doi:10.5128/ERYa19.05

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 19, 75–92

A REGISTER APPROACH TO ESTONIAN EFL LEARNERS' UNIVERSITY WRITING

Jane Klavan

Abstract. The present study assesses the alleged informality of academic texts written by Estonian learners of English, which to date has yet to be empirically tested. It relies on the purpose-built Tartu Corpus of Estonian Learner English and applies Multidimensional Analysis (MDA) to situate the Estonian EFL learner texts relative to other spoken and written registers and L1 English university writing. In addition, the study describes the linguistic features characteristic of Estonian EFL learners' writing in higher education. The MDA indicates that although there are some differences between learner and L1 English university writing, the two data samples are similar and align on several dimensions with the written register of academic prose. At the same time, the differences between learner production and L1 English professional writing imply that Estonian students of English would benefit from more explicit instructions to raise their language and register awareness in terms of academic writing in English.

Keywords: register awareness, learner writing, expert writing, multidimensional analysis, English

1. Introduction

One of the major problems that learners face in their university writing is their lack of register awareness. English as a Foreign Language (EFL) learners tend not to achieve an appropriate level of formality compared to L1 writing. EFL learners tend to use features more typical of speech than academic prose. According to Larsson and Kaatari (2019: 54), many of the previous studies have commented precisely on learners' informal style by looking at the overuse of speech-like features. Currently, there are no studies that target Estonian EFL learners and their writing at the university level. The study aims to assess the alleged informality of academic texts written by Estonian learners of English. Using the Multidimensional Analysis (MDA), learner writing will be situated relatively to other spoken and written registers in L1 English and L1 English university writing.

Studies investigating the "spoken-like" nature of learner writing have demonstrated that learners of different L1 backgrounds either over- or under-use specific features. For example, French-speaking learners overuse features such as first and second person pronouns or short Germanic adverbs (*also*, *only*, *so*, *very*, etc.) typical of speech but underuse many characteristics of formal writing, e.g., nominalisation and prepositional phrases (Granger, Rayson 1998). In a more comprehensive overview, Gilquin and Paquot (2008) extend the study of spoken-like features in learner academic prose beyond one or two L1 populations and show that this problem is general for learners of English from many L1 backgrounds (in total, 14 L1 subcorpora were studied, Estonian was not included in the data); see Gilquin et al. (2007) for a detailed description of the project.

Although learners' tendency to use informal writing is relatively well established for EFL learners from various L1 backgrounds, no study has looked at Estonian learners of English. This is problematic since the informality of Estonian EFL learners' university writing is taken as a given. The present study aims to fill this gap. Differently from the studies reported above, which have looked at specific speech-like features in learner writing, the present study attempts to locate EFL and L1 English university writing holistically on the speech-writing continuum. Seminal work conducted by Biber (1988, 1989) has shown the vast extent of variation in language – different registers display different characteristics. Learners, however, are not attuned to the differences between speech and writing, as has been repeatedly noted in the literature.

Following Biber et al. (1999: 15), I use the term register to distinguish between situational characteristics of different texts. See also Lee (2001) for a comprehensive discussion of the difference between the terms register, genre, and text type. The present study avoids taking a dichotomous view of formality where student production is either seen as formal or informal. Following Larsson and Kaatari (2020, 2019), a more nuanced picture of formality is assumed with two specific assumptions: 2) formality is viewed as a continuum rather than a dichotomy; 2) registers can be placed along this continuum based on their situational characteristics (Biber et al. 1999: 16). Registers are placed on the informal-to-formal continuum based on their a priori situational characteristics (Larsson, Kaatari 2020: 2). Following Biber et al. (1999: 16), an ordering of the registers from more formal to less formal looks as follows: academic prose, popular science, news, and fiction. Register will be used as a proxy to see where students' writing can be placed on the (in)formality continuum.

Based on previous research on other L1 learners of English, the learners are expected to write in a more "informal" manner; both differences and similarities are expected between learner and L1 English student writing. In the context of the "spoken-like" nature of learner writing, it is essential to keep in mind that some studies have shown the "informality" of the learner academic texts to depend on text type (Larsson, Kaatari 2019). Informality of student writing may indicate insufficient register awareness – learners use language structures more strongly associated with one of the non-academic registers. They are proficient language users but do not pick up on the register differences to the same extent as L1 users.

The following research questions are used to guide the analysis:

- Which of the L1 English registers is the Estonian EFL learners' writing closest to, and what can this tell us about learner university writing?

- Which of the L1 English registers is L1 English students' writing closest to, and what can this tell us about L1 English university writing?
- What differences and similarities exist between L1 English students' university writing and Estonian EFL learners?

In what follows, I will summarise previous register-based research on learner writing (Section 2), followed by an overview of the corpus data, the situational variables, and the methodology used (Section 3). I will present the results of the MDA in Section 4. In the final part (Section 5), I will discuss and contextualize the findings and outline implications for EFL education in Estonia.

2. Register-based studies of academic writing in L1 and L2 English

The multi-dimensional approach to studying textual variation, as developed by Biber (1988, 1989), analyses a vast set of linguistic features across different registers. Critical to this approach to register is the understanding that no one dimension is equal to a straight-forward distinction between speech and writing (cf the non-dichotomous approach to register as discussed in Larsson, Kaatari 2020, 2019). All the dimensions are needed to characterise the difference between the registers in the study. The strength of the MDA is in its powerful, radically corpus-based, and statistically advanced approach to problems in text and language analysis. A key methodological aspect is that the subsequent studies do not repeat the process that produced the dimensions on the set of new data, but rather the dimensions established by Biber (1988, 1989) are treated as given. The new set of texts is positioned on those dimensions.

More recent work within the MDA framework has focused on language as it is used in the context of universities. This research paradigm aims to assist non-native speakers in English-speaking higher educational settings (Biber 2006: 2). More broadly, a register perspective has been taken on learner language in general (Larsson 2019, Larsson, Kaatari 2019, 2020). Larsson and Kaatari (2020) found some support for the claim that learners tend to be somewhat informal in their writing when comparing the learner texts to expert academic writing. However, Larsson (2019), who examined sets of grammatical stance markers that are morphologically and semantically related across five registers in apprentice (i.e., learner) and expert production, found very little evidence to support previous claims of the "spoken-like" nature of learner writing. When Larsson (2019) added native-speaker student data to the analysis, it became clear that both sets of apprentice writers exhibited surprisingly similar behaviour concerning the stance markers studied. Two crucial aspects follow from these previous studies that are important from the perspective of the present study. First, learner data should not be only compared to expert writers but also native speaker student writing (hence the inclusion of the BAWE data sample); and second, the traditional view of (in)formality as a binary factor is not sufficient (hence the MDA approach to register).

