UNIVERSITY OF TARTU
DEPARTMENT OF ENGLISH STUDIES

MULTI-DIMENSIONAL ANALYSIS OF ACADEMIC WRITING IN ESTONIAN LEARNER ENGLISH

BA thesis

JANELY RÜDEIN

SUPERVISOR: *Assoc. Prof.* JANE KLAVAN, PhD

TARTU
2023

ABSTRACT

Academic writing is essential in academic communities worldwide and necessary for all university students; however, the specificity within that genre often makes it difficult for students to master it. Describing the academic writing of a specific learner population could have practical pedagogical applications, and therefore, smaller learner populations, such as learner populations of Estonian EFL speakers at the university level, should be researched to improve the teaching of academic writing for those populations. The multi-dimensional analysis (MDA) method is based on quantitative analysis of co-occurring variables in texts, thus providing information about multiple characteristics of texts simultaneously. To provide this information, this thesis aims to conduct an MDA of academic writing done by L2 and L1 English speakers and thus determine their specific characteristics.

To this end, a multidimensional analysis was conducted on two samples of academic writing, one consisting of BA theses written by Estonian speakers of EFL and the other a sample from British Academic Written English (BAWE) corpus depicting L1 academic writing at the university level. The thesis consists of an introduction, two sections and a conclusion. The first section includes a literature review, discussing previous research involving English as a Foreign Language (EFL) and academic writing, research conducted with learner corpora and multidimensional analysis (MDA), and defines central concepts of the method of MDA. The second section forms the empirical analysis, including the methodology, data, results, and discussion of the results.

TABLE OF CONTENTS

ABSTRACT 2
LIST OF ABBREVIATIONS 4
INTRODUCTION 5
1. LITERATURE REVIEW 7
1.1 English as a Foreign Language (EFL) and Academic Writing 7
1.2 Learner Corpus Research 8
1.3 Multidimensional Analysis 10
2. EMPIRICAL ANALYSIS 18
2.1 Data 18
2.2 Methodology 20
2.3 Data analysis and results 21
2.4 Discussion 27
CONCLUSION 33
REFERENCES 36
APPENDICES 40
Appendix 1 40
Appendix 2 41
RESÜMEE 42

LIST OF ABBREVIATIONS

EFL – English as a Foreign Language

ELE – Estonian Learner English

ESL – English as a Second Language

MAT – Multi-Dimensional Analysis Tagger

MDA – Multi-Dimensional Analysis

POS – Part of Speech

INTRODUCTION

Academic writing, practised regularly by university students worldwide, can be considered one of the invaluable skills obtained during academic studies. Academic writing can also be considered one type of university writing, defined by Biber et al. (2002: 19) as a comprehensive term for written works associated with academic life. According to Cameron et al. (2009: 269), many students find the academic writing style challenging since it is often regarded as technical, formal, and elaborate; yet this style is essential for communicating in the scientific domains.

As academic writing can be considered an essential aspect of academic discourse, and university students practice communicating within this sphere, the students thus form a learner population. Granger (2008: 260) defines learner language as language that is not considered the speaker's first language or does not exist as an official second language in their country; thus, the English spoken by Estonian university students can be considered learner language. As the diverse background of each learner population is reflected in their writing, these specific populations must be studied to gain an overview of unique challenges and thus further help their acquisition of this complex style to foster scientific discourse and contribution by learners. Although academic writing done by L1 English speakers has been researched extensively, the same cannot be said for Estonian learner English (ELE).

This thesis aims to fill a gap in the current knowledge of the learner population of Estonian learners of English and analyse the academic writing done by EFL learners among the BA students of English Language and Literature at the University of Tartu, intending to provide insight into the students' use and proficiency of academic English and contribute information which could be applied in teaching academic writing to EFL learners in Estonia. This thesis focuses on answering the following research questions: (1)

does the academic writing in the BA theses by Estonian learners of English and the academic writing of L1 English speakers correspond with the text type of learned exposition and the genre of academic prose established by Douglas Biber (1988: 171), (2) whether there are any distinct differences between the samples of learner academic writing and L1 academic writing, and if there are, then what differentiates these populations in terms of academic writing.

This thesis answers these questions through a method of empirical analysis known as Multi-Dimensional Analysis (MDA), established by Douglas Biber, on samples of academic writing from two corpora: a sample of defended BA theses of the Department of English Studies at the University of Tartu and a reference sample compiled of L1 English speakers' academic writing from the *British Academic Written English Corpus* (BAWE). The analysis uses the Multi-Dimensional Analysis tagger (Nini 2015), replicating Biber's (1995: 18-20) analysis. Biber's (1988: 13) established method of MDA uses factorial analysis on corpora to assign scores reflecting specific linguistic dimensions to a collection of texts by measuring the frequency of specific co-occurring features in the text. Biber (1989: 13) also states that text types and genres are assigned based on calculated dimensional scores. The specific dimensional scores for academic prose and variations in scores that seem to describe academic prose in the humanities, as described by Biber (1989: 27-29), will be observed in the analysis, specifically whether these established scores and, thus text types correlate with the dimensional scores of the chosen corpora.

This thesis is structured into two sections. The first section and sub-sections form a literature review, providing an overview of learner English, academic writing, learner corpus research, previous MDA research, and the method used for the following analysis. The second section forms the empirical analysis, and its sub-sections present the data, methodology, analysis and results, concluding with a discussion.

1. LITERATURE REVIEW

1.1 English as a Foreign Language (EFL) and Academic Writing

To clarify the terms, Deshors (2014: 277) and Kachru (1992: 54-55) contrast English as a Second Language (ESL) and English as a Foreign Language (EFL) based on the status and usage of the English language in the speakers' country, as ESL is used in a broader range of contexts in comparison to EFL. These two types are distinctly different: while Deshors (2014: 302) and Laporte (2012: 286) suggest the similarities between these two in terms of grammatical features, Schauer (2006: 311) notes significant differences due to different learning environments. Gilquin and Granger (2011: 74) support this view whilst arguing against the opposition of EFL and ESL and suggesting a continuum of variation for their description instead. In the Estonian context, the term EFL is used for Estonian speakers of English, as is visible in Raud and Orehhova (2017) and Peter et al. (2016). The term, therefore, applies to the population under review in the current thesis.

The academic writing of EFL speakers has been studied extensively, as shown by analyses of texts written by speakers with varying backgrounds and different types of tasks, from learners with European backgrounds (Gilquin and Paquot 2008; Larsson et al. 2020) to Asian ones (Park 2020; Wang and Xie 2022). The task types included in these analyses range from theses, term papers (Larsson et al. 2020), and essays (Gilquin and Paquot 2008) to freewriting (Park 2020) and rubrics (Wang and Xie 2022). Regarding Estonian EFL research, Sundh (2018) and Vassiljev et al. (2015) have focused on EFL amongst younger learners. While Yallop and Leijen (2018) and Yevchuk (2021) have researched the academic writing of Estonian EFL speakers at the university level, they have not analysed the language used in theses, having focused on feedback and vocabulary.

1.2 Learner Corpus Research

Learner corpus research, a subfield of corpus linguistics, focuses on describing patterns apparent in learner language based on analysis of learner corpora (Granger 2019: 1). As a computerised field, learner corpus research is a relatively recent field of research still in development (Granger 2019: 1; Tono 2003: 802). Awareness of the specific limitations found in learner corpus research is essential for constructing and analysing learner corpora. As the following discussion cannot sufficiently address all possible limitations of the field, it will focus on describing the limitations related to the specific characteristics of learner corpora, such as the range of variables and proposed applications, establishing a gap in the current knowledge on this topic and proposing recommendations for future research.

Flowerdew (2009: 343), Gilquin et al. (2007: 327), and Granger (2009: 14) define learner corpora as a type of corpora, a computerised collection of data which is compiled from data which can be used to describe the language use of foreign and second language learners, and language patterns characteristic to learner populations. Flowerdew (1998: 340), Gilquin et al. (2007: 322), and Tono (2003: 800) emphasise the distinctly different and more complex nature of learner corpora compared to general corpora, particularly the number of specific variables included in the data of learner corpora, such as variables describing task type and learner population. The contrast is further strengthened by Granger (2009: 25), who views learner corpus research as a multi-disciplinary field requiring analysis that combines perspectives from various research fields. According to Gilquin and Paquot (2008: 44) and Granger (2009: 4), a range of variables affects the criteria set for any learner population under observation; therefore, the specific range of variables, such as task variables (including types and whether the production is timed, for example) and learner variables, such as age, sex, language background, education level,

has to be set and observed during all research processes involving learner corpora.

Gilquin et al. (2007: 323), Gilquin and Paquot (2008: 42), and Granger (2019: 3-4) state that learner corpus analysis reveals patterns in language use which characterise either a more general or a more specific population of learners. Connecting the aspects of populations and data, Gilquin et al. (2007: 321) and Granger (2009: 15; 2019: 3) claim that the sizeable data sets describing large populations of learners are one of the proposed advantages of learner corpus research. Nevertheless, this proposed advantage seems to be criticised by Pendar and Chapelle (2008: 204) and McEnery et al. (2019: 79-80), who found that a lack of data and variety is conversely a significant problem for the field of learner corpus research, as the basis of a wide range of variables in research data often contributes to a lack of data representing specific learner populations, which could influence any results of research or even inhibit research due to a lack of information.

