Alkalmazott Nyelvészeti Közlemények, Miskolc, XIV. évfolyam, 2. szám (2019), pp. 209–221.

LINGUISTIC DIFFERENCES IN ABSTRACTS WRITTEN BY NATIVE AND NON-NATIVE SPEAKERS OF ENGLISH: A PILOT STUDY

NYELVI ELTÉRÉSEK ANGOL ÉS NEM ANGOL ANYANYELVŰ SZERZŐK ÖSSZEFOGLALÓIBAN: EGY KÍSÉRLETI TANULMÁNY

XIAOYUN LI[1] – VINCZE VERONIKA[2]

Abstract: In this paper, we analyze a corpus of abstracts collected from MA theses written by English major students in the United Kingdom, China and Hungary. We make use of several features derived from the linguistic analysis of abstracts; moreover, we also perform some machine learning experiments in order to see what features are the most useful for distinguishing the language use of the three groups of English speakers. It is revealed that it is primarily morphological and semantic features that distinguish native and non-native language use. As for the two groups of non-native speakers, Hungarians tend to use their personal viewpoint more frequently while Chinese prefer to be more objective. All in all, morphological features seem to contribute the most to the automatic distinction of the three groups of speakers.

Keywords: *second language acquisition, learner corpus, academic English, corpus linguistics*

Absztrakt: Ebben a tanulmányban nagy-britanniai, kínai és magyar diákok angol nyelven írt szakdolgozatainak absztraktjait elemezzük korpusznyelvészeti szemmel. Az összefoglalók automatikus nyelvi elemzéséből számos nyelvi jellemzőt nyertünk ki, s ezek statisztikai vizsgálatán felül gépi tanulási kísérleteket is végeztünk annak eldöntésére, hogy mely nyelvi jellemzők képesek a leghatékonyabban elkülöníteni az angol nyelvet beszélők három csoportját. Eredményeinkből kiderül, hogy leginkább a morfológiai és szemantikai jellemzők térnek el statisztikailag az anyanyelvi és nem anyanyelvi beszélők nyelvhasználatában. A kínai és magyar beszélők nyelvi sajátságainak összevetéséből pedig fény derül arra is, hogy a magyar diákok gyakrabban utalnak személyes nézőpontjukra, míg a kínaiak az objektivitást részesítik előnyben. Gépi tanulási kísérleteink pedig azt mutatják, hogy elsődlegesen a morfológiai jegyek különítik el a három beszélőcsoportot egymástól.

Kulcsszavak: *idegennyelv-elsajátítás, tanulói korpusz, angol tudományos írás, korpusznyelvészet*

[1] XIAOYUN LI
University of Szeged
Department of Theoretical Linguistics
6722 Szeged, Egyetem u. 2.
635197504@qq.com

[2] VINCZE VERONIKA
MTA-SZTE Research Group on Artificial Intelligence
6720 Szeged, Tisza Lajos krt. 103.
vinczev@inf.u-szeged.hu

INTRODUCTION

Over the past 200 years or so, English has grown to be the biggest world Lingua Franca due to the people's need for communication in a globalizing world. The consequence of the globalization of English is the emerging of a number of English varieties (LAPORTE 2012) or indigenized varieties of English. Those varieties have attracted enormous interest from researchers for a number of reasons including social, cultural, historical, spatial considerations or a combination of these. In studying English varieties, linguistic description takes the lion's share (XIAO 2009) and in a certain sense provides a basis for further socio-cultural study. As a linguistic study, the present paper contributes to that field of research by focusing on the characteristics of academic writing in English taking place in Hungary and China compared to native English (British English). Due to the limited space of this study, we limit our research scope to academic language.

In this paper, we would like to answer the following research questions:

- Are there any statistically significant differences among the academic language usages of native and non-native speakers?
- Are there any differences among the academic language usages of non-native speakers from different countries and cultures?
- What linguistic features are effective in the automatic identification of the students' native language?

