in 2000, differing only slightly (the relative changes are: 2% for *abrupt*, -26% for *unexpected*, 16% for *surprising* and 22% for *rapid*). The frequency of *sudden* falls (-49%). The feature clearly stands in opposition to one of the previously mentioned news values, i.e. *consonance*, as it refers to extraordinary events.

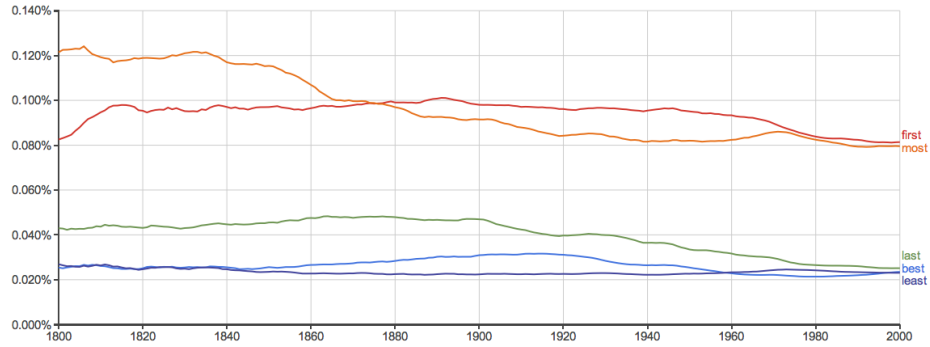

Figure 7. The frequency of the five words representative of *superlativeness* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

Surprisingly, the frequency in the observed lexical items representative of *superlativeness*, which refers to high newsworthiness of the most spectacular occurrences, does not rise at all. It falls substantially in case of *last* (-41%) and

most (-34%). It also decreases in the other cases: the relative change in the frequency of *first* in the studied period is -1%, of *best* it is -7% and of *least* it is -14%.

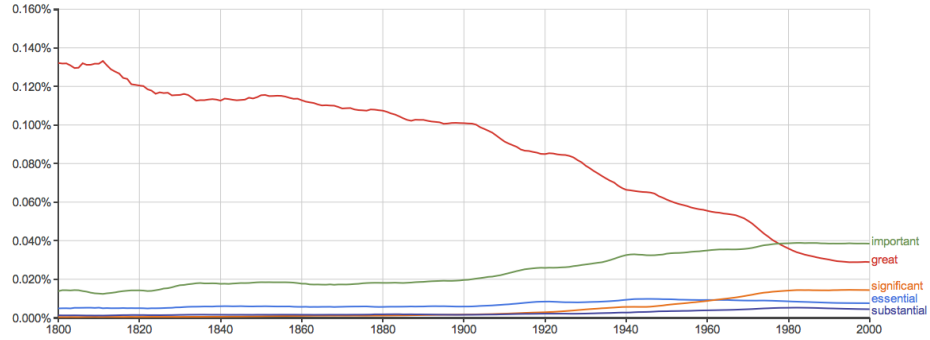

Figure 8. The frequency of the five words representative of *relevance* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

Apart from *great*, whose frequency falls between 1800 and 2000 (relative change -78%), all words representative of *relevance*, a value referring to high newsworthiness of occurrences important to the readers, rise. The relative change in the frequency of *essential* is 53%, and the frequency of *important* and *substantial* increases over three times and of *significant* as many as over 35 times.

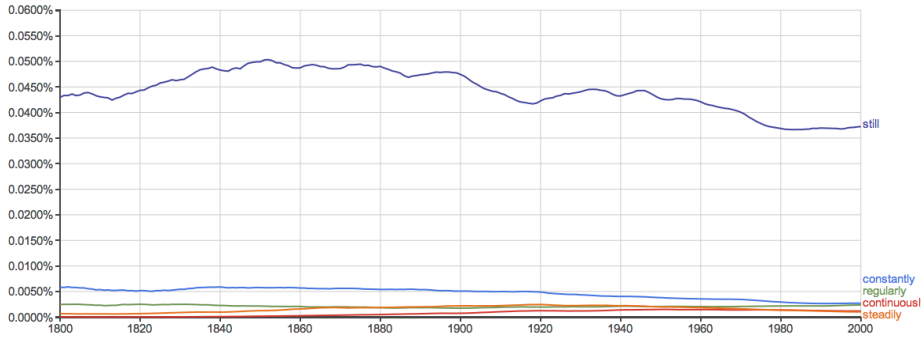

Figure 9. The frequency of the five words representative of *continuity* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

The next news value studied was *continuity*, a value underlining the relevance of information referring to previous news. The frequency of the most common word representative of this value, i.e. *still*, does not change much in the studied period. The relative change in frequency of the word is -13%. The frequency of *regularly*,

whose relative change is -4% does not fall significantly either. However, the frequency of *constantly* decreases by half and the frequency of *continuously* rises over 60 times. The frequency of *steadily* also increases (relative change 67%).