Perhaps the most relevant previous work in the context of the present study is Larsson and Kaatari (2019), who investigate to what extent register and text type can be used to explore learners' reportedly "informal" use of the subject extraposition construction (e.g., *it is important to remember*). Their results show important differences across both registers and text types. Even though the learners' use is very similar to expert academic writing, certain similarities to the non-academic registers were also noted. Larsson and Kaatari (2019) stress that the earlier claims about the informal status of learner writing (cf. the literature reported in Section 1 of this paper) seem to have been driven by the text types included in the corpora previously investigated. The comparison of learner data from the ALEC and VESPA to the BNC-15 data showed that the learners used the construction primarily in an expert-like manner (Larsson, Kaatari 2019: 53).

Larsson and Kaatari (2019: 54) conclude that dismissing all learner writing as "informal" appears to be an oversimplification. Text type is one of the factors to be taken into account in the discussion of (in)formal language use. According to Larsson and Kaatari (2019: 52), factors likely to play a role in the language used in ALEC/VESPA (theses) and SWICLE/LOCNESS (argumentative essays) include differences concerning text length (theses are considerably longer than essays), number of drafts permitted (redrafting was allowed for the theses) and whether the texts are timed or not. In terms of finding comparable data, Larsson and Kaatari (2019: 53) point out that expert academic writing is characterized by longer, untimed academic writing; it would seem, therefore, that learner corpora such as ALEC and VESPA are more suitable for comparisons with longer expert texts than SWICLE.

Researchers criticise the use of professional writing in learner corpus research as a point of comparison, going as far as claiming the "expert writer" model to be an "unrealistic standard" (Hyland, Milton 1997: 184) and the comparison with it to be "both unfair and descriptively inadequate" (Lorenz 1999: 14). Although Gilquin and Paquot (2008: 5) propose that native student writing is arguably a better type of comparable data for EFL learner writing, they seriously doubt whether findings from such a comparison will find their way into the classroom. L1 English-speaking students do not necessarily serve as good role models for learners to imitate. The question of the norm is a challenging one and can only be settled by taking the aim of the comparison into account. From the perspective of the present study, both types of data – L1 English professional writing and L1 English student writing – are needed to compare Estonian EFL student writing in the university setting. Advanced foreign learners strive for the norm represented by professional writing, while L1 English student writing is needed to provide a fairer evaluation of EFL learner writing. In any case, whatever differences between student writing and expert writing we may find may reflect the differences in their communicative goals and settings.

3. Data and method

In this section, an overview of both the data and method used is given. The corpora used will be introduced in Section 3.1, and the method will be described in Section 3.2.

3.1. Data used in the study

The study uses data from two corpora: the Tartu Corpus of Estonian Learner English (TCELE) and the British American Written English (BAWE). Currently, TCELE is not publicly available, and a project is underway to collect Estonian learner English data for both the spoken and written registers. The learner data sampled for the present study comprises 76 texts from TCELE. These texts are untimed BA theses written by English language and literature students at the University of Tartu whose self-reported first language is Estonian and who are, on average, in their third year of university studies. Although it would have been preferable to have more detailed information about the student's level of proficiency in English, no such tests were applied; level of proficiency will therefore not be included as a factor in the present study. Entry to the English language and literature BA programme at the University of Tartu requires students to demonstrate their ability of English at level C1 in the Common European Framework of Reference for Languages (CEFR) level. Therefore, the study targets the writing of advanced English learners.

In order to ensure comparability to the greatest extent possible, the BAWE corpus was carefully sampled to be as similar as possible to the learner corpus with regard to text type and discipline. Nonetheless, unavoidable differences remain between the learner and L1 English corpora, mainly pertaining to the length of the texts, with the L1 texts being shorter than the learner texts. Since there are no differences across levels of study in BAWE, all four levels were included in the analysis (see Gardner et al. 2019 for the analysis). The following disciplines were sampled from BAWE: Comparative American Studies, English, and Linguistics; from among the 13 genre families, the following were deemed most appropriate for the present study: critique, essay, methodology recount, literature survey, narrative recount, proposal, research report. In total, 295 texts were sampled from the BAWE corpus. Table 1 gives an overview of the data included in the present study.

Table 1. Overview of the data included

Data	Nr of texts	Word count
TCELE sample (Estonian L1)	76	491,198
BAWE sample (L1 English)	295	692,683

It is important to stress, once again, that from the perspective of the present study, the student essays from BAWE are not necessarily seen as a norm for Estonian learners to strive for; they provide a point of comparison. Following Larsson and Kaatari (2019), who studied (untimed) BA theses written in English by Swedish L1 students, the text type of both student corpora is referred to as academic prose. The situational characteristics of the texts included in this study are taken to be most similar to the text type "academic prose" proposed by Biber et al. (1999). Academic prose is "a very general register, characterized as written language that has been carefully produced and edited, addressed to a large number of readers who are separated in time and space from the author, and with the primary communicative purpose of presenting information about some topic" (Biber, Conrad 2019: 32).

3.2. Method: the MDA approach

The approach taken in the present study is register-based (Biber, Conrad 2019) – parameters of the situation of use of a particular text variety (learner and L1 English university writing) as well as linguistic features commonly occurring in this text variety are analysed and related to the communicative functions and purposes of linguistic features in texts of this register. Quantitative MDA, as developed in the seminal works of Biber (1988, 1989), enables to contrast student writing with other (established) registers and provides insight into the (in)formal nature of both learner and L1 English university writing.

To conduct the quantitative MDA, I used the Multidimensional Analysis Tagger (MAT; Nini 2018, 2019), a Java-based tool that is freely available for general use. According to Nini (2019: 71), MAT is a replication of the tagger developed for Biber's (1988) study with its programming based on the information on the algorithms used in the original work (see Biber 1988: Appendix II). MAT creates a frequency profile of the target text type based on multiple linguistic features and locates this text variety both along Biber's (1988) dimensions and relative to other text types (academic prose, official documents, etc.) established in Biber (1989). MAT locates the overall input along the six dimensions (according to Biber 1988) and assigns a text type (according to Biber 1989) to each input text file. In addition to output in numbers (token frequencies for the linguistic features, z-scores, dimension scores), the "Analyser" also creates the input for visualizations of 1) dimension scores of the input register (vs. Biber's (1988) scores), including mean and range if the input corpus consisted of multiple texts, and 2) the location of the texts analysed relative to the Biber (1989) text types. In the following section, the dimension scores for Estonian EFL writing and L1 English student writing will be explored to see which of the registers their use is closest to on the speech-writing continuum.

