DISSERTATIONES
LINGUISTICAE
UNIVERSITATIS
TARTUENSIS
47

KRISTIINA VAIK

Beyond genres: A Dimensional Text Model for text classification

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS
47

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS
47

KRISTIINA VAIK

Beyond genres: A Dimensional Text Model for text classification

UNIVERSITY OF TARTU
Press
1632

Institute of Estonian and General Linguistics, Faculty of Arts and Humanities, University of Tartu, Estonia.

Dissertation accepted for the defence of the degree of Doctor of Philosophy on 5 November 2024 by the Committee of the Institute of Estonian and General Linguistics, Faculty of Arts and Humanities, University of Tartu.

Supervisors: Associate Professor Kadri Muischnek (University of Tartu)
Associate Professor Kairit Sirts (University of Tartu)

Opponent: Professor Veronika Laippala, PhD (University of Turku)

Commencement: December 19, 2024 at 14.15 at the Senate's Hall (Ülikooli 18-204, Tartu)

This study has been supported by the Graduate School of Linguistics, Philosophy and Semiotics; funded by the European Regional Development Fund (University of Tartu ASTRA Project PER ASPERA).

ISSN 1406-5657 (print)
ISBN 978-9916-27-750-8 (print)
ISSN 2806-237X (pdf)
ISBN 978-9916-27-751-5 (pdf)

University of Tartu Press
www.tyk.ee

ACKNOWLEDGEMENTS

What a strange journey it has been. A simple task, perhaps, but only those who've walked the same path can truly understand the liberation from completing your thesis. First and foremost, I express my immense gratitude to supervisors Kadri Muischnek and Kairit Sirts for their guidance and support. Without them, this thesis would not have been completed. I would like to thank all my friends. Without them, I probably would have finished this a bit earlier. I would also like to honour the Hamburg family – Triinu, Ets, Elias and Emil. Your kindness and generosity have always shown me that I have friends for life. My deepest gratitude goes to my mother and brothers, who have supported me morally, although I'm sure none of them know what I have been up to. I dedicate this thesis to my father, my greatest supporter, whose memory I will cherish forever. Finally, thank you, Priit, for being the only thing that truly matters.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Acronyms

AGI Automatic Genre Identification. 23, 30

DTM Dimensional Text Model. 15–17, 29, 43, 56, 104

FTD Functional Text Dimensions. 15, 23, 29, 104

MDA Multidimensional analysis. 15, 22, 23, 29, 104

1. INTRODUCTION

Even for those who are not experts in linguistics, it is evident that individuals modify their language depending on the circumstances. They may not consciously consider their word choices in daily communication, but they often possess an intuitive understanding of the typical linguistic features associated with specific contexts. For example, a casual conversation with a friend might involve slang and contractions, while a formal presentation would require more technical vocabulary and grammatically complex sentences. This shows that we inherently shape our language according to the communicative function and situation.

In modern times, corpus and computational linguists use the vast resources of the Web to study language patterns. These automatically collected text collections, known as Web corpora, serve as valuable resources, as they are large and cost-effective (Schäfer & Bildhauer 2013; Jakubíček et al. 2020). Although Web corpora only reflect a subset of the content of the Web (Mehler et al. 2010: 16), their extensive coverage of online content allows one to quantitatively analyze linguistic trends and patterns across a wide range of genres, registers, and styles. Moreover, the relatively low cost associated with obtaining and processing Web corpora makes them accessible for a wide range of linguistic studies. As technology continues to evolve, language is likely to adapt further, incorporating new forms of expression and communication tools into everyday interactions. Web corpora allow researchers to conduct cross-linguistic and diachronic studies, which helps to explore linguistic trends and variations on a global scale (Kulkarni et al. 2015; Kutuzov et al. 2018).

Although Web corpora are a valuable resource for many research fields, they are not without challenges. Unlike traditional text corpora (e.g., British National Corpus, Brown), where the sources for all texts are carefully curated, well described, and categorized, texts automatically collected from the Web are handled differently since the only available metadata is typically the URL. The lack of metadata restricts contextual understanding of texts, making it difficult to extract important information. In addition to traditional written content (e.g. news articles, prose, etc.), Web corpora also accommodate user- and computer-generated content (e.g. blogs, posts on social networks, etc.) which can vary greatly in quality, ranging from well-written and informative to informal, and grammatically incorrect. Consequently, this has presented a challenge in categorizing texts compared to traditional text compilation methods (Mehler et al. 2010; Schäfer & Bildhauer 2013). The unknown composition of the Web corpora also poses a major obstacle to linguistic analysis, as researchers may not have a clear picture of the types of texts represented in the corpora. The lack of transparency surrounding the content within Web corpora introduces uncertainties and complexities in interpreting and generalizing findings, highlighting the need for cautious and thorough validation in using Web-based linguistic data for research purposes. To enhance the utility of Web corpora, it is important to understand their composition.

1.1. Research object

The research object of this dissertation is to propose a method used for categorizing (Web) texts. Over the years, numerous classification taxonomies and schemes have been developed. These existing taxonomies differ from each other in terms of methodology, level of granularity, and functionality. This section briefly outlines how taxonomies were created in existing studies, with a more detailed description provided in Section 2.3.

In terms of methodology, classification taxonomies have been created using either the top-down or bottom-up method. The top-down method is based on theoretical principles and predefined taxonomies and categories, which are frequently derived from traditional corpus-building strategies (Santini et al. 2010; Jakubíček et al. 2020; Kuzman et al. 2022). In contrast, the bottom-up method is a more costly approach that involves annotators generating categories independently (Meyer zu Eissen & Stein 2004; Egbert et al. 2015). With the bottom-up method, categorization emerges from ground-level observations and input, although annotators are often provided with some general predefined categories by the researcher as a starting point.

Taxonomies can also be differentiated based on their nestedness, i.e., hierarchical or flat taxonomies. Hierarchical taxonomies are akin to library catalogue systems, with categories arranged in a parent-child relationship, forming a tree-like structure where general categories, such as fiction, include more specific subcategories, such as romance, mystery (Berninger et al. 2008; Kuzman et al. 2022). Although hierarchical taxonomies offer the ability to make finer distinctions, they also contribute to the overall complexity of the taxonomy. In contrast, flat taxonomies categorize texts at a single level with no structural relationships between the categories, such as politics, entertainment (Vidulin et al. 2009; Asheghi et al. 2016). Both taxonomies can present challenges when representing large or diverse corpora. For example, too many categories may contribute to the overall complexity, while too few may introduce ambiguity.

Taxonomies can also be distinguished by the labeling method, i.e., single- or multi-labeling. In single-label taxonomies, a text is assigned to only one category (Asheghi et al. 2016; Jakubíček et al. 2020), whereas in multi-labeled taxonomies, a text can belong to several categories simultaneously, e.g., a blog post can be a product review, news reporting, an opinion piece etc. (Biber & Egbert 2018; Sharoff 2018; Madjarov et al. 2019). Both strategies have their advantages and disadvantages. Single-label taxonomies, similar to flat taxonomies, simplify classification, but may not entirely capture the complexity of textual variation. In contrast, multi-label taxonomies can be more complex but are capable of capturing the hybrid nature of texts.

The application of existing taxonomies to the classification of Web corpora has presented a number of challenges, primarily due to the difficulty encountered in capturing the hybrid and evolving nature of Web texts. This has resulted in

different methodologies and granularities. Firstly, the existing taxonomies tend to range from broad categories (e.g., *journalism*, *fiction*) to more specific subcategories (e.g., *opinion pieces*, *sports reports*, *advertisements*), or from functional or situational styles to dimensional distinctions (e.g., *explanation*, *information*). Some taxonomies are also created focusing only on specific sub-languages with specialized categories, e.g., academic language, the Web language. Taxonomies often include arbitrary and miscellaneous categories, which are used as a catch-all category for texts that do not fit into other predefined categories. This naturally leads to the second challenge associated with the existing taxonomies, namely the difficulty of achieving a high level of agreement between the annotators (Meyer zu Eissen & Stein 2004; Crowston et al. 2010; Sharoff et al. 2010; Egbert et al. 2015; Suchomel 2020). This raises the question whether taxonomies created for classifying Web texts are even feasible at all, given that even expert and non-expert annotators are unable to unanimously categorize texts by register or genre (Suchomel 2020: 107).

This dissertation introduces the Dimensional Text Model (DTM) to address these challenges by building on these existing foundations and taking a step back to try to find the characteristics that define texts by their purpose, i.e., communicative function. The theoretical framework is based on the Multidimensional analysis of Biber (1988) and the Functional Text Dimensions of Sharoff (2018; 2021), which represent pivotal methodological approaches that have significantly advanced corpus-based research on language variation.

Multidimensional analysis (MDA) (Biber 1988) is a methodological framework that employs quantitative techniques to compare different registers through their co-occurring linguistic features. The foundation of MDA is rooted in identifying linguistic features and their associated communicative functions. The MDA framework utilizes factor analysis to identify the dimensions by examining the communicative purposes realized by the co-occurring linguistic patterns. This approach is predicated on a fundamental assumption that linguistic features co-occur because they share similar communicative functions. For instance, a high frequency of nouns and longer words often indicates an intent to integrate complex, precisely defined information — a characteristic frequently observed in academic discourse.

Similar to Biber's approach to studying language variation, Sharoff (Sharoff 2018) proposed the Functional Text Dimensions (FTD) framework. In this approach, each dimension represents a functional category that describes a text's communicative purpose — such as, *to what extent is the text concerned with expressing feelings or emotions?* The FTD framework employs a top-down method that enables multi-labelling, which allows describing texts through a combination of functional dimensions. The taxonomy offers a set of 18 dimensions, which are divided into primary (mandatory) and secondary (optional) dimensions. Since the presence of a dimension is measured on a scale (ranging from none to being strongly present), the FTD introduces the *multidimensional space* where each

text can be represented as a vector of functions. These vectors allow mapping functionally similar texts across corpora.

The objective of this dissertation is to reconcile these two approaches by creating a framework, the Dimensional Text Model (DTM), which could be used to automate the register classification task while also exploring how the texts differ linguistically and functionally. Both, the MDA and FTD frameworks focus on the duality of form and function: linguistic features are employed to express a particular function, and vice versa, but they have different starting points. While the MDA identifies underlying dimensions of linguistic variation from a wide range of spoken and written registers of English, the FTD follows traditional methods for automatic text classification by developing a taxonomy where each label is motivated by a function and then classifying large Web corpora into registers/genres by those functional dimensions. The DTM borrows concepts from MDA and FTD, such as *dimension* and *multidimensional space*, and adopts the term *register* from Biber & Conrad (2009) as a general cover term in which purpose and function are associated with linguistic features.

The DTM adapts the MDA's concept of a *dimension* as a quantifiable linguistic quality measured by co-occurring lexical, grammatical, and textual features. Unlike the MDA, which views dimensions as scales with two opposing functional poles, DTM treats dimensions as a continuum of salience representing a singular communicative function. The DTM's dimensions were generated by systematically breaking down the functionally complex MDA dimensions into more elementary, single-function dimensions. This resulted in a 12-dimensional space: *abstractness*, *affectivity*, *argumentativity*, *impersonality*, *interactivity*, *instructability*, *formality*, *complexity*, *spontaneity*, *information density*, *temporality* and *subjectivity*. Furthermore, the DTM extends the FTD's *multidimensional space* concept by representing texts as a vector in a multidimensional space of all dimensions. Whereas FTD limits its multidimensional representation through optional dimensions, the DTM maps texts in a 12-dimensional unified space where dimensions can be present to a limited extent or gravitate towards being more strongly present. Similarly to FTD, the spatial proximity between texts in the multidimensional space suggests that those texts share similar register characteristics.

This dissertation presents a theoretical framework that aims to show that the proposed dimensions represent different communicative functions, but its application to real-world data remains a topic for future research. Furthermore, while the DTM focuses on the Estonian language, it does not imply exclusivity towards other languages and is not limited to only Web corpora. It can be applied to a wide variety of texts, including printed texts and (written) spoken language. The DTM is thoroughly introduced in Chapter 3.

1.2. Research purposes and questions

The main objective of the dissertation is to propose the Dimensional Text Model (DTM), which addresses the challenges of the lack of a comprehensive taxonomy for text categorization and ways to deal with the evolving and unknown composition of web corpora, which have led to current taxonomies failing to achieve higher inter-annotator agreement measures. This dissertation aims to address these challenges by building upon existing research, but also by taking an alternative perspective to identify the fundamental elements that distinguish various types of text.

Regardless of the particular approach adopted, all models and frameworks must provide quality and reliability. This ensures that the results generated by the models are accurate, consistent, and generalizable to new data sets. This dissertation seeks to address the following questions.

(RQ1) To what extent is it feasible to annotate the data using the dimensions outlined in the Dimensional Text Model (DTM)?

The goal of RQ1 is to assess whether the theoretically proposed dimensions are recognizable and distinguishable to the annotators. The DTM represents the texts as vectors in a multidimensional space, where the presence of the dimension can be represented as a continuum of salience. For example, academic discourse is typically more abstract than a poem or a conversation between friends, and a poem is typically more abstract than a conversation, etc. To measure the continuous nature of the dimensions and whether the dimensions are distinguishable, an annotation study was conducted in which the annotators were asked to rate the level of salience of the proposed dimensions on a four-point Likert scale (*strong*, *moderate*, *weak* or *non-existent*) in a collection of texts gathered from the EnTenTen Web corpus (Koppel & Kallas 2022). While the annotation study was designed to measure the non-existence of dimensions, the underlying framework posits that all dimensions are inherently salient in every text, albeit with varying levels.

The agreement among the annotators was measured with a Krippendorff α (2004) agreement, which is an inter-annotator agreement that takes into account chance, different types of error, and other factors that can influence agreement levels. The analysis also included correlation and exploratory factor analysis. The purpose was to assess the collinearity between the dimensions, i.e., a high correlation indicates that the dimensions may be measuring similar communicative functions. Exploratory factor analysis was used to identify potential underlying patterns between the dimensions. These methods were not the primary focus of this dissertation, but they provided knowledge on how the proposed dimensions form macro dimensions by showing how the dimensions are similar to each other based on the annotations. The annotation process and the results of the inter-annotator agreement, correlation and exploratory factor analysis are discussed in Chapter 4.

While the objective of RQ1 is to validate whether the proposed dimensions are distinguishable, the next step is to further investigate whether and how these dimensions differ from one another in terms of their linguistic profiles. The Dimensional Text Model (DTM) assumes that the dimensions manifest through co-occurring linguistic features and that the feature distribution in different contexts is not random, but is motivated by a communicative function. Following this assumption, RQ2 emerges:

(RQ2) How do the dimensions defined by the Dimensional Text Model (DTM) differ from one another in terms of their statistically significant features (e.g., vocabulary choices, and grammatical structures)?

In natural language processing, two techniques have been fundamental for different text classification tasks: bag-of-words and word embeddings. Both are good at capturing statistical patterns in large-scale corpora, though they often fall short in representing fine-grained linguistic phenomena and grammatical relationships that are crucial for addressing RQ2 - the task of determining whether and how the dimensions proposed by the Dimensional Text Model differ from one another linguistically. Consequently, this dissertation adopted a quantitative approach based on the frequencies of lexico-grammatical features to establish more informative linguistic profiles for each dimension. Given the limited research in Estonia on the linguistic features that contribute to the discrimination of registers/genres, there is not much previous knowledge from which to draw. Accordingly, the choice of linguistic features was based on the output of Stanford's Stanza (Qi et al. 2020) parser. The objective was to include a diverse range of different features. For example, it included parts of speech, different grammatical categories of nouns and verbs, and syntactic features, and also incorporated language-independent textual features, such as average word length and type/token ratio. The list of the extracted features and their potential functional relationships underlying dimensions are discussed in Chapter 5.

To assess the statistical relevance of each feature on the level of salience in a given dimension, a one-way non-parametric analysis of variance (ANOVA) was performed. This analysis compared the variances of the medians of the relative frequencies between the three groups representing different levels of dimensional salience: *strong/moderate*, *weak*, or *non-existent*. A statistically significant ANOVA result would suggest that there is a difference between the groups in question, but would not identify which specific groups are different. For this reason, ANOVA is usually followed by *post hoc* tests. In the case of non-parametric data, Dunn's multiple comparison test was used to identify which specific pairwise comparisons of the groups showed statistically significant differences.

Establishing the co-occurring linguistic patterns allows for an analysis of the differences between the proposed dimensions. These findings are analyzed in Chapter 6.

1.3. Contributions

The contributions of this dissertation are to propose a theoretical framework which can be used for automatic register classification, while also exploring how the texts differ linguistically and functionally. In summary, the contributions of this dissertation are as follows:

I A theoretical framework of the Dimensional Text Model (DTM) describing texts through

   a) *dimensions* which capture the core characteristics of the communicative function of a text, manifesting through the co-occurring linguistic features;

   b) modelling texts in a *multidimensional space*. The assumption is that texts which are in proximity in the multidimensional space will share similar functions. These similar functions may be a representation of a register.

II A corpus-linguistic study:
Traditionally in Estonia, researchers have focused on a small set of specific handpicked linguistic features within limited registers/genres such as fiction, academic writing, or journalism. However, the Dimensional Text Model proposes a model capable of accommodating an extensive list of linguistic features in any text, thereby enabling the examination of language variation on a broader scale.

III Dataset:
The outcome of this dissertation is a manually annotated dataset comprising 120 Web texts. These texts are rich in dimension-specific annotations, which are based on the set of dimensions proposed in the Dimensional Text Model. The annotation scheme used a four-point Likert scale to capture the level of salience of each dimension within each text. The scale was interpreted as follows:

   – **strong**: the dimension is a prominent and defining characteristic of the text.
   – **moderate**: the dimension is present to some extent, but not the most defining characteristic of the text.
   – **weak**: the presence of the dimension is weak and cannot be regarded as a defining characteristic of the text.
   – **non-existent**: the dimension is not a relevant characteristic of the text.

1.4. Chapters outline

The remainder of this dissertation is structured as follows:

- Chapter 2 gives an overview of how text corpora have been used to study language variation in relation to registers and genres worldwide and in Estonia. The chapter further explores the existing ambiguity surrounding the terminology. Lastly, this chapter introduces various existing taxonomies to facilitate automatic text classification.
- Chapter 3 presents the Dimensional Text Model, a framework developed within the scope of this dissertation. This chapter elaborates on the dimensions defined within this framework.
- Chapter 4 explains how the data were annotated using the set of dimensions proposed by the Dimensional Text Model. The chapter describes the annotation process and the methods used to assess the validity of the annotation results, as well as discusses the results and their implications.
- Chapter 5 gives an overview of the linguistic features used to study the dimensional variation through the set of dimensions proposed by the Dimensional Text Model.
- Chapter 6 gives an overview of the methods and results of distinguishing the dimensions from one another in terms of their statistically significant features, highlighting the similarities and differences between the dimensions. The last section summarizes the key takeaways.
- Chapter 7 summarizes the Dimensional Text Model and discusses future work.

2. BACKGROUND

This chapter provides some background to the central themes of this dissertation. Section 2.1 gives an overview of how language variation has been studied over the years, both globally and in Estonia, and examines the role that the Web has played in shaping this field of research. Section 2.2 gives a brief overview of the diverse and confusing terminology used in this area of research, often described as a "jungle". Finally, Section 2.3 describes the various methods employed to create classification schemes.

2.1. The evolution of using corpora for language research

From Aristotle onward, rhetoricians have attempted to classify communication into "genres" based on similarities in their form or purpose. Investigating the linguistic variation systemically commenced during the mid-20th century, when researchers started advocating the notion that speakers possess knowledge of the grammatical structure of a language, as well as the ability to adapt their language use according to specific contexts and situations. For example, Reid (1956) introduced the concept of situation-related language variation, which meant that a speaker/writer makes different linguistic choices depending on the situation, referring to the variations as *registers*. The exploration of how language varies across different social contexts and among diverse groups of speakers was gaining traction in many different research disciplines.