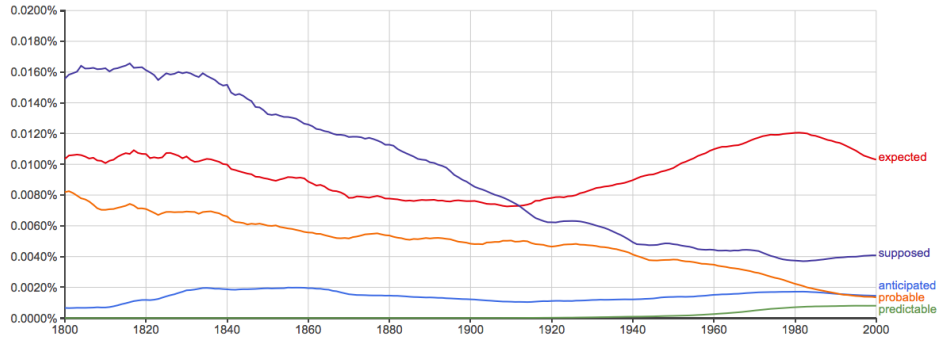

Figure 10. The frequency of the five words representative of *predictability* from 1800 to 2000 (prepared with the use of Google Books Ngram Viewer and retrieved from https://books.google.com/ngrams)

Among the words illustrated in Figure 10 and representative of *predictability*, a value referring to high newsworthiness of occurrences that are easy to foretell, the most dramatic change in frequency was observed in *predictable*. Its frequency rises over 880 times. Another word whose frequency increases is *anticipated* (over two times). The frequency of *expected* in 2000 is very similar to that in 1800 (-0.5%) and the frequency of *probable* and *supposed* falls from 0.0082% to 0.0013% and from 0.0156% to 0.0040% respectively.

To confirm that the frequency of selected words corresponds the socio-cultural change in question we should expect an increase in all or at least most of the items. Yet, as illustrated in Table 2 an increase in the frequency of the lexical items is noted only in 20 instances (which make up 40% of all lexical items). Moreover, the frequency of the selected words falls in 18 cases (36% of all items) and does not change significantly in 12 (24%). Even if we take into account the fact that the most rapid changes concern only increases (*predictable* which rose over 880 times, *nearby* over 200 times, *today* 130 times, and *normal* over 100 times), the results for the 30 items which do not present any major increase still undermine the discussed hypothesis.

Table 2. The trends in frequency change of the selected lexical items between 1800 and 2000

news value	number of words whose frequency increased	number of words whose frequency decreased	number of words whose frequency did not change significantly (less than 30%)
negativity	1	3	1
recency	3	2	0
proximity	3	2	0
consonance	3	2	0
unambiguity	2	2	1
unexpectedness	0	1	4
superlativeness	0	2	3
relevance	4	1	0
continuity	2	1	2
predictability	2	2	1
total	**20**	**18**	**12**

4. Discussion

It seems that the Google Books Ngram Viewer though providing an extensive database and enabling a fast collection of data does not give clear evidence of the influence that social changes have on word frequency. The results of the study may be surprising, as it has been argued for long that the relationship between values fostered in a society and its language is close. It seems reasonable to assume that if culture and language are linked, one should have an impact on the other.

Undoubtedly, Google Books Ngram Viewer will find application in linguistic studies. Easy access to digitalised texts offers incomparable opportunities in lexicography, chronologization of units of language and datation of textual objects (area thoroughly studied by Wierzchoń (2008)). Its latest feature, i.e. 'wildcards', used for retrieving popular collocations can be beneficial both to foreign language teaching (e.g. the writing component), and translation, as finding the right collocations improves the naturalness of texts.