4. Results

MAT provides an assessment of the data in relation to the text types identified in Biber (1989), both for the overall corpus and for each text contained therein, taking account of the individual text as a basic unit in an MDA. The program classifies each text according to its closer text type using Euclidean distance. In line with the research questions, this provides a general idea of the "register identity" (Nini 2019: 91) of learner and L1 English university writing compared to the established English text types. For the TCELE data, MAT assigns 'scientific exposition' as the closest type overall, as this is the category assigned to the majority of individual texts (60%, 45 out of 76). The second most frequent category is 'general narrative exposition' (30%, 23 out of 76). The other two text types identified in the sample are 'involved persuasion' (5%, 4 out of 76) and 'learned exposition' (5%, 4 out of 76). The individual variation between texts indicates that although there are strict guidelines for writing a BA thesis at the Department of English Studies at the University of Tartu and the students are required to take a course on academic writing, individual texts differ on the variables identified by Biber (1988). For the BAWE data, MAT also assigns 'scientific exposition' as the closest type overall – this is the category assigned to most individual texts (43%, 127 out of 295). The second most frequent category is 'learned exposition' (33%, 96 out of 295), and the third 'general narrative exposition' (18%, 53 out of 295). The fourth text type – 'involved persuasion' – is much less frequent, with 6 per cent of the texts (19 out of 295) assigned to this category. As with TCELE texts, considerable individual variation between the texts included in the BAWE sample can be seen.

Table 2 presents the mean score, standard deviations, and the range or the difference between maximum and minimum scores for both TCELE and BAWE texts in every six dimensions. The standard deviations indicate how tightly the scores within a genre are grouped around the mean score. The large numbers in the columns 'Range' and 'SD' in Table 2 are witnesses to the large variability in the texts included in the BAWE sample. Overall, the scores for the BAWE texts are more spread out than the scores for the TCELE texts. This is to be expected since texts in the TCELE sample are all of the same type (untimed BA theses), while texts in the BAWE sample represent different text types (critiques, essays, methodology recounts, etc.). Another observation is that depending on the dimension, the data can be either more or less spread out compared to the mean. For example, Dimension 2 and Dimension 6 exhibit the least variability, while the range of values for Dimension 1, Dimension 3, and Dimension 5 are fairly wide. Table 2 shows that the MAT analyser has assigned the same text type for both TCELE and BAWE texts for three out of six dimensions: Dimensions 4, 5, and 6. The closest text type genre differs across the two data samples in the first three dimensions.

Table 2. Dimension scores and text type assignment for the TCELE and BAWE texts across the six dimensions

Dimension	TCELE				BAWE			
	Mean score	Range	SD	Closest genre	Mean score	Range	SD	Closest genre
D1	−9.3	26.6	4.4	Broadcasts	−12.3	44.6	6.8	Acedemic prose
D2	−0.9	7.6	1.4	Conversations	−1.6	12.9	1.8	Academic prose
D3	5.6	11.1	2.1	Academic prose	6.7	19.1	2.6	Official documents
D4	−1.5	9.3	2.0	Press reportage	−1.6	16.6	2.5	Press reportage
D5	5.8	15.0	2.9	Academic prose	5.2	24.5	4.2	Academic prose
D6	0.1	5.7	1.1	Conversations	0.2	8.1	1.6	Conversations

Following is a graphical representation of the analysis results for every six dimensions. All of the figures were created with RStudio (RStudio Team 2016) using the package "lattice" (Sarkar 2008). The source code in Sönning (2016) was tweaked to create the plots used in this paper. Dimension scores for registers other than TCELE and BAWE are taken from Biber (1988). From these graphs, it becomes clear that even though the closest text type assigned for the TCELE and BAWE samples may differ in some dimensions (as indicated by the scores in Table 2), they are still relatively closely positioned on the graphs. For each dimension, a brief discussion of the key linguistic features associated with the dimension is given. This will shed light on the students' linguistic patterns as described by the register analysis conducted in the present study.

4.1. Dimension 1

Figure 1 presents the scores for Dimension 1 (Biber 1988: 129–135) and maps involved vs. informational production. Dimension 1 is of particular interest for the present study as it has been found to differentiate between “discourse with interactional, affective, involved purposes, associated with strict real-time production and comprehension constraints [and] discourse with highly informational purposes, which is carefully crafted and highly edited” (Biber 1988: 115). The distinction can be understood in terms of involved real-time production versus informational, edited production. Low scores on Dimension 1 indicate that the text is informationally dense, e.g., academic prose; high scores indicate that the text is affective and interactional.

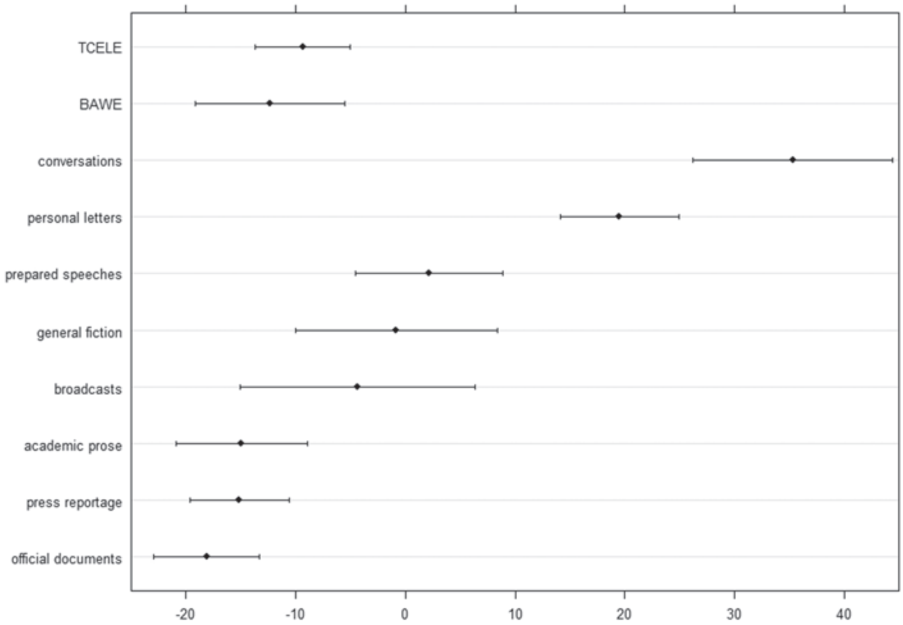

Figure 1. Scores for Dimension 1: involved vs. informational production
(dot = average; bars = +/– one standard deviation)