Tono (2003: 802) and Granger (2009: 25) also describe the field as a recent development in corpus linguistics, which could further account for the quantity of information and data available about the field. One proposed solution to the apparent limitations of data sparsity and relevancy in learner corpus research is the creation and analysis of corpora representing specific local learner populations, as recommended by Mukherjee and Rohrbach (2006: 217,219), Gilquin and Paquot (2008: 57) and Granger (2009: 19). On the other hand, this solution should be approached with care, as Tono (2003: 806) and Granger (2009: 14) emphasise the interdisciplinary expertise needed for correct conduct and interpretation of such research.

Based on the suggestions mentioned above, it appears that research must be conducted on the mentioned target audience to get an accurate overview of the patterns apparent in the language use of Estonian learners of English. Tammekänd and Torn-Leesik (2022: 264) call for further research on Estonian learner English. This call has thus far

been answered by theses written in the Department of English Studies at the University of Tartu. BA theses by Vaher (2021), Kraak (2021), Konso (2021), Haamer (2022) and MA theses by Daniel (2015) and Savchenko (2022) use learner corpora to investigate learner language; however, they focus on single aspects of ELE such as the use of adjectives, adverbs, or pronouns. Although Kraak (2021), Haamer (2022) and Vaher (2021) use theses in their research, their work investigates singular aspects of learner language or its development; others (Konso 2021; Savchenko 2022) rely on spoken data from LOCNEC (Louvain Corpus of Native English Conversation) and LINDSEI-EST (the Estonian subcorpus of the Louvain International Database of Spoken English Interlanguage), or data from entrance exams (Daniel 2015) which does not describe academic writing at the university level.

In conclusion, a significant limitation of learner corpus research is the wide range of specific variables appearing in the learner data. This makes finding data sets describing specific learner populations difficult, despite the proposed sizable datasets advertised as an advantage of learner corpora. Although the contribution to learner corpora research and academic writing research in Estonia is growing via various theses, specific populations should still be researched to provide information and describe learner language more effectively.

1.3 Multidimensional Analysis

This section describes the methodological approach utilised in the thesis at hand, multi-dimensional analysis, and provides an overview of the central concepts included, such as dimensions, text types, and previous research using that method.

The Concept of a Dimension in MDA

In the context of MDA, Biber (1988: 6) established the dimension as a measurement of "linguistic co-occurrence". MDA relies on six different dimensions to describe the linguistic variation in a language, with each dimension characterising a specific function of a text (Biber 1988: 57). Each dimension incorporates two extremes; however, a dimension still functions as a scale rather than a binary opposition and each analysed text is assigned a score along the scale of each dimension. The statistical basis of the method involves factorial analysis of the occurrence of multiple variables in texts, a more thorough description of which is available in Biber (1988: 79). This method shares several advantages of corpus analysis, as it has the advantage of providing qualitative data describing language use, and as a computerised method, enables the analysis of large amounts of data. Despite that, Biber (1988: 92) also stresses that a qualitative analysis of the analysed data is necessary for an accurate interpretation of the functions of a text.

The specific characteristics of each dimension are further discussed below, mainly focusing on the three relevant dimensions for this thesis: the first, third and fifth dimensions. There are also multiple variants for the names of the dimensions; however, this thesis utilises the original names from Biber (1988) for consistency. As Dimensions 1, 3 and 5 are involved in assigning the text type of academic prose, they are relevant for addressing the research question regarding academic writing and are thus described more specifically in the following sections: other dimensions included in MDA are also summarised below. The following descriptions of dimensions are based on the book *Dimensions of Register Variation: A Cross-Linguistic Comparison* written by the method's creator, Douglas Biber (1995); the same source is used throughout the thesis for consistency.

Dimension 1 – Involved vs Informational Production

Dimension 1 assigns texts along a scale of involved and informational production. Involved production, marked by the high frequency of positive variables combined with the low frequency of negative variables on Dimension 1, is described as personal, interactive and direct. Some positive features of the first dimension include private verbs, contractions, that-deletion, and present tense verbs, which characterise a direct and emphatic style in a text with an active involvement of participants. In contrast, Informational production, marked by the high frequency of negative variables and the low frequency of positive variables, is described by its precision and informational density. The negative features of this dimension – high noun density, high frequency of prepositional phrases, use of attribute adjectives, a longer word length, and a high type-token ratio can be associated with a goal to present and convey specific information precisely (Biber 1995: 141-43). Some examples of involved production are direct and phone conversations, whilst informational production is usually characteristic of academic texts and official documents (Biber 1995: 145).

This dimension can describe the influence of the circumstances and purpose of the text production, ranging from controlled and careful production with a possibility for revisions for informational texts, such as documents, to a largely spontaneous production without revisions, for involved texts, such as conversations. Thus, it can describe the opposition between spontaneous and edited texts (Biber 1995: 145-47).

Dimension 3 – Explicit vs Situation-Dependent Reference

Dimension 3 assigns dimensional scores along the scale of explicit and situation-dependent reference. This dimension addresses the importance of contextual information for the texts and can also be used (to a lesser extent) to describe the mode differences

between spoken and written registers. Explicit reference, marked by the high frequency of negative features combined with the low frequency of positive values on Dimension 3, is described as explicitly identifying referents of the text via included information and is less ambiguous (Biber 1995: 155-6).

In contrast, situation-dependent reference on the other end of the scale is described as dependent on references to the text's external context/situation (time, place). The positive features of this dimension include a high frequency of adverbials and adverbs, but the negative features are WH-relative clauses on subject/object positions, pied-piping constructions, phrasal coordination and nominalisations. Some examples of situation-dependent reference are broadcasts, conversations and letters, whilst official documents and academic prose fall under the explicit reference category (Biber 1995: 155-159).

Dimension 5 – Abstract vs Non-Abstract information

The scale of dimension 5 differentiates abstract information and non-abstract information. A score on this dimension reflects the use of passive voice, whether the focus is on either referents or agents, and thus is often prominent in assigning the text types of academic prose and official documents, which often rely on passive constructions. Dimension 5 differs from the previously mentioned dimensions as it includes only negative co-occurring linguistic features, such as conjuncts, agentless passives, and by-passives, often used to mark the passive voice. Dimension five features do not occur in spoken registers, and Biber concludes that texts with high negative scores in dimension 5 are technical in vocabulary and topics (Biber 1995: 163-164).

Other Dimensions included In MDA

Dimension 6 (*Online Informational Elaboration)* describes stance marking in texts based on the occurrences of dependent clauses, demonstratives and stranded prepositions, which are often found in informational text types, mainly in spoken registers, such as speeches, public conversations, but also in editorials and professional letters (Biber 1995: 166-167).

Dimension 2 (*Narrative vs Non-Narrative Concerns)* illustrates the scale of past and present reference in texts, measured via the occurrence of past and present tense, personal pronouns, and public verbs – high frequencies of which can often be found in narratives (such as in fictional texts) for referring and describing. In contrast, academic prose and conversations are considered non-narrative (Biber 1995: 152-54).

Scores on Dimension 4 (*Overt Expression of Persuasion)* reflect the presence of the author's stance in texts, measured via the occurrence of infinitives, various modals, suasive verbs and conditional subordination. Such features are not found in most registers but often occur in editorials, reflecting argumentation and persuasion (Biber 1995: 159, 162-3).

Biber's Classification of Text Types

Biber (1989: 6) classifies text types based on frequently co-occurring sets of linguistics variables, which he calls dimensions, which can be used to describe the characteristics and functions of different types of texts. According to Biber (1989: 6); 1988: 170), this classification contrasts with the classification of text genres, as the latter does not rely on a linguistic and empirical basis; however, a type can include several genres of texts, and a genre can belong to several types as illustrated below. Biber (1988: 67-68) established 23 types in total, including 15 for written texts and six for spoken texts,

with some types also containing sub-types; however, the thesis at hand can adequately address only the types and genres appearing in the analysis of the samples: scientific exposition, learned exposition and the genres of academic prose, official documents and broadcasts, which are described below.

Scientific Exposition and Learned Exposition

The text types of scientific and learned exposition are quite similar in terms of their dimensional scores across the majority of dimensions in MDA, as both convey information, lack narrative, and feature abstract style (Biber 1989: 27). These characteristics appear in texts as a varied vocabulary with high counts of nouns, and prepositions, long words and attributive adjectives, but a lack of verbs (Biber 1989: 27-28). The main difference between those types is the degree of abstraction or technicality included in the texts, and Biber (1989: 27-29) illustrates this with the contrast of academic texts from natural sciences and academic texts from humanities, claiming that academic texts from the humanities are less technical and prefer active voice more than those in the natural sciences, while still acknowledging that there are some exceptions.