For this purpose, we analyze a corpus of abstracts of MA theses written by students in the United Kingdom, in China and in Hungary. Such a focus is of great importance for three reasons: First, studies which take both the language usage of Hungary and of China into consideration are quite rare. The present study therefore is expected to enrich the comparison of the two English varieties. Second, there is a scarcity of research on the genre of the master's thesis, as well. In the literature, a large number of language studies on academic language centre around research articles (RA) (BHATIA 1993; HYLAND 2000; SWALES 1990, 2004). This study is thus expected to narrow the research gap between RA abstracts and thesis abstracts. The third reason is that compared with RA abstracts, which receive relatively more limitations on content and length of abstract and are generally internationally peer-reviewed, thesis abstracts have more freedom in content and length and are commonly reviewed by local supervisors who share the same language or cultural backgrounds with thesis authors. It is thus safe to say that thesis abstracts are in a better position in reflecting the authors' language and cultural backgrounds in comparison with RA abstracts.

In our data, the students' major was English language or related fields like English literature; hence they are supposed to be advanced speakers of English who are familiar with the norms of English academic writing. We make use of several features derived from the linguistic analysis of abstracts. Moreover, we also perform some machine learning experiments in order to see what features are the most useful for distinguishing the language use of the three groups of English speakers. We

conclude our paper with some remarks regarding the applicability of the results and some suggestions for future work.

1. THE CORPUS

The corpus we utilized in this study is self-constructed since there was no existing corpus available for using. For each language, we selected 45 abstracts of approximately 12,000 tokens, altogether reaching almost 40,000 tokens. For the Hungary subcorpus, given that public access to MA theses in most of the universities in Hungary is limited and that this is a pilot study, from the repositories of the University of Szeged we randomly selected 45 abstracts from the theses on English language or related studies. The 45 abstracts included in the China subcorpus were randomly retrieved from CNKI (China National Knowledge Infrastructure) with English language study as the search term. The remaining 45 abstracts from the United Kingdom subcorpus were downloaded from several university repositories in the same way. *Table 1* shows the basic statistical data on the corpus.

Below, we include three examples from the corpora.

An abstract from a UK student:

A small body of recent research on vocabulary explanations (VEs) in second language (L2) classrooms (e.g. Mortensen, 2011; Waring et al., 2013) has attempted to provide the sequential descriptions of the key elements of VEs and investigate how teachers draw on their linguistic and semiotic resources to construct the VE sequences (e.g. Smotrova and Lantolf, 2013). Nevertheless, more work is needed in order to allow educators to better understand how VEs are provided in L2 classrooms. In particular, there is a shortage of studies (e.g. Tai and Brandt, 2018) illustrating the nature of VEs in beginning-level English for Speakers of Other Languages (ESOL) classrooms, where learners all share different first languages (L1s) and have limited English proficiency. Moreover, the shared linguistic resources between the teacher and learners are typically limited in beginning-level ESOL classrooms. To date, there is no longitudinal study which will allow for tracking the impact of VEs on contributing to learners' conceptual understandings of the meanings of target vocabulary items. The vast majority of the studies, which identified learners' display of understanding of L2 word meanings in classroom interactions, were based on one-off analyses of the classroom discourse (e.g. Waring et al., 2013). This prevents educators and researchers from observing the learner's change of conceptual understandings over time. This MSc dissertation contributes to the identified research gaps by employing Conversation Analysis (CA) to 1) investigate the nature of VEs in a beginning-level ESOL classroom and 2) conduct a 4-month longitudinal analysis to explore the potential for employing CA as the methodological tool for tracking learners' development of the conceptual understandings of the meanings of particular vocabulary items which are previously explained. The classroom data is taken from a corpus of video-data collected in a beginning-level adult ESOL classroom in the United States. The key findings demonstrate that other than verbal

sources, teacher's use of embodied resources in explaining vocabulary items in the classroom plays an important role in facilitating the learners' understandings of the meanings of different vocabulary items. The learner's use of gestures allows her to externalise her understandings of the L2 word meanings and also allows teachers to evaluate the learner's current knowledge states. These findings also suggest that CA provides some, albeit incomplete, evidence of the learner's developing conceptual understandings of L2 word meanings and it allows researchers to investigate how these developmental changes occur in each interactional context of L2 vocabulary use.