The scores for TCELE and BAWE are also relatively low (–9.3 and –12.3 respectively) and can thus be characterised as informational text variety closer to the more ‘literate’ end located toward the low end of the scale in Figure 1. Though there is substantial overlap of the bars, rather than aligning with academic prose, the closest text type for TCELE is broadcasts, a moderately involved but written register (Biber 1988: 132). Some of the typical features of informational texts are the use of nouns, prepositions, long words, more varied vocabulary, and attributive adjectives (Biber 1988: 129–131). Characteristic to the texts with a low score on Dimension 1 is high informational density: “there are many quite long words and a careful selection of vocabulary, resulting in a high type/token ratio” (Biber 1988: 131).

4.2. Dimension 2

Dimension 2, narrative vs. non-narrative concerns, distinguishes between "active, event-oriented discourse and more static, descriptive or expository types of discourse" (Biber 1988: 109). Scores for TCELE and BAWE and the other text types are shown in Figure 2. Low scores on this variable indicate that the text is non-narrative, whereas high scores indicate that the text is narrative.

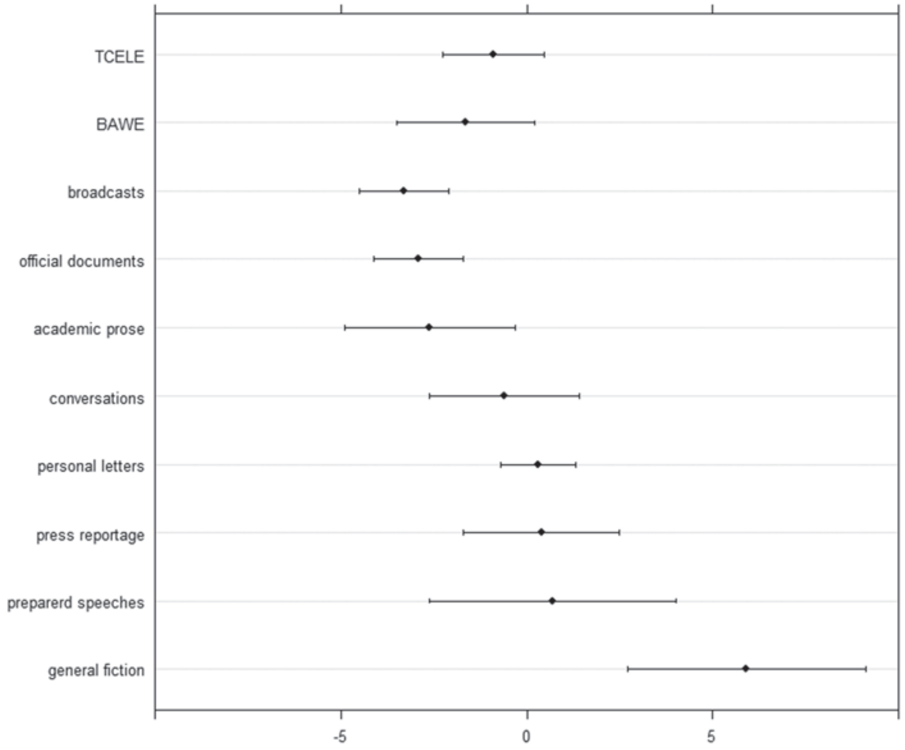

Figure 2. Scores for Dimension 2: narrative vs. non-narrative concerns (dot = average; bars = +/– one standard deviation)

The score for Dimension 2 for TCELE and BAWE texts is neither particularly high nor low. The mean scores (–0.9 for TCELE and –1.6 for BAWE) indicate that the text is slightly more non-narrative than narrative, but the value is only marginally lower than 0. In Dimension 2, both student writing samples align with many other text varieties, including, for example, academic prose (closest genre assigned to BAWE texts) and conversations (closest genre assigned to TCELE texts). The fact that both TCELE and BAWE texts do not have a particularly low or high score on Dimension 2 indicates that both learners and L1 English students use narrative and non-narrative features in their university writing production. Listed among the features that can be viewed as "markers of narrative action" (Biber 1988: 108), are verbs in the past tense and perfect aspect, third-person personal pronouns, public verbs, synthetic negation, as well as present participial clauses together with markedly infrequent occurrences of present tense verbs and attributive adjectives (Biber 1988: 135).

4.3. Dimension 3

The scores for Dimension 3, explicit versus situation-dependent reference (Biber 1988: 142–148), are displayed in Figure 3. This dimension has been found to discern textual varieties that more or less explicitly identify referents in the discourse (Biber 1988: 115). TCELE and BAWE texts have high scores on this dimension (5.6 and 6.7, respectively). This means that they primarily contain linguistic features characteristic of 'highly explicit, text-internal reference' (Biber 1988: 142), such as relativization, pied-piping, phrasal co-ordination and nominalisations together with infrequent use of place and time adverbials and other adverbs. In Dimension 3, TCELE texts align closely with academic prose and BAWE texts with official documents.