Additionally, the tool provides information on the popularity of topics of discussion in the digitalised material, yet it does not explain why the values increase or decrease. An example could be the word *family*, whose frequency rose from 0.02% to 0.03% in the studied period, even though it seems that the declining marriage rates, lower number of children being born and a growing number of divorces could suggest a devaluation of this institution. In this case one might assume that such a fundamental change is worth many a discussion and hence the increase in the frequency of the word. Certainly the data could not lead to a conclusion that there were more families in 2000 than in 1800. Moreover, the topic could also be discussed with no mention of *family*, as other words concerning this

idea could be used (*relatives*, *children*, *husband*, *wife*, etc.), which further diminishes the value of the data.

Therefore, despite the fact that the tool may be helpful in developing certain theories concerning socio-cultural phenomena, we claim that the data obtained with Google Books Ngram Viewer is not reliable enough to confirm these theories.

First, the material selected by Michel et al. includes only 4 percent of all books ever published. Even though 5 million books is a considerable number, it is only a fraction of all printed texts and hence inferences should not be drawn on this basis, as they could lead to false statements.

It also appears that Google Books Ngram Viewer does not take into account the different contexts in which the analysed words are set in, even though such contexts seem essential in any study concerning semantics. Contexts carry meaning. The fact that the frequency of a word rises does not necessarily mean that the concept is valued more, but, as mentioned above, that it is discussed extensively.

Omitting context means ignoring various lexical senses of words. Greenfield states that the decrease in the frequency of the word 'give' is symptomatic of a social change from gemeinschaft into gesellschaft values. However, if we type into the program the phrases: *give back*, *give away*, *give priority*, *give a hand* or *give birth* the trend in frequency is actually rising. Greenfield claims that selecting words with narrow range of semantic interpretations prevents incorporating into the study words in contexts irrelevant to the analysis. However, it seems that all instances should be taken into account to ensure the reliability of the study, especially as the proposed selection would be random unless done manually, which appears implausible. Moreover, it appears that the narrower the range of semantic interpretations the lower the frequency of the word, which in turn affects the analysis, as changes in low-frequent words could seem less meaningful. For the same reason Greenfield chooses high frequency words for her analysis, and for the same reason it seems difficult to explain the observed trends in word frequencies that occurred in our study of the lexical items representative of news values. The fact that the most rapid changes concerned only increases seems meaningful, but without the contexts and vast etymological knowledge we are unable to determine the cause. It is possible that the rising influence of media on societies, culture and language plays a role, but it might as well be caused by other factors: political, economic, linguistic or psychological.

Admittedly, Google Books Ngram Viewer enables viewing the excerpts from which the analysed words come, however, as collecting such data has not been automated yet, and would have to be done manually for all 50 words in millions of contexts, it seems implausible to incorporate such information into the study, even if for reasons of time and space.

As Lakoff (2013) states in her criticism towards the approach examined in this paper, we are rarely able to say whether the changes in frequencies carry meaning or are just accidental. She claims that even though there are words whose

appearance, or increase in frequency, can be easily explained by the socio-cultural phenomena, such words are scarce and usually limited to technological innovation or political transformations.

Nevertheless, Lakoff agrees that even though the presence of most words and the changes in their frequency do not tell much about the values ascribed to certain phenomena it may be a sign of recognition of a problem. A case in point might be the appearance, and/or increasing frequency of words such as *homophobia*, *racism* or *sexism*. As Lakoff aptly notes, the fact that these terms were not used (or used incidentally) in the 19th and early 20th century does not mean that the phenomena did not exist, only that in the 20th century they were noticed and became worthy of naming and changing. And indeed, the original studies presented in the article in Science concern such terms. The occurrence in the corpus and the significance of the fluctuations in frequencies of words as e.g. *netiquette* or *World War I*, as well as names of well-know people, seem self-explanatory and therefore may entice researchers to apply data obtained via Google Books Ngram Viewer to more complex studies.

However, one should be circumspect in such undertakings especially as Hilpert and Gries (2009) warn that since the trends in frequencies are rarely unidirectional or strong enough to be intuitively clear, a statistical measure that would help determine if the observed frequencies differ from the mean more than it could be expected, should be incorporated in more complex studies.