An abstract from a Chinese student:

China is playing an increasingly important role on the international stage. News is a major channel for the world to understand China's development – and news translation develops rapidly in recent years and is crucial for China's international exchange and publicity.

In this report – the author, based on her own online news translation practice – explores methods of translating the headline and body of Chinese news under the guidance of translation variation theory. Chinese and English news reports differ in their headlines and bodies. Chinese news reports usually have headlines that convey the complete message of the reports and start with background information, while English news reports have headlines that grasp the gist and start with the main idea of the reports. Taking these differences into account – the translator chooses to adopt various methods of translating under the guidance of translation variation theory in translating Chinese news into English. Moreover, online news reports have to be posted within limited time – rendering full translation inefficient – and even unnecessary in the case of translating Chinese news into English. Compared with full translation, translating not only contributes to intercultural communication, but also meets the requirements on timeliness and efficiency of news reports translation. The author believes that in translating Chines news the translator has to consider the difference between Chinese and English news reports – respect the need and expectation of foreign readers – and try to produce smooth and natural translation by translating.

An abstract from a Hungarian student:

The 'porpoise' of this work is to show the famous and fascinating world of Charles Dodgson in his works Alice in Wonderland and Through the Looking Glass.

Firstly, I introduce the author, look into how his pen-name, Lewis Carroll, was created and take a closer look at the importance of his Child-friends.

In the next step, I discuss Charles Dodgson's first meeting with Alice Liddell and how this meeting was marked as important in further writings.

In the following part I show how the main character represents the Victorian era child, I discuss ways in which Wonderland and Looking Glass are in fact using Bildungsroman through Alice's character to show the passage from childhood to adulthood, and I also examine the beginnings of the two books.

In the next part of my work, I show observations and criticism by the author of Victorian education, morality, logic and how they are represented in both books.

I then examine how and why Alice in Wonderland and Through the Looking Glass were created, look at the characters, and also compare them, based on biographies, with how the author represents himself as many characters in each work.

In the final part I analysis Victorian society in the world represented by the author.

I conclude my thesis establishing that the Alice stories were far more than children's tales. The purpose of this thesis is to demonstrate and support this idea by using references taken from biographies of the author and comparing them with both books Alice in Wonderland and Through the Looking Glass.

Table 1

Basic statistical data on the corpus

Subcorpus	**Documents**	**Sentences**	**Tokens**
China	45	601	14,402
Hungary	45	468	12,510
United Kingdom	45	478	12,600
Total	135	1547	39,512

2. EXPERIMENTS

In our experiments, we employed a rich feature set extracted from the abstracts and the results of the automatic linguistic analyses (lemmatization, morphological and dependency parsing) performed with the tool UDpipe (STRAKA – STRAKOVÁ 2017). Altogether, the feature set consisted of 81 features, listed below.

We extracted basic statistical features, namely:

- The number of sentences;
- The number and frequency of tokens;
- The number of words;
- The number and frequency of distinct lemmas compared to the number of words;
- The average sentence length.

As for morphological features, we extracted the following features:

- Part-of-speech features:
  - The number and frequency of nouns, verbs, adjectives, pronouns, numerals, adverbs, proper nouns and conjunctions;
  - The number of punctuation marks;
  - The number and frequency of words that could not be analyzed by the POS tagger, i.e. those with an "unknown" POS tag.
- Deep morphological features:
  - The number of first person singular pronouns;

- The number of first person plural pronouns;
- The number and frequency of past and present tense verbs;
- The number and frequency of demonstrative pronouns.

As for syntactic features, we extracted the following characteristics:

- The number and frequency of active and passive subjects and objects;
- The number and frequency of attributes and adverbials;
- The number and frequency of coordination occurrences.