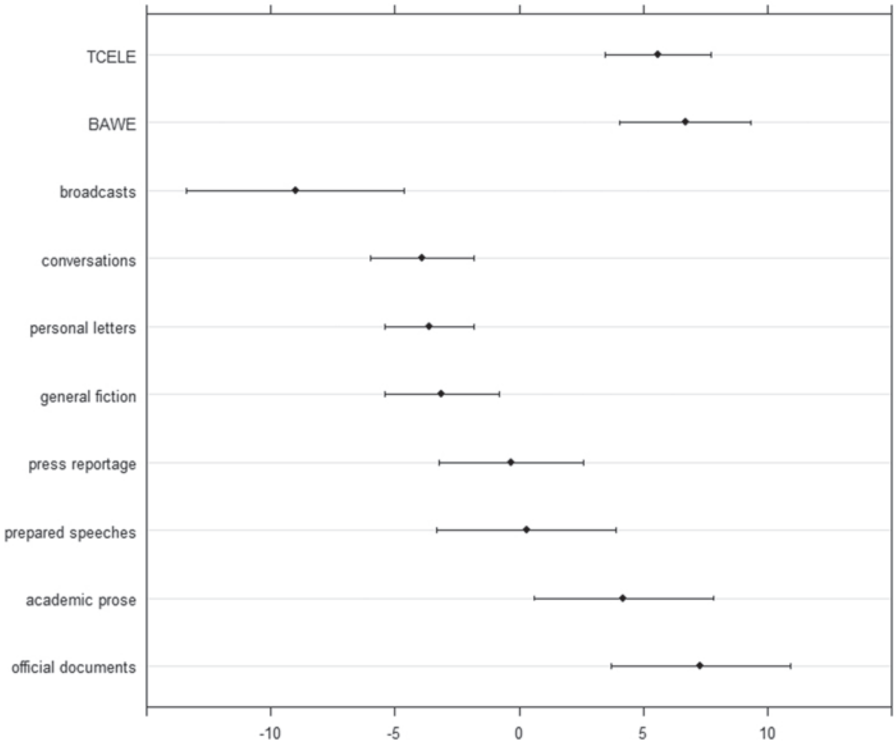

Figure 3. Scores for Dimension 3: explicit vs. situation-dependent reference
(dot = average; bars = +/– one standard deviation)

4.4. Dimension 4

Dimension 4, labelled overt expression of persuasion, is non-polar as it is only characterized by items with positive factor loadings. The higher the scores for this dimension, the more features "associated with the speaker's expression of own point of view or with argumentative styles intended to persuade the addressee" (Biber 1988: 115) are present. In addition, high scores indicate the author's assessment of the likelihood and/or certainty, cf., personal letters. Figure 4 displays the relevant values. Both TCELE and BAWE yield a moderately negative mean value (–1.5 and –1.6, respectively) for this dimension. For both data samples, the nearest general register is press reportage, while both align with several other text types, including academic prose. Characteristic of press reportage is the direct presentation of the author's opinion to be accepted or rejected as the reader chooses (Biber 1988: 151). Figure 4 shows that the ranges for almost all text varieties appear extensive for this dimension, partly due to the scope of Dimension 4 being comparatively narrow. Ultimately, the genres are relatively undistinguished along this dimension, suggesting that there is no general characterization as persuasive; instead, some texts are persuasive, while others are not (Biber 1988: 151).

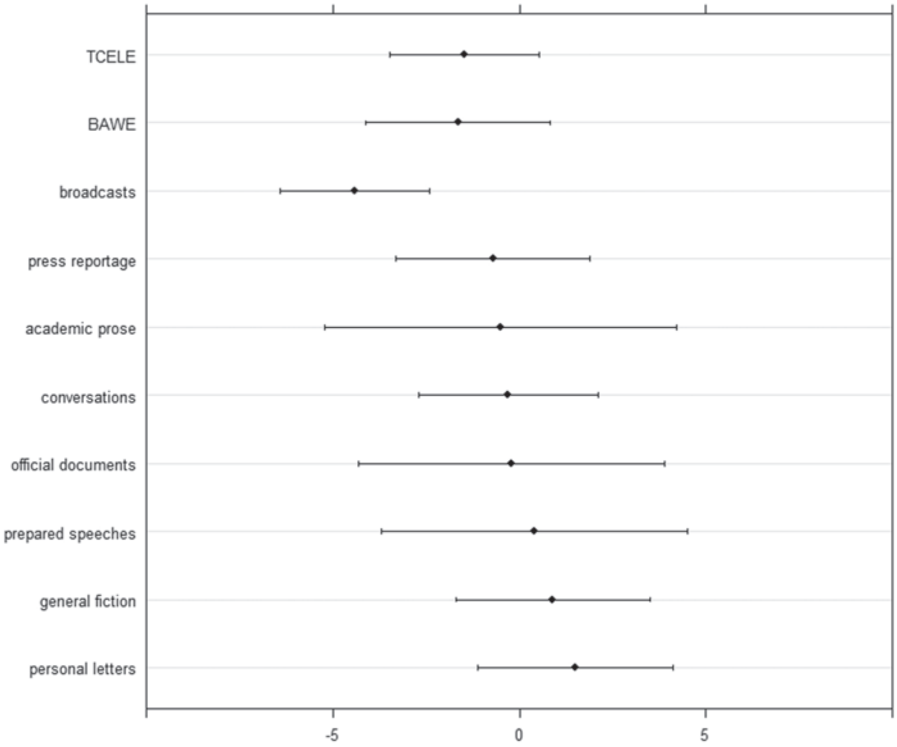

Figure 4. Scores for Dimension 4: overt expression of persuasion (dot = average; bars = +/– one standard deviation)

4.5. Dimension 5

The results for Dimension 5 – abstract versus non-abstract information (Biber 1988: 151–154) – are displayed in Figure 5. In text varieties with high scores for this dimension, concepts rather than agents are highlighted in the discourse (Biber 1988: 151–153). The results for this dimension are displayed in Figure 5. Both TCELE and BAWE texts yield a comparatively high score (5.8 and 5.2, respectively), indicative of an overall abstract type of language, like academic prose (the closest register for both TCELE and BAWE texts) or official documents. Characteristic linguistic features for texts that receive a high score on Dimension 5 are the occurrence of conjuncts, passives, past participial clauses, and adverbial subordination, while non-abstract discourse lacks those features (Biber 1988: 151–153).