It could be concluded in Lakoff's words that the relationship between language and the reality it refers to is complicated and difficult to embrace by merely following changes in word frequencies. The study of single lexical items (or even phrases or sentences) answers the questions posed by researchers interested in the relationship between word frequency and specific socio-cultural phenomena only partially. Not only does it fail to address the context, but also omits the meaning conveyed at the text level. To judge whether certain phenomena are represented in a language or to follow trends one should perform a thorough analysis incorporating whole texts, not just single words, into the study. Employing the examples provided above, i.e. *homophobia*, *racism* or *sexism*, we should see that a text with homophobic, racist or sexist contents would rarely include the words naming the phenomena. Therefore, if the researcher studying these phenomena took under consideration only the frequency of these items, the results would be far from reliable, as the meaning of texts may be implicit. It may lie in metaphors, intertextual allusions or even images.

Therefore, the conclusions drawn from the study by Greenfield even though reasonable, seem far-fetched as the research is based on few words indexing the contrasting values and, more importantly, it does not take into account the contexts in which the words occurred, nor does it include any sort of thematic analysis based on texts as units thereof.

5. Conclusion

It is a fact, that a research based on a 361 billion word corpus is prone to produce valuable results. It should help answer the questions posed by scholars both in humanities and social sciences and show the co-dependencies between certain phenomena and language at the lexical level or reveal the patterns of grammatical changes in irregular forms of verbs (as suggested by Michel et al. (2010)).

Google Books Ngram Viewer enables the researcher to put word frequency in a historical context as it shows the changes in the frequency of any selected word or group of words in time. Furthermore, its digital form is yet another reason to acknowledge it as a valuable tool as it allows the researcher to spend more time on the analysis of data than on their collection.

Nonetheless, the usage of Google Books Ngram Viewer should be limited to uncomplicated studies related to word frequency. It cannot be treated as the only tool in a research into complex socio-cultural transformations, as it does not provide extensive information on the contexts in which the words occurred and may lead to superficial, inaccurate, or less precise descriptions of studied phenomena if not confronted with a comprehensive textual analysis.

Our goal was to verify if Google Books Ngram Viewer is indeed a valuable tool in socio-cultural research as suggested both by its creators and Greenfield. Thus, we decided to perform a study similar to Greenfield's *Ecological Analysis* and follow the trends in changes in word frequency throughout the 19th and 20th centuries to observe if these changes correspond to one of the major socio-cultural transformations that took place in the studied period, i.e. mediatization. The data obtained in the course of the study suggest that the changes in word frequency do not directly depend on the rise of the role of the news values in modern societies. Additionally, a close examination of the methodology suggested by Greenfield demonstrated its shortcomings: most importantly disregarding the importance of contexts in which the words occurred and their different semantic interpretations. It seems that in a research of the relationship between socio-cultural transformations and word frequencies, a multifaceted study based on etymological considerations and incorporating thematic analysis should be performed. So far a tool enabling such an undertaking on a scale of millions or even billions of texts remains unknown.

References

Alcock, Joe. 2012. Emergence of Evolutionary Medicine: Publication Trends from 1991-2010. *Evolutionary Medicine, 1*. doi:10.4303/jem/235572

Atkins, Sue. 2010. The DANTE Database: Its Contribution to English Lexical Research, and in Particular to Complementing the FrameNet Data. In: Gilles Maurice de Schryver (ed.), *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks*, 267-97. Kampala: Menha Publishers.

Atkinson, Maxine P. and Stephen P. Blackwelder. 1993. Fathering in the 20th Century. *Journal of Marriage and the Family*,*55*(4), 975–986.

Bell, Allan. 1991. *The Language of News Media*. Oxford: Blackwell Publishers Ltd.

Berelson, Bernard. 1971 [1952]. *Content Analysis in Communication*. New York: Hafner Publishing Company.

Berry, David M. 2012. The Social Epistemologies of Software. Social Epistemology: A Journal of Knowledge. *Culture and Policy, 26*(3-4), 379–398. doi:10.1080/02691728.2012.727191

Cabrera, Natasha, Tamis-LeMonda, Catherine S., Bradley, Robert H., Hofferth, Sandra, & Michael E. Lamb. 2000. Fatherhood in the twenty-first century. Child development, 71, 127–136. doi: 10.1111/1467-8624.00126

Carroll, John B., Davies, Peter and Barry Richman. 1971. *The American Heritage Word Frequency Book*. Boston: Houghton Mifflin.

Castells, Manuel. 1996. *The Rise of the Network Society, The Information Age: Economy, Society and Culture*. Malden, Oxford: Blackwell.