We also carried out an analysis of the following semantic features of the texts:

- The number and frequency of fillers and uncertain words compared to the total number of tokens;
- The number and frequency of words belonging to several classes of linguistic uncertainty based on Vincze (2014);
- The number and frequency of words belonging to the emotions described in Mohammad (2017);
- The number and frequency of negation words;
- The frequency of content words and function words;
- The number and rate of private, public and suasive verbs (QUIRK et al. 1985).

In the following, we describe the analysis of these linguistic features from a statistical point of view and some machine learning experiments based on these features.

2.1. Statistical analysis of data

In order to examine what features act as characteristics of each group of abstracts, we carried out a statistical analysis of the features (pairwise t-tests for each pair of groups, as well as for native- non-native abstracts). The significant p values are listed in *Tables 2* and *3*.

Table 2

Significant statistical and morphological features

	China-Hungary	**China-UK**	**Hungary-UK**	**native-nonnative**
Statistical features	0.0151			
number of tokens	0.0011	0.0048		
number of sentences	0.0191	0.0005		0.0092
rate of lemmas				
Morphological features				
number of unknown words	0.0004		0.0050	
rate of unknown words	0.0013		0.0190	
number of nouns	<0.0001	0.0013		
number of adjectives	<0.0001	0.0252		
number of pronouns	<0.0001		0.0017	
number of conjunctions	0.0431	0.0354		
number of numerals	0.0018		<0.0001	0.0061

	China-Hungary	**China-UK**	**Hungary-UK**	**native-nonnative**
number of punctuation marks	0.0007		0.0098	
rate of verbs	0.0374			
rate of nouns	<0.0001	0.0004		
rate of adjectives	0.0001		0.0128	
rate of pronouns	<0.0001	0.0072	0.0002	
rate of numerals	0.0398	0.0203	0.0001	0.0006
number of Sg1 pronouns	<0.0001	0.0027	0.0002	
number of present tense verbs		0.0118		0.0295
rate of past tense verbs		0.0172		0.0377
rate of Sg1 pronouns	<0.0001	0.0048	0.0001	0.0460
number of demonstrative pronouns		0.0073	0.0007	0.0008
rate of demonstrative pronouns	0.0032		0.0109	

Table 3

Significant syntactic and semantic features

	China-Hungary	**China-UK**	**Hungary-UK**	**native-nonnative**
Syntactic features				
number of objects	0.0145	0.0191		
number of attributes	0.0002	0.0260		
number of coordination occurrences	0.0093	0.0046		
rate of subjects	0.0006		0.0396	
Semantic features				
number of negation words	0.0398		0.0244	
rate of negation words	0.0215		0.0140	
number of weasel words		0.0154		0.0337
number of peacock words		0.0042		0.0040
number of hedge words	0.0390	0.0183		
rate of epistemic words		0.0037		0.0186
rate of investigation words		0.0039		0.0055
rate of weasel words			0.0059	0.0117
rate of peacock words		0.0168	0.0234	0.0081
number of anger words	0.0275			
number of sorrow words				0.0176
rate of joy words			0.0389	0.0386
rate of fear words	0.0396			
rate of anger words	0.0195			
rate of content words	<0.0001	0.0172	<0.0001	
rate of function words	<0.0001	0.0155	<0.0001	
number of public verbs	<0.0001	0.0446	0.0062	
number of suasive verbs		0.0141		
rate of private verbs		0.0156		
rate of public verbs	0.0001		0.0144	
rate of suasive verbs		0.0048		0.0464

As the above tables illustrate, almost all linguistic features exhibit significant differences among the groups of speakers. It is also salient that the native - non-native differences mostly cover the morphological and semantic levels of the language, meaning that syntax does not seem to be a distinctive factor here. All these significant differences will be analyzed in detail in Section 4.