Academic Prose

The text genre of academic prose, characterised by a formal and specific style and use of specialised terminology, is featured in academic texts such as research articles, dissertations and Bachelor's theses. Academic prose texts are well-structured and intended for a specialist audience; their goal is to convey theories and results of research; thus, they must be concise and understandable (Biber 1988: 69).

Biber shows academic prose to have a significant internal variation across all dimensions, possibly due to various fields and sub-fields within that genre, which differ in

style and content (1988: 171).

Official Documents

Although official documents are similar to academic prose in their dimensional scores in MDA, their contents are distinctly different, as official documents are often more limited in style (Biber 1988: 178). This genre, along with academic prose, is characterised by its informational goals, which appear in texts via a high count of nouns and prepositions and a low count of verbs (Biber 1988: 11). Documents can also be considered similar to the genre of professional letters, although they differ in terms of stance markers and persuasion (Biber 1988: 167)

Broadcasts

Broadcasts, which Biber (1988: 34) defines as a spoken genre, are distinct due to their situated reference, which appears in texts via descriptive variables explaining an ongoing process. Broadcasts have numerous unique characteristics, such as a lack of informational or involved production, lack of verbs, and high counts of nouns and prepositions, which assist in conveying fast-paced descriptions (Biber 1988: 135).

Previous research utilising MDA

Although the MDA method was first established by Biber (1988) to compare written and spoken productions of text, it has since been used to research various genres of language production, such as song lyrics (Werner 2021), language use in television (Al-Surmi 2012), or for researching differences between the language productions of different genders (Biber and Burges 2000), thus highlighting the vast range of the method. Academic writing research has benefitted from this method, as it has been utilised

numerous times to describe the academic texts of various populations. In his longitudinal study, Crosthwaithe (2016: 176) analyses multiple types of academic texts for a single learner population and recommends MDA for measuring progress in academic writing.

Similar uses appear in Staples et al. (2016: 179), where the variation of academic writing across genres and disciplines is emphasised. This implies that personalised feedback is only available if the writing of a specific population is under review, as writing varies with genres, levels, and other variables, such as task types. The differences across multiple variables related to the production modes and producers of the texts are also evident in Biber et al. (2002: 24), which focused on describing the language used across various genres related to academic life at universities in the US. A shared aspect of MDA appearing in multiple works (Crosthwaithe 2016: 176; Biber et al. 2012: 12; Staples et al. 2016: 179) is the emphasis on possible pedagogical implications of research, which regarding academic writing, could be used to improve training.

Thus it appears that the significant advantages of MDA are its various possible pedagogical implications; through analysing the texts of specific learner populations, the feedback and implications derived from it can be personalised to the needs of specific learner populations.

2. EMPIRICAL ANALYSIS

This thesis aims to analyse the text types used in the Bachelor's theses written by students at the University of Tartu, specifically in the English department, and to compare them with academic writing done by L1 English-speaking university students. The thesis hypothesises that the academic writing of both Estonian and English university students is classified as an academic prose text type and the learned exposition sub-type according to MDA. To achieve the aims of the thesis, quantitative research in the form of MDA was conducted using secondary data in the form of samples from two different Academic English corpora, a learner sample from TCELE and an L1 English speakers' sample from BAWE for reference purposes.

The following sections provide a detailed description of the methodology and empirical analysis process used in this thesis. The first sub-section (Section 2.1) explains the data selection for the analysis, and the methodology of the analysis process is described in the second sub-section (Section 2.2). The data analysis and results are provided in the third section (Section 2.3), and the section concludes with a discussion of the results and limitations of the analysis (Section 2.4).

2.1 Data

To examine and compare the academic writing of L1 English speakers and learners, the MDA process was performed on two corpus samples comprised of texts written in academic English. The first sample ("BA theses sample" below) from the *Tartu Corpus of Estonian Learner English* (TCELE) consists of defended Bachelor's theses written at the Department of English Studies at the University of Tartu. This sample includes 76 theses written by EFL-speaking university students on various topics in academic English. Before the MDA process, the texts included in the BA theses sample

were prepared for the MDA process following the steps as described in Kaljuste (2021: 16-17): the texts were collected from the University of Tartu Dspace website, downloaded and processed using R (R Core Team 2023), and converted to TXT format to be suitable for the following MDA process.

The reference corpora chosen for this analysis is a sample from the British Academic Written English (BAWE) corpus ("BAWE sample" below). The BAWE corpus features more than six million words and consists of texts classified as assignments, written by students studying various disciplines at three universities in the United Kingdom at undergraduate and master's levels (Alsop and Nesi 2009: 79). The corpus is freely available on the Internet for research purposes (Coventry University n. d.). To make the samples more comparable, the texts chosen for the BAWE sample include 294 texts from the Arts and Humanities disciplines featuring assignments for BA. English studies, American Studies, and Linguistics courses. Additionally, all texts included in the sample were written by students with English as a first language (L1). The texts included in the BAWE sample were collected and processed similarly to the BA thesis sample. As a collection of various types of academic assignments, the BAWE sample can be considered a good representation of academic writing; however, it is worthwhile noting that texts included in the BAWE sample are not perfect equivalents to the texts included in the theses sample, as BAWE does not include any texts that could be classified as BA theses. This aspect must be acknowledged during the analysis and interpretation of the results. This is due mainly to a lack of a direct equivalent to a thesis type of assignment at the undergraduate level at the universities of the United Kingdom. Nevertheless, both samples can be considered examples of university-level academic writing.

2.2 Methodology

The Multidimensional Analysis Tagger (MAT) (Nini 2015) computer program was used for the analysis of the samples from the corpora mentioned above, as it replicates the analysis of the Biber (1988) tagger. The MAT is freely available online and incorporates the Stanford tagger (Toutanova et al. 2003), a part-of-speech tagger. The MAT works in two stages: first, the input is grammatically tagged, and then the tagged data undergoes the MDA process (Nini 2019: 74).

The data from both samples were converted into plain text format (.txt) to suit the Multidimensional Analysis Tagger (MAT) program. Upon opening the program, the user can choose between the options to POS tag, conduct the MDA process, or do both on the input. Additionally, an option to inspect a text for dimensional features is available. The option to both grammatically tag and conduct MDA was selected for the analysis of both samples. Additionally, the user must decide whether to enable Z-score correction for the analysis. As the manual for MAT does not recommend Z-score correction for the analysis of longer texts, this option was disabled for analysing both samples (Nini 2019). Another decision required of the user before the analysis process is whether the program counts “all tags” in the input or “only VASW tags”, meaning tags only counted in Biber (1988) (Nini 2019). For the input processing, the user must also set the number of tokens for the type/token ratio, which is set as 400 by default, as in Biber’s (1988) original tagger, to preserve consistency with Biber’s results (Nini 2019). The default option was enabled for the analysis of both samples as the chosen type/token ratio was initially set by Biber, and the counting of all tags provides more detailed data from the analysed input. Finally, the user can decide to generate graphs for each dimension.

The output of the program includes a folder of grammatically annotated text files and a folder of text files annotated with Biber’s (1988) labels for MDA. Additionally, a

folder of statistics is generated, which includes graphs visualising each chosen dimension, a graph visualising the text types of the input across the six dimensions (both in Portable Network Graphic format (.png)) and Excel files containing dimensional scores, assigned text types, z-scores and statistics measuring the occurrence of each tag for each file.

Although it is possible to generate graphs for all dimensions of each analysed corpora using the MAT program, the figures presented herein were generated independently of the program based on the data provided by the MAT program to facilitate a more effective comparison between the two samples. All figures visualising the dimensions were created using the RStudio (Posit Team 2023) program, free software for using R (R Core Team 2023), and Xldotplotter (Sönning 2016), an extension which enables the creation of plots visualising the data. The figures include data provided by the MAT for the samples; data for the reference text types originate from (Biber 1988: 122-125).

2.3 Data analysis and results

This section includes the results of the MDA. The statistics of the dimensional scores, including mean scores, minimums, maximums, range of the dimensional scores, and standard deviation of the dimensional scores, were calculated to assist in further data analysis and are included in Table 1. The figures were calculated using corresponding Microsoft Excel formulas and are based on the data provided by MAT. The MAT program provided the closest text genres for both samples across the first, third and fifth dimensions. Figures 1-3 illustrate the calculated ranges for the dimensional scores and the mean dimensional scores across the three dimensions for both samples and comparisons to set ranges and mean scores for specific text types by Biber (1988).

The visualisations of the assignment of text types across the six dimensions of MDA and the assignments of the text subtype for the BAWE sample and the BA theses sample are available as figures generated by the MAT program and included respectively in Appendix 1 and Appendix 2. Based on the assigned closest text types across all six dimensions, both samples were assigned the text subtype of scientific exposition by MAT (Appendix 1; Appendix 2). As visible from Table 1, the mean score of the sample is the most similar, and the assigned closest genres overlap on dimension five but differ for the third and first dimensions. It is also noteworthy that the range for the BAWE sample is more extensive across all dimensions under review, suggesting a higher degree of variation within that sample.