These early observations paved the way for subsequent work in the 1970s-80s when systematic theoretical frameworks emerged. The researchers were interested in how the social and situational context of communication shaped language use and began to analyze linguistic features and contextual and communicative purposes. For instance, Irvine (1979) compared formal and informal interaction; Hymes' SPEAKING model (1974) showed how we modify our language use according to different contexts and participants; Brown & Fraser (1979) studied how speech varies in different situations; Halliday (1985) proposed a three-key model (Field-Tenor-Mode) which analyzes the choices one makes based on the language context; and Chafe (1982) proposed underlying parameters of speech and writing. The study of the relationship between language variation and context was attracting more scientific interest, but research was sporadic. Studies only examined linguistic variation through a single parameter/characteristic (e.g., written vs. spoken, a few participants) or a handful of linguistic features. Research was strongly affected by technological limitations. Any kind of data gathering, annotation, and analysis had to be done manually and on a much smaller scale; therefore, a large comprehensive analysis was unthinkable at the time.

The utilization of large corpora in corpus and computational linguistics was initially introduced in the late 1980s, and the connection between corpora, corpus studies, and computational linguistics was soon established. One of the first

corpus linguists was Douglas Biber (1986: 384-387; 1988: 52-53) who made use of the advances in technology to overcome the challenges of earlier work, such as researchers making generalizations based on a few texts, focusing on English, studying a small handful of linguistic characteristics, and hand-picking specific spoken or written varieties to fit their needs. Biber introduced the Multidimensional analysis (MDA), which was one of the first approaches using quantitative methods to compare different registers concerning different co-occurring linguistic features, which are analyzed as underlying dimensions of linguistic variation (1986; 1988). The MDA is based on the assumption that linguistic features co-occur in texts due to shared communicative functions, e.g. using pronouns and direct questions are related to a more interactive context. It uses factor analysis to identify the groups of co-occurring linguistic features on a larger scale, and then the linguistic patterns are interpreted by researchers in functional terms. Over the years, multidimensional analysis has become widely known and used[1], with more studies in other languages, such as Nukulaelae Tuvaluan (Besnier 1988), Somali (Biber & Hared 1992), Korean (Kim & Biber 1994), Mandarin (Song et al. 2021), Gaelic (Lamb 2002), Spanish (Biber et al. 2006; Parodi 2007), Dagbani (Purvis 2008), American English (Grieve et al. 2011; Passonneau et al. 2014; Grieve 2014), Brazilian Portuguese (Sardinha et al. 2014), Russian (Katinskaya & Sharoff 2015), Urdu (Shakir & Deuber 2019), and Czech (Cvrček et al. 2020). The MDA has also been successfully applied to study specialized domains, e.g., outsourced call centres (Friginal 2008), Reddit data (Liimatta 2019), social media (Sardinha 2022), Trump's tweets (Clarke & Grieve 2019).

By the turn of the millennium, researchers began using the Web as a corpus, although the Web was initially perceived as anarchic and unfamiliar to research domains focused on studying language (Kilgarriff & Grefenstette 2003: 335). After a while, more researchers started turning their attention to the Web due to its unrestricted nature, vast data, and instant accessibility. Unlike traditional corpora with predetermined genre and text distributions, Web corpora featured a diverse and unpredictable variety of texts. Genre labels originating from traditional corpora were used to categorize texts from the Web systematically. However, this approach soon revealed several challenges, e.g., text boundaries being unclear; texts not fitting into the existing taxonomies of traditional text corpora; texts belonging into multiple categories and being hybrid (see Mehler et al. (2010) for a comprehensive overview). Furthermore, instead of trying to structure the entire Web, researchers focused on a limited number of predefined genres, such as Beaudouin et al. (2002) who looked at personal and commercial Web pages, Roussinov et al. (2001) who selected five major groups of genres that could be used in an interactive search tool, and Rehm (2002) who concentrated on academic Web pages.

Given the challenges posed by the unstructured and diverse nature of Web texts, researchers sought a more systematic approach to categorize them. One

[1] In case of interest, see `https://larissagoularts.wordpress.com/bibliography/`

strategy for classifying the data involves labelling the new electronic genres based on their virtual location (which is still an applicable practice today, for example, TenTen Corpus Family (Jakubíček et al. 2020) on the Web), but this approach has not provided desirable results due to the wide-ranging situational and linguistic variations of genres. Marina Santini and colleagues (2007; 2010) drew attention to the need for a comprehensive empirical analysis and a systematic approach to identify Web genres. This marked the emergence of a new tradition known as Automatic Genre Identification (AGI)[2], using machine learning methods to automatically assign labels or categories to texts based on the underlying patterns and features present in the data. Studies in AGI required a predetermined classification scheme and could utilize complex features such as part of speech tags, parse trees or rhetorical relations, or surface-level features such as frequently occurring words or *n*-grams (Santini et al. 2010; Kanaris & Stamatatos 2009; Madjarov et al. 2019).

However, the classification approach of AGI also had limitations. For instance, taxonomies are too limited in terms of labels or too exhaustive and do not take into account the high degree of hybridism. To address these issues, Serge Sharoff (2018; 2021) and Biber et al. (2021) proposed novel approaches to model the hybrid nature of registers/genres in a topological register space. Sharoff introduced the Functional Text Dimensions FTD[3], where dimension (= genre) represents a functional category that describes the communicative purpose of a text, for example, *To what extent is the text concerned with expressing feelings or emotions?* or *To what extent does the text provide information to define a topic?* In this way, texts are described through a combination of different dimensions, which allows one to measure the distance of a text from its prototypical genres. Biber et al. (2021) proposed a continuous situational space in which texts are characterized by a set of situational parameters[4] and analyze register variation using a combination of linguistic and situational characteristics in the space of MDA. Although both methods aim to capture the hybridism of texts, their motivations differ. Sharoff's FTD aligns with the AGI tradition, seeking to automatically categorize large Web corpora into genre categories. The goals of Biber et al. (2021) are similar to previous MDA studies, namely to analyze the underlying situational dimensions of the registers.

Shifting from traditional frequency-based methods, research has embraced

[2]This is used synonymously with *automatic text classification*.

[3]The 18-dimensional taxonomy: argumentation, emotive, fictive, flippant, informal, instructive, news, legal, personal, commercial presentation, ideological presentation, science and technology, specialized, information or encyclopedic, evaluation, poetic, dialogue, and appellative.

[4]E.g., (1) "Text is ..." *a) a spoken transcript, b) lyrical or artistic, c) pre-planned and edited, d) interactive*; (2) "The author/speaker ..." *a) is an expert, b) focuses on himself/herself, c) assumes technical background knowledge, d) assumes cultural/social knowledge, e) assumes personal knowledge about himself/herself*, etc. These situational characteristics were used to define the register categories used for the Corpus of Online Registers of English (CORE) corpus (`https://www.english-corpora.org/core/`).

neural networks as a more powerful tool for automatic text classification, better able to incorporate word similarities and context than models using count-based feature vectors (Mikolov et al. 2013). Recent research has explored the potential of neural networks for cross-lingual settings. For example, using machine translation for English and Russian corpora to classify Arabic genres (Bulygin & Sharoff 2018); using transfer learning and the register scheme of the Corpus of Online Registers of English (CORE) (Egbert et al. 2015; Biber et al. 2020) to predict Finnish, Swedish, and French registers (Laippala et al. 2019; Repo et al. 2021; Rönnqvist et al. 2021; Skantsi & Laippala 2023), and to predict registers for Arabic, Catalan, Chinese, Hindi, Indonesian, Portuguese, Spanish, and Urdu languages (Laippala et al. 2022).

There is no comprehensive research comparing multiple registers based on a wide range of linguistic and textual characteristics in the Estonian language. Research has been mainly qualitative, analyzing a single register or comparing two or more based on a handful of linguistic features, e.g., analyzing the complexity (Kerge 2009) and formality (Kerge et al. 2007) of the text, or how certain linguistic features are distributed between different registers (Hennoste et al. 2022). Nonetheless, a wide range of language domains have been examined for their linguistic and textual characteristics, such as spoken every day language (Hennoste et al. 2015, 2021), news articles (Kasik 2008; Hennoste et al. 2015, 2021, 2022), religious texts (Hennoste et al. 2015), essays and short stories (Meier 2003), online chat rooms and dialogues (Salla 2002; Hennoste et al. 2021, 2022), advertisements (Kasik 2000), interviews (Kasik 2004), online comments (Kerge 2004; Hennoste et al. 2015, 2021, 2022), spoken institutional dialogues (Hennoste et al. 2021, 2022), official and personal letters (Kerge 2009), telephone conversations (Rääbis 2009), author's style across science, everyday conversations and news (Kerge 2003a), medication package leaflets (Sirel 2013), editorials (Sarapuu 2008), regulations (Mandra 2009), literary works (Lepajõe 2011), academic and scientific research (Hennoste et al. 2021, 2022), personal ads (Kongot 2013), and decisions, directives, orders, statements, information requests, complaints, contracts for services and agency agreements (Reinsalu 2019).

2.2. Terminology

In automatic genre classification, confusion around terminology has often been described as a jungle (see Lee (2001) for a broader view). The situation is complicated by the fact that in the literature, the terms *genre* and *register* are used interchangeably and/or synonymously. Over time, researchers have shown a preference for either genre or register (see Swales 1990; Bhatia 2002), while some use both simultaneously, e.g., Systemic Functional Linguistics, SFL, (Halliday 1985; Matthiessen 1993). In SFL, the term *genre* was viewed as a system of cultural and social processes, and *register* as a situational context of a particular configuration of Field (the subject of the text), Tenor (the relationship between the author and

the audience), and Mode (how the text is constructed, particularly whether it is written-like or spoken-like) choices. The terms *genre* and *register* differ slightly in different theoretical approaches, therefore researchers have attempted to clarify their usage more transparently (see Biber 1988: 70; Biber 1995: 7-10; Lee 2001; Biber & Conrad 2009: 21-23).

Biber and Conrad (2009: 2) view *genre* as a text variety that has a socially recognized purpose and shares common structures with texts within the same variety (e.g., the structure of an email) and *register* as a text variety where the common core linguistic features within a variety are associated with the communicative purpose and situational context. Lee (2001: 46) sees *genre* and *register* as concepts covering the same ground: "register is used when we view a text as language: as the instantiation of a conventionalized, functional configuration of language tied to certain broad societal situations, that is, variety according to use", and "genre is used when we view the text as a member of a category: a culturally recognized artefact, a grouping of texts according to some conventionally recognized criteria". Therefore, Lee (2001: 47) prefers to use genre to describe groups of texts because they are conventionally more recognizable as text categories and are more associated with power/social purposes reflecting the dynamic nature of language use. This demonstrates that due to their inconsistent definitions, *genre* and *register* are often used interchangeably, leading to a certain degree of vagueness and overlap.

The vague use of terms is also characteristic in studying language variation in Estonian, especially with the terms *type* and *genre*. Although straightforward definitions have been proposed by Reet Kasik (2007)[5], studies have used the terms knowingly very loosely (Kerge 2000; Kerge et al. 2007, 2008), used them as a general cover term (e.g. Meier 2003; Vaik & Muischnek 2018; Vaik et al. 2020), or focused on studying language use within the context of text type, such as narrative (Aava 2004; Lindström 2005), instructive (Reinsalu 2019), or argumentative (Lepajõe 2011; Kuldnokk 2011).

In much of the existing research within the field of computational linguistics, the distinctions between the terms genre, register, and type are not considered relevant, and the choice between terminology depends on a researcher's preference or tradition. Most studies in automatic text classification have preferred to use the term *genre* to refer to a category of texts which share a communicative function manifested through linguistic realizations (Santini 2007; Wu et al. 2010; Crowston et al. 2011; Sharoff 2018; Madjarov et al. 2019). The term *register* is preferred by

[5]Kasik (2007: 29-30) defines *type* as a variety associated with function and form (*descriptive type*, *narrative* and *argumentative type)* which is broadly in line with the tripartite division of de Beaugrande & Dressler (1981) - *descriptive*, *narrative*, and *argumentative* -, as well as the quintuple division of Werlich (1983) - *descriptive*, *narrative*, *argumentative*, *expository*, and *instructive type*. In addition, Kasik (2007: 35) associates genre as a culturally bound variety with a schematic structure consisting of more general or more strictly defined parts (e.g. a recipe has a different structure than an article).

those who are more interested in the linguistic focus, where the situational contexts and functions are analyzed in terms of their pervasive lexico-grammatical linguistic characteristics (e.g., Biber 1995; Biber & Conrad 2009; Egbert et al., 2015; Biber & Egbert 2018; Laippala et al. 2019).

Given the overlap and ambiguity between the terms and the lack of a clear consensus on their definitions, the choice between genre and register often depends on the researcher's preference. This dissertation acknowledges the ongoing debate surrounding these terms and seeks to contribute to the field by analyzing the pervasive linguistic features found in different texts. For the sake of consistency, this dissertation uses the term *register* even when discussing works that follow the *genre* perspective.

2.3. Register taxonomies for Web corpora

Taxonomies for Web corpora have been created in many ways, varying in methodology, level of granularity, and functionality. The subsequent sections explore these approaches in more depth.

There are two basic methods for creating taxonomies for Web documents: top-down vs. bottom-up method (Crowston et al. 2011: 73-75). With the top-down method, the researcher designs an initial taxonomy and conducts a pilot study to assess its validity (Berninger et al. 2008; Vidulin et al. 2009). The initial taxonomy is usually a combination of existing taxonomies, registers from traditional text corpora (e.g., *reportage*, *novel*, *popular lore*), and novel Web registers (e.g., *e-shop*, *blog*, *homepage*, *FAQ*) (Rehm et al. 2008; Santini et al. 2010; Sharoff et al. 2010; Jakubíček et al. 2020; Kuzman et al. 2022). With the bottom-up method, the annotators are asked to label texts using tutorials or decision tree surveys provided by the researcher (Dewe et al. 1998; Meyer zu Eissen & Stein 2004; Crowston et al. 2011; Egbert et al. 2015). For instance, annotators can be given questions such as *What type of Web page would you call this?* (Crowston et al. 2011) or *Do you consider this genre useful?* (Meyer zu Eissen & Stein 2004). Both methods offer different perspectives and can be used in combination to achieve more reliable results. For example, Asheghi et al. (2016) used the top-down approach to create a preliminary taxonomy, conducted a pilot study using bottom-up methods to gather feedback from annotators, and used the feedback to refine the initial taxonomy.

Taxonomies can differ based on their nestedness. One distinction is made between flat taxonomies and hierarchical taxonomies. In flat taxonomies, each register is treated as an independent and equal category, and cross-category relationships are not prioritized (Stubbe et al. 2007; Vidulin et al. 2009; Crowston et al. 2011; Asheghi et al. 2016; Jakubíček et al. 2020; Sharoff 2018). With hierarchical taxonomies, registers are nested within broader main categories. For example, one of the main categories of the Berninger et al. (2008) taxonomy is *Serial*, which includes registers such as *periodical*, *essay*, *fictional piece*, etc. Similarly, the main

category of Kuzman et al. (2022) *Objective Informative* includes registers such as *news*, *announcement*, and *information/explanation*. Although hierarchical taxonomies offer the ability to make finer distinctions between categories, they also contribute to the overall complexity of the taxonomy and may not offer much improvement over a flat list. Both can present challenges when representing large or diverse corpora. For example, too many categories may contribute to the overall complexity, while too few may introduce ambiguity.

Another distinction can be made between single and multiple labelling. Single labeling refers to the practice of assigning only one register label to a text, based on its dominant characteristics (Asheghi et al. 2016; Jakubíček et al. 2020). This approach assumes that Web content can be neatly classified into distinct categories while overlooking the dynamic and hybrid nature of registers on the Web. Today, most Web taxonomies have adopted the notion of hybridism, allowing texts to belong to multiple registers (Egbert et al. 2015; Biber & Egbert 2018; Sharoff 2018; Madjarov et al. 2019; Kuzman et al. 2022). Furthermore, although most taxonomies assume that each text must belong to one of the predefined categories, there are taxonomies which recognize the obscurity of a Web page and purposely use an unknown category, such as *misc*, *noise*, *unknown*, *nothing*, as a way to accommodate content that does not fit into the existing predefined categories (Stubbe et al. 2007; Santini 2010; Asheghi et al. 2016; Pritsos & Stamatatos 2018; Jakubíček et al. 2020; Kuzman et al. 2022).

In addition, categories within different register taxonomies have varying levels of specificity. Taxonomies consist of only broad categories (e.g., *Blogs*, *News*, *Literature*, *Communication*), subcategories (e.g., *promotion of a product*, *CV*), functional/situational styles and dimensions (e.g., *information*, *explanation*, *instruction*), arbitrary text classes (e.g., *link collections*, *download pages*, *error message*); or as a combination of broad categories and subcategories. As a consequence, existing taxonomies vary widely, with the number of categories ranging from seven (Santini 2010), twenty (Vidulin et al. 2009) to 292 (Crowston et al. 2011). In the past, there have been attempts to create consistent Web register taxonomies, such as an initiative called Web-as-corpus by Rehm et al. (2008). Their idea was to establish the ground rules for assigning register labels to Web content, but their initiative did not thrive.

Some taxonomies have been developed with the sole intention of examining variations between specific subsets of registers, but there have also been efforts to create comprehensive taxonomies that can cover unrestricted samples of the Web, such as CORE (Egbert et al. 2015; Biber & Egbert 2018), Functional Text Dimensions (Sharoff 2018), and Ginco (Kuzman et al. 2022). These efforts recognize the need for a taxonomy that goes beyond individual handpicked registers and encompasses the entire spectrum of the Web's content. Although these initiatives aim to establish a classification scheme that can effectively categorize and organize a diverse range of content found across the Web, without limiting it to specific registers or topics, a significant challenge lies in achieving high inter-

annotator agreements (Suchomel 2020; Kuzman et al. 2022).

Obtaining suitable labels for Web corpora proves to be challenging because of the same eternal problems: the content of the Web corpora demonstrates a degree of variability, and there is no full consensus among researchers regarding the foundational principles upon which taxonomies should be constructed. These problems show that automatic text classification is not as universal, and there does not seem to be a one-size-fits-all solution, suggesting the need to explore alternative approaches.

3. DIMENSIONAL TEXT MODEL

This section introduces the theoretical foundations of the Dimensional Text Model (DTM) which are based on the Multidimensional analysis (MDA) of Biber (1988) and the Functional Text Dimensions (FTD) of Sharoff (2018; 2021). It also describes how the deconstruction of dimensions from different MDA studies resulted in a 12-dimensional space comprising *abstractness*, *affectivity*, *argumentativity*, *impersonality*, *interactivity*, *instructability*, *formality*, *complexity*, *spontaneity*, *information density*, *temporality* and *subjectivity*. All dimensions are individually introduced in Section 3.2.

3.1. Theoretical framework

The Dimensional Text Model (DTM) is a hierarchical framework that identifies the characteristics that differentiate various texts based on their social context and communicative functions. These characteristics are defined through dimensions that manifest through co-occurring linguistic features. This model is a synthesis of the Multidimensional analysis (MDA) of Biber (1988) and the Functional Text Dimensions (FTD) of Sharoff (2018; 2021) which have made significant contributions by providing pioneering methodologies for studying register variation.

The MDA (Biber 1986; 1988) was the first approach to use quantitative methods to compare different registers concerning co-occurring linguistic features that are analyzed as underlying dimensions of linguistic variation. The MDA is based on the assumption that linguistic features co-occur in texts due to shared communicative functions, e.g. the assumption that the use of pronouns and direct questions is related to a more interactive context. The MDA computes the frequencies of the linguistic features of each text in a corpus and uses factor analysis to identify the factors which are interpreted as functional dimensions by assessing the communicative functions realized by the linguistic patterns. The original study of MDA (Biber 1988) established seven dimensions for English: D1 *informational vs. involved text*; D2 *narrative vs. non-narrative text*; D3 *elaborated reference, independent from the context versus non-specific situation-dependent reference*; D4 *being persuasive (public, expressed on the textual level)*; D5 *the abstractness of the used language*; D6 *communicating information in real time*; D7 *academic argumentation together with hedging*. The first five dimensions are more significant. The MDA is a corpus- and feature-dependent methodology, but different languages and corpora (see Section 2.1) using MDA have revealed some universal traits, e.g., distinguishing speech-like and written, narrative and non-narrative texts.

The FTD (Sharoff 2018; 2021) follows the AGI tradition and seeks to automatically categorize large Web corpora into registers[6]. The taxonomy was established through an annotation study, using a list of questions that represent each functional dimension which defines the communicative purpose of a text and gives examples of its prototypical registers; for example, *To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view? (e.g., argumentative blogs, opinion pieces)* or *To what extent does the text aim at teaching the reader how something works? (e.g., tutorial, FAQ)* As a result of several classification attempts (Sharoff 2010; Sorokin et al. 2014; Katinskaya & Sharoff 2015), the final taxonomy offers a set of 18 dimensions. These dimensions are divided into primary (mandatory) where the presence of a dimension in a text is mandatory, and secondary optional dimensions, which are intended to offer some internal variation. The primary dimensions are *argumentation* (e.g., editorials, opinion pieces), *fictive* (e.g., novels, stories), *instructive* (e.g., tutorials, manuals), *news* (e.g., newswires), *legal* (e..g., laws, contracts), *personal* (e.g., diaries, personal letters), *commercial presentation* (e.g., advertisements), *ideological presentation* (e.g., manifests, propaganda), *science and technology* (e.g., articles, essays), *information or encyclopedic*, *evaluation* (e.g., product review), *poetic* (pays attention to text aesthetics), *appellative* (requests, small ads). The secondary dimensions are *emotive* (texts expressing feelings, emotions), *flippant* (light-hearted and entertaining texts), *informal* (communication deviating from the standard), *specialist* (requires background knowledge), *dialogue* (interaction between several participants). In addition, rather than proposing a taxonomy in which the text either is or is not a member of a register, FTD introduces the notions of *multidimensional space* and *distance*. In FTD, texts are described through a combination of dimensions and the presence of a dimension in a text is measured on a scale[7] which allows modelling the texts in a multidimensional space of registers.

MDA is one of the first theoretical approaches that systematically studied the register variability through linguistic parameters. When MDA explores language variability in corpora where registers are predefined, the FTD was motivated by the need to classify corpora for which the register of a text is unknown. Both represent important methodological approaches in the modern corpus-based research field, but their inherent limitations motivate the development of the Dimensioned Text Model.

The foundation of the MDA is based on the duality of form and function. This implies that the communicative functions are expressed by linguistic features, which serve communicative functions. The DTM adopts the MDA's concept of a *dimension* as a quantifiable linguistic quality, measured by co-occurring lexical, grammatical, and textual features. However, unlike the MDA, which views

[6]Sharoff (2018: 4) prefers to use the term *genre* when lexico-grammatical features are not included in the analysis.

[7]None or hardly at all - 0; slightly - 0.5; somewhat or partly - 1; strongly or very much so - 2.

dimensions as scales with two opposing functional poles, the DTM treats a dimension as a continuum of salience representing a single communicative function. The MDA also has a central role in generating the set of dimensions proposed by the DTM. Since the DTM aims to identify the universal characteristics that differentiate various texts based on their communicative functions, the set of dimensions was generated based on different MDA studies (e.g., Kim & Biber 1994, Biber et al. 2006, Passonneau et al. 2014, Sardinha et al. 2014, Katinskaya & Sharoff 2015), with a focus on the original study (Biber 1988). The MDA dimensions, which are inherently functionally complex by incorporating many communicative functions at once, were broken down into smaller elementary dimensions, each potentially realizing a single communicative function. For example, all MDA studies have identified an oral and written discourse, which can be associated with two communicative parameters. The first parameter reflects whether the text aims to be informative or interactive, affective and participatory. The second parameter characterizes circumstances where the editing possibilities allow using more sophisticated means compared to circumstances where time is limited. These parameters were broken down into smaller dimensions, such as information density, interactivity, complexity, affectivity, formality, subjectivity, and spontaneity. Another cross-lingual dimension *narrative vs. non-narrative* was further subdivided into smaller dimensions that prioritize time-, information-, or participation-focused aspects, such as temporality, and impersonality. Many MDA studies (Besnier 1988, Biber & Hared 1992, Passonneau et al. 2014, Sardinha et al. 2014) identified argumentativity, instructability, and abstractness as relevant dimensions, leading them being included in the final list of dimensions.

Compared to MDA, the FTD is a method which is oriented towards the automatic categorization of large web corpora into registers, which is the future objective of this dissertation. However, in contrast to the linguistically grounded approach, FTD employs a surface-level, label-driven classification taxonomy that is an eclectic combination of different dimensions. The taxonomy includes supercategories (e.g. fiction, news, research), communicative methods (e.g. dialogue), and different ways to organize the language (e.g. poetic) or literary norms of the language (e.g. informal). In addition, when MDA and DTM define a dimension as a quantifiable linguistic quality which manifests through co-occurring linguistic features, then in FTD, a dimension is viewed as representative of a set of registers and thus, a single text can be an example of a hybrid register because it can be characterized by multiple dimensions. The DTM borrows the FTD's concept of *multidimensional space* by expanding it to represent all texts within a comprehensive dimensional space. Unlike the FTD, where the multidimensional space describes the texts through the primary dimensions (mandatory dimensions), while the secondary dimensions (optional) may or may not be present, the DTM expands the multidimensional space to view the texts as having all dimensions simultaneously. This ensures that all texts are modelled in the same space, while dimensions

differ in terms of the levels of salience, i.e., some dimensions are salient in a text to a limited extent or gravitate towards being more strongly present. By modelling the texts in a unified space, spatial proximity can be measured suggesting that texts which are closer together in the multidimensional space share similar register characteristics.

Based on the decomposition of the MDA dimensions into smaller dimensions motivated by a singular function, twelve dimensions for the DTM were defined: *abstractness*, *affectivity*, *argumentativity*, *impersonality*, *interactivity*, *instructability*, *formality*, *complexity*, *spontaneity*, *information density*, *temporality* and *subjectivity*. These dimensions constitute the core of the DTM. However, the framework includes two additional levels, making it a hierarchical model (see Figure 1):

I **Features** are represented at the lowest level. These are directly quantifiable, e.g. the number of nouns, vocabulary size, number of relative clauses, abstract words, etc.

II Latent **dimensions** are at the intermediate level. They are continuous and described by sets of co-occurring lexico-grammatical features that serve a communicative purpose. Dimensions are cross-lingual, although the sets of features may be language-specific.

III **Registers** are at the top level. Registers are defined by the patterns of co-occurring dimensions. Texts that are spatially close in the multidimensional space share similar functions and could belong to the same register.

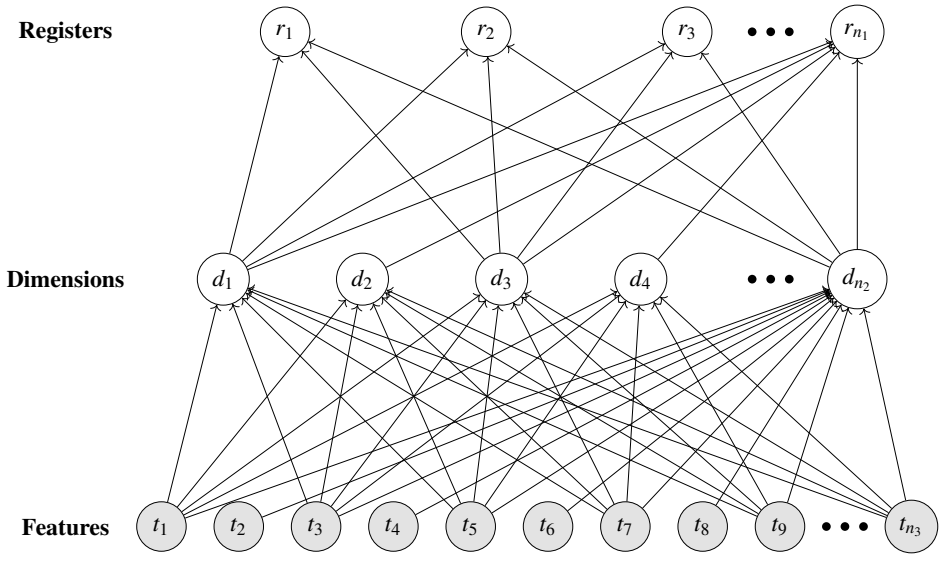

Figure 1: Hierarchical structure of the Dimensional Text Model.

Features and dimensions are the core concepts of the DTM. A dimension is a quantifiable quality expressed by a set of co-occurring linguistic features. For this reason, the dimensions of the DTM satisfy the following conditions:

I a dimension manifests through a language-specific set of linguistic features which comprise its linguistic profile;

II a dimension is continuous, it is not dichotomous;

III a dimension has a distinctive linguistic profile;

IV a dimension alone does not describe a register since registers consist of the patterns of dimensions that occur together. However, a single dimension can describe a particular aspect of the characteristics of a register, e.g. academic language use can be described as linguistically complex, informational, or abstract.

3.2. Dimensions

The following subsections introduce the twelve dimensions defined in the Dimensional Text Model. Each of them is illustrated with a text sample listed in Table 1. In addition, dimensions are qualitatively mapped to text samples to show which dimensions are more or less characteristic within individual samples. This is presented in a separate column, "More (↑) & less (↓) salient dimensions". The assessment of whether the theoretically proposed dimensions are distinguishable to the annotators was based on Estonian text samples. The text samples provided in this section are in English for illustrative purposes.

3.2.1. Abstractness

Abstractness (abs) in a text is concerned with mental, non-concrete phenomena that cannot be perceived by the senses. Abstractness can be found in a variety of areas, such as scientific research papers, philosophical essays, academic articles, and even certain forms of literature.

Abstractness is illustrated by Text 1 and Text 9 (see 1). Both examples contain words and phrases that refer to concepts that are not immediately observable, e.g. organization names, *abundance*, *relative frequency*, *population*, *biodiversity*. Abstract words and phrases make it difficult to understand the meaning of a text because they are not easily interpreted (fewer concrete words). The opposite of abstractness is illustrated by Text 2 and Text 4. These examples contain many concrete words that are easily perceived and interpreted, e.g. *two-room flat*, *older couple*, *cold living room*, *onion*, *boiling water*, *stir*, *peel*.

3.2.2. Affectivity

Affectivity (aff) refers to the personal experiences and feelings of emotions and reactions expressed in a text. Affective text evokes emotional responses and reac-

tions from the writer and can resonate strongly with the reader and evoke empathy on their part.

Affectivity is illustrated by Text 6. In this example, the writer has combined lexical items (e.g. *no artistry*, *f****, *manufactured skanks*), rhetorical questions, and punctuation marks to express their emotional attitude towards a particular topic. As the opposite of high affectivity, Text 4 and Text 9 are more neutral and do not convey emotional or attitudinal preferences.

3.2.3. Instructability

Instructability (inst) in a text provides the reader with instructions on how to perform certain activities or describes the stages of a process. Instructable texts are often found in manuals, user guides, tutorials, recipes, DIY articles, and technical documentation.

Instructability is illustrated by Text 4 and Text 10, where both texts are about giving some advice/instructions to the reader. This is achieved by (mainly) using the second-person imperative verb form. The opposite end of instructability can be illustrated via Text 1 and Text 2 where the narrative discourse makes the degree of instructability fully absent.

3.2.4. Information density

Information density (info) in a text aims at conveying more information with fewer words and clauses, i.e., information is packaged in an economical way. In addition to encyclopedic entries, information density can be high in public law texts, scientific and journalistic texts, and information providers, such as pharmaceutical package leaflets, documents, etc.

Information density is illustrated by Text 4 and Text 9. The information density of both examples is mainly manifested by the richness of the noun phrases, e.g., *7-8 medium-sized yellow summer squash*, *squash into thin slices*, *10 minutes*, *measure of the abundance and relative frequency of species*, *purposes of biodiversity conservation*.

3.2.5. Spontaneity

Spontaneity (spont) in a text represents unplanned and unedited language. It can be illustrated in spoken discourse, where what is written cannot be taken back, but the possibility of correction remains. Spontaneous text can be (written) spoken (transcripts of conversations, monologues) and written (comments, forums, online conversations).

Spontaneity is illustrated by Text 5 and Text 6. Both examples contain features such as abbreviations (e.g., *kinda*), and acronyms (e.g., *lol*), that are characteristic of net speak. Both also contain errors due to the time pressure or the communication environment, e.g., lacking proper punctuation and capitalization, and using periods/exclamation marks for emphasis. In Text 5, the first comment contains an

insertion of **pouts**, which is a new way to express emotion in online conversations. Text 3 and Text 9 are the opposite of spontaneity, as examples of polished texts.

3.2.6. Impersonality

Impersonality (imp) is characteristic of texts in which the focus is on the action or the recipient, and the information about the doer (agent) of the action is suppressed. Impersonal texts typically prioritize clarity, accuracy and objectivity, making them suitable for academic papers, news articles, scientific reports, or technical manuals.

Impersonality is illustrated by Text 3 and Text 9, where an educated guess can be made, but the exact agent of the action is still not explicitly expressed, e.g. *sharks are killed*, *populations have been decimated*, or *species diversity is often used*. On the opposite scale of the impersonality dimension are Text 2 and Text 6, where personal expressions or statements are acceptable and agents are not suppressed; on the contrary, agents play a big role in storytelling.

3.2.7. Formality

Formality (form) can be described by texts that aim to maintain a professional and objective tone. Highly formal texts avoid colloquialisms, slang and overly casual language.

Formality is illustrated by Text 1 and Text 12. When comparing Text 1 and Text 12 to Text 5, the former examples are coherently more formal due to the use of nominalization and polished sentences. Text 5 contains user reviews that appear less formal. The degree of formality does not depend on the anonymity of the auditorium, but rather on what kind of communication channel is being used.

3.2.8. Complexity

Complexity (comp) in a text refers to the degree of difficulty or intricacy of its structure and language. Texts of high complexity require additional effort from the reader.

Complexity is illustrated by Text 1 and Text 11. The former is very complex because of its syntactic structure: a long sentence containing multilevel complex sentences with many phrases and non-finite embedded clauses, etc. The latter is less syntactically complex, but the use of specialized and foreign vocabulary makes it more difficult for a non-expert to understand.

3.2.9. Temporality

Temporality (temp) refers to the sequential order of events or ideas within a text following a chronological structure. Usually, temporality is important in a narrative discourse (e.g., news, fiction, etc.). At the other end of the dimension, there

are texts in which the events have no temporal dimension, the activity is related to the present moment of the speech, or there is no reference to time.

Temporality is illustrated by Text 2 and Text 4. The former example is about someone describing their past while showing the temporal linearity of different events. The latter involves temporality because for one event to take place, another event must be completed; for example, the temporal linearity of different events in a recipe is important. On the opposite end of the temporality dimension are Text 1 and Text 3 in which the activity does not include an apparent reference to time.

3.2.10. Interactivity

Interactivity (inter) characterizes communicative and reactive texts in which participants actively engage with each other. Interactivity can take different forms: from traditional print media (fiction, interviews, etc.) to online mediums (e.g., online forums, newsroom commentaries, etc.) where interactivity is present through dialogue.

Interactivity is illustrated by Text 5 and Text 7. In Text 7, there is more direct communication between the subjects; in Text 5, all the participants are talking about the same topic without talking directly to each other, but there is still some hidden communication between them. Comparing the two, Text 5 is less interactive than Text 7, but both can still be considered as interactive as there are several participants.

3.2.11. Subjectivity

Subjectivity (subj) refers to the presence of personal opinions, feelings, and biases expressed by the author. It can be contrasted with objectivity, which aims to present information without personal interpretations. Examples of subjective texts include opinion pieces, personal blogs, or editorials.

Subjectivity is illustrated by Text 8, where the author expresses their personal view using opinion verbs (e.g. *take the view*, *believe*), subjective adverbs (e.g., *personally*, *constantly*) and personal pronouns (e.g., *I*, *we*). The use of personal pronouns misleadingly suggests that the opinion of one person is the opinion of the majority. At the other end of the subjectivity is Text 9, which does not express any judgement or personal preferences but appears to describe things as they are, that is, objectively.

3.2.12. Argumentativity

Argumentativity (arg) is present in texts in which the author presents his or her point of view on a topic or phenomenon, but as opposed to subjectivity, the points of view are supported by objective arguments. For example, scientific literature can be considered argumentative, as opposed to descriptive or narrative discourse.

Argumentativity is illustrated by Text 12, where the author presents a view on a certain topic while remaining neutral. This is achieved by using syntactic tools:

stating a fact and supporting it with clausal subordination. At the opposite end of argumentativity is Text 2 which expresses a narrative discourse without the need to take a position on an issue.

Table 1: Text Samples with dimensional salience.

	Text[8]	More (↑) & least (↓) salient dimensions
Text 1	Before any meeting of the Commission or a subsidiary body of the Commission, a Regional Economic Integration Organisation that is a member of the Commission or its member States that are members of the Commission shall indicate which, as between the Regional Economic Integration Organisation and its member States, has competence in respect to any specific question to be considered in the meeting and which, as between the Regional Economic Integration Organisation and its member States, shall exercise the right to vote in respect of each particular agenda item.	↑ D1 abs, D7 form, D8 comp; ↓ D3 inst, D9 temp
Text 2	We lived in an upstairs two-room apartment in an older couple's house. It was on the edge of town. He had several cows and a garden, and they gave us milk, vegetables and fruit. We lived there three years and probably paid little or no rent most of that time. I had started piano lessons previously and was able to practice in their cold living room, which was kept closed off so they didn't have to heat it. The fourth year we lived in this small town, we moved closer to the school but had nearly the same arrangement: two upstairs rooms and a small kitchen next to the owner's kitchen.	↑ D9 temp, D4 info; ↓ D1 abs, D3 inst, D6 imp

[8]Source: English Web 2021 (enTenTen21) via SketchEngine.

Text 3	As essential as sharks are to the oceans, they are being dramatically overfished, primarily to fill market demand for their valuable fins. Shark fins are used for shark fin soup, a luxury food item sometimes served at weddings and banquets. Tens of millions of sharks are killed each year, many of which are targeted for their fins. Many shark populations have been decimated by as much as 90%. The International Union for Conservation of Nature (IUCN) reports that a full 1/3 of shark species face extinction.	↑ D6 imp, D4 info; ↓ D5 spont, D9 temp
Text 4	You will need 7-8 medium sized yellow summer squash. Wash the squash & cut off ends. Slice the squash into thin slices. Take one large onion (preferably Vidalia, but any onion will do) peel & slice, then cut each slice into 4ths. Place squash & onion slices into boiling water & boil about 10 minutes. While squash & onions are boiling mix other ingredients in large bowl. Mix 1 can Cream of Chicken Mushroom Soup 1 cup Sour Cream 2 cups Shredded Sharp Cheddar Cheese 1 cup crushed Ritz crackers. Stir all ingredients together in large mixing bowl. Once squash & onions have finished boiling remove from stove & drain. Add squash & onions to other ingredients in bowl & stir together gently.	↑ D3 inst, D4 info, D9 temp; ↓ D1 abs, D12 arg

Text 5	-This was a super cute pilot and I was very impressed with Rupert's accent. I remember him saying in an interview that he loves to do accents so that might be a distinguishing skill for him. The only thing I would have wanted to see in this pilot is how Clyde would put his own personal spin on the "lost wallet" gimmick. Maybe that would have come in later episodes. Of which there aren't any. *pouts* Oh well. I will look forward to the next thing. -we are men and mom have been cancelled lol . -My fav thing about the show were Stephen and Rupert. the siblings were kinda eh. but I love how the show has heart unlike so much crap put on tv these days -I watched the pilot online and was reading the comments. It has very positive feedback. I hope CBS considers picking up this show for a mid-season replacement or give it to another network. This show needs to be on t.v!!! -I hadn't heard Mom was cancelled? We Are Men definitely is, and that probably will give Mom some breathing room unless it starts to break down again. (Was steady at a 2.0 two weeks in a row.) Honestly, the only comedy pilot I liked from CBS was The Crazy Ones. The Millers was just painful to watch	↑ D5 spont, D10 inter; ↓ D7 form, D8 comp
Text 6	If I was an artist/singer, I'd want to break album records, tour records, performance records etc....and not just "singles/digital records"... It's funny how most of you crazy ass stans are making all kinds of excuses as to why the CD is flopping claiming she's not promoting it right now. f*** THAT! She shouldn't had released a new CD while she was still promoting the last one. WHO THE f*** DOES THAT ???? Rihanna has no artistry growth. Her music all sounds the same and manufactured. I'm ready to have Christina back to show these manufactured skanks how it's done!	↑ D2 aff, D5 spont; ↓ D6 imp, D1 abs