2.2. Machine learning experiments

We also carried out some machine learning experiments on the data in order to check how effectively machine learning methods can identify the native language of the author of the abstracts. We implemented the above features and we trained an SVM model (CORTES AND VAPNIK 1995) on the data, using Weka's (HALL et al. 2009) default settings, applying tenfold cross validation. As an evaluation metric, we used accuracy score and precision, recall and F-measure per class. We use majority labeling as a baseline result, which yields an accuracy score of 33.33%, i.e. a third of the data can be correctly identified.

First, we trained our system with all the features, which resulted in an accuracy score of 67.71% with SVM. This result is well above our baseline (33.33%). Then we wanted to examine what the effect of each feature group can have on the results. Thus, we retrained the system without a specific group of features and we compared the results obtained in this way to those provided by applying all of the features.

As the results in *Table 4* show, we can distinguish the three groups with an accuracy of 67.71% and an F-score of 67.4, which is well above the baseline. We also wanted to examine the added value of each feature set separately, hence we carried out an ablation analysis, rerunning the experiments with the omission of one specific feature set at a time. The ablation analysis (see *Table 5*) highlights the importance of morphological features, since their aggregated value is over 8 percentage points concerning the F-score. Nevertheless, the semantic and statistical features seem to improve performance by 1-2 percentage points, the only exception being the syntactic features: the overall performance is harmed, as is the case for the Chinese data, but their individual contribution is visible in the case of the UK and Hungarian datasets. Thus, it appears that syntactic features are less valuable in identifying the Chinese native speakers.

Analyzing the effects of each feature group for each group of students separately, it can be seen that morphological features are especially important for identifying abstracts written by Hungarian and UK students. On the other hand, syntactic and semantic features seem to be insignificant for identifying the Chinese abstracts, which probably means that there are no extraordinary syntactic and semantic features that are characteristic of this group of data from a machine learning point of view. Finally, statistical features appear to be less important for distinguishing Hungarian data as removing the statistical features does not result here in a loss of efficiency, compared to the other two sets of data.

Table 4

Precision, recall and F-measure for identifying the three groups of speakers

Class	Precision	Recall	F-Measure
China	0.707	0.644	0.674
Hungary	0.689	0.689	0.689
UK	0.633	0.689	0.66
all	0.676	0.674	0.674

Table 5

Results of ablation analysis

Results without statistical features				**Difference**		
Class	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
China	0.675	0.6	0.635	-0.032	-0.044	-0.039
Hungary	0.696	0.711	0.703	0.007	0.022	0.014
UK	0.612	0.667	0.638	-0.021	-0.022	-0.022
all	0.661	0.659	0.659	-0.015	-0.015	-0.015
Results without morphological features				**Difference**		
Class	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
China	0.636	0.622	0.629	–0.071	–0.022	–0.045
Hungary	0.605	0.578	0.591	–0.084	–0.111	–0.098
UK	0.542	0.578	0.559	–0.091	–0.111	–0.101
all	0.594	0.593	0.593	–0.082	–0.081	–0.081
Results without syntactic features				**Difference**		
Class	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
China	0.756	0.689	0.721	0.049	0.045	0.047
Hungary	0.674	0.689	0.681	–0.015	0	–0.008
UK	0.625	0.667	0.645	–0.008	–0.022	–0.015
all	0.685	0.681	0.682	0.009	0.007	0.008
Results without semantic features				**Difference**		
Class	**Precision**	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
China	0.714	0.667	0.69	0.007	0.023	0.016
Hungary	0.7	0.622	0.659	0.011	–0.067	–0.03
UK	0.547	0.644	0.592	–0.086	–0.045	–0.068
all	0.654	0.644	0.647	–0.022	–0.03	–0.027

3. DISCUSSION

As for the main differences between native vs. non-native language users, we could see that native-speaking writers use more past tense verbs than the non-native speakers, which might reflect a difference in how the function of the abstract is regarded: for natives, it is a summary of the work already done while for non-natives, the abstract appears to be seen as an introduction to the MA thesis, where students report what they are going to write about. Demonstrative pronouns are also preferred by native speakers, which might strengthen the coherence of the abstract by adding more co-referential elements. Native speakers also employ more investigation words, i.e. words related to studying, investigating and exploring certain phenomena (such as *analyze*, *explore*, *investigate* etc.), which suggests that UK students more

explicitly state what their research question is while the abstracts of these two groups of non-native students convey this information more implicitly.