Table 1. The assigned dimension scores and text types for both samples on the first, third and fifth dimensions in MDA

Dimension	Sample	Mean score	Minimum score	Maximum score	Range	Standard Deviation	Closest Genre
D1	BA theses	-9.31	-18.41	8.15	26.56	4.323	Broadcasts
	BAWE	-12.31	-29.8	14.7	44.5	6.807	Academic Prose
D3	BA theses	5.58	1.25	12.32	11.07	2.103	Academic Prose
	BAWE	6.69	-2.93	16.21	19.14	2.639	Official Documents
D5	BA theses	5.82	-0.55	14.43	14.98	2.835	Academic Prose
	BAWE	5.186	-1.92	22.61	24.53	4.164	Academic Prose

Results of the MDA across the Dimension 1 (Involved *vs* Informational Production)

The closest text genres assigned to the samples differed on the first dimension, as the BA theses sample was assigned the text genre of broadcasts with a mean score of -9.31, while the BAWE sample was assigned the text genre of academic prose with a mean score of -12.31. For the BA theses sample, the dimensional scores ranged from -18.41 to 8.15, resulting in a range of 26.56 and a standard deviation measuring 4.323, which, compared to the BAWE sample, is distinctly different, as the range of the BAWE sample was almost doubled, ranging from -29.8 to 14.7, measuring 44.5 for range and 6.807 for standard deviation. With results leaning towards the negative end of the scale compared with the other presented text types included in the figure, both samples are marked as including informational production (Figure 1). According to Biber (1988: 104), texts featuring this assignment are precise in structure due to their lack of time constraints during the production process and feature a high count of nouns, longer word length and prepositional phrases used to convey information. These characteristics seem descriptive regarding academic writing in theses and written assignments at the university level.

Figure 1 also shows that both samples' mean dimensional scores were somewhat higher than that of the academic prose; the range of the BAWE sample was also more prominent than that of the academic prose, and the BA theses sample measures the opposite with a range smaller than the academic prose type. The BA theses sample was possibly assigned the broadcasts text type due to its higher mean dimensional score, which is closer to that of the broadcasts text type, as the mean broadcast score was -4.3, whilst the equivalent for academic prose was -14.9 (Biber 1988: 123-125).

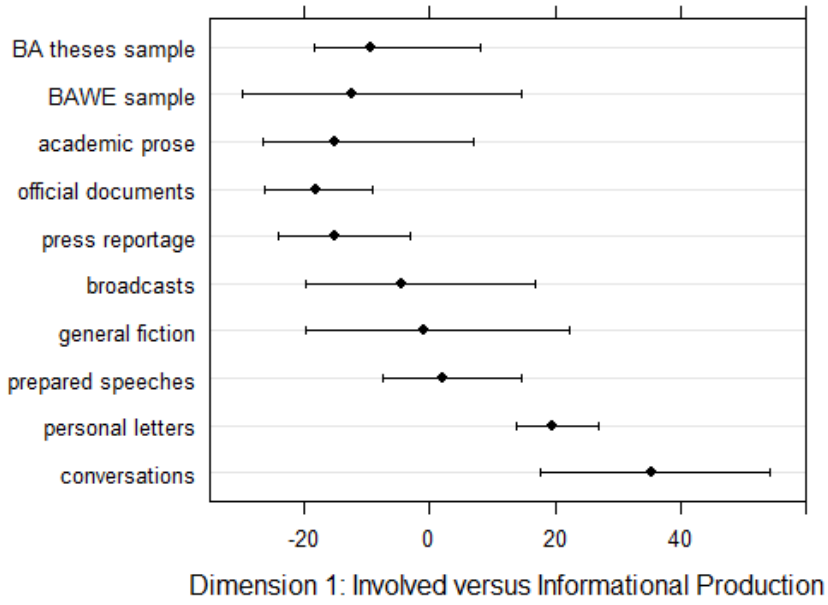


Figure 1. Dimensional scores for Dimension 1. The dots represent the mean dimensional score of the samples; the bars illustrate the range of the dimensional scores for the samples.

Results of the MDA across Dimension 3 (Explicit *vs* Situation-Dependent Reference)

The closest text genres assigned to the samples also differed on the third dimension: the BA theses sample was assigned the text genre of academic prose with a mean dimensional score of 5.58, and the BAWE sample was assigned the genre of official documents with a mean score of 6.69. The scores range from 1.25 to 13.32 for the BA theses sample, and the standard deviation across the dimensional scores was 2.103. For the BAWE sample, the dimensional scores ranged from -2.93 to 16.21, with a standard deviation of 2.639 and a 19.14 range. As seen from Figure 2, both samples were mainly located on the positive end of the scale, marking it as including texts with mostly situation-dependent reference, and according to Biber (1988: 110), this label would suggest that texts in both samples feature a high use of relative clauses and nominalisations, which are

also used to convey information. Figure 2 also illustrates that the range of academic prose across the third dimension is more significant than that of the two samples. Whilst the mean scores of the samples were quite similar, the mean score of official documents across this dimension is 7.3 compared to 4.2 for academic prose, which explains the assignment for the BAWE sample (Biber 1988: 123)

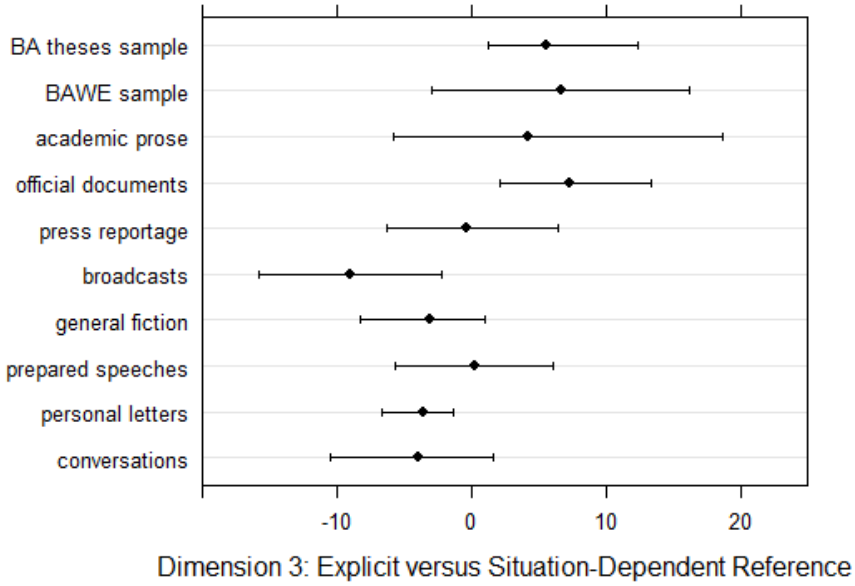


Figure 2. Dimensional scores for Dimension 3. The dots represent the mean dimensional score of the samples; the bars illustrate the range of the dimensional scores for the samples.

Results of the MDA across Dimension 5 (Abstract *vs* Non-Abstract Information)

For this dimension, academic prose was assigned as the closest text genre for both samples. The BA theses sample featured a mean dimensional score of 5.82, ranging from -0.55 to 14.43, with a standard deviation of 2.835. The mean score of 5.186 for the BAWE sample was quite similar to the other sample; however, the range was more comprehensive, reaching from -1.92 to 24.53 with 4.164 as the standard deviation. As seen

in Figure 3, the BAWE sample features the broadest range in the figure, while the range of the BA theses sample is narrower than the range for the academic prose sample. It is also noticeable that the mean scores for the academic prose type and both samples are the highest in the figure. Both samples are located mainly on the positive end of the scale and are thus labelled non-abstract information. According to Biber (1988: 112), texts under this label include high counts of passive structures used in technical texts to divert attention from the agent.

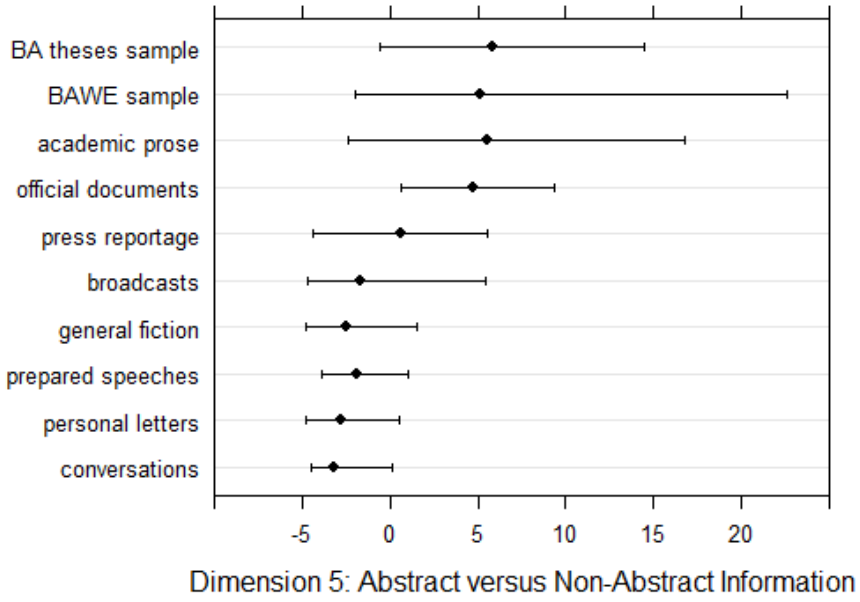


Figure 3. Dimensional scores for Dimension 5. The dots represent the mean dimensional score of the samples; the bars illustrate the range of the dimensional scores for the samples.