Chow, Esther Ngan-ling. 2003. Gender Matters Studying Globalization and Social Change in the 21st Century. *International Sociology*, *18*(3), 443–460.

Cockerill, Kristan. 2013. A Failure Reveals Success. *Journal of Industrial Ecology*, *17*, 633–641. doi: 10.1111/jiec.12049

Cowan, Ruth Schwarz. 1976. The "Industrial Revolution" in the Home: Household Technology And Social Change in the 20th Century. *Technology and Culture*, *17*(1), 1–23.

Crasto, Chiquito J. 2011. Bioinformatics for Biological Researchers – Using Online Modalities. In: Eta Berner (ed.), *Informatics Education in Healthcare*, 147–165. Birmingham: Springer.

Davies, Mark. 2005. The Advantage of Using Relational Databases for Large Corpora: Speed, Advanced Queries, and Unlimited Annotation. *International Journal of Corpus*, *10*(3), 307–334. doi:10.1075/ijcl.10.3.02dav

Davies, Mark. 2010. The Corpus of Contemporary American English as the First Reliable Monitor Corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–465. doi:10.1093/llc/fqq018

Davis, Mark. 2014. Making Google Books n-grams Useful for a Wide Range of Research on Language Change. *International Journal of Corpus Linguistics 19*(3), 401–16.

Edmunds, June and Bryan S. Turner. 2005. Global Generations: Social Change in the Twentieth Century. *The British Journal of Sociology*, *56*, 559–577. doi: 10.1111/j.1468-4446.2005.00083

Fellbaum, Christiane. 2005. WordNet and Wordnets. In: Keith Brown (ed.), *Encyclopedia of Language and Linguistics, Second Edition*, 665–670. Oxford: Elsevier.

Fuchs, Christian. 2008. *Internet and Society: Social Theory in the Information Age*. London: Routledge.

Greenfield, Patricia M. 2013. The Changing Psychology of Culture From 1800 Through 2000. *Psychological Science*, *24*(9), 1722-1731. doi:10.1177/0956797613479387

Grigonyte, Gintare, Rinaldi, Fabio and Martin Volk. 2012. Change of Biomedical Domain Terminology Over Time. In: Arvi Tavast, Kadri Muischnek and Mare Koit (eds.), Human Language Technologies – The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012 (Vol. 247). IOS Press.

Hill, Felix. 2012. *Beauty Before Age?: Applying Subjectivity to Automatic English Adjective Ordering*. Proceedings of the NAACL HLT '12 2012 Student Research Workshop, 11–16. Stroudsburg, PA: Association for Computational Linguistics.

Hilpert, Martin and Stefan Gries. 2009. Assessing Frequency Changes in multistage Diachronic Corpora: Applications for Historical Corpus Linguistics and the Study of Language Acquisition. *Literary and Linguistic Computing*, *24*(4), 385–401. doi: 10.1093/llc/fqn012

Hjarvard, Stig. 2008. The Mediatization of Society. A Theory of the Media as Agents of Social and Cultural Change. *Nordicom Review*, *29*(2), 105–134.

Hjarvard, Stig. 2013. *The Mediatization of Culture and Society*. Oxon: Routledge.

Hsieh, Hsiu-Fang and Sarah E. Shannon. 2005. Three Approaches to Qualitative Content Analysis. *Qualitative Health Research*, *15*(9), 1277–1288.

Johnson, Clay A. 2011. *The Information Diet: A Case for Conscious Consumption*. Beijing, Cambridge, Tokyo: O'Reilly.

Kesebir, Pelin and Selin Kesebir. 2012. The Cultural Salience of Moral Character and Virtue Declined in Twentieth Century America. *Journal of Positive Psychology*, *7*(6), 471–480.

Krippendorff, Klaus. 1980. *Content Analysis: An Introduction to its Methodology*. London: Sage.

Kumar, Nitu and Manish Sahu. 2010. The Evolution of Marketing History: a Peek Through Google Ngram Viewer. *Asian Journal Of Management Research*, *1*, 415–426.

Lakoff, Robin. 2013. *What Words Don't Tell Us*. Retrieved May 20, 2014 from http://blogs.berkeley.edu/author/rlakoff/

LaRossa, Ralph, Gordon, Betty A., Wilson, Ronald J., Bairan, Annette and Charles Jaret. 1991. The Fluctuating Image of the 20th Century American Father. *Journal of Marriage and Family*, *53*(4), 987–997.