As for uncertainty cues, there are fewer weasels, hedges and peacocks in the native abstracts than in the other two groups. Events with no obvious sources are called weasels in Wikipedia (GANTER – STRUBE 2009) while hedges blur the exact meaning of some qualities or quantities (LAKOFF 1973). Words that express unprovable qualifications or exaggerations are called peacock by Wikipedia editors. The lack of such linguistic devices in native abstracts again might indicate that natives present their results more confidently whereas non-natives are a bit more cautious when reporting their results obtained. However, non-natives often employ peacock, i.e. subjective devices as the following excerpt from a Hungarian student's work illustrates:

> *I found **very interesting** that there are more than ten years between the two works, but the similarity among them is **very surprising**, not just in the plot but in the characters, too.*

As shown above, it is primarily morphological and semantic features that distinguish native and non-native language use, which is in harmony with the results of the ablation analyses: those two sets of features seem to be most essential in identifying the mother tongue of the students.

Focusing on the Hungarian data, it can be seen that there are fewer nouns but more pronouns in this subcorpus than in the other two. This might be in connection with the fact that Hungarian students employ more function words and fewer content words than the other two groups of students. As conjunctions, pronouns, linking words and the like help connect different parts of the text, the logical structure and the internal coherence of abstracts seem to achieve a high level here. Hungarians use more emotion words in their abstracts whereas there are fewer public words and more private verbs, which might suggest a cultural difference between Chinese and Hungarian students: Hungarians tend to use their personal viewpoint more frequently while Chinese prefer to be more objective. However, there are more adjectives and adverbs in the Chinese subcorpus than in the Hungarian subcorpus, meaning that Chinese students add more details and circumstances to the description of their work, while Hungarian students seem to focus just on the essential points, without paying too much attention to the details. As a Chinese student, who happens to analyse academic abstracts in his or her work, states:

> *An abstract, as a **fully self-contained**, capsule description of a research, also plays an **indispensable** role in MA thesis, so graduate students need to try their utmost to compose **meaningful**, **logical**, and **clarified** abstracts for their theses.*

Special attention should be paid to the use of first person singular pronouns as there are significant differences among the three groups of authors here. Hungarian authors seem to use the highest number of such pronouns whereas Chinese students

use the fewest of them, native speakers being in the middle. Just to cite a Hungarian example:

> *In **my** effort, **I** would argue that behind the narrative voice there is a real presence of the author. From **my** point of view, Murdoch's way of impersonating a male narrator serves her as possibility to get rid of any categorization of being a 'woman writer' […].*

An example of "impersonification" from a Chinese abstract:

> ***It is concluded** that these findings would shed some light on future study on the syntactic acquisition of second language. **It is also hoped** that the findings of this study would provide some pedagogical implications for the purpose of improving the syntactic teaching in China.*

All this might be explained by sociocultural norms: Hungarian students seem to emphasize what they achieved and what their contribution is, the individual achievements being in the focus, while in China, the results are told in a highly impersonal way, the individual remaining in the background (see also the case of public and private verbs, mentioned above).

CONCLUSIONS

In this paper, we attempted to distinguish three classes of novice writers of academic English, i.e. native English speakers in the UK, speakers from China and speakers from Hungary, based on their MA thesis abstracts. We employed statistical significance tests as well as machine learning methods while relying on several linguistic features. Results showed that it is primarily morphological and semantic features that distinguish native and non-native language use. As for the two groups of non-native speakers, Hungarian writers tend to use their personal viewpoint more frequently while Chinese writers prefer to be more objective. All in all, morphological features seem to contribute the most to the automatic distinction of the three groups of speakers.