2.4 Discussion

According to the results of the MDA, both samples were mainly assigned the text genres of academic prose with some exceptions, as predicted by the research question; however, the scientific exposition text type assigned to both samples was surprising, as Biber (1989: 29) assigned that type to texts from natural sciences, which he characterised as technical and featuring high counts of passive voice.

Upon a closer look, calculations based on the data from the statistics file the MAT provided show that most of the texts in the BAWE sample were primarily assigned the scientific exposition (43%) and learned exposition text type (33%). The text type assignment for the BA theses was similar, with scientific exposition (59%) and learned exposition (5%) accounting for a large proportion of the texts. The high percentage of scientific exposition challenges Biber's association of humanities texts with the learned exposition type. Notably, the percentage of learned exposition in the BA theses sample is significantly lower than in the BAWE sample, which implies that the assignment type of theses resembles writing in natural sciences more than other assignments in the humanities.

Whilst Biber (1989: 27-28) mentions that the learned and scientific expositions genres share some characteristics of academic prose, such as the informational purpose and explicit reference, which appear as high counts of nouns and prepositions, long word lengths and a variation in vocabulary; he contrasts the subtypes on the grounds of variations in technicality and precision which is apparent in abstract style featuring high counts of passive voice. However, it is still possible that the text types of theses and assignments could also be more technical and detailed as carefully planned and edited productions of texts and research.

The text type of general narrative exposition, which was assigned to 18% of the

BAWE sample and 31% of the BA theses sample, combines elements of informational writing, which is visible in high counts of nouns and prepositional phrases, and narrative, which appears as high counts of past tense, (e.g. *whether it was possible for a woman),* perfect aspect (e.g. *Estonia had just gained independence)*, and third person pronouns (e.g. *his mother tells him)*. Whilst these elements do not overlap with the variables included in the classification of academic prose, the examples from the BA sample show that these elements can be used in academic writing to retell a narrative or describe/analyse previous events, which can occur in theses or assignments focusing on literature or narrative analysis.

Whilst Biber (1989: 35-38) classifies the type of involved persuasion, which was assigned to 6% of the BAWE sample and 5% of the BA theses sample, as a mostly spoken text type, he acknowledges the existence of written informational texts which could have that label due to their strongly argumentative goals, apparent in texts as high counts of modals (e.g. *It could be said),* private verbs (*e.g. protagonist who feels alienated),* hedges (e.g. *more or less the same*, emphatics (*e.g. discussed in more detail)* and pronouns (e.g. *find one's true identity)*. Biber (1988: 194) also suggests that humanities texts can be assigned as persuasive due to argumentation found within those texts, which would apply well to both samples under review.

Salager-Meyer (2011: 35), Crompton (1997: 271), and Hyland (1994: 240) consider hedging as a distinct element of academic discourse and mention modals as one aspect of hedging. Private verbs such as *assume, believe, imply* are defined by Quirk et al. (1985: 1181) and Biber (1988: 242) as markers of unseen intellectual acts, which suggests that they could be suitable in academic writing; however, Biber marks them with characteristics of interactive and emotional texts on the first dimension (1995: 105), which contrast informational texts such as academic prose. Hinkel (2003: 292) has also observed

a higher count of private verbs in L2 academic texts than those of L1 speaking students, suggesting that it could hint towards a smaller range of vocabulary. However, this variable should be examined more closely to make further suggestions. Brown and Aull (2017: 403) found that student essays with more emphatics received lower grades; however, the same cannot be said for the BA theses sample, as no comparable data is available. However, this shows that emphatics, which received a mean count of 0.6 across the theses sample, are not usually associated with the standards of academic writing and thus avoided in academic writing.

The mean dimensional scores were similar for both samples across all dimensions, with differences ranging from 0.6 to 3 points. As is also visible from Figures 1-3, the mean scores provided by Biber (1988: 122-125) for the text type of academic prose across these dimensions (-14.9 for D1, 7.3 for D3, and 5.5 for D5) resemble those calculated by MAT for both samples explaining the similar categorisation of text types for the majority of dimensions. In terms of standard deviation, it is notable that the results for both samples were generally similar across the three dimensions, with differences ranging from 0.5 to 2.5, while the BAWE sample had a consistently higher standard deviation across all dimensions. As relatively "low" scores, these scores could imply that the general characteristics of the texts included in both samples were similar, which could be expected from the BA thesis sample, as the texts included feature the distinct assignment type of a thesis. However, for the BAWE sample, this result could be interpreted as a sign of general similarity found among school assignments; however, this suggestion would require further analysis of specific assignment types, such as essays or summaries.

Although the ranges for the dimensional scores differ across the dimensions and between the samples, it is worth noting that the ranges for the BAWE sample are greater across all three dimensions. This difference could be another hint towards the similar

characteristics of the texts included in the BA theses sample resulting from a single assignment type, which in turn makes the range of the BAWE sample seem more significant, as it includes texts from several fields related to humanities and different task types; however, this suggestion should be examined further.

The mean dimensional scores for both samples across the first dimensions are negative and mark the texts as including informational production, which Biber (1988: 107-108) suggests is characterised by the mutual influence of the purpose and circumstances of the content on one another, as highly informational texts are usually carefully edited and do not involve time constraints to allow for maximum communication in terms of information. These characteristics also apply to texts included in both samples. The standard deviations found across D1 for both samples are the greatest, which could indicate that many texts from both samples differ in their degree of conveying information. However, as the type of academic prose is located towards the negative end of the scale, this range could imply that there is a variation in terms of the use of negative characteristics on D1, such as nominalisations, noun count, prepositional phrases, and attributive adjectives, which assist in conveying information and are thus necessary elements in academic writing.

Biber (1988: 134-135) suggests that although broadcasts are an oral genre, they are somewhat anomalous due to their similarity with several written types regarding the shared goals of conveying information. Thus, the broadcast label of the BA theses sample could be explained through the characteristics of this text type, as the text types of academic prose share a lack of involvement with the audience and a lack of narrative (Biber 1989: 35). The MAT statistics for this dimension show that the BA theses sample featured higher counts of use of present tense (e.g. *the book contains numerous examples...*) which is consistent with the broadcast text type, and using *be* as a main verb (e.g. *protagonist was*

truly free…), whilst the BAWE sample had a higher type-token ratio (207 *vs* 181 / per 400 tokens), and a higher count of agentless passives (e.g. *it is assumed that children…*), which are both characteristic of academic prose.

For the third dimension, the range of the BA theses sample was the narrowest, while the BAWE sample has a broader range in terms of dimensional score, similar to the type of academic prose. Whereas both samples were assigned the label of situation-dependent reference, which Biber (1988: 145) described as including elaborate reference through a high count of nouns, the BAWE sample was assigned the text type of Official Documents due to its higher mean dimensional score and thus a higher degree of elaborate reference, which appears in the data as a higher count of piped-piping relative clauses (e.g. *the way in which they used it*), nominalisations (*e.g. definitions of modality*). The BA theses sample measured higher in terms of adverbs (e.g. *quite a* phenomenon).

As seen in Figures 1-3, regarding dimensional scores, the text genre of official documents is similar to academic prose across all three dimensions under review. This is also confirmed by Biber (1988: 178), who relates that result to the informational nature of both texts but emphasises the lack of personal style and the existence of stricter rules for documents. Thus, the assigned type can be related to the similarities between those two types. Biber's (1988: 180) suggestion on the connection between wide ranges and small standard deviations applies well for both samples across the first dimension, showing that variation is allowed in terms of both samples, which is surprising for theses, which should adhere strictly to the academic writing conventions.

For the fifth dimension, both samples had similar mean scores and were thus assigned the text type of academic prose. Notably, the BAWE sample featured higher counts of conjuncts, passives, adverbial subordinates and a higher type-token ratio than the BA theses sample. As seen in Figure 3, the BAWE sample also featured the widest range,

while the BA theses sample resembled the academic prose text type range. Biber (1988: 179) has also marked academic prose as an outlier in terms of D5, as it has a wide range on that dimension. This variety within the sample could be explained through a higher amount of assignment types included in the sample or the fact that samples originate from various humanities fields, which require different presentations of information. As other types are located considerably more on the negative end of the scale, it is clear that both samples belong to the group of academic prose and official documents, differing from the others, which rely on explicit reference.