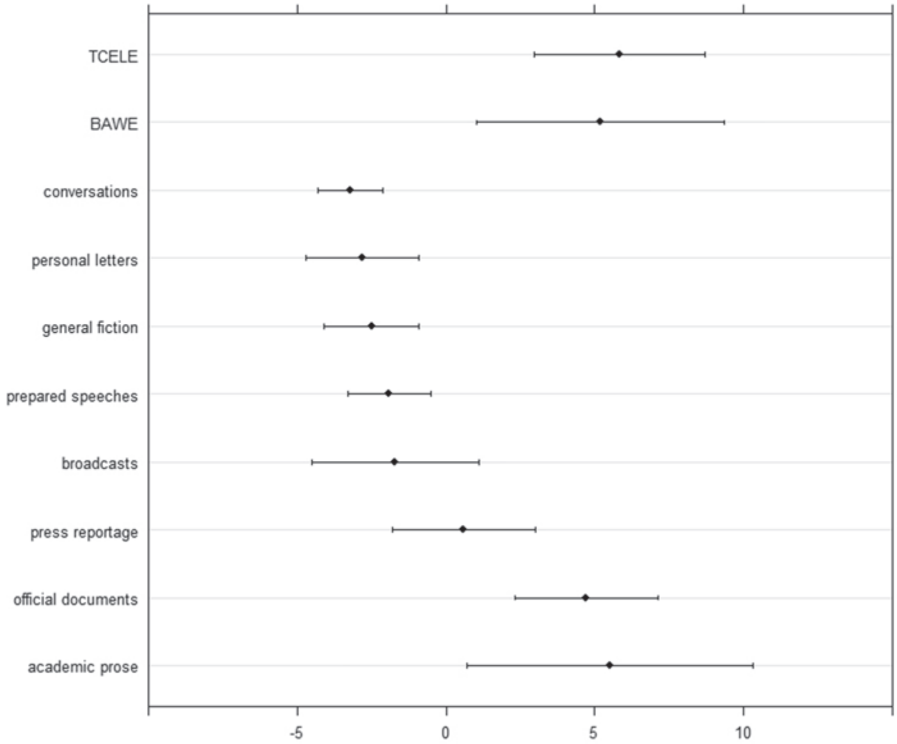

Figure 5. Scores for Dimension 5: abstract vs. non-abstract information
(dot = average; bars = +/– one standard deviation)

4.6. Dimension 6

Dimension 6, on-line informational elaboration, “distinguishes between informational discourse produced under highly constrained conditions, in which the information is presented in a relatively loose, fragmented manner, and other types of discourse, whether informational discourse that is highly integrated or discourse that is not informational” (Biber 1988: 115). High scores on this dimension mean that the text is informational but associated with time constraints, e.g., speeches and conversation. The relevant values are given in Figure 6 – the mean value of student writing, with a moderately positive score (0.1 for TCELE and 0.2 for BAWE), aligns most closely with conversations. Almost all text varieties cover a relatively broad range of values in Dimension 6, and the scores overlap. Characteristic text features for texts with a high score on this dimension include *that*-complements to verbs and adjectives, *that*-relative clauses, and demonstratives (Biber 1988: 154). Features with lesser positive weights include final prepositions, existential *there*, demonstrative pronouns, and *wh*-relative clauses (*ibid.*).

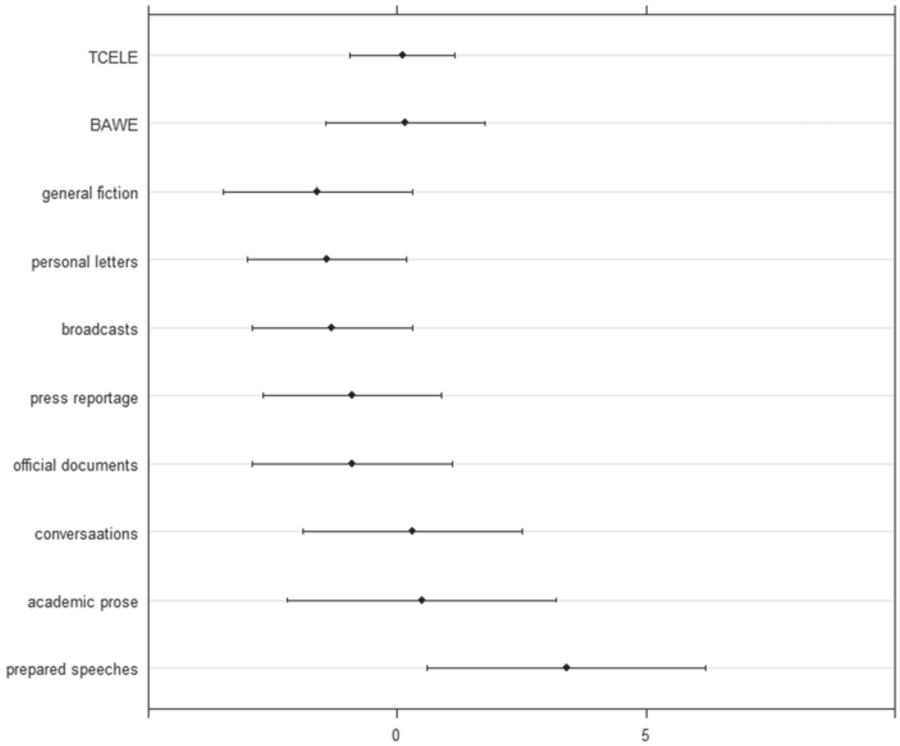

Figure 6. Scores for Dimension 6: On-line informational elaboration (dot = average; bars = +/– one standard deviation)

5. Concluding discussion

The primary goal of this study was to use the Multidimensional Analysis Tagger (MAT) that replicates the tagger used by Biber (1988) to analyse the textual relations in Estonian EFL university writing and L1 English university writing. These textual relations help us to see the extent of linguistic similarities and differences among learners' and L1 English writing. Samples from two corpora were used for these purposes – the untimed BA theses from the TCELE corpus and a sample of British students' university writing from the BAWE corpus. Biber's (1988) six parameters of variation have been used as underlying textual dimensions. Factor scores were computed by summing the frequency of each of the features on a factor for each text in the two data samples. The factor scores for each text provided by the Biber tagger are averaged across all texts to compute a mean dimension score for the sample. These mean dimension scores were compared to specify the relations among different sets of texts, e.g., comparing Estonian learner English to L1 English. One of the tenets of the study, supported by previous register-based studies (Larsson, Kaatari 2019, 2020), is that register and the related concepts of (in)formality are not binary, but rather a multi-dimensional perspective should be taken which allows us to place registers on a continuum.

The study addressed three specific research questions. The first research question – which of the L1 English registers is the Estonian EFL learners' writing closest to, and what can this tell us about learner university writing? – can be answered by looking at the results of three dimensions. According to Biber (1988: 160), it is Dimensions 1, 3, and 5 that present the oral/literate dimensions with the poles characterizing academic exposition and conversation, respectively. The closest genres assigned for the TCELE texts are broadcast for D1 and academic prose for D3 and D5. The scores for learner university writing along these three dimensions show that the texts are informational, use explicit reference, and provide abstract information. A multi-dimensional perspective of Estonian EFL learners' university writing exhibits characteristics of academic exposition of L1 English.