Our findings can be applicable in several fields. For instance, native and non-native differences in language use might be pointed out in English academic writing classes, thus having implications in language teaching. From a natural language point of view, these differences might be applied in authorship attribution and plagiarism detection, where the task is to identify the author of certain texts. As future work, we would like to extend our corpus with more material: on the one hand, more abstracts from the speaker groups examined here, on the other hand, we would like to add abstracts from speakers of other languages too. Finally, we would also like to improve our automatic methods to identify the native language of the author of the texts.

REFERENCES

[1] Bhatia, Vijay K. (2014). *Analysing Genre: Language Use in Professional Settings*. Routledge.

[2] Cortes, Corinna; Vapnik, Vladimir (1995). *Support Vector Networks* volume 20, Kluwer Academic Publishers.

[3] Ganter, Viola; Strube, Michael (2009). Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pp. 173–176, Suntec, Singapore. Association for Computational Linguistics.

[4] Hall, Mark; Frank, Eibe; Holmes, Geoffrey; Pfahringer, Bernhard; Reutemann, Peter; Witten, Ian H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations* 11 (1), pp. 10–18.

[5] Hyland, Ken (2004). *Disciplinary Discourses. Social Interactions in Academic Writing*. University of Michigan Press.

[6] Lakoff, George (1973). Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic* 2 (4), pp. 458–508.

[7] Laporte, Samantha (2012). Mind the Gap! Bridge between World Englishes and Learner Englishes in the Making. *English Text Construction* 5 (2), pp. 264–291.

[8] Mohammad, Saif M. (2017). *Word Affect Intensities*. arXiv preprint. arXiv.

[9] Quirk, Randolph, Greenbaum, Sidney, Leech, Geoffrey, Svartvik, Jan (1985). *A Comprehensive Grammar of the English Language*. London: Longman.

[10] Straka, Milan, Straková, Jana (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada.

[11] Swales, John (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

[12] Vincze, Veronika (2014). *Uncertainty Detection in Natural Language Texts*. PhD thesis, Szeged: University of Szeged.

[13] Xiao, Richard (2009). Multidimensional Analysis and the Study of World Englishes. *World Englishes* 28 (4), pp. 421–450.

[14] Neslihan Onder Ozdemir, Bernadette Longo (2014). Metadiscourse Use in Thesis Abstracts: A Cross-cultural Study, *Procedia – Social and Behavioral Sciences* Volume 141, pp. 59–63, https://doi.org/10.1016/j.sbspro.2014.05.011.

[15] T. Dahl (2004). Textual metadiscourse in research articles: a marker of national culture or of academic discipline? *Journal of Pragmatics* 36 (2004), pp. 1807–1825, https://doi.org/10.1016/j.pragma.2004.05.004.

[16] Xiangdong Li (2020). Mediating cross-cultural differences in research article rhetorical moves in academic translation: A pilot corpus-based study of abstracts. *Lingua* Volume 238, p. 102795, https://doi.org/10.1016/j.lingua.2020.102795.

[17] Jianping Xie (2016). Direct or indirect? Critical or uncritical? Evaluation in Chinese English-major MA thesis literature reviews. *Journal of English for Academic Purposes* Volume 23, pp. 1–15, https://doi.org/10.1016/j.jeap.2016.05.001.

[18] Burneikaitė, Nida (2009). Metadiscoursal connectors in linguistics MA theses in English L1 & L2. *Kalbotyra* Vilnius : Vilniaus universiteto leidykla. t. 61, pp. 36–50, http://doi.org/10.15388/Klbt.2009.7636. https://epublications.vu.lt/object/elaba:59659494/.

[19] Karim Sadeghi & Arash Shirzad Khajepasha (2015). Thesis writing challenges for non-native MA students. *Research in Post-Compulsory Education* 20, 3, 357–373, http://doi.org/10.1080/13596748.2015.1063808.