It is also crucial to highlight the limitations of the current analysis, such as the previously mentioned fact that the samples are not directly comparable, making direct comparison difficult. However, the samples included in the analysis can still be described and analysed separately, providing helpful information for future research, especially regarding the sample of ELE academic writing in theses, for which there is currently little information available. Further qualitative analysis of current samples could also provide further implications for improving the teaching of academic writing to Estonian learners of English.

For further research, several improvements and modifications are also recommended to improve the quality and comparability of the data. One possibility is to include a better equivalent for the text type of theses in the analysis, allowing direct comparison with BA theses written by ELE. Some possible equivalents could be available from Scandinavian universities, where the task type of thesis is used and written by undergraduates. Another avenue for further research could be a longitudinal study tracking the possible progression in academic writing from a first-year undergraduate to a graduate student; however, a similar text type should be used to warrant a direct comparison.

CONCLUSION

Academic writing, as the language of scientific discourse, is a necessary aspect of academic life for students and researchers alike and, thus, an important skill to master as a BA student. Previous research in the EFL and learner corpora fields show that while the topic of academic writing has been researched regarding various learner populations worldwide, there is currently little information available about ELE and, more specifically, the EFL used in the BA thesis task type of academic writing. However, as established by previous research in learner corpora, information about specific learner populations is necessary, as every population is characterised by various task- and learner-related variables and could be used to improve the teaching of academic writing. Although there is currently not much information describing the academic writing occurring in the BA theses of ELE, various MA and BA theses at the University of Tartu have investigated several specific aspects of academic writing of Estonian speakers of EFL.

The method of MDA, established by Douglas Biber, uses factorial analysis and therefore has a quantitative data basis characterising various co-occurring grammatical aspects of texts simultaneously to characterise texts via the assignment of dimensional scores and text types. This method has been used in researching various types of texts, including academic writing; however, no information is currently available about MDA research on academic writing by ELE. Thus, the current thesis intended to fill the research gap by conducting an MDA on academic writing by ELE at the university level to provide potentially helpful information for future research.

To this end, two research questions were proposed: (1) does the academic writing both in the BA theses by Estonian learners of English and the academic writing of L1 English speakers correspond with the text type of learned exposition and the genre of academic prose established by Douglas Biber (1988: 171), (2) whether there are any

distinct differences between the samples of learner academic writing and L1 academic writing, and if there are, then what differentiates these populations in terms of academic writing. A multi-dimensional analysis was conducted to answer these questions.

The empirical analysis in the form of an MDA analysis was conducted with the MAT tool (Nini 2015) across three dimensions (1,3 and 5) on two samples from learner corpora: one depicting the BA theses written at the Department of English Studies at the University of Tartu and the other sample compiled of L1 English speakers' academic writing from BAWE, whilst noting that the differences between the task types and learner populations affect any comparisons between the samples. Data provided by the MAT program was analysed and compared to Biber's (1988: 122-125) set values for other text genres and types. Analysis and comparison of the data regarding dimensional scores and assigned text genres, which the MAT provided, enabled answering both research questions.

Contrary to Biber's (1989: 27) assignment of the learned exposition text type to academic prose in the humanities, neither sample was assigned that type, with scientific exposition being the type of the majority of texts in both samples. This leads to the conclusion that the texts included in both samples were more technical and thus more alike to academic prose in the fields related to natural sciences. This label could apply to the academic writing found in the BA theses, as these are planned and edited texts; however, as the BAWE sample included different task types from various fields of the humanities, this exact characterisation cannot be applied to that sample in general and must be researched further. Notably, other types included in the samples (learned exposition, involved persuasion, general narrative exposition) were shown to have possible academic functions, such as retelling narrative or strong argumentation. The genre of academic prose was assigned to both samples on two dimensions out of three: the first and fifth

dimensions for the BAWE sample and the third and fifth dimensions for the BA theses sample. Other genres assigned were official documents, which relate to academic prose via technicality and informational goals, and broadcasts, which have descriptive goals.

Some differences between the samples were found, as the BAWE sample had a broader range and standard deviation across the three dimensions, suggesting a greater variety within the sample. Regarding academic writing, the type of scientific exposition suggests a high degree of passive voice used in both samples. More BA theses were assigned the label of involved persuasion, usually considered a spoken genre, and thus could feature a higher count of emphatics and modals, which are not associated with academic writing.

In conclusion, the current thesis illustrates one of the first attempts at using MDA analysis to research language use and academic writing by ELE at the university level. Further qualitative and quantitative research of the current population is necessary to draw more far-reaching conclusions and could provide more helpful feedback to university lecturers teaching academic writing to Estonian speakers of EFL.

REFERENCES

Alsop, Sian and Hilary Nesi. 2009. Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora*, 4:1, 71-83.

Al-Surmi, Mansoor. 2012. Authenticity and TV Shows: A Multidimensional Analysis Perspective. *TESOL Quarterly*, June 2012, 671-694.

Biber, Douglas. 1988. *Variation Across Speech and Writing.* Cambridge and New York: Cambridge University Press.

Biber, Douglas. 1989. A typology of English texts. *Linguistics,* 27:1, 3-44.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison.* Cambridge: Cambridge University Press.

Biber, Douglas. 2012. Register as a Predictor of Linguistic Variation. *Corpus Linguistics and Linguistic Theory,* 8:1, 9–37.

Biber, Douglas and Jena Burges. 2000. Historical Change in the Language Use of Women and Men Gender Differences in Dramatic Dialogue. *Journal of English Linguistics*, 28, 21-37.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, and Marie Helt. 2002. Speaking and Writing in the University: A Multidimensional Comparison. *TESOL Quarterly,* 36:1, 9-48.

Brown, David West and Laura L. Aull. 2017. Elaborated Specificity versus Emphatic Generality: A Corpus-Based Comparison of Higher and Lower-Scoring Advanced Placement Exams in English. *Research in the Teaching of English,* 51:4, 394-417.

Cameron, Jenny, Karen Nairn and Jane Higgins. 2009. Demystifying Academic Writing: Reflections on Emotions, Know-How and Academic Identity. *Journal of Geography in Higher Education*, 33:2, 269-284.

Coventry University. n. d. British Academic Written English Corpus (BAWE). Available at https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/, accessed April 10, 2023.

Crompton, Peter. 1997. Hedging in academic writing: Some theoretical problems, *English for Specific Purposes,* 16:4, 271-287.

Crosthwaite, Peter. 2016. A longitudinal multidimensional analysis of EAP writing: Determining EAP course effectiveness. *Journal of English for Academic Purposes*, 22, 166-178.

Daniel, Anna. 2015. *The Use of Adjectives and Adverbs in Estonian and British Student Writing: A Corpus Comparison.* MA thesis. Department of English, University of Tartu, Tartu, Estonia.

Deshors, Sandra. 2014. A case for a unified treatment of EFL and ESL: A multifactorial approach. *English Worldwide,* 35:3, 279-307.

Flowerdew, Lynne. 1998. Integrating `Expert' and `Interlanguage' Computer Corpora Findings on Causality: Discoveries for Teachers and Students. *English for Specific*

Purposes, 17:4, 329–45.

Flowerdew, John. 2009. Corpora in Language Teaching. In Michael H. Long and Catherine J. Doughty (eds), *The Handbook of Language Teaching*, 327-350. Malden, Massachusetts: Wiley-Blackwell.

Gilquin, Gaëtanelle, Sylviane Granger and Magali Paquot. 2007. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes.* 6, 319-335.

Gilquin, Gaëtanelle and Magali Paquot. 2008. Too Chatty: Learner Academic Writing and Register Variation. *English Text Construction* 1:1, 41–61.

Gilquin, Gaëtanelle and Sylviane Granger. 2011. From EFL to ESL: Evidence from the International Corpus of Learner English. In Marianne Hundt and Joybranto Mukherjee (eds), *Exploring Second-Language Varieties of English and Learner Englishes: Bridging a Paradigm Gap*, 57-80. Amsterdam: John Benjamins Publishing Company.

Granger, Sylviane. 2008. Learner Corpora. In Anke Lüdeling and Merja Kytö (eds), *Corpus Linguistics. An International Handbook. Volume 1*, 259-275. Berlin, New York: Walter De Gruyter

Granger, Sylviane. 2009. The Contribution of Learner Corpora to Second Language Acquisition and Foreign Language Teaching: A Critical Evaluation. Studies in Corpus Linguistics, 33, 13–32.

Granger, Sylviane. 2019. Learner Corpora. The Encyclopedia of Applied Linguistics, 1–8.

Haamer, Kadi. 2022. *Similarities and differences in the syntactic complexity of bachelor's and master's thesis: a comparative study using L2SCA.* Unpublished BA thesis. Department of English Studies, University of Tartu, Tartu, Estonia.

Hinkel, Eli. 2003. Simplicity Without Elegance: Features of Sentences in L1 and L2 Academic Texts. *TESOL Quarterly*, 37, 275-301.