The second research question of the study asks which of the L1 English registers is L1 English students' writing closest to, and what can this tell us about native university writing? Very similar results are obtained for L1 English learner writing. The closest genres assigned for the BAWE texts are academic prose for D1 and D5 and official documents for D3. Similarly to learner university writing, L1 English university writing is also informational, makes use of explicit reference, and provides abstract information. According to the multi-dimensional analysis applied in the study, L1 English university writing also exhibits characteristics of academic exposition of L1 English. The BAWE texts showed considerable variation across the six dimensions assessed, as suggested by the comparatively wide standard deviations. This may be due to the fact that the sample used includes material from various sub-genres or that individual texts have a broad range of communicative concerns.

Based on the discussion above and the analysis of the graphs presented in Section 4, the third research question can be addressed: what differences and similarities can be found between L1 English students' university writing and that of the Estonian EFL learners? For most of the dimensions covered in the paper, the differences between the TCELE and BAWE texts are rather small. The mean scores for both sets of samples are positioned very closely on the graphs. The calculation of differences in the mean scores indicates that the largest differences are observed for Dimension 1 (difference in mean scores is 3.0), Dimension 3 (difference in mean scores is –1.1), and Dimension 2 (difference in mean score is 0.75). These are also the dimensions for which the MAT tagger has assigned for the TCELE and BAWE texts a different text type as the closest register.

Both learners and L1 English student writers exhibit characteristics of academic prose. Biber and Conrad (2019: 114–129) list fifteen features common to academic prose – all of these features are also present in the text samples of Estonian EFL learners' writing. Nominal features are one of the most obvious ways in which academic prose differs from, say, conversation. Sentences tend to be long, often containing only one finite verb but many nouns, resulting in a much higher number of nouns than verbs. The referents are very specific since nouns are modified by adjectives and prepositional phrases. Related to the linguistic features characteristic of academic prose are the situational characteristics of this text type: the specific purpose, production and comprehension circumstances, and the physical setting. Biber and Conrad (2019: 118) point out that academic prose has the general purpose of informing with plenty of time for planning, revising, and editing the language. This is certainly true for the TCELE texts in the study: how much revising and editing of the text has been done by the learners themselves, and how much input has been received by outside sources, e.g., the supervisors of the theses? This untimed aspect of student writing allows them to formulate more dense noun phrases, fascilitating precise identification of the referents.

The precision of noun phrases with their various modifiers leads to academic prose having a high "type-token ratio": a measure of how many different words are used in a text (*ibid.*). Biber and Conrad further emphasise that the use of present tense rather than past tense verbs in academic prose is related to analyzing and explaining, not just reporting. The relatively high frequency of nominalization in academic prose is connected with the discussion of general (sometimes abstract) patterns and concepts, while the common use of linking adverbials is needed since interpretations must be made and conclusions drawn. Finally, academic prose is characterized by the dense use of passive verbs, which allows writers to structure dense, informational prose – the use of active voice would render it difficult for the readers to quickly see the main points of the sentences. (Biber, Conrad 2019: 118–123) Of course, these different features are present to a higher or lesser degree in both TCELE and BAWE texts. The mean scores discussed in the paper only represent an average text. Individual texts (and students) demonstrate a considerable degree of variation. Further research might focus on a specific set of features to determine the extent individual learners differ in using these features and compare their use to L1 English student writing.

Two issues are worth pointing out in light of the results discussed above. First, there are both similarities and differences in written production when Estonian EFL learner writing is compared to L1 English academic prose and L1 English student writing. Secondly, there are differences when comparing L1 English student writing to L1 English academic prose. This raises the question of what data should be considered when investigating learner university writing – should learners be compared to professional writers (L1 English academic prose in the present study) or apprentice writers (L1 English student writing in the present study)? The answer depends on one's research question. However, in light of the present study, it is clear that both sets of apprentice writers (learners and L1 writers) exhibit similar behaviour and diverge from expert academic writing. When studying the over- or underuse of specific linguistic features in learner writing, as is often done in learner corpus research, a valid base for comparison needs to be considered. Larsson and Kaatari (2019: 53), for example, state that learner corpora such as ALEC and VESPA are more suitable for comparison with longer expert texts rather than SWICLE since expert academic writing is characterized by longer, untimed, academic writing. Comparisons between the ICLE sub-corpora commonly used to represent learner academic writing and reference corpora that are not comparable to ICLE in terms of text type, text length, etc. should be treated with care. Readers interested in corpus comparability in learner corpus research should see the discussions in Ädel (2006) and Callies (2013). The question of what constitutes "academic writing" in the context of comparability between corpora is a vital methodological question for researchers not only in the field of learner corpus research but also in English for Academic Purposes (EAP), English for Specific Purposes (ESP) and Second Language Acquisition (SLA). The present study showed that at least according to the textual relations studied based on Biber's (1988) six parameters of variation, a considerable degree of linguistic similarity was present among Estonian EFL learners and L1 English student writing making these two sets of data comparable for the intents and purposes of potential follow-up studies.

The results of the present study confirm those of Larsson (2019) – when taking a broad, multi-dimensional register perspective, there seems to be little evidence pointing towards the "spoken-like" nature of learner writing. Both EFL learners' and L1 English students' university writing was mapped fairly closely to L1 English academic prose along the six dimensions covered in the study. Of course, we must keep in mind that the learner data in the present study comprised untimed BA theses. As shown by previous research (Ädel 2006, Callies 2013: 363–364, Larsson, Kaatari 2019), learner writing exhibits different characteristics depending on the type of text, making text type a more powerful predictor than native speaker status. Future studies need to look at Estonian EFL learners' written production in other text types besides untimed BA theses, e.g., (timed) argumentative essays. It is hoped that the results of the present study will contribute to a more nuanced discussion of register variation, (in)formality, and complexity in Estonian EFL university writing, thereby benefitting both Learner Corpus Research and English for Academic Purposes (EAP) instruction and theories.

References

Ädel, Annelie 2006. Metadiscourse in L1 and L2 English. Studies in Corpus Linguistics, 24. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.24

Biber, Douglas 1988. Variation across Speech and Writing. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, Douglas 1989. A typology of English texts. – Linguistics, 27 (1), 3–43. https://doi.org/10.1515/ling.1989.27.1.3

Biber, Douglas 2006. University Language: A Corpus-Based Study of Spoken and Written Registers. Studies in Corpus Linguistics, 23. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.23

Biber, Douglas; Conrad, Susan 2019. Register, Genre, and Style. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, Douglas; Johansson, Stig; Leech, Geoffrey; Conrad, Susan; Finegan, Edward 1999. Longman Grammar of Spoken and Written English. Harlow: Longman.