Lilleker, Darren. 2008. *Key Concepts in Political Communications*. London: SAGE

Lucier, Paul. 2012. The Origins of Pure and Applied Science in Gilded Age America. *ISIS*, *103*(3), 527–536.

Mazzoleni, Gianpietro and Winfried Schulz. 1999. "Mediatization" of Politics: A Challenge for Democracy? *Political Communication*, *16*(3), 247–261.

Michalski, Brian, Krishnamoorthy, Mukkai and Tsz-Yam Lau. 2012. *Temporal Analysis of Literary and Programming Prose*. Retrieved September 23, 2014 from Cornell University Library http://arxiv.org/pdf/1202.2131.pdf

Michel, Jean-Baptiste, Shen, Yuan Kui, Aiden, Aviva P., Veres, Adrian, Gray, Matthew K., The Google Books Team, Pickett, Joseph P., Hoiberg, Dale, Clancy, Dan, Norvig, Peter, Orwant, Jon, Pinker, Steven, Nowak, Martin A. Erez Lieberman Aiden. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, *331*(6014), 176–182.

Mowery, David C. and Nathan Rosenberg. 1998. *Paths of Innovation: Technological Change in 20th-Century America*. Cambridge: Cambridge University Press.

Murray, Denise E. 2000. Protean Communication: The Language of Computer-Mediated Communication. *TESOL Quarterly*, *34*, 397–421. doi: 10.2307/3587737

Oishi, Shigehiro, Graham, Jesse, Kesebir, Selin and Iolanda C. Galinha. 2013. Concepts of happiness across time and cultures. *Personality and Social Psychology Bulletin*, *39*(5), 559–577.

Ong, Walter J. 2002. *Orality and Literacy: The Technologizing of the Word*. London, New York: Routledge.

Phani, Shanta, Lahiri, Shibamouli and Arindam Biswas. 2012. Culturomics on a Bengali Newspaper Corpus. *International Conference on Asian Language Processing*, 237–240. doi: 10.1109/IALP.2012.68

Roseneil, Sasha and Shelley Budgeon. Cultures of Intimacy and Care beyond 'the Family': Personal Life and Social Change in the Early 21st Century. *Current Sociology*, *52*(2), 135–159.

Rutten, Ellen, Fedor, Julie and Vera Zvereva. 2013. *Memory, Conflict and Social Media*. Abingdon: Routledge.

Schoen, Robert and Vladimir Canudas-Romo. 2006. Timing Effects on Divorce: 20th Century Experience in the United States. *Journal of Marriage and Family*, *68*, 749–758. doi: 10.1111/j.1741-3737.2006.00287

Stemler, Steve. 2001. An Overview of Content Analysis. *Practical Assessment, Research & Evaluation*, *7*(17). 137–146.

Thurlow, Crispin, Lengel, Laura and Alice Tomic. 2004. *Computer Mediated Communication*. London, New Delhi, London: Sage.

Ullmann, Stephen. 1962. *Semantics: an Introduction to the Science of Meaning*. Blackwell: Oxford.

Volti, Rudi. 1988. *Society and Technological Change*. New York: St. Martin 's Press.

Weber, Robert P. (ed.). 1990. *Basic Content Analysis*. London, New Delhi, London: Sage.

Wellman, Barry, Quan-Haase, Anabel, Boase, Jeffrey, Chen, Wenhong, Hampton, Keith, Díaz, Isabel and Kakuko Miyata. 2003. The Social Affordances of the Internet for Networked

Individualism. *Journal of Computer-Mediated Communication, 8*. doi: 10.1111/j.1083-6101.2003.tb00216

Wierzchoń, Piotr. 2008. *Fotodokumentacja, chronologizacja, emendacja: teoria i praktyka weryfikacji materiału leksykalnego w badaniach lingwistycznych*. [*Photo-documentation, chronologization, emendation: theory and practice of lexical material verification in linguistic studies*] Poznań: Instytut Językoznawstwa Uniwersytetu im. Adama Mickiewicza.

Wood, Andrew F. and Matthew J. Smith. 2005. *Online Communication: Linking Technology, Identity, and Culture (Second Ed.)*. Mahwah, NJ: Lawrence Erlbaum & Associates.