Hyland, Ken. 1994. Hedging in Academic Writing and EAF Textbooks. *English for Specific Purposes*, 13:3, 239-256.

Kachru, Braj B. 1992. *The Other Tongue: English across Cultures.* Urbana, Chicago: University of Illinois Press.

Kaljuste, Karl August. 2021. *Error Rate of Automated Part-of-Speech Tagging of Estonian Academic Learner English.* Unpublished BA thesis. Department of English Studies, University of Tartu, Tartu, Estonia.

Konso, Johanna. 2021. *A corpus-based study of like in Estonian EFL learners' speech.* Unpublished BA thesis. Department of English Studies, University of Tartu, Tartu, Estonia.

Kraak, Liisi. 2021. *The comparison of the usage of prefabs in the academic writing of Estonian EFL learners and native English speakers.* Unpublished BA thesis. Department of English Studies, University of Tartu, Tartu, Estonia.

Laporte, Samantha. 2012. Mind the Gap! Bridge between World Englishes and Learner Englishes in the Making. *English Text Construction* 5, 265-292.

Larsson, Tove, Marcus Callies, Hilde Hasselgård, Natalia Laso, Sanne Vuuren, Isabel Verdaguer, and Magali Paquot. 2020. Adverb Placement in EFL Academic

Writing: Going Beyond Syntactic Transfer. *International Journal of Corpus Linguistics*, 25, 155-184.

McEnery, Tony, Vaclav Brezina, Dana Gablasova and Jayanti Banerjee. 2019. Corpus Linguistics, Learner Corpora, and SLA: Employing Technology to Analyse Language Use. *Annual Review of Applied Linguistics*, 39, 74-92.

Mukherjee, Joybrato and J.M. Rohrbach. 2006. Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research. In B. Kettemann, G:Marko (ed). *Cognitive Linguistics Bibliography (CogBib)*, 205-232. Frankfurt: Peter Lang.

Nini, Andrea. 2015. *Multidimensional Analysis Tagger* (Version 1.3). Available at http://sites.google.com/site/multidimensionaltagger, accessed January 2023.

Nini, Andrea. 2019. The Multi-Dimensional Analysis Tagger. In Tony Berber Sardinha, and Marcia Veirano Pinto (eds), *Multi-Dimensional Analysis: Research Methods and Current Issues*, 67-94. London; New York: Bloomsbury Academic.

Nini, Andrea. 2019. MAT v.1.3.2. Manual. Available at https://sites.google.com/site/multidimensionaltagger, accessed January 2023.

Park, Jeongyeon. 2020. Benefits of Freewriting in an EFL Academic Writing Classroom. *ELT Journal*, 74:3, 318-26.

Pendar, Nick and Carol A. Chapelle. 2008. Investigating the Promise of Learner Corpora: Methodological Issues. *CALICO Journal* 25:2, 189–206.

Peter, Viktoria, Liljana Skopinskaja and Sulikko Liiv. 2016. Using Authentic Cultural Materials in Estonian Secondary EFL Instruction. *Eesti Rakenduslingvistika Ühingu Aastaraamat*, 12, 187-200.

Posit team. 2023. *RStudio: Integrated Development Environment for R.* Posit Software, PBC, Boston, MA. Available at http://www.posit.co/, accessed April 2023.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. New York: Longman.

R Core Team. 2023. R*: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. Available at https://www.R-project.org/, accessed April 2023.

Raud, Nina and Olga Orehhova. 2017. In-service Training of Teachers of English as a Foreign Language in Estonia: Mapping of Trends and Opportunities. *Problems of Education in the 21st Century*, 75, 194-203.

Salager-Meyer, Françoise. 2011. Scientific discourse and contrastive linguistics: Hedging. *European Science Editing*, 37, 35-37.

Savchenko, Denys. 2022. *A corpus-based study of adjective intensification among native speakers and learners of English.* MA thesis. Department of English, University of Tartu, Tartu, Estonia.

Schauer, Gila A. 2006. Pragmatic Awareness in ESL and EFL Contexts: Contrast and Development. *Language Learning*, 56, 269-318.

Staples, Shelley, Jesse Egbert, Douglas Biber and Bethany Gray. 2016. Academic Writing Development at the University Level: Phrasal and Clausal Complexity Across Level of Study, Discipline, and Genre. *Written Communication* 33:2, 149–183.

Sundh, Stellan. 2018. Estonian, Latvian, Lithuanian, Russian and Swedish Young Learners' Written Production in EFL - Descriptions and Comparisons of Their Use of Vocabulary. *International Journal of Language and Linguistics*, 5:4, 17-27.

Sönning, Lukas. 2016. The dot plot: a graphical tool for data analysis and presentation in Hanna Christ, Daniel Klenovšak, Lukas. Lukas Sönning and Valentin Werner (eds), *A Blend of MaLT: Selected Contributions from the Methods and Linguistic Theories Symposium 2015*, 101–32. Bamberg: University of Bamberg Press.

Tammekänd, Liina and Reeli Torn-Leesik. 2022. POS-Tagging Tartu Corpus of Estonian Learner English with CLAWS7. *Eesti Rakenduslingvistika Ühingu Aastaraamat (Estonian Papers in Applied Linguistics)* 18: 263-278.

Tono, Yukio. 2003. Learner corpora: design, development and applications. In *Proceedings of the Corpus Linguistics 2003 conference*, 800-809.

Toutanova, Kristina, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In *Proceedings of HLT-NAACL 2003*, 252-259.

Vaher, Anna. 2021. *The use of first-person pronouns in academic texts.* Unpublished BA thesis, Department of English Studies, University of Tartu, Tartu, Estonia.

Vassiljev, Liina, Liljana Skopinskaja, and Suliko Liiv. 2015. The Treatment of Lexical Collocations in EFL Coursebooks in the Estonian Secondary School Context. *Eesti Rakenduslingvistika Ühingu Aastaraamat* 11, 297-311.

Wang, Yumin and Qin Xie. 2022. Diagnosing EFL Undergraduates' Discourse Competence in Academic Writing. *Assessing Writing*, 53, 100641

Werner, Valentin. 2021. Catchy and conversational? A register analysis of pop lyrics. *Corpora*, 16, 237-270.

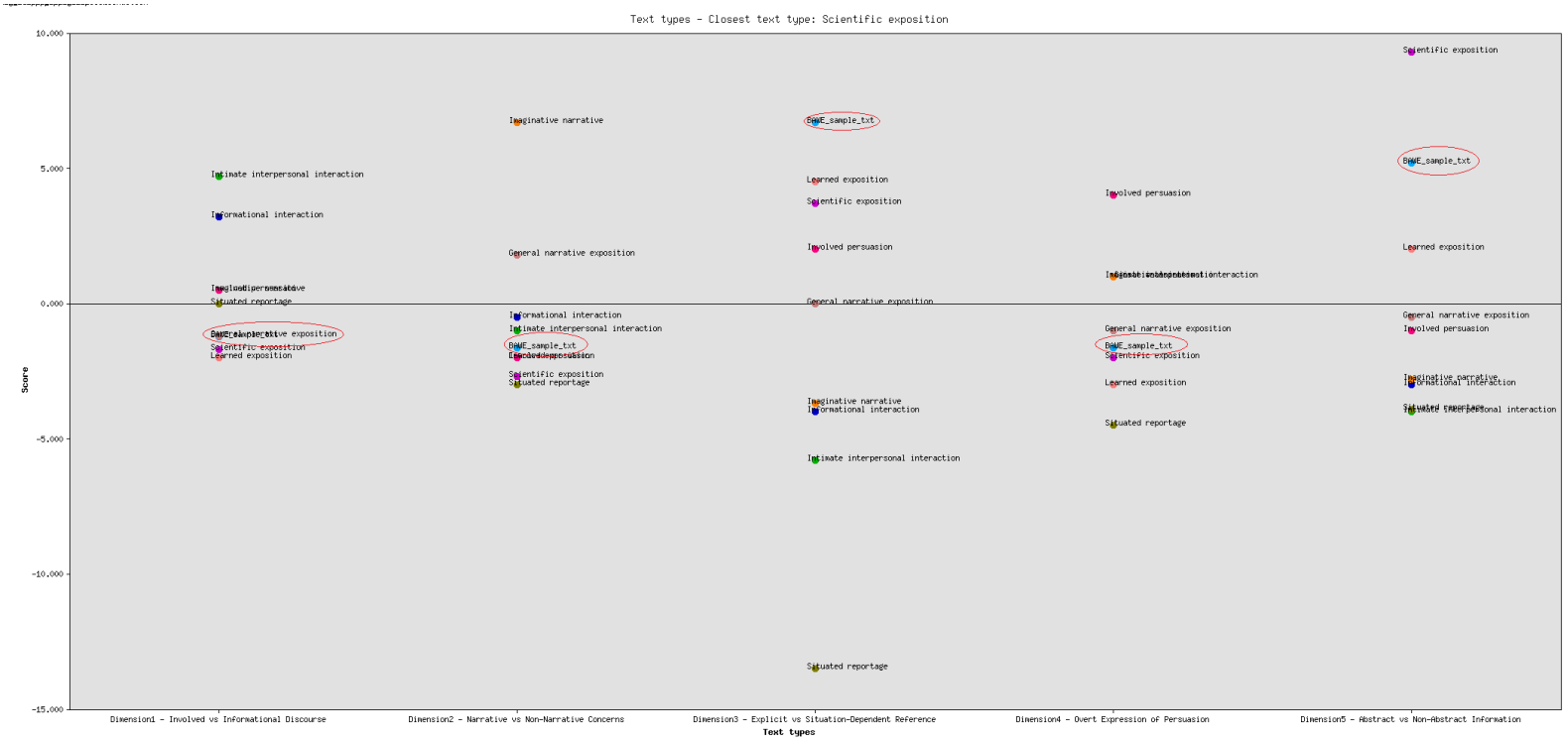
Yallop, Roger M. A., and Djuddah A. J. Leijen. 2018. The Perceived Effectiveness of Written Peer Feedback Comments within L2 English Academic Writing Courses. *Eesti Rakenduslingvistika Ühingu Aastaraamat.* 14, 247-271.