Callies, Marcus 2013. Agentivity as a determinant of lexico-grammatical variation in L2 academic writing. – International Journal of Corpus Linguistics, 18 (3), 357–390. https://doi.org/10.1075/ijcl.18.3.05cal

Gardner, Sheena; Nesi, Hilary; Biber, Douglas 2019. Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. – Applied Linguistics, 40 (4), 646–674. https://doi.org/10.1093/applin/amy005

Gilquin, Gaëtanelle; Granger, Sylviane; Paquot, Magali 2007. Learner corpora: The missing link in EAP pedagogy. – Journal of English for Academic Purposes, 6 (4), 319–335. https://doi.org/10.1016/j.jeap.2007.09.007

Gilquin, Gaëtanelle; Paquot, Magalli 2008. Too chatty: Learner academic writing and register variation. – English Text Construction, 1, 41–61. https://doi.org/10.1075/etc.1.1.05gil

Granger, Sylviane; Rayson, Paul 1998. Automatic profiling of learner texts. – Sylviane Granger (Ed.), Learner English on Computer. London: Addison Wesley Longman, 119–131. https://doi.org/10.4324/9781315841342-9

Hyland, Ken; Milton, John 1997. Qualification and certainty in L1 and L2 students' writing. – Journal of Second Language Writing, 6 (2), 183–205. https://doi.org/10.1016/S1060-3743(97)90033-3

Larsson, Tove 2019. Grammatical stance marking across registers: Revisiting the formal-informal dichotomy – Register Studies, 1 (2), 243–268. https://doi.org/10.1075/rs.18009.lar

Larsson, Tove; Kaatari, Henrik 2019. Extraposition in learner and expert writing: Exploring (in)formality and the impact of register. – International Journal of Learner Corpus Research, 5 (1), 33–62. https://doi.org/10.1075/ijlcr.17014.lar

Larsson, Tove; Kaatari, Henrik 2020. Syntactic complexity across registers: Investigating (in)formality in second-language writing. – Journal of English for Academic Purposes, 45, 100850. https://doi.org/10.1016/j.jeap.2020.100850

Lee, David 2001. Genres, registers, text-types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. – Language Learning and Technology, 5 (3), 37–72.

Lorenz, Gunter 1999. Learning to cohere: Causal links in native vs. non-native argumentative writing. – Wolfram Bublitz, Uta Lenk, Eija Ventola (Eds.), Coherence in Spoken and Written Discourse: How to Create it and How to Describe It. Pragmatics & Beyond New Series, 63. Amsterdam: John Benjamins, 55–75. https://doi.org/10.1075/pbns.63.07lor

Nini, Andrea 2018. Multidimensional Analysis Tagger 1.3.1 https://sites.google.com/site/multidimensionaltagger/ (30.11.2022).

Nini, Andrea 2019. The Multi-Dimensional Analysis Tagger. – Tony Berber Sardinha, Marcia Veirano Pinto (Eds.), Multi-dimensional Analysis: Research Methods and Current Issues. London: Bloomsbury, 67–94. https://doi.org/10.5040/9781350023857.0012

RStudio Team 2016. RStudio: Integrated Development for R. Boston, MA: RStudio, Inc. http://www.rstudio.com (30.11.2022).

Sarkar, Deepayan 2008. Lattice: Multivariate Data Visualization with R. New York: Springer. https://doi.org/10.1007/978-0-387-75969-2

Sönning, Lukas 2016. The dot plot: A graphical tool for data analysis and presentation. – Hanna Christ, Daniel Klenovšak, Lukas Sönning, Valentin Werner (Eds.), A Blend of MaLT: Selected Contributions from the Methods and Linguistic Theories Symposium 2015. Bamberg: University of Bamberg Press, 101–132.

EESTI EMAKEELEGA INGLISE KEELE ÕPPIJATE KEELEKASUTUS REGISTRI PERSPEKTIIVIST

Jane Klavan
Tartu Ülikool

Artikkel keskendub eesti inglise keele õppijate akadeemilisele kirjutamisele ülikooli kontekstis ja analüüsib õppijate keelekasutust registri perspektiivist. Varasemad uurimused on järjepidevalt näidanud, et kui õppijad, sh edasijõudnud õppijad, inglise keeles kirjutavad, kipuvad nad kasutama keele struktuure, mis on pigem omased inglise keelt emakeelena kõnelejate suulisele keelele. Kuna eesti inglise keele õppijate keelekasutust ei ole sellest aspektist uuritud, siis puudub empiiriline alus väita, et ka eesti inglise keele õppijate kirjalikus eneseväljenduses peegelduvad suulise keelekasutuse mustrid. Artiklis on kasutatud multidimensionaalset analüüsi TCELE korpuse tekstide peal. Analüüsitud on inglise keele ja kirjanduse üliõpilaste bakalaureusetööde keelekasutust. Analüüsi eesmärgiks on võrrelda eesti inglise keele õppijate kirjalikke tekste erinevate suuliste ja kirjalike tekstitüüpidega, mis põhinevad inglise keelt esimese keelena kõnelejate keelekasutusel. Lisaks professionaalsetele kirjutajatele võrreldakse Eesti üliõpilaste ingliskeelseid tekste ka Briti üliõpilaste tekstidega. Analüüsist selgub, et hoolimata mõningatest erinevustest Eesti ja Briti üliõpilaste akadeemilises kirjutamises, on need kaks andmestikku oma registrikasutuse poolest üsna sarnased. Siiski näitab võrdlus professionaalsete kirjutajatega, et Eesti üliõpilastele tuleks rõhutada registritevahelisi erinevusi, pidades silmas, et erinevad registrid eeldavad erinevat keelekasutust.

Võtmesõnad: registrianalüüs, õppijakeel, akadeemiline kirjutamine, multidimensionaalne analüüs, inglise keel

Jane Klavan (Tartu Ülikool) on inglise keele kaasprofessor. Tema peamiseks teadustöö huviks on keeleteaduslikud meetodid, lingvistiline varieerumine ja õppijakeel.
Jakobi 2, 51005 Tartu, Estonia
jane.klavan@ut.ee