Text 7	He finishes off his meal and says, "As for the snakes, where do you plan to find them?" Aubrey looks thoughtful. "I have not heard that anyone is that close, but not everyone talks to me, so perhaps... I will ask my grandfather, and my Aunt. They would certainly know." He smiles. "Snakes are everywhere. I will look by the river, and in the marshland, and in our back yard here...the little striped snakes count, too." Lights-the-Path gives a small nod of his head and he says, "Aah, very good." He gives a small yawn and asks, "Perhaps tomorrow we shall go hunting for snakes then?"	↑ D10 inter; ↓ D6 imp, D1 abs, D3 inst
Text 8	I personally take the view that we could explore this route a little more, provided it does not lead to further bureaucracy, especially in our host country. We are constantly being faced with all kinds of states of affairs, blockades and such like, and I believe it would be of great benefit if more objective information on future changes could be provided so that we could perhaps prevent a few mistakes from being made, as might have happened in the case of the Titanic.	↑ D11 subj, D8 comp; ↓ D6 imp, D9 temp
Text 9	Measure of the abundance and relative frequency of species in a specified area. Species diversity is often used with respect to animal or plant populations in a single stand, but can also be thought of on regional and global scales. For the purposes of biodiversity conservation, spatial scales of species diversity are hierarchical: global diversity is a higher conservation priority than regional diversity, and both are more important than local or stand-level diversity.	↑ D1 abs, D4 info, D6 imp; ↓ D2 aff, D5 spont, D11 subj

Text 10	With your support, we hope to make Earth Hour a self- perpetuating drive that inspires us to correct the way we live, every single day. 1.Carry your dinner and snacks (bring some extra so you can share with others)(There is no food available around so u must bring your food) 2. One must carry warm clothing as it will get cold near the lake. Temperatures can go to 10°C 3. Carry chocolates: they are very helpful as they keep you awake and give you energy 4. VERY IMPORTANT!! always carry a torch covered with red cellophane sheet, any light source(torches etc)without a cellophane sheet is strictly prohibited . 5. Carry your equipment with you(even if you don't have telescope you please join us anyways, there will be a couple of telescopes around you) 6. cover you camera flash with black tape so no light comes out and always keep the flash on off mode 7. We are not providing any transport so you will have to come on your own 8. If you bring a camera bring a tripod with it to shoot	↑ D3 inst, D5 spont; ↓ D7 form, D1 abs
Text 11	For those who are a little rusty at math, Pi is the ratio of the circumference of a circle to its diameter. Most people who are not mathematicians will simply say that Pi is approximately 3.14 for short, or perhaps 3.1415 "and then some." However, this ratio is special in that it has never been completely calculated to completeness by humans or computers – the numbers after the decimal go on and on. Thanks to modern computer technology, Pi is calculated to thousands of digits with relative ease, and to trillions of digits by dedicated experts, but this is still not the complete number. Pi is called an irrational number. The complete numerical representation of Pi is incomprehensible, and yet, this ratio exists for every circle that we see. Pi is a key component of the mathematical subsystem underlying the universe. Whether you knew it or not, Pi is part of your life.	↑ D8 comp, D4 info; ↓ D5 spont, D3 inst

Text 12	For so long, government ministers have treated biodiversity as way down the to-do list, beneath winning the next election and ensuring asset markets and public services are not in meltdown. Plurality and integrity of natural life, of everything from parasites to parakeets, is no more objectionable to a politician than the latest Attenborough documentary. But doing much about it has never seemed a high enough priority. Until now, maybe. While the world is following every twist and turn in the battle against Covid-19, environmental experts from the UN to the World Health Organization have been busy pointing out that this pandemic's root cause is humanity's slow ravaging of nature. Annihilating forests to create farms and build roads, bringing wildlife into contact with people and their livestock, is fundamentally how this lethal virus spilled from wildlife into humankind. The result is 2.2 million deaths and rising fast. As Prof Dasgupta observes, the destruction of nature means that there will be another pandemic – and another, with all the devastation to life and economic value they could bring.	↑ D12 arg, D7 form; ↓ D5 spont, D10 inter

4. DATA ANNOTATION

This chapter addresses RQ1 by assessing the feasibility of annotating the data using the dimensions outlined in the Dimensional Text Model (DTM). A test dataset with dimension-specific annotations was collected through an annotation study using LimeSurvey to evaluate the DTM. Section 4.1 outlines the data selection process, while Section 4.2 describes the annotation procedure. The results of the annotation study are presented in Section 4.3, and Section 4.4 discusses how the proposed dimensions relate to each other using correlation and factor analysis. Finally, Section 4.5 summarizes the overall results observed within the annotated dataset.

4.1. Data

A total of 120 texts were selected from the etTenTen corpus (*etTenTen: Corpus of the Estonian Web* 2021; Koppel & Kallas 2022). The etTenTen[9] corpus is divided into six categories: *government*, *forum*, *religion*, *unknown*, *blog*, *periodicals* and *informative*, with the category *unknown* having the highest proportion of texts in the corpus. When selecting texts for the annotation study, the goal was to maintain the proportions of the existing categories in the corpus. For example, if 10% of the texts in the corpus are labelled as *periodical*, then 12 (10% of 120) texts with a label *periodical* were selected for the annotation study. A Python script was used to randomly select texts for each category. The texts were manually examined, and if a text was deemed unsuitable (e.g., too many URLs from the same source, or a text lacking length and intelligibility), the script selected alternative texts for each category.

The final dataset used in the annotation study can be found in GitHub[10]. It consists of 120 texts in separate files. The average text length is 186 words (including punctuation) and the average sentence length is 19 words. The median text length is 161 words, while the shortest is 67 words, and the longest is 423 words long.

4.2. Annotation study

This section gives an overview of the annotation study, which was designed to measure whether the proposed dimensions were recognizable to the annotators. Section 4.2.1 describes the annotation process and shows the guidelines provided to the annotators. Since the objective was also to measure the continuity of a dimension in a text, Section 4.2.2 describes how the judgements collected from

[9] This study used the first edition of the EtTenTen which contained Estonian websites crawled in the year 2013 (see `https://www.keeleveeb.ee/dict/corpus/ettenten/about.html`).

[10] `https://github.com/kristiinavaik/phd_project_vaik/tree/5a0db8228b5da5ae6d52111b98d3c49806a19213/data`.

the annotators were harmonized to suit the purposes of this dissertation. Section 4.2.3 gives an overview of the annotation results.

4.2.1. Annotation process

The objective was to assess whether the theoretically proposed dimensions were recognizable to the annotators. Accordingly, an annotation study was conducted using the LimeSurvey platform. The introduction page of a session in Estonian is shown in Figure 2. The English translation of the introduction page is given in Appendix A and the translations for the dimensions given to the annotators are presented in Appendix B.

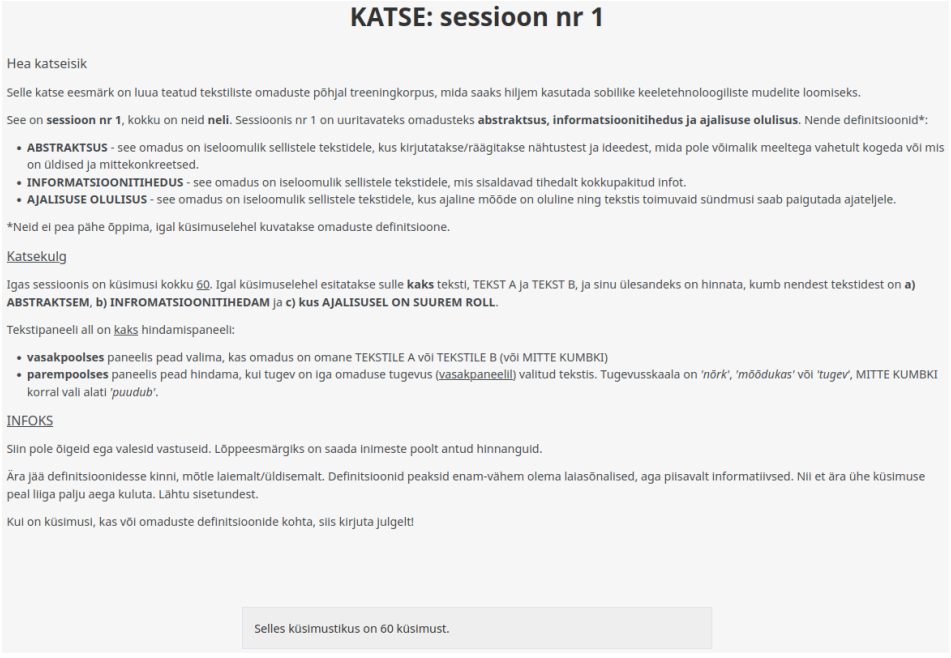

Figure 2: View of the introduction page for Session 1.

Ten annotators with a background in linguistics were involved in the annotation study. As it can be challenging for annotators to assess the 120 texts and twelve dimensions simultaneously, the task was structured to compare two texts at a time while assessing the level of salience of three dimensions per session on a four-point Likert scale. The goal was to determine the presence or absence of each dimension in a given text. The task was to determine which text displayed the dimension more prominently. The chosen text had to be scored from 1 to 3: 'dimension is weakly present' (1), 'dimension is moderately present' (2), and 'dimension is strongly present' (3). The unchosen text received a score of -1 indicating that there is no knowledge about the level of salience of a dimension. If the dimension was not considered to be present in either text, both texts received a score of 0, indicating that the dimension is uniformly absent in both texts. This

Table 2: Dimensional triplets for the annotation study.

Session nr	Dimensional Triplet
1;3	abstractness, information density, and temporality
4;8	instructability, complexity, and subjectivity
5;2	affectivity, formality, and interactivity
7;6	spontaneity, impersonality, and argumentativity

label was used for annotation purposes only, to highlight the extreme values for further analysis.

The annotation was carried out in two groups where each group had four independent sessions with a unique set of dimensions – Group 1 (G1) had sessions labelled 1, 4, 5, and 7; and Group 2 (G2) had sessions labelled 3, 8, 2, and 6, resulting in a total of eight sessions. Each session consisted of 60 questions (120 texts and 2 texts per question), and to ensure unbiased judgements, the three dimensions for each annotation session were chosen to be as independent as possible. The assessable dimensional triplets used in each session are presented in Table 2.

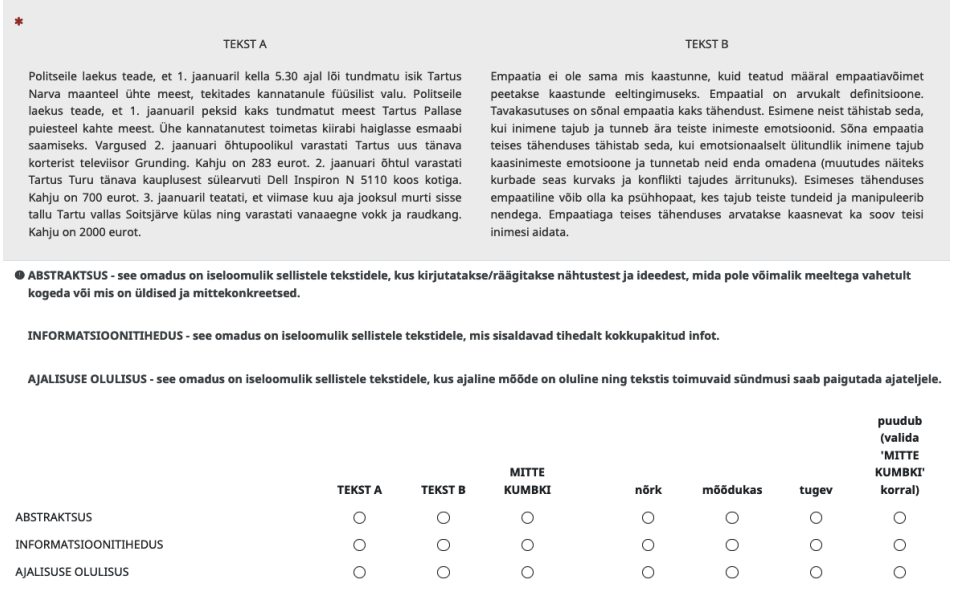

Figure 3: A question from Limesurvey.

The texts were displayed side by side, as seen in the screenshot in Figure 3. Below the grey text box, the definitions of the dimensions were reiterated. At the bottom of the page was the scoring panel, where the annotators were asked to assess the presence or absence of the dimensions in the texts on the left and to rate the texts on a 4-point Likert scale on the right: *weakly* 'nõrk', *moderately* 'mõõdukas', *strongly* 'tugev', *not present* 'puudub'.

4.2.2. Harmonizing the judgements

The annotation study was designed to measure the continuity of a dimension across texts, utilizing a four-point scale (*strongly*, *moderately*, *weakly*, *not present*) to assess the level of salience of each dimension. The annotation study generated multiple judgments per dimension for each of the 120 texts, necessitating a harmonization process. This harmonization was achieved by calculating the average Likert scale scores across all twelve dimensions, thereby creating unified judgements for subsequent analysis.

In some cases, the number of judgements provided by either group was insufficient to calculate a group average because the annotators had to choose one of the two texts or none at all. This resulted in a situation where knowledge of one text was available, while that of the other was lacking. To illustrate, if the annotator selected TEXT B as the one representing more of the dimension and judged it as *strong*, the level of salience of that dimension in TEXT A would be unknown (the dimension could be present *moderately*, *weakly* or not present at all). The TEXT A would then receive a score of -1 indicating that there is no knowledge about the salience of a dimension.

To address this issue, the group-based average was calculated using the majority voting scheme. This means that if a text received at least three judgements, the group-based average was calculated by dividing the sum of the judgements by the number of judgements available (`sum(judgements)/len(judgements)`). This resulted in dividing the texts into three groups based on their ratings, as illustrated in Figure 4. All texts were included in the final dataset, even though the judgements of G1 or G2 may not be as reliable as those of the $G1 \cap G2$ grouping.

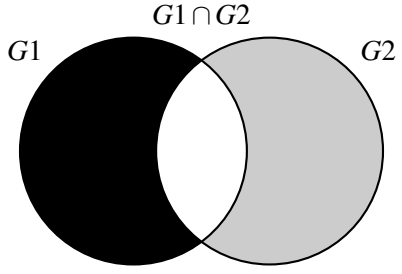

Figure 4: Grouping the judgements collected from the annotators.

- G1: a text received a judgement (min 3 judgements, max 5 judgements) only from Group 1 (black colour in Figure 4), or
- G2: a text received a judgement (min 3 judgements, max 5 judgements) from only Group 2 (grey colour),
- $G1 \cap G2$: a text received a judgement (min 6 judgements, max 10 judgements) from Group 1 and Group 2 (white colour).

The final rating for each text was computed by taking the average of the group-based averages. Due to the uneven sizes for the judgements of *strong* and *moderate* compared to texts labelled as *weak* or *not present*, the *strong* and *moderate* were combined into a single label. These final ratings were mapped as follows:

- *strong/moderate*: avg ≥ 2
- *weak*: $2 <$ avg ≥ 1
- *not present*: $1 <$ avg ≥ 0

The final dataset with the group-based averages is available in GitHub[11]. The CSV file includes the filenames and the final rating for each dimension by taking the average of the group-based averages.

4.2.3. Results of the annotation study

This section gives an overview of the annotation results based on the final dataset. The overall results are presented in Table 3. The rows show for each dimension how many texts were labelled as *strong/moderate*, *weak* or *not present* (including the proportions of texts that did not receive a score of -1). The column 'Total per Dimension' shows the sum of the columns 'strong/moderate', 'weak', and 'not present' and the percentage of these texts out of 120.

The column 'Total per dimension' shows great variability between the dimensions, with proportions (%) ranging from 62% to 83%. The highest proportions were for interactivity (83%), spontaneity (81%), subjectivity (81%) and abstractness (80%). The lowest proportions were for temporality (62%), information density (63%), and argumentativity (66%).

When comparing 'strong/moderate', 'weak' and 'not present' column proportions, the results show that, for certain dimensions, the distribution is skewed towards being *present* vs. being either *strongly/moderately* or *weakly* present. For example, for spontaneity, 81% of the texts were not considered spontaneous at all, while 12% of the texts were considered strongly/moderately and 6% of the texts were weakly spontaneous. The same applies to interactivity (75% not present vs. 10% strongly/moderately & 15% weakly interactive), instructability (64% not present vs. 16% strongly/moderately & 20% weakly instructive), complexity (61% not present vs. 10% strongly/moderately & 29% weakly complex).

Some dimensions show different distribution patterns compared to the other dimensions. For example, most annotators rated information density as strongly/moderately present for 59% of the texts, which shows that the annotators found that a significant proportion of the texts contained a substantial amount of information. In contrast, for temporality, the majority (53%) of the texts were rated as having temporality weakly present. The temporality is defined as being characteristic of texts in which time plays an important role and events can be plotted on a timeline (see Section 3.2). However, determining the linearity of time can be challenging.

[11] `https://github.com/kristiinavaik/phd_project_vaik/blob/5a0db8228b5da5ae6d52111b98d3c49806a19213/annotation_phd_project_vaik.csv`

Table 3: Results of the annotation study: the distribution of judgements on 4-point Likert scale.

Dimension	strong/ moderate	weak	not present	Total per dimension
subj	30 (31%)	15 (15%)	52 (54%)	97 (81%)
aff	28 (31%)	18 (20%)	45 (49%)	91 (76%)
form	24 (15%)	28 (31%)	49 (54%)	91 (76%)
spont	12 (12%)	6 (6%)	79 (81%)	97 (81%)
inst	15 (16%)	18 (20%)	58 (64%)	91 (76%)
inter	10 (10%)	15 (15%)	75 (75%)	100 (83%)
imp	37 (43%)	21 (24%)	28 (33%)	86 (72%)
temp	16 (21%)	39 (53%)	19 (26%)	74 (62%)
comp	9 (10%)	25 (29%)	53 (61%)	87 (73%)
arg	21 (27%)	27 (34%)	31 (39%)	79 (66%)
abs	6 (6%)	35 (36%)	56 (57%)	97 (80%)
info	45 (59%)	26 (34%)	5 (7%)	76 (63%)

Thus, the predominant rating of the temporality dimension as weakly present may indicate that the degree of temporality may not be apparent and clearly expressed to the annotator.

4.3. Inter-annotator agreement

Given the subjective nature of the judgments collected from the annotators, it was crucial to evaluate the reliability of each dimension. To this end, the inter-annotator agreement was employed as a metric for assessing the consistency and reliability of the annotations. This section describes the measure used for the evaluation and provides a concise overview of the results.

The majority of the known inter-annotator agreement metrics are based on the following common formula (Artstein & Poesio 2008):

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)}, \tag{1}$$

where p_o is the observed agreement (the proportion of items where all annotators agree), and p_e is the chance agreement (expected agreement if annotations were made randomly). κ is a numerical value that ranges from -1 to 1, where 1 indicates perfect agreement between annotators, 0 conveys agreement due to randomness, and -1 indicates systemic disagreement between annotators.

The most common measures of inter-annotator agreement include Cohen's κ (1960), Fleiss' κ (1971), and Krippendorff's α (2004). Cohen's κ is most appropriate when there are nominal observations (i.e., categories), each item has one label, and there are two annotators. The observed agreement, p_o, can be considered the percentage of agreement between annotators. In contrast, p_e represents the chance agreement between categories when assuming random annotation. In the event of there being more than two annotators, Fleiss' κ is applicable, whereby extending Cohen's κ to handle more than two annotators. Given the structure of the annotation study conducted in this dissertation, a more general measure was required. The study yielded two types of data: interval (the level of dimensional salience measured on a 4-point Likert scale) and missing data (texts not evaluated by the annotators). To accommodate these different kinds of data, Krippendorff's α was deemed as a more appropriate metric, as it is capable of handling both interval scales and missing values, while also accommodating the inclusion of multiple annotators.

According to Krippendorff (2004), a reliability threshold of $\alpha = 0.8$ indicates strong agreement, although preliminary conclusions can still be drawn with $\alpha \geq 0.667$. In this dissertation, this standard was considered to be too conservative and therefore, the conventions proposed by Landis & Koch (1977) were adopted, where acceptable agreements start at 0.4 (see Figure 5).

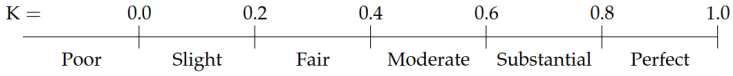

Figure 5: Interpretations for κ values according to Landis and Koch (1977).

The results of the inter-annotator agreement for each dimension are presented in Table 4. The inter-annotator agreements ranging from moderate to substantial ($0.4 < \alpha < 0.8$) are shown in the first and second column blocks. These include dimensions such as subjectivity, affectivity, formality, spontaneity, instructability, interactivity, impersonality, and temporality (8 out of 12). Fair agreement with $0.2 < \alpha < 0.4$ was observed for dimensions listed in the third column block, such as complexity, argumentativity, abstractness, and information density.

Table 4: The inter-annotator agreements for each dimension.

Dimension	α	Dimension	α	Dimension	α
subj	.76	inst	.47	comp	.38
aff	.74	inter	.46	arg	.38
form	.63	imp	.42	abs	.33
spont	.6	temp	.4	info	.25

The results demonstrate a promising trend in the inter-annotator agreement measures. It is noteworthy that subjectivity and affectivity exhibit a relatively high degree of agreement, with $\alpha \geq 0.70$. In general, eight out of twelve dimensions exhibit moderate to high levels of substantial agreement, which demonstrates a notable consistency between annotators. Additionally, there were dimensions where the annotators did not achieve high agreement, these included abstractness, complexity, argumentativity, and information density. These dimensions require further refinement, and additional training sessions for annotators could be beneficial in achieving more consistent annotations.

4.4. Co-occurrence patterns between dimensions

Despite the Dimensional Text Model's assumption of dimensional independence, further investigation through correlation and factor analyses was needed to examine potential relationships between dimensions. The correlation analysis aimed to verify that the dimensions did not measure similar communicative functions, and exploratory factor analysis was employed to identify potential underlying patterns among dimensions. This section proceeds by first examining pairwise correlations between dimensions to establish direct relationships, which is followed by demonstrating the broader latent patterns through exploratory factor analysis.

4.4.1. Correlation analysis

Correlation determines the strength and direction of the relationship between two dimensions collected from the annotation study.

The most commonly used metrics to measure the relationship between two variables are Pearson's and Spearman's correlation coefficients (Akoglu 2018). The Pearson correlation coefficient is a parametric test measuring the strength of linear association between two continuous variables. It is suitable for interval or ratio, and normally distributed data, assuming a linear relationship between the variables. Since the interest was to demonstrate the monotonic relationship (whether linear or not) between the dimensions, and the data were ordinal and not normally distributed, Spearman's correlation coefficient was considered more appropriate. The coefficient values range from -1 to 1 where the value of -1 indicates a perfect negative correlation, which means that as the salience of one dimension increases, the salience of another dimension decreases. In contrast, a value of 1 indicates a perfect positive correlation, which shows that if the salience of one dimension increases, so does the salience of the other dimension. A highly positive correlation may be an indication of collinearity, requiring some additional measurements. In cases where the strength and direction between the dimensions do not show a discernible pattern or relationship, the coefficient is 0, indicating the absence of correlation. The Spearman's correlation was measured using the SciPy (Virtanen et al. 2020) statistical *stats*[12].

Table 5: Spearman's correlations between dimensions.

	abs	info	temp	aff	inter	inst	form	comp	subj	spont	imp	arg
abs	**1**	-0.2	**-0.41**	0.24	0.09	0.12	-0.18	0.29	0.24	0.08	-0.03	0.29
info		**1**	0.14	**-0.57**	**-0.45**	0.13	**0.61**	**0.41**	**-0.55**	**-0.41**	**0.62**	-0.12
temp			**1**	-0.01	-0.01	-0.16	0.05	0.13	-0.05	-0.007	-0.09	-0.25
aff				**1**	**0.64**	-0.11	**-0.65**	-0.15	**0.64**	**0.54**	**-0.64**	0.26
inter					**1**	0.05	**-0.44**	-0.12	**0.43**	**0.41**	**-0.61**	0.27
inst						**1**	0.05	0.001	-0.15	0.07	0.18	0.19
form							**1**	0.35	**-0.62**	**-0.63**	**0.61**	-0.1
comp								**1**	-0.19	-0.16	0.36	0.11
subj									**1**	**0.58**	**-0.55**	0.36
spont										**1**	**-0.46**	0.01
imp											**1**	-0.05
arg												**1**

Table 5 presents the pairwise correlations between the dimensions. Correlations in bold represent those which demonstrate at least a moderate relationship between the dimensions (the threshold was set between 0.4 and -0.4). The overall results show that there is no indication of strong collinearity among the dimensions. Table 5 demonstrates that affectivity, interactivity, subjectivity and spontaneity have a positive correlation (ranging between .41 and .64). Conversely,

[12] `https://docs.scipy.org/doc/scipy/reference/stats.html#statistical-functions-scipy-stats`

the second bundle, consisting of formality, impersonality and information density, also showed a positive correlation (ranging between .61 and .62). Notably, these two bundles appear to have an adversarial relationship, suggesting that texts high in affectivity, interactivity, subjectivity and spontaneity tend to be low in formality, impersonality and information density, and vice versa.

While the two primary bundles show moderate correlations (either positive or negative) with each other, other dimensions show fewer associations with other dimensions. Complexity shows a moderate positive correlation with information density (r = .41), while abstractness and temporality only correlate with each other (r = .41). The lack of significant correlations between instructability and argumentativity with other dimensions suggests that these dimensions have unique characteristics.

4.4.2. Exploratory factor analysis

Exploratory factor analysis is a statistical technique used to identify underlying factors or dimensions in a dataset with numerous variables. By reducing the dimensionality of the data, exploratory factor analysis helps to uncover hidden patterns and relationships. Factor loadings, which represent the correlation between a variable and a factor, are used to interpret the results. A high loading (close to 1 or -1) indicates a strong association between the variable and the factor, while a low loading (close to 0) suggests a weak relationship. Additionally, the percentage of variance (% Var) is reported to assess the significance of each factor. A higher % Var signifies a more important factor that explains a larger portion of the total variation in the data. The Python package FactorAnalyzer (2022) (inspired by the R packages *psych* and *sem*) was used for conducting the factor analysis. The analysis was performed using the default settings - the minimal residual method and the varimax rotation.

The correlation matrix presented in Table 5 was used as an input for the exploratory factor analysis. The scree plot (Figure 6) demonstrated that three was the most appropriate number to explain the variance in the data (eigenvalue ≥ 1).

Table 6 shows which dimensions load to which factors. The table only reports dimensions with a loading ≥ 0.4. As can be seen, the dimensions load on three factors but are primarily associated with one factor each, with minimal cross-loading on other factors. The only exception is argumentativity, which, although mainly loaded on Factor 2, is also significantly loaded on Factor 1.

The three factors together explain about 99% of the total variance in the data (67% Factor 1, 23% Factor 2 and 9% Factor 3), suggesting that the dimensions are not entirely independent. Factor 1 grouped eight dimensions. From one end, Factor 1 represents a shift towards a more expressive and interactive style of communication, where texts are more focused on verbal communication and participants play a more prominent role. From the other end, Factor 1 captures more formal and information-rich texts. Factor 2 combines three dimensions. Argumentativity

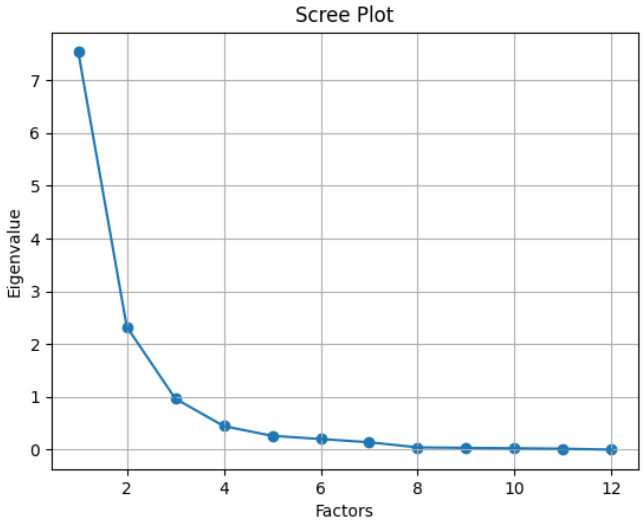

Figure 6: **Scree Plot for Exploratory Factor Analysis.** Typically, the eigenvalues ≥ 1 are used to determine the number of factors.

and abstractness had a positive load on Factor 2, and temporality had a negative load on Factor 2. One end of Factor 2 represents language that is used to offer viewpoints or perspectives on ideas and concepts without emphasizing time, and the other end describes language where the linearity of actions is most important. Instructability emerges as the dimension that differs from the others, as it was the only dimension that loaded positively on Factor 3.

These results provide a valuable understanding of how these dimensions relate to each other within the Dimensional Text Model. However, it is advisable to interpret the results with caution because of the low agreement for certain dimensions, such as information density, abstractness, argumentativity, and complexity.

4.5. Summary

This chapter addressed RQ1 by assessing whether the dimensions proposed by the Dimensional Text Model were recognizable and distinguishable to the annotators. For this purpose, an annotation study was conducted in which the ten annotators were asked to rate the level of salience of the proposed dimensions on a four-point Likert scale (*strong*, *moderate*, *weak* or *non-existent*) in a collection of 120 texts gathered from the EtTenTen Web corpus (Koppel & Kallas 2022). The analysis was further enriched by performing correlation and exploratory factor analysis (EFA). Correlation analysis was used to assess the collinearity between the dimensions, i.e., a high correlation indicates that the dimensions may be measuring similar communicative functions. Additionally, although the dimensions proposed by the Dimensional Text Model are assumed to be distinct and inde-

Table 6: Results of the factor analysis. This analysis was performed using the default minimal residual method and the varimax rotation.

Dimensions	Factor 1	Factor 2	Factor 3
spontaneity	0.90		
affectivity	0.94		
interactivity	0.91		
subjectivity	0.92		
formality	-0.95		
information density	-0.94		
impersonality	-0.97		
complexity	-0.80		
temporality		-0.93	
instructability			0.78
argumentativity	0.41	0.60	
abstractness		0.85	
Eigenvalues	6.99	2.36	0.91
Var	0.58	0.20	0.08
% Var	0.67	0.23	0.09

pendent, EFA was used to identify whether the proposed dimensions form macro dimensions by showing how the dimensions are similar to each other based on the annotations.

The inter-annotator agreement, measured with Krippendorff's α, demonstrated that subjectivity, affectivity, formality, and spontaneity were relatively easy for the annotators to identify (α ranged between 0.6 and 0.76). This suggests that subjectivity, affectivity, formality, and spontaneity dimensions are well-defined, recognizable and distinguishable within the Dimensional Text Model. The results demonstrate that instructability, interactivity, impersonality and temporality were generally identifiable to the annotators (α ranged between 0.4 and 0.47), although there was some disagreement among the annotators for certain dimensions (such as temporality, and impersonality). The fair agreement (α ranged between 0.25 to 0.38) for abstractness, complexity, argumentativity, and information density shows that these were less identifiable and distinguishable for the annotators.

The overall results of the correlation analysis show that, although some dimensions exhibited strong positive correlations, there is no indication of strong collinearity among the dimensions. The analysis revealed two different dimensional groups. One comprising affectivity, interactivity, subjectivity and spontaneity, and another comprising formality, impersonality and information density. These groups showed that there were positive correlations within the group, e.g. when a text was considered subjective, it was also more affective, interactive and spontaneous; when a text was considered impersonal, it was also more information dense and formal. There was also an inverse relationship between these

groups. For example, when a text was considered affective, it was less impersonal, complex and information-dense, or when a text was considered information-dense, it was less spontaneous, subjective and interactive. Other dimensions showed fewer connections with other dimensions. Complexity shows a moderate positive correlation with information density, while abstractness and temporality show a moderate correlation. The lack of significant correlations between instructability and argumentativity with other dimensions suggests that these dimensions have distinctive characteristics within the Dimensional Text Model framework.

Factor analysis provided further insight into the dimensional structure, revealing three primary factors that collectively explained 99% of the total variance. The dominant factor, accounting for 67% of variance, represented a spectrum between informal, interactive communication (characterized by affectivity, interactivity, subjectivity, and spontaneity) and formal, information-focused communication (marked by formality, impersonality, information density, and complexity). The second factor united argumentativity and abstractness in opposition to temporality, suggesting a distinction between conceptual and temporal aspects of text. The third factor, uniquely loaded by instructability, indicated this dimension's distinctive nature within the framework.

This chapter concludes that while some dimensions of the Dimensional Text Model were recognizable to the annotators, others were less so. The distinct emergence of dimensional groups and factors reveals underlying patterns in how these dimensions are manifested across texts. These results demonstrate that while the model provides an applicable taxonomy suitable for text classification, further refinement and clarification may be necessary for certain dimensions to improve the models' generalizability and applicability.

5. LINGUISTIC FEATURES

The Dimensional Text Model (DTM) assumes that the latent dimensions are realized by unique sets of co-occurring features, without specifying the compositions of the feature sets. To address RQ2, namely whether and how the dimensions defined by the Dimensional Text Model (DTM) exhibit statistically significant differences in their linguistic profiles, it was necessary to automatically extract a wide range of different lexical and grammatical features.

The Estonian text analysis has not fully explored the extent to which linguistic features contribute to the discrimination of registers. The objective of the next overview is not to provide an exhaustive list of all relevant research but rather to highlight some of the most significant studies conducted in this field using Estonian data. Most studies have done qualitative research using handpicked features and have studied how these features vary across various registers. For example, Kerge and her colleagues used part-of-speech and Heylighen and Dewaele's (Heylighen & Dewaele 2002) index[13] to examine the formality of texts (Kerge et al. 2007; Kerge 2009; Kerge & Pajupuu 2010; Puksand & Kerge 2011). Their work found that the lower frequency of pronouns, verbs, adverbs, and interjections makes a text more contextual and ambiguous, whereas the higher frequency of nouns, adjectives, and prepositions reduces contextuality and makes the text more formal. The role of nominalization in complex texts has been extensively studied in Kerge's dissertation (2003b), also in Puksand & Kerge (2011). Texts with a higher degree of nominalization are more difficult to understand, as the abundance of nouns makes the chain of referential concepts denser and less comprehensible. Experimental methods for identifying the emotional valence of texts have been pursued by Pajupuu and colleagues, whose emotion detector is based on a lexicon pre-annotated for emotional valence (Pajupuu et al. 2012, 2016). Meier (2002; 2003) used correlational analyses of linguistic features to study the characteristics of various registers (mainly essays, scientific texts, and literary texts). Meier highlighted, among other things, that narrative texts (where the plot and narrator are crucial) use more personal pronouns, conjunctions, participles, and shorter words (2 to 4 characters) compared to documents and more formal texts. She found that pronouns such as *see* 'it' and *keegi* 'somebody' are more common in subjective texts. Meier also notes that formal texts (such as academic papers and documents) have more *mine*-nominalization[14], impersonal voice, and subordinating conjunctions. Gailit (2023) studied the spontaneity and formality of Web texts by using an extensive set of linguistic features (e.g. different lexicons for slang and loanwords, type/token ratio, average lemma length, number of repetitions in

[13] The idea was that the use of nouns acting as referentials, in conjunction with adjectives and prepositions, would enhance the unambiguity of the text, particularly in formal contexts. Conversely, the presence of other parts of speech was posited to increase ambiguity, necessitating a greater degree of contextualization.

[14] *mine* is a deverbal suffix used for nominalization in Estonian (Erelt 2017).

a word). Reinsalu (2019) examined how instructivity is expressed in directives, orders, statements, requests for information, complaints, service contracts, and agency agreements. The results show that instructivity is mainly expressed by modal verbs (*kohustama* 'to oblige', *võima* 'to be able to', *pidama* 'to hold'), impersonal voice, *da*-infinitives, and lexical markers (*palun* 'please', *soovin* 'I would like to').

The Dimensional Text Model does not dictate which features are essential to extract. Instead, the linguistic features serve as a means to an end in highlighting the distinct linguistic profiles of the dimensions, thus the relative frequencies of 85 lexical and grammatical features were automatically extracted from the final dimensionally annotated dataset (see Section 4.1). The dataset was automatically tagged using Stanford's Stanza (Qi et al. 2020) parser, which can handle many languages with consistent performance in different tasks. It uses a language-agnostic neural pipeline for text analysis, which includes tokenization, lemmatization, parts of speech and morphological feature tagging, and dependency parsing. Stanza's design is rooted in the Universal Dependencies (Nivre et al. 2017, UD) framework, which is based on the concept of dependency grammar which views sentence structure as a network of directed links between words, where each link represents a grammatical relationship between two words. These relationships can include subject-verb, object-verb, modifier-head, and many others. Scripts used for automatic tagging and feature extraction can be found at GitHub[15].

5.1. Lexical features

Lexical features are typically represented as lists of strings, such as bags of words (Finn & Kushmerick 2006; Fang & Cao 2010), specialized lexicons (Zhang 2016; Biber 1988), or *n*-grams (Crossley & Louwerse 2007; Santini 2007; Sharoff 2010; Gries et al. 2011; Pritsos & Stamatatos 2018). Lexical features typically also include simpler derivative features to measure lexical diversity, e.g., type/token ratio, *hapax legomena* and average word length. These derivative features are prone to be sensitive to text length. The following lexical features were extracted:

1. pronoun/noun ratio
2. type/token ratio (TTR)
3. average word length
4. average sentence length
5. *hapax legomena*
6. filled pauses, e.g., *jaa*, *mhmh*, *mh*
7. *ioon*-suffix, e.g. *organisats-ioon*
8. *mine*-suffix, e.g., *magamine* 'sleeping',
9. abbreviations, e.g., *nt* (*näiteks*) 'for example', *vt* (*vaata*) 'look'
10. core verbs, e.g., *olema* 'be', *saama* 'get'
11. *see* 'it' as a pronoun
12. *see* 'it' as a determinative
13. 1st person singular pronoun, e.g.,

[15] https://github.com/kristiinavaik/phd_project_vaik/tree/5a0db8228b5da5ae6d52111b98d3c49806a19213.

mina/ma ‘I, me’

14. 1st person plural pronoun, e.g., *meie/me* ‘we’
15. 2nd person singular pronoun, e.g., *sina/sa* ‘you’
16. 2nd person plural pronoun, e.g., *teie/te* ‘you’
17. 3rd person singular pronoun, e.g., *tema/ta* ‘he/she’
18. 3rd person plural pronoun, e.g., *nemad/nad* ‘they’
19. 1st person verb forms, e.g., *mina sain* ‘I got’
20. 2nd person verb forms, e.g., *sina said* ‘you got’
21. 3rd person verb forms, e.g., *tema sai* ‘he/she got’

Some features were inspired by Biber (1988). For example, a high type/token ratio **(2)** may be characteristic of formal, academic texts, but not of spoken discourse, which is constrained by temporal and memory limitations. *Hapax legomena* **(5)** characterize unique vocabulary and could provide useful insights into discriminating dimensions by identifying lexical richness. In written and edited discourse, temporal constraints are less prevalent and participants may be motivated to invest time in improving lexical variation by employing longer words **(3)**, and constructing longer sentences **(4)**. For example, Biber (1988: 238) observed that the use of shorter words was characteristic of presenting information, which aims to convey the exact content with as few words as possible. A lower pronoun/noun ratio **(1)** may suggest a more general and impersonal approach. The higher use of abbreviations **(9)**, 3rd person verb forms and pronouns **(17, 18, 21)** could be more common in formal, legal, scientific, and technical texts. The *ioon*-suffix **(7)** is typically used in load words, and *mine*-suffix **(8)** is for deverbal nominalization, thus they could be used for expressing abstract concepts and ideas, and be more characteristic of the formal and more complex style of writing.

The feature extraction script also included features that may be more characteristic for the written spoken discourse where the participants face temporal and cognitive restrictions, such as a higher use of core verbs[16] **(10)** which are frequently occurring verbs that perform different grammatical functions and/or express very general concepts. Furthermore, the need to establish a direct connection or address the reader directly may lead to the use of more first- and second-person (**13**-**16**) and demonstrative pronouns (**12**, **13**), filled pauses **(6)**, and first- and second-person **(20**-**21)** verb forms. These features can be used in discourses where personal narratives and persuasion aim to engage the reader on a personal level.

[16]The list of Estonian core verbs is taken from (Tragel 2003); it includes 18 verbs: *olema* ‘to be’, *saama* ‘to get’, *tulema* ‘to come’, *pidama* ‘to keep’, *tegema* ‘to do, to make’, *minema* ‘to go’, *võima* ‘to may’, *jääma* ‘to remain’, *võtma* ‘to take’, *hakkama* ‘to begin’, *andma* ‘to give’, *tahtma* ‘to want’, *panema* ‘to put’, *käima* ‘to walk’, *tooma* ‘to bring’, *viima* ‘to take (to a place)’, *laskma* ‘to let’, *ajama* ‘to drive’.

5.2. Grammatical features

The extracted grammatical features are divided into four subclasses: Parts of speech (Section 5.2.1), grammatical categories of verbs (Section 5.2.2), grammatical categories of nouns (Section 5.2.3) and syntactic features (Section 5.2.4).

5.2.1. Parts of speech

22. common nouns, e.g., *loom* 'animal'
23. adjectives, e.g., *kallis* 'dear' or 'expensive'
24. numerals, e.g., *kolm* 'three',
25. pronouns, e.g., *meie* 'we', *sama* 'same'
26. adverbs, e.g., *kõrvuti* 'beside'
27. interjections, e.g., *aitäh* 'thank you'
28. coordinating conjunctions, e.g., *ning*, *ja* 'and'
29. subordinating conjunctions, e.g., *et* 'that', *sest* 'because'
30. adpositions, e.g., *enne* 'before', *kolmas* 'third'
31. symbols, e.g., @, *www.ut.ee*
32. punctuation
33. proper nouns, e.g., *Edgar*
34. determiners, e.g., *kõik* 'all', *muu* 'other'
35. nominals = nouns + adjectives + numerals + pronouns

This subclass contains nine Estonian parts of speech tags – common nouns **(22)**, adjectives **(23)**, numerals **(24)**, pronouns **(25)**, adverbs **(26)**, interjections **(27)**, coordinating **(28)** and subordinating **(29)** conjunctions, adpositions **(30)**. In Estonian grammar descriptions, the determiner **(34)** is a novel category that provides information about the noun, such as definiteness, quantity and possession. Determiners operate as instruments for marking coreference[17]. The list of parts of speech tags was complemented with three additional categories that are not considered parts of speech in traditional grammar descriptions, but that are marked as such by the Stanza tagger – symbols **(31)**, punctuation marks **(32)** and proper nouns **(33)**. In UD, a symbol is a word-like entity that differs from ordinary words in form and function. Many symbols contain non-alphanumerical characters, such as mail addresses, websites, and emojis. The difference between a symbol and punctuation is that a symbol can be replaced with a real word. For example, the '$' in *10$* can be replaced with 'dollar' in *10 dollars*. A proper noun is a subclass of nouns (or nominal content words) used for the names of specific people or places, e.g., *New York*, *Mary*. The acronyms of proper nouns, such as *NATO*, are tagged by Stanza as abbreviated proper nouns (*PROPN Abbr=Yes*). The subclass of Parts of speech is supplemented with the nominals **(35)** which aggregates the normalized counts of nouns, adjectives, numerals and pronouns. Adding nominals

[17]Determiners are thoroughly described in the comprehensive overview of Estonian syntax (Pajusalu 2017: 382-384).

to the feature list was motivated by the recognition that individual nouns, adjectives, numerals, and pronouns may not have a statistically significant influence on their own.

5.2.2. Verbs

36. personal voice, e.g., *tema sai* 'he/she got'
37. impersonal voice, e.g., *saadi* 'got'
38. verb type ratio = number of finite verb forms/number of infinite verb forms
39. finite verb forms, *mina sain* 'I got'
40. infinite verb forms, *magama*, 'to sleep', this is a combined feature of 42 to 45
41. *da*-infinitive, e.g., *saada* 'to get'
42. converb, e.g., *saades* 'getting'
43. supine, e.g., *saama* 'to get'
44. verb participle, e.g., *saanud* 'got', *võetud* 'taken'
45. present tense, e.g., *saame* '(we) get'
46. past tense, e.g., *saime* '(we) got'
47. indicative mood, e.g., *saite* '(you) got'
48. conditional mood, e.g., *saaksin* '(I could) get'
49. imperative mood, e.g., *saa* 'get'
50. quotative mood, e.g., *saavat* 'get'
51. negative polarity, e.g., *ei saa* 'can not get'

Due to the limited research on Estonian verbal inflectional features in text analysis, verbal features as a separate subclass reflect an educated guess. For example, personal voice **(36)** could be more common in texts where the communication and the transfer of information are more important. The impersonal voice **(37)** may be more common in texts in which the agent is less prioritized or intentionally obscured. The impersonal voice is used for the deliberate suppression of the syntactic subject, and this feature is characteristic of Estonian academic writing.

The past tense **(46)** might be characteristic of texts where narration is important, and the focus is on past events. The present tense **(45)** might be more characteristic of texts that more commonly engage directly in a conversation or dialogue. The indicative mood **(47)** might be more characteristic of situations where the primary aim is to convey information clearly and concisely. The functional significance of other moods – conditional **(48)**, imperative **(49)**, quotative **(50)** – is less obvious and remains to be revealed.

The subclass also includes different verb-related features, such as finite **(39)** and infinite verb **(40)** forms, and the verb type ratio **(38)** (finite verb forms/infinite verb forms)[18]. The higher the use of finite verb forms, the more emphasis is placed on actions, events and specific temporal references, thereby indicating a

[18] A smaller ratio indicates that there are more non-finite constructions, or more modal constructions (such as *pean tegema* '(I) must do') present. The former could be more characteristic for more complex and formal texts.

more dynamic or narrative context. On the other hand, infinite verb forms, such as *da*-infinitive **(41)**, converbs **(42)**, supine **(43)** and verb participles **(44)**, which could indicate more complex sentence structures. In Estonian, converbs are much more common in the written language than in the spoken language (Erelt 2017: 808). *Da*-infinitive in Estonian has no grammatical categories other than being a verb. It has many functions, e.g., the main predicate in certain types of clauses, part of a complex predicate in modal constructions, and the subject or object of a clause. It also plays the role of the subject in experiential clauses where the subject is absent or not present at the beginning of the clause. A higher frequency of *da*-infinitives can indicate a more descriptive, elaborative and informative style of writing.

Polarity is a binary morphological category, but since positive polarity is semantically unmarked, only **(51)** negative polarity can be used for further investigation.

5.2.3. Nouns

52. nominative, e.g., *maja* 'house', 'building', *õde* 'sister'
53. genitive, e.g., *maja*, *õe*
54. partitive, e.g., *maja*, *õde*
55. illative, e.g., *majja*, *õesse*
56. inessive, e.g., *majas*, *ões*
57. elative, e.g., *majast*, *õest*
58. allative, e.g., *majale*, *õele*
59. adesive, e.g., *majal*, *õel*
60. ablative, e.g., *majalt*, *õelt*
61. transitive, e.g., *majaks*, *õeks*
62. terminative, e.g., *majani*, *õeni*
63. essive, e.g., *majana*, *õena*
64. abesive, e.g., *majata*, *õeta*
65. comitative, e.g., *majaga*, *õega*

The existing research on the variation of cases between registers in Estonian is relatively limited. However, the case frequency analyses conducted on the Balanced Corpus of Written Estonian and its sub-corpora have demonstrated that the frequency of cases differs between fictional, news and scientific texts.[19]. Therefore, this dissertation can contribute to a better understanding of how the cases influence textual variation.

5.2.4. Syntax

66. nominal subject, e.g., *kass nägi koera* 'cat saw the dog'
67. nominal copular subject, e.g., *kass on triibuline* 'cat is stripy'
68. modal verb/ auxiliary, e.g., *pean tegema* 'I must do'
69. relative clause modifier, e.g., *naine, kes mind üles kasvatas, ei ole minu ema* 'woman who raised me isn't my mother'

[19] https://cl.ut.ee/ressursid/gram-kat/tabel9,10,11,12,13/

70. clausal subject, e.g., *oodata on igav* ‘it is boring to wait’ (lit. ‘to wait is boring’)
71. clausal copular subject, e.g., *naerda on tervislik* ‘it is healthy to laugh’
72. object, e.g., *kass nägi koera* ‘cat saw the dog’
73. clausal complement, e.g., *ma ei teagi, kas see film mulle meeldis* ‘I’m not sure whether I liked that movie’
74. open clausal complement, e.g., *julgeks seda filmi soovitada kõigile* ‘I would dare to recommend this movie to everyone’
75. oblique nominal, e.g., *ta teeb selles filmis head tööd* ‘he is doing a great job in this movie’
76. nominal modifier, e.g., *rumala ja vaese rahva lollitamine* ‘fooling people who are stupid and poor’
77. appositional modifier, e.g., *ta tuli oma poegadega Mati ja Jüriga* ‘he came with his sons Mati and Jüri’
78. numeric modifier, e.g., *vahemaa oli 20 kilomeetrit* ‘distance was 20 kilometres’
79. adjectival modifier, e.g., *triibuline kass jõi vett* ‘stripy cat drank water’
80. adverbial clause modifier, e.g., *võrreldes eelmise aastaga on asi rahulik* ‘compared to last year things are peaceful’
81. vocative particle, e.g., *Peeter, ära jama nüüd* ‘Peeter, stop messing around’
82. copula, e.g., *see on minu maja* ‘this is my house’
83. conjunct, e.g., *kass ja koer* ‘the cat and dog’
84. coordinating conjunction, e.g., *kass ja koer* ‘the cat and dog’
85. discourse element, e.g., *appi* ‘jesus’, ‘help’; *nojah* ‘oh well’

As with other subclasses, the existing research on whether and how syntactic features differentiate between registers in Estonian is relatively limited. This might be because syntactic features are complex, their function in a text may not be as transparent as with other feature classes. Also, they have the potential to overlap with other linguistic elements mentioned above, such as part of speech.

Since nominal **(76)**, adjectival **(79)**, appositional **(77)**, and numeric modifiers **(78)** contribute to the construction of noun phrases, they can be associated with texts where the emphasis is on conveying information. The relative clause modifier **(69)** serves the purpose of modifying a nominal, while maintaining a coreferential relation with a constituent within the relative clause, and is often used in descriptive and argumentative texts. The copula **(82)**, and other features related to it, such as the nominal copular subject **(67)**, are used to express a state of being or a relationship between two elements. Copula sentences and their counterparts follow a clear structure, i.e., they map subjects to their complements, which describe or rename the subject. Thus, copula and other related features could be associated with texts that convey precise and accurate information within limited time constraints.

The vocative **(81)** and discourse **(85)** particles both have a role in engaging with an addressee and explicitly expressing emotions. Therefore, they may be characteristic of texts where dialogue and communication serve the main purpose. The modal/auxiliary verbs **(68)** express the probability, obligation, recommendation, permission, etc., so they might be associated with texts where it is important to convey attitude or perspectives.

5.3. Summary

The Dimensional Text Model (DTM) postulates that the proposed twelve dimensions can be represented and distinguished from each other by a distinct set of co-occurring linguistic features that act as a linguistic profile. However, the DTM does not prescribe which features are essential to extract, nor the composition of the feature sets of the dimensions. These features serve only as a tool to distinguish the linguistic profiles of dimensions.

As there has been little research in Estonian on which linguistic features contribute to register discrimination, there is little existing knowledge from which to draw. Consequently, the selection of linguistic features was based on the output of Stanford's Stanza (Qi et al. 2020) parser. The aim was to represent a diverse range of features. The final list contained 85 linguistic features. These features were grouped into two classes. The first class included lexical features, such as lexicons and various derived features, which were used to measure lexical diversity. The second class included grammatical features such as parts of speech, various noun and verb categories and syntactic categories.

6. RESULTS OF THE LINGUISTIC PROFILES OF THE DIMENSIONS

This chapter addresses RQ2 by determining which linguistic features, if any, are significantly associated with the level of salience of the twelve dimensions proposed by the Dimensional Text Model. Section 6.1 gives an overview of the methodological approaches used to identify statistically significant features and determine their relationships with the dimensions. Section 6.2 reports the results for each dimension individually. Section 6.3 offers a comparative analysis by discussing the emerging patterns among the dimensions. Finally, Section 6.4 summarizes this chapter by recapping the key results.

6.1. Methods

This section outlines the analysis of variance (ANOVA) method (Section 6.1.1), which identifies linguistic features that have a statistically significant association with the level of dimensional salience. Following this, Section 6.1.2 introduces the *post-hoc* test used to explore the specific relationships between these significant features and the dimensions.

6.1.1. Analysis of variance

ANOVA is a method used to compare whether two or more groups are significantly different from each other. Given the non-normal distribution of the feature's relative frequencies, the non-parametric ANOVA Kruskal-Wallis test (Kruskal & Wallis 1952) was used. Unlike standard parametric ANOVA, which analyzes mean differences, non-parametric ANOVA evaluates disparities in median values across the groups. The ANOVA test outputs an H-statistic and its corresponding p-value for each variable.

When conducting multiple tests, the likelihood of committing a Type I error increases. To address the issue of multiple hypothesis testing, the Holm-Bonferroni correction method (Holm 1979) is used. This correction method adjusts the p-value based on the α (usually .05), the ranks of the Kruskal-Wallis p-values, and the number of tests (see Formula (2)) for each variable.

$$P_{HB} = \frac{\alpha}{m+k+1}, \tag{2}$$

where α is the target significance level, m is the number of tests, and k is the rank of the p-value in a sorted list. Ranks are assigned sequentially, with 1 given to the smallest p-value, 2 to the next smallest, etc.

The Kruskal-Wallis tests were performed using Python's SciPy *stats* library (Virtanen et al. 2020). In total, 12 x 85 tests were performed, where 12 represents the

number of dimensions and 85 is the number of features (introduced in Chapter 5). In each test, the medians of the relative frequencies of the linguistic features were compared between the *strong/moderate*, *weak*, and *not present* groups. The relative frequencies were extracted from the final dataset presented in Chapter 4. For each test, the Kruskal-Wallis p-value was compared with the Holm-Bonferroni corrected p-value. Given the small dataset, the α in the Holm-Bonferroni correction method was increased to 0.1 to decrease the likelihood of false negatives (Type II errors). Features that had a lower Kruskal-Wallis p-value than the Holm-Bonferroni corrected p-value were considered to have a statistically significant association with the levels of dimensional salience.

6.1.2. *Post hoc* testing

ANOVA is used to determine which linguistic features have a statistically significant association with the dimension, but it does not specify the relationship between the dimension and the feature, e.g., does the use of nouns increase as the text becomes more informative or does the use of relative clause modifiers decrease as the text becomes more subjective? To determine the specific relationships between linguistic features and the levels of dimensional salience, Dunn's test (Dunn 1961) was used to do pairwise comparisons of the medians of the strong/moderate (S/M), weak (W), and non-existent (NE) groups.

The primary goal was to identify monotonic changes in the medians of relative feature frequencies across the dimension, meaning a consistent increase or decrease as the dimension progresses from NE to W to S/M. Thus, the median of a feature's frequency should either:

a. *decrease* along the NE $\rightarrow$ W $\rightarrow$ S/M vector or

b. *increase* along the NE $\rightarrow$ W $\rightarrow$ S/M vector.

A monotonic change is characterized by a consistent increase or decrease in a feature's relative frequency as the dimension transitions from non-existent to strong. In contrast, a non-monotonic change involves a less predictable pattern, where a feature's relative frequency does not exhibit a gradual increase or decrease across the NE, W, and S/M groups. For instance, a non-monotonic relationship is observed when a feature is more frequent in the W group compared to the NE group but less frequent in the S/M group compared to the NE group.

While *post hoc* tests can identify both monotonic and non-monotonic relationships, this analysis focused exclusively on features exhibiting consistent, statistically significant trends across all pairwise comparisons. The primary goal was to identify monotonic changes in the relative frequencies of features across the dimension, meaning a consistent increase or decrease from NE to W to S/M. Based on that, observations for the following distinctions between the groups were considered to be relevant:

- The feature has statistically significant monotonic differences between all three groups, i.e., from S/M to W, from S/M to NE, and from W to NE.
- The feature has a statistically significant difference between the S/M and NE groups and, additionally, one other significant difference between either S/M and W or W and NE groups.
- The feature has statistically significant differences between the S/M and NE groups, and non-significant differences between the pairs of S/M and W, and W and NE groups.

All features that demonstrated a non-monotonic change (e.g., the feature exhibited a statistical significance in differentiating the S/M and W group but not the S/M and NE group) were excluded from further analysis for that particular dimension.

6.2. Results

The following sections will report the results for each dimension separately (a more comprehensive analysis is given in Section 6.3). The results for each dimension will be accompanied by a figure illustrating how the statistically significant features were distributed across the *strong/moderate* (S/M) and *non-existent* (NE) judgements. The *weak* (W) category is excluded from further discussion, as significant differences between NE and S/M judgements likely imply similar trends for the W judgement. Features listed in a white box indicate a **higher frequency** when the dimension is strongly or moderately present, while features in a dark grey box suggest a **lower frequency** when the dimension is strongly or moderately present. The raw data are presented in Appendix C. The table only includes those linguistic features which were considered to be statistically significant (51 out of 85).

The order of results presentation is based on the inter-annotator agreement score for each dimension (Section 4.3). Subjectivity, with the highest agreement, is presented first, while information density, with the lowest agreement, is presented last.

6.2.1. Subjectivity

Subjectivity refers to the presence of personal opinions, feelings, and biases expressed by the author. It can be contrasted with objectivity, which aims to present information without personal interpretation. A total of 36 statistically significant linguistic features were identified (see Figure 7).

When texts become more subjective, the significance of the authors and the participants' centrality becomes more pronounced. This is evident through the heightened utilization of personal voice, finite and supine verb forms, modal verbs, core verbs, indicative mood, adverbs, and first- and second-person verb forms. There is a tendency to use fewer nouns, adjectives, genitive case forms,

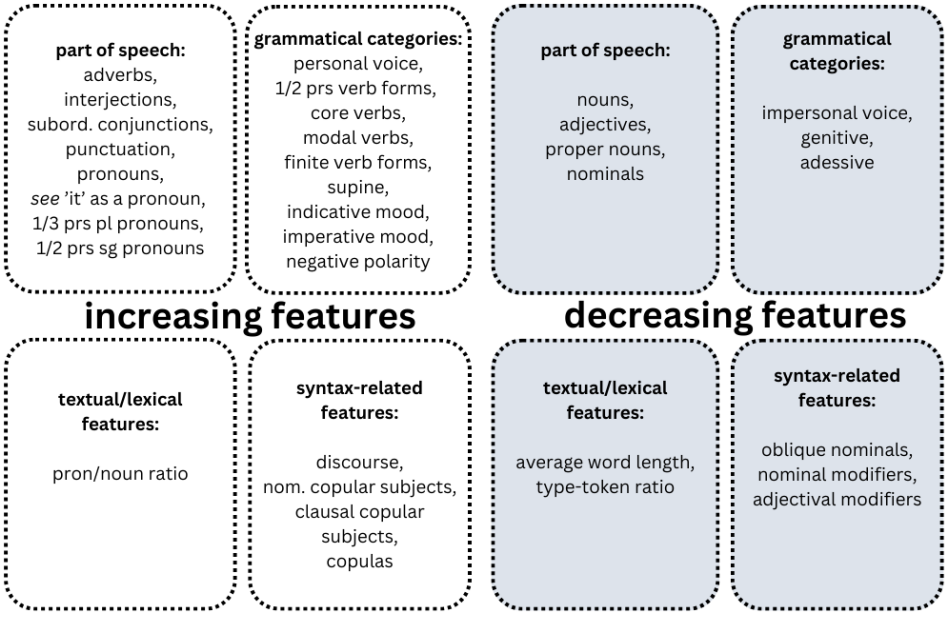

Figure 7: **Salient features for subjectivity.** Features on the left (in white) side of the table are associated with increasing subjectivity (their frequency increases as the text becomes more subjective). Features on the right (in grey) side of the table are associated with decreasing subjectivity (their frequency decreases as the text becomes more subjective).

and modifiers, both nominal and adjectival, which contribute to the formation of noun phrases. Moreover, using less adjectival modifiers is highly unique for subjectivity. Interestingly, the use of first singular pronouns decreases as subjectivity increases. This suggests that expressing opinions and biases is done subtly without making the author overtly present. The use of oblique nominals also decreases, which could be explained by the fact that oblique nominals are nominals functioning as non-core arguments which attach to a verb. Furthermore, such texts exhibit a reduced presence of proper nouns.

Notable features for subjectivity also include using more pronouns, which consequently causes a higher pronoun/noun ratio. Interestingly, the first- and third-person plural (*we*, *they*) and the first- and second-person singular pronouns (*I*, *you*) were considered statistically significant, but the second-person plural and third-person singular pronouns were considered not to be significant. These findings suggest a theme of community or alliance, emphasizing a collective identity of "you against them" or "we against them", reflecting group dynamics and solidarity among individuals.

Subjective texts can be characterized by using more subordinate conjunctions, discourse particles, interjections, punctuation, and shorter words. Syntactically, copular sentences, nominal and clausal copular subjects contribute texts to being more subjective, jointly suggesting that subjectivity prefers to use simpler sen-

tence structures. In addition, clausal copular subjects are distinctive features that are exclusively associated with subjectivity. A noteworthy aspect involves the presence of a demonstrative pronoun, such as *see* it as a pronoun, which denotes a specific situation, attribute, or action that is evident from either the speech or the context.

In light of the aforementioned observations, it can be argued that in subjective texts, there is often a presence of individuals in the roles of a speaker and a listener within the text. These individuals employ first-, second-, and third-person pronouns and verb forms, and fewer proper nouns. The statistically significant features collectively indicate that the role of 'listener' is characterized by the use of adverbs, subordinating conjunctions, interjections, and indicative speech mode. The presence of core verbs, active voice, copular constructions, a reduced number of nouns and building blocks for noun phrases (such as genitive case forms, adjectives, nominal/adjectival modifiers) and shorter words indicates the presence of simpler clauses that could be similar in structure to written spoken language. This could be intended to facilitate comprehension and accessibility for the intended audience.

6.2.2. Affectivity

Affectivity refers to personal experience, feelings, and reactions expressed in a text. A total of 34 statistically significant linguistic features were identified (see Figure 8).

As the text becomes more affective, the significance of verbality and the articulation of authorship or participation increases. The expression of verbality and personal experience can be achieved through the use of a minimalist approach, whereby fewer proper nouns and an impersonal voice are employed. A variety of verbal and syntactic features, as well as parts of speech, may be significant in this regard. These include increased use of interjections, adverbs, different grammatical categories of verbs (finite verb forms, supine, core verbs, etc.), copulas and discourse particles, and pronouns. It is noteworthy that all pronouns, except second-person plural pronouns, demonstrated statistical significance in differentiating texts with a moderate or strong salience of affectivity from those where affectivity was deemed absent. Moreover, third-person singular pronouns were found to be exclusively associated with affectivity. Concerning pronouns, it is reasonable to anticipate a higher pronoun/noun ratio when the use of pronouns increases and the use of nouns decreases. In the context of affectivity, a higher pronoun/noun ratio indicates a greater emphasis on personal involvement and emotional expression.

Notably, determiners occur with greater frequency in affective texts, potentially serving to link the noun phrase with the referents from the context. This indicates that the author or writer assumes a certain degree of familiarity with the content being discussed on the part of the recipients, thereby suggesting a

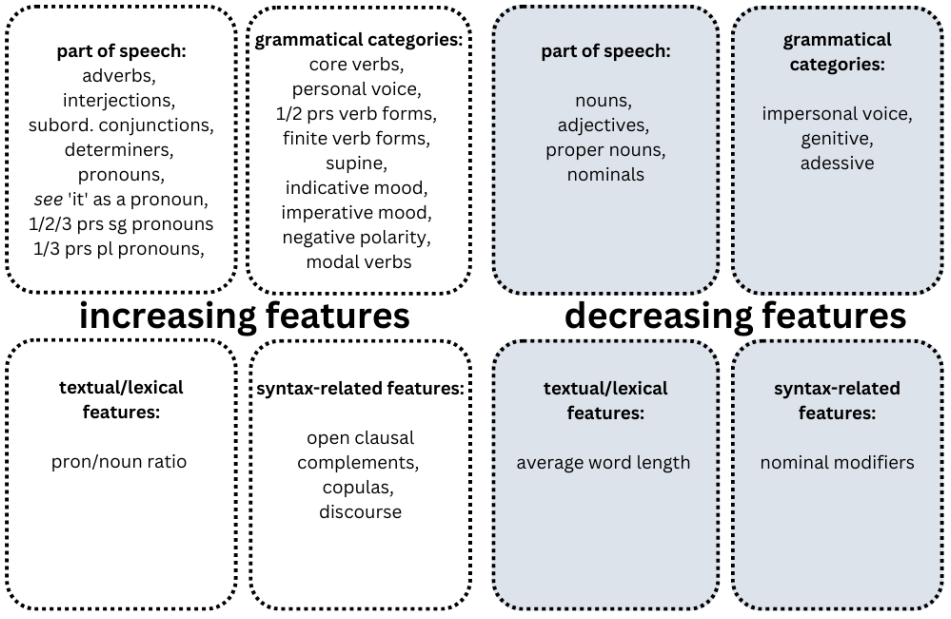

Figure 8: **Salient features for affectivity.** Features on the left (in white) side of the table are associated with increasing affectivity (their frequency increases as the text becomes more affective). Features on the right (in grey) side of the table are associated with decreasing affectivity (their frequency decreases as the text becomes more affective).

shared understanding between the author and the readers. This assumption of shared knowledge can be viewed as a strategic linguistic choice to enhance the communication and comprehension of the affective message conveyed in the text. Additionally, the use of imperative mood, negative polarity, adessive case and indicative mood are distinctive features that are also present in affectivity. The functionality of some features, such as negative polarity, adessive case, and open clausal complements, is complex and requires further investigation.

Texts showing a moderate or strong salience of affectivity tend to utilize shorter words and exhibit a reduced number of linguistic elements associated with nouns and noun phrases. These include adjectives, genitive forms, nominal modifiers and oblique nominals.

In consideration of the aforementioned evidence, affective texts show heightened verbality through personal engagement, and emotional expression through the utilization of linguistic elements such as adverbs, pronouns, and distinctive verbal and syntactical structures. The deployment of determiners in affective texts suggests a mutual comprehension between the author and the readers, thereby facilitating communication.

6.2.3. Formality

Formality can be defined as a style of written communication that aims to maintain a professional and objective tone. It is characterized by adherence to established norms and the avoidance of colloquialisms, slang and overly casual language. A total of 26 statistically significant features emerged from the analysis (see Figure 9).

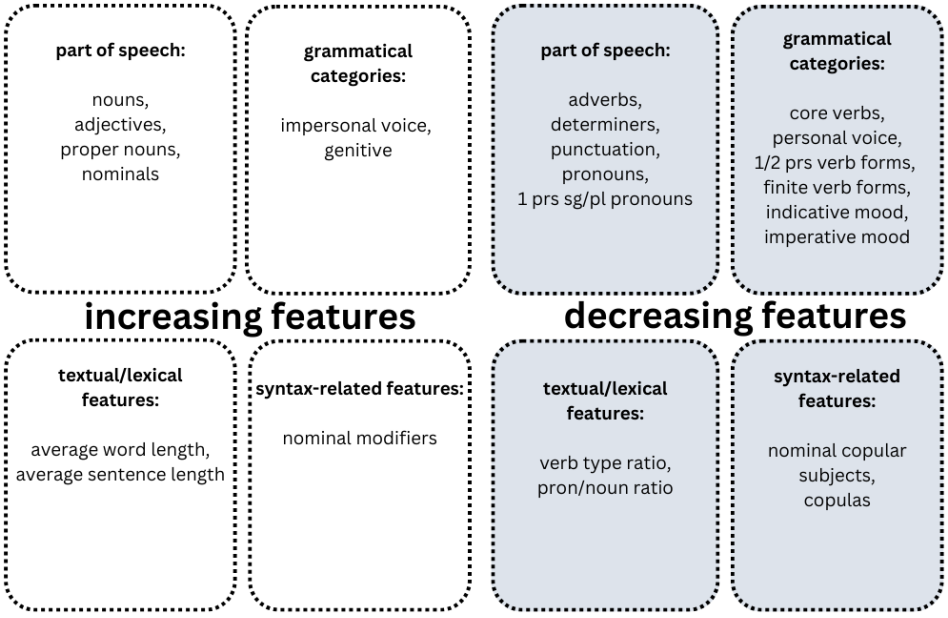

Figure 9: **Salient features for formality.** Features on the left (in white) side of the table are associated with increasing formality (their frequency increases as the text becomes more formal). Features on the right (in grey) side of the table are associated with decreasing formality (their frequency is lower as the text becomes more formal).

Texts that exhibit a moderate or strong degree of formality tend to utilize a greater number of nouns, nominals in general, proper nouns, adjectives, the genitive case and nominal modifiers. Texts that are more formal utilize longer words and complex sentence structures, adopt a more impersonal tone, and exhibit a limited range of different verb forms. It is noteworthy that longer sentences were exclusively associated with formality. The features in the right panel (in dark grey) demonstrate that formal texts tend to avoid constructing simple sentence structures (e.g., reduced use of copulas, and punctuation) and to incorporate personal involvement through pronouns, particularly first-person singular and plural pronouns, determiners, or different verb forms.

These features mentioned above show that formality is related to nominal language by emphasizing the subject over personal involvement or subjective experience, which contributes to a more detached and objective tone. The results show

that more formal texts use longer words and sentences to convey information precisely and clearly, making the content authoritative, credible, and appropriate for professional contexts such as administration or academia.

6.2.4. Spontaneity

Spontaneity in texts reflects immediate thoughts or emotions created without prior planning. Although spontaneity is most evident in spoken discourse, where words cannot be retracted but can be corrected post-speaking, it can manifest itself also in written forms (online comments, forums, online conversations). A total of 25 statistically important features emerged (see Figure 10).

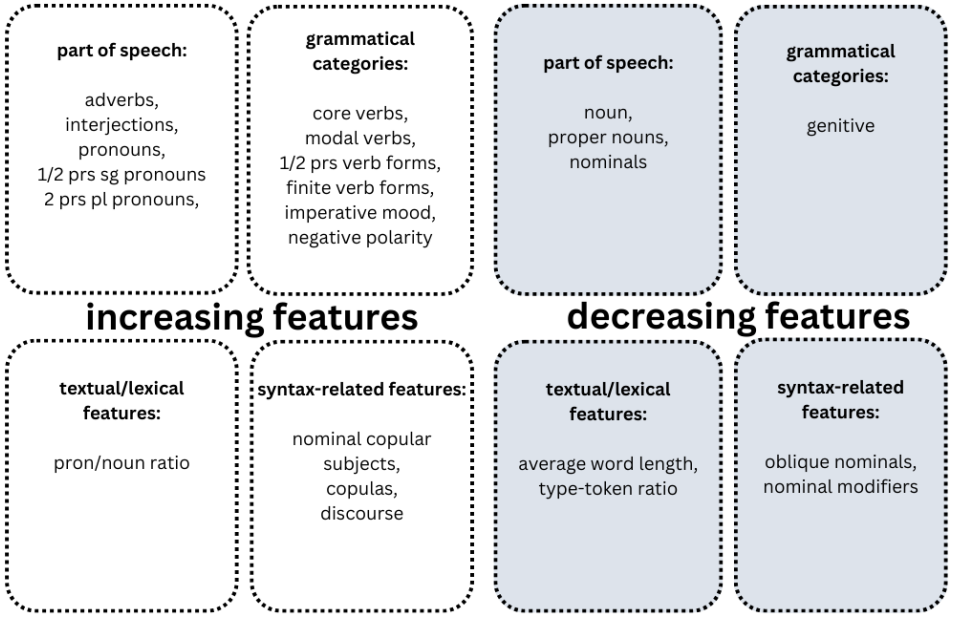

Figure 10: **Salient features for spontaneity.** Features on the left (in white) side of the table are associated with increasing spontaneity (their frequency increases as the text becomes more spontaneous). Features on the right (in grey) side of the table are associated with decreasing spontaneity (their frequency decreases as the text becomes more spontaneous).

As texts become increasingly spontaneous, they become more conversational in nature, utilizing a range of verb-related features, including core and modal verbs, finite verb forms, and parts of speech such as adverbs and interjections. This is in contrast to less frequent features, such as nouns, genitive case forms, and nominal modifiers.

Furthermore, the centrality of the participant is also emphasized by the use of pronouns, first- and second-person singular, and second-person singular plural pronouns, and first- and second-person verb forms. These features indicate that the discussion revolves around the participants. This assertion can be substanti-

ated by two factors: firstly, the use of proper nouns is less prevalent, and secondly, the use of third-person pronouns and verb forms is not statistically significant. As with subjectivity and affectivity, the use of the imperative mood and negative polarity is considered a significant indicator of spontaneity.

6.2.5. Instructability

Instructability embodies texts that provide the reader with instructions on how to perform certain activities or describe the stages of a process.

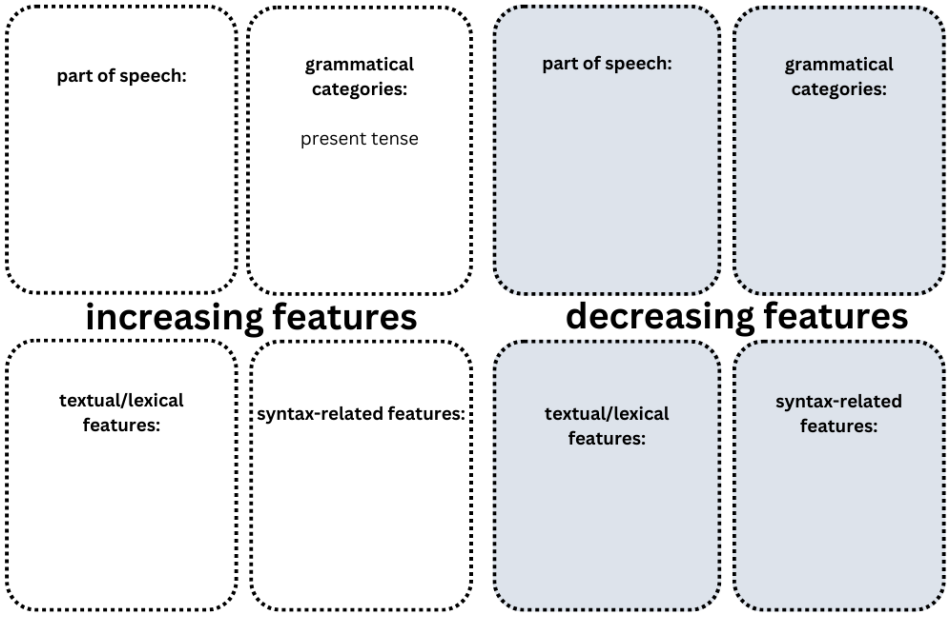

Figure 11: **Salient features for instructability.** Features on the left (in white) side of the table are associated with increasing instructability (their frequency increases as the text becomes more instructable). Features on the right (in grey) side of the table are associated with decreasing instructability (their frequency decreases as the text becomes more instructable).

For instructability, the only notable feature was the present tense (see Figure 11). The use of the present tense can indicate immediacy, as it denotes actions that are occurring in the present moment. Alternatively, it can be employed to convey timeless truths or to present direct commands or step-by-step procedures, thereby facilitating comprehension and adherence. Relying on a single feature to interpret the functionality of a dimension can be misleading, thus instructability needs some further research.

6.2.6. Interactivity

Interactivity is communicative and reactive, participants are actively engaging with each other. It involves creating an interactive experience that encourages

participation, allows feedback, and allows for a dynamic exchange between the participants. Interactivity can exist in various forms, from traditional printed media to online mediums. A total of 22 statistically significant features were identified, see Figure 12.

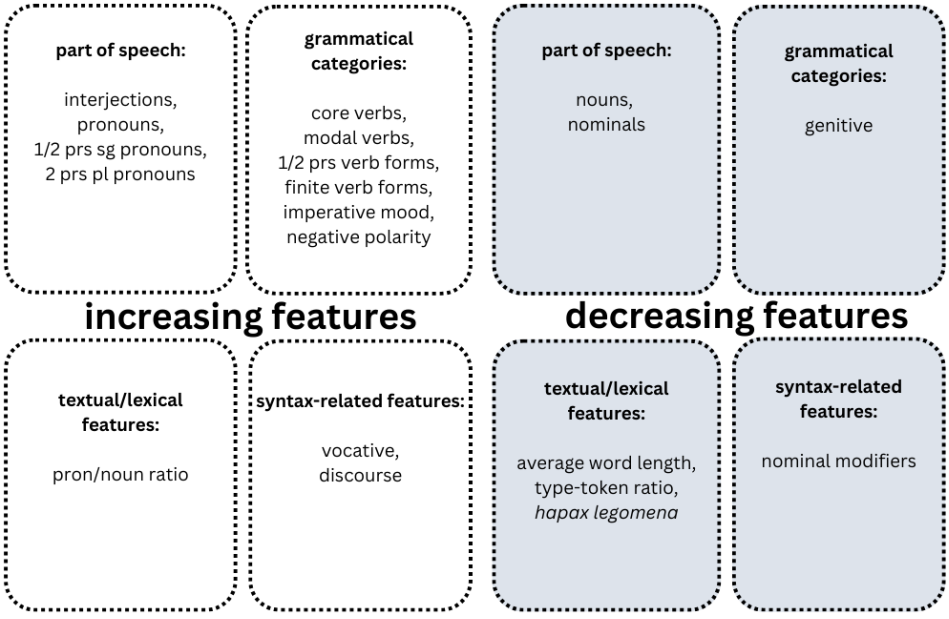

Figure 12: **Salient features for interactivity.** Features on the left (in white) side of the table are associated with increasing interactivity (their frequency increases as the text becomes more interactive). Features on the right (in grey) side of the table are associated with decreasing interactivity (their frequency decreases as the text becomes more interactive).

The results for interactivity are in close alignment with those for subjectivity, affectivity, and spontaneity. The significant features of interactivity indicate that interactive texts are dynamic and expressive. Pronouns can help to establish a reference within a discourse, whereas interjections add emotional or emphatic elements. As interactivity is concerned with dialogue, communication, and active participation, the features of first-person singular, second-person singular, and plural pronouns, first- and second-person verb forms were to be expected. Furthermore, active participation is also expressed by other verb-related features, including finite verb forms, modal verbs, and core verbs. Similarly to subjectivity, affectivity, and spontaneity, moderately or strongly interactive texts employ the imperative mood, negative polarity, shorter words, and a smaller vocabulary (except affectivity, where the vocabulary is bigger).

The results show that *hapax legomena* is a unique feature for interactivity. While shorter texts tend to have higher occurrences of *hapax legomena*, a lower occurrence in the given dataset might indicate a more limited lexical diversity or a specialized language domain. It is also worth noting that the vocative, a linguistic

device that directly addresses or calls upon the intended recipient, is a distinctive feature of interactivity. The vocative serves to establish a direct connection, convey emotions, or emphasize a sense of immediacy within the interaction. However, due to the limited size of the dataset, the generalizability of the findings may be constrained. Further research with a larger dataset is therefore necessary.

6.2.7. Impersonality

Impersonality is characteristic of texts where the focus is on the action or the recipient and the information of the doer (agent) of the action is suppressed. A total of 25 statistically important features emerged, see Figure 13.

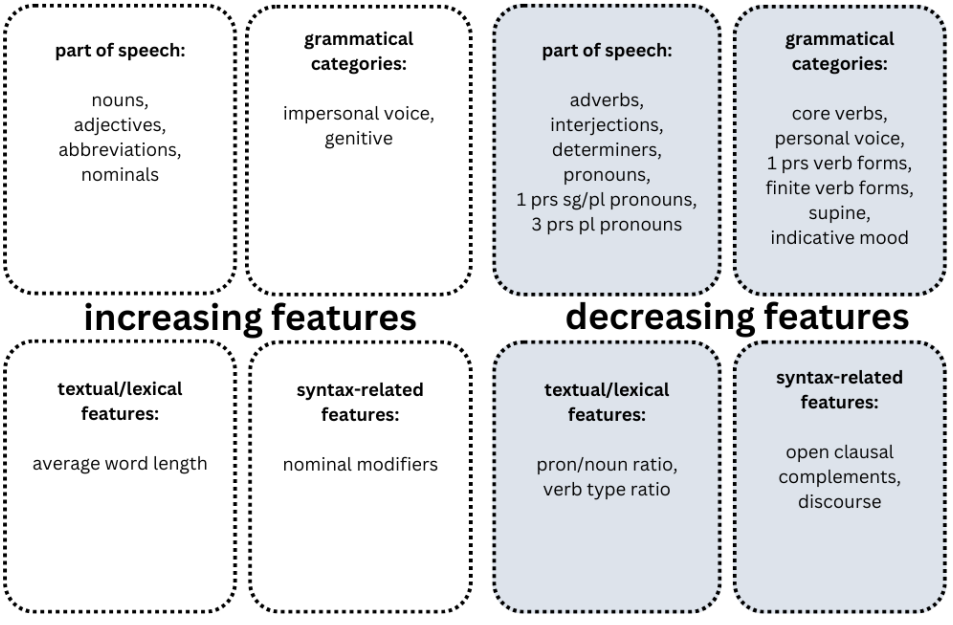

Figure 13: **Salient features for impersonality.** Features on the left (in white) side of the table are associated with increasing impersonality (their frequency increases as the text becomes more impersonal). Features on the right (in grey) side of the table are associated with decreasing impersonality (their frequency decreases as the text becomes more impersonal).

As with formality, impersonality is expressed through a variety of linguistic means, including nouns, adjectives, the genitive case, and nominal modifiers. Texts that were considered to be more impersonal exhibited a greater use of longer words, an impersonal voice, and a reduction in the number of verb forms. Impersonal texts may avoid personal involvement through the use of fewer finite verb forms and pronouns, particularly first-person verb forms, the use of *see* 'it' as a pronoun, first-person singular and first-person and third-person plural pronouns. This may be a deliberate strategy to avoid using pronouns that directly involve the writer or reader, thereby maintaining objectivity and detachment.

In comparison to formality, more impersonal texts tend to exhibit a greater prevalence of abbreviations and a lesser use of interjections, supine forms, open clausal complements, and discourse particles. The use of more abbreviations streamlines the text which facilitates brevity, while the limited use of interjections, supine forms, open clausal complements, and discourse particles contributes to a more straightforward and impersonal tone, in keeping with the text's aim of presenting information in a detached manner.

6.2.8. Temporality

Temporality refers to the coherence and sequential order of events or ideas within a text. It demonstrates how the author or speaker structures his discourse in a chronological or logical progression. For *temporality*, five statistically important features emerged, see Figure 14.

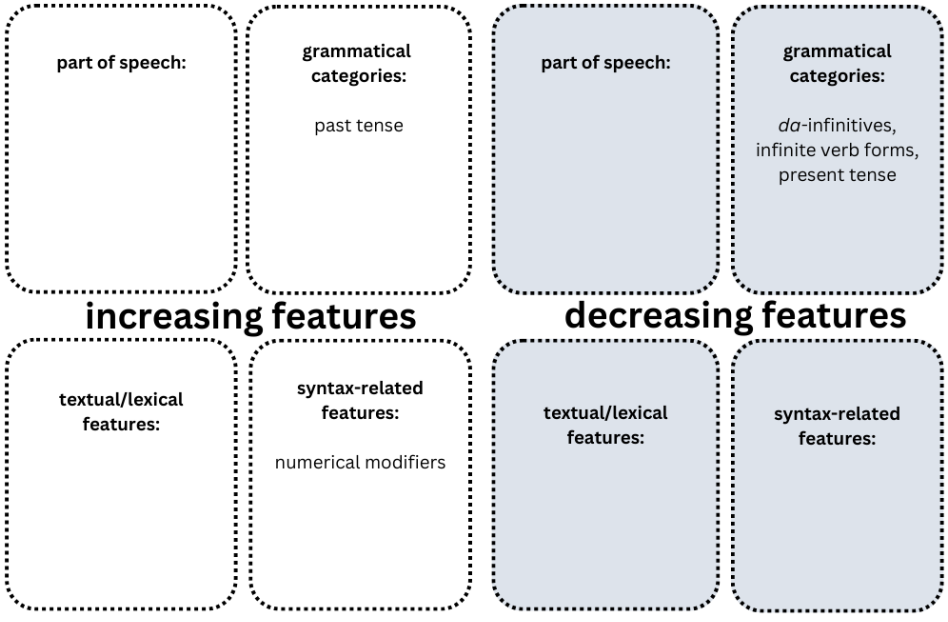

Figure 14: **Salient features for temporality.** Features on the left (in white) side of the table are associated with increasing temporality (their frequency increases as the text emphasizes the significance of time). Features on the right (in grey) side of the table are associated with decreasing temporality (their frequency decreases as the text emphasizes the significance of time.).

Temporality is expressed through the use of more past tense and numerical modifiers. This finding suggests the importance of discussing past events and including quantitative details to convey temporal information. Furthermore, numerical modifiers were a distinctive feature only for temporality. The results also showed that texts with a moderate or strong salience of temporality used fewer *da*-infinitives and present tense, which may indicate a need to explicitly outline the

temporal context of actions. The utilization of fewer infinitive verb forms is also a characteristic exclusive to texts where temporality is considered to be moderately or strongly present.

6.2.9. Complexity

Complexity in a text refers to the level of difficulty or intricacy of its structure and language. It requires additional effort from the reader. For *complexity*, four statistically important features emerged (see Figure 15).

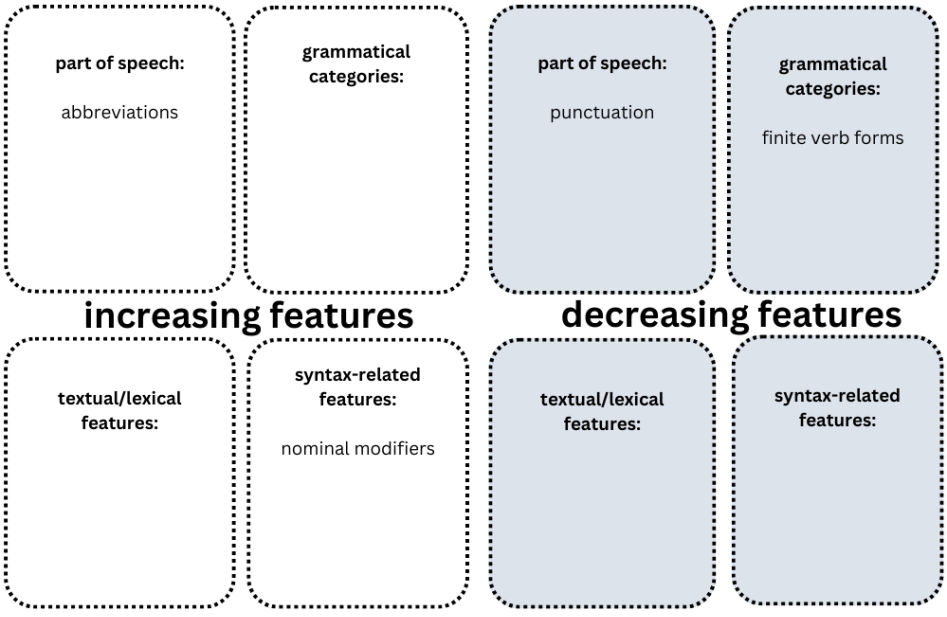

Figure 15: **Salient features for complexity.** Features on the left (in white) side of the table are associated with increasing complexity (their frequency increases as the text becomes more complex). Features on the right (in grey) side of the table are associated with decreasing complexity (their frequency decreases as the text becomes more complex).

As texts become more complex, the use of abbreviations and nominal modifiers tends to increase, while the use of finite verb forms and punctuation tends to decrease. These findings may indicate that as complexity increases, the language becomes more condensed and streamlined, with a direct writing style that focuses on conveying key information with minimal linguistic elaboration.

6.2.10. Argumentativity

Argumentativity is present in texts where the author(s) introduces their viewpoint about a topic or phenomenon. The scientific literature can be considered argumentative, as opposed to descriptive or narrative discourse. For *argumentativity*, five statistically important features emerged, see Figure 16.

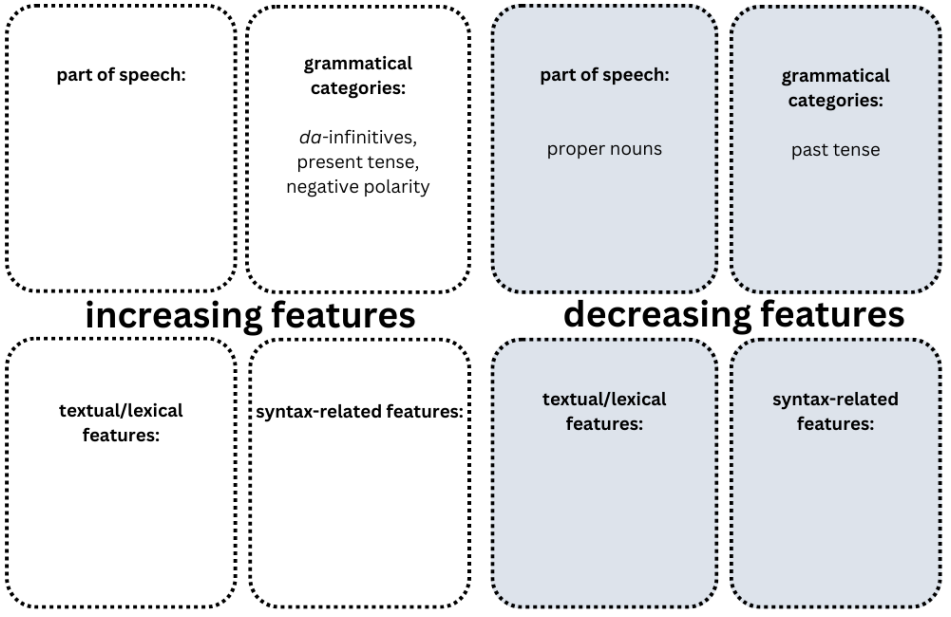

Figure 16: **Salient features for argumentativity.** Features on the left (in white) side of the table are associated with increasing argumentativity (their frequency increases as the text becomes more argumentative). Features on the right (in grey) side of the table are associated with decreasing argumentativity (their frequency decreases as the text becomes more argumentative).

As texts become more argumentative, the more frequent use of *da*-infinitives, the present tense and negative polarity becomes more prevalent, while the use of proper nouns and the past tense becomes less frequent. The results may indicate that the use of the present tense in argumentative texts conveys immediacy by emphasizing the current situation and the validity of the points being made. The use of negative polarity can be employed to challenge opposing viewpoints, refute claims, or introduce critical perspectives within arguments. Furthermore, the use of fewer proper nouns may assist in maintaining focus on the central arguments and ideas, rather than on specific individuals, places, or entities. These linguistic choices can be employed to enhance the persuasive and confrontational aspects of argumentative discourse, facilitating a more forceful and decisive presentation of ideas.

6.2.11. Abstractness

Abstractness refers to the level of generalization and conceptualization of a given text. It describes the degree to which some information is removed from specific details and concrete realities, instead focusing on broader concepts and ideas. Abstractness can be found in various domains, such as scientific research papers, philosophical essays, academic articles, and even certain forms of literature.

Figure 17 shows that the sole statistically significant feature associated with

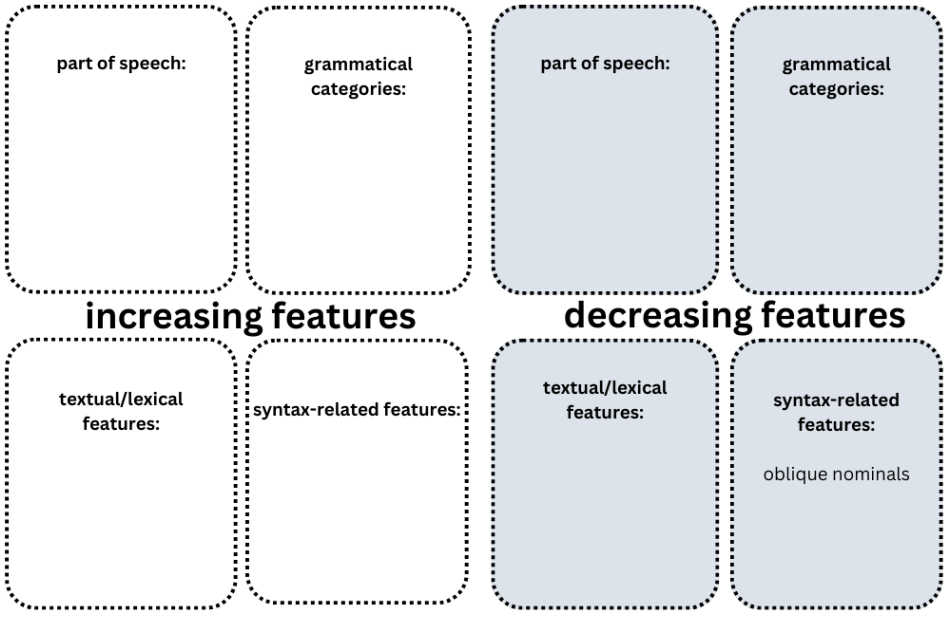

Figure 17: **Salient features for abstractness.** Features on the left (in white) side of the table are associated with increasing abstractness (their frequency increases as the text becomes more abstract). Features on the right (in grey) side of the table are associated with decreasing abstractness (their frequency decreases as the text becomes more abstract).

abstractness is the use of fewer oblique nominals. Oblique nominals are noun phrases that typically function as complements or adjuncts in a sentence, providing additional information through the addition of layers of meaning and nuance to the text. As with instructability, relying on a single feature to interpret the functionality of a dimension can be misleading. Therefore, further research is required to gain more understanding of abstractness.

6.2.12. Information density

Information density is primarily focused on providing factual details, explanations, or analysis of a specific topic. They aim to educate or inform the reader about a subject rather than to entertain or persuade. In addition to encyclopedic entries, information density can be high in law texts, science and journalism, and information providers, such as pharmaceutical package leaflets, documents, etc. A total of nine statistically important features emerged, see Figure 18.

More informative texts showed a preference for nouns and longer words. They also used fewer pronouns to express information. In particular, first-person singular and plural pronouns, which can detract from the objectivity and formal tone necessary to convey informative content, were used less frequently. As a result, there is a lower pronoun/noun ratio, emphasizing factual content over personal or

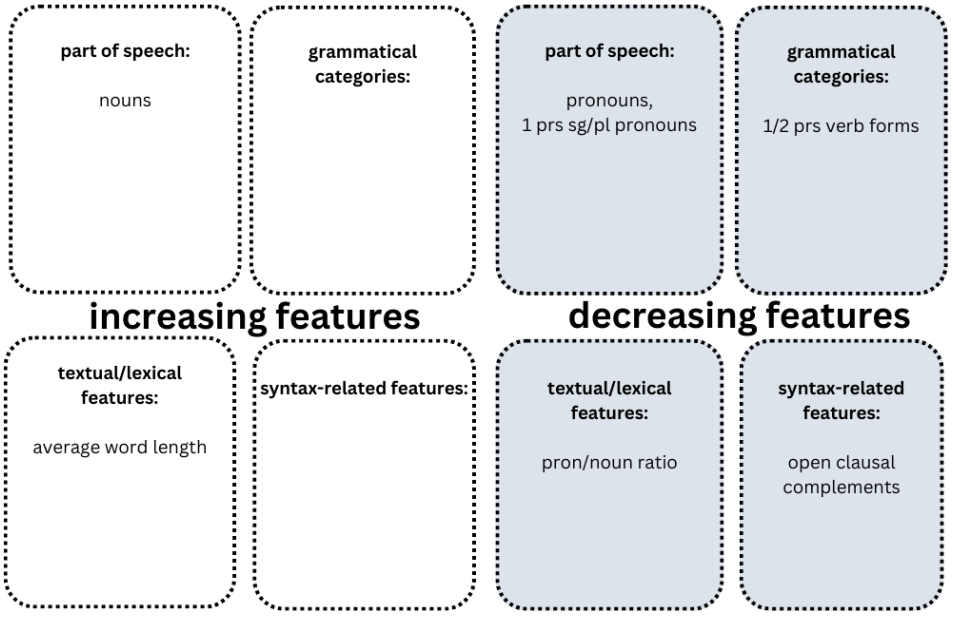

Figure 18: **Salient features for information density.** Features on the left (in white) side of the table are associated with increasing information density (their frequency increases as the text becomes more informative). Features on the right (in grey) side of the table are associated with decreasing information density (their frequency decreases as the text becomes more informative).

subjective elements. In addition, the use of first- and second-person verb forms is minimized in favour of a more neutral and authoritative tone that prioritizes the content itself, contributing to the overall informativeness of the text.

Interestingly, open clausal complements were regarded as negative features for moderately or strongly informative texts. By minimizing the use of open clausal complements, the narrative can be made more concise and focused, avoiding unnecessary complexity that could potentially detract from the informative content. Thus, in informative texts, the emphasis on precision and coherence may lead to a reduced frequency of open clausal complements to ensure a clearer communication strategy.

6.3. A comparative analysis between dimensions

The objective of the RQ2 was to ascertain to what extent, if any, the dimensions proposed by the Dimensional Text Model (DTM) differ from one another in terms of their linguistic profiles. Although the DTM posits that each proposed dimension is unique and distinct, Section 4.4 revealed that the dimensions do not exist independently. In contrast, the findings demonstrate that the dimensions proposed in the DTM can be categorized into two macro dimensions: the *unplanned-verbal* and *planned-nominal* dimensions. The remaining dimensions, which do not fit

into the two macro dimensions, can be categorized as the *complementary dimensions*.

The **unplanned-verbal** dimensions include dimensions that represent spontaneous and unscripted language use, reflecting the dynamic nature of spoken language, such as subjectivity, affectivity, spontaneity, and interactivity. This type of language can often be observed in everyday conversations, informal interactions, and informal writing. The **planned-nominal** dimensions include dimensions representing language that adhere to prescribed grammar and vocabulary, such as formality, impersonality, and information density. It is typically found in formal contexts such as academic, newsrooms, and judiciary settings. It can be argued that the planned-nominal dimensions are more standardized by showing less variation, while the unplanned-verbal dimensions show greater variation in terms of linguistic features. Furthermore, the results demonstrate that the planned-nominal and unplanned-verbal dimensions are accompanied by a set of **complementary dimensions**. These complementary dimensions include instructability, temporality, complexity, argumentativity, and abstractness.

The following sections will examine the similarities and differences between the unplanned-verbal, planned-nominal, and complementary dimensions in terms of linguistic features. It should be noted that the aim of this dissertation is not to analyze the dimensions as macro-dimensions per se. Rather, the objective is to determine whether and how the linguistic profiles between the dimensions differ. While the dimensions are assumed to be distinct, they are not independent. To gain a better overview of the proposed dimensions, viewing the variation among the dimensions with similar communicative functions enables a clearer distinction of the overall results. The conclusions of this analysis are speculative. To definitively establish the causal relationship between a linguistic feature and a dimension, it would be necessary to conduct qualitative studies involving manual analysis.

6.3.1. Unplanned-verbal dimensions

The unplanned-verbal dimensions include dimensions representing spontaneous and unscripted language use, reflecting the dynamic nature of spoken language, which includes subjectivity, affectivity, spontaneity, and interactivity. This section examines the linguistic variation between these dimensions.

As described in Section 3.2, affectivity refers to the domain of emotions and expressive responses communicated through language. It can evoke a variety of emotional responses, adding depth and richness to our interactions by tapping into feelings and sentiments. Subjectivity, on the other hand, incorporates individual viewpoints, emotions, and biases that shape how we perceive and communicate information. Subjectivity recognizes the inherent personal perspectives and experiences that influence our understanding. Spontaneity characterizes the development of narratives in a natural, unpremeditated way, without extensive planning or organization. Spontaneity in language captures the raw, immediate flow

Table 7: The distribution of features between the unplanned-verbal dimensions.

AFF	INTER	SPONT	SUBJ	INCREASING USAGE	DECREASING USAGE
aff	inter	spont	subj	- intj, pron, pron/noun ratio (higher), 1st/2nd prs sg pron, - core verbs, 1st/2nd prs verb, finite verb, imp mood, negative polarity, modal verb - discourse	- noun, nominal - avg word length (shorter) - genitive case - nominal mod.
aff		spont	subj	- adverb - copula	- proper noun
	inter	spont	subj		- TTR
	inter	spont		- 2nd prs pl pron	
		spont	subj	- nsubj cop.	-oblique nominal
aff			subj	- sub. conjunction, 1st/3rd prs pl pron - personal voice, supine, indicative mood - *see* as pron	- imp voice - adessive - adjective
UNIQUE FEATURES					
			subj	- punct - csubj cop	- adjectival mod.
aff				- determiner, 3rd prs sg pron - open clausal comp.	
	inter			- voc	- *hapax legomena*

of thoughts and ideas and reflects authentic expression. Interactivity enhances the level of participation and communication between individuals involved in the exchange of information. These dimensions intertwine to shape a language use characterized by dynamic, action-oriented language, which varies in its degrees of personal involvement, emotion, and interactive dialogue, creating a complex network of written communication. The general results are presented in Table 7 which summarizes how the features were distributed across the dimensions representing the unplanned-verbal dimensions. It includes two key aspects: how dimensions cluster