Yevchuk, Alina. 2021. Plesionyms as a Vocabulary Teaching Tool: The Case of Estonian EFL Learners. *Sustainable Multilingualism*, 19:1, 203–26.

APPENDICES

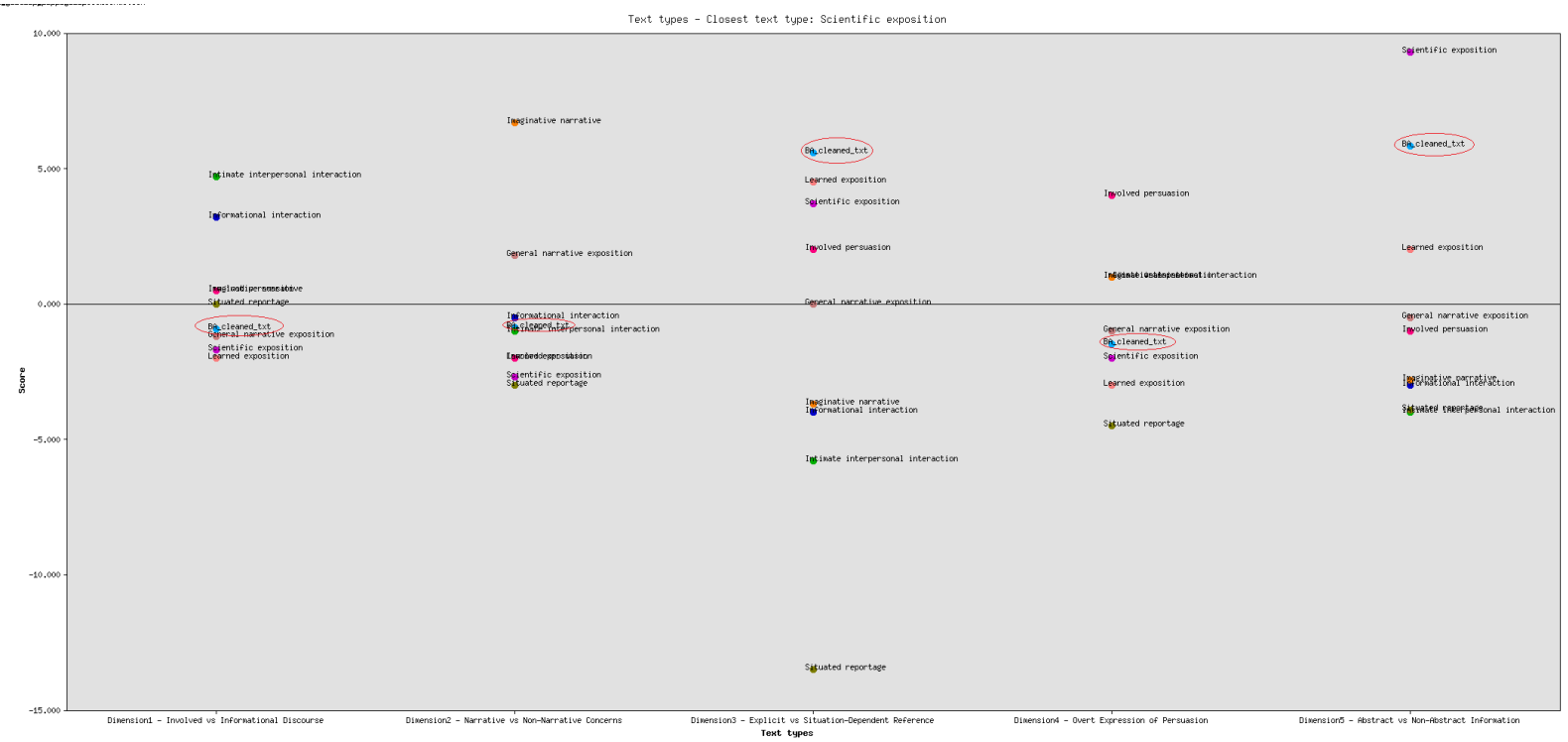
Appendix 1

The assignment of text types across the six dimensions in MDA for the BAWE sample according to the MAT tagger (Nini 2019)



Appendix 2

The assignment of text types across the six dimensions in MDA for the BA theses sample according to the MAT tagger (Nini 2019)



RESÜMEE

TARTU ÜLIKOOL
ANGLISTIKA OSAKOND

Janely Rüdein
Multi-dimensional Analysis of Academic Writing in Estonian Learner English. Eesti inglise keele õppijate akadeemilise kirjutamise multidimensionaalne analüüs.
bakalaureusetöö
2023
Lehekülgede arv: 42

Annotatsioon:
Antud bakalaureusetöö uurib Eesti inglise keele võõrkeelena (EFL) õppijate akadeemilist kirjutamist bakalaureusetöödes multidimensionaalse analüüsi (MDA) meetodi abil. Töö eesmärgiks oli illustreerida õppijakeelt multidimensionaalse analüüsi tulemuste kaudu, uurida valitud akadeemiliste tekstide valimite klassifitseerumist multidimensionaalse analüüsi dimensioonide, tekstitüüpide ja -žanrite läbi, ning vaadelda, kuidas erineb inglise keelt emakeelena kõnelejate (L1) akadeemiline kirjutamine õppijate poolt kirjutatud akadeemilistest tekstidest.

Käesolev töö on jaotatud kahte osasse, millest esimene on kirjanduse ülevaade. Sellele järgneb empiiriline analüüs teises osas. Kirjanduse ülevaates selgitatakse multidimensionaalse analüüsi meetodi keskseid mõisteid ja vaadeldakse varasemaid uuringuid, milles metoodikat kasutatud on, ning käsitletakse õppijakorpuste uuringute eripärasid, ning varasemaid uurimusi, mis käsitlevad akadeemilise kirjutamist inglise keelt võõrkeelena rääkivate populatsioonide kontekstis. Empiirilise analüüsi osa selgitab MDA protsessi läbiviimist, kasutatud andmeid ja valimeid (BAWE ja bakalaureusetööd), ning analüüsi tulemusi, mille kirjeldustele järgneb arutelu.

Tekstide analüüs näitas, et mõlemaid valimeid klassifitseeriti *scientific exposition* (teaduslik esitlus) tekstitüübina, mis viitab passiivse kõneviisi tihedale kasutusele tekstides. Tekstižanri *academic prose* (akadeemiline proosa) alla klassifitseerusid mõlemad valimid kahel dimensioonil kolmest. Andmete analüüsist selgus, et kuigi Eesti EFL õppijate akadeemiline kirjutamine erineb L1 inglise keeles kirjutatud akadeemilisest kirjutamisest modaalverbide kasutuse ning rõhumäärsõnade (*emphatics)* kasutuse poolest, on mõlemad valimid pigem tehnilise ning akadeemilise keelekasutusega.

Märksõnad:

Inglise keel, õppijakeel, korpusuuring, akadeemiline kirjutamine, multi-dimensionaalne analüüs,

Lihtlitsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Janely Rüdein,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Multi-dimensional Analysis of Academic Writing in Estonian Learner English
mille juhendaja on Jane Klavan,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Digiallkirjastatud
Janely Rüdein

Tartus, 23.05.2023

Autorsuse kinnitus

Kinnitan, et olen koostanud käesoleva bakalaureusetöö ise ning toonud korrektselt välja teiste autorite panuse. Töö on koostatud lähtudes Tartu Ülikooli maailma keelte ja kultuuride kolledži anglistika osakonna bakalaureusetöö nõuetest ning on kooskõlas heade akadeemiliste tavadega.

Digiallkirjastatud
Janely Rüdein

Tartus, 23.05.2023

Lõputöö on lubatud kaitsmisele.

Jane Klavan

Tartus, 23.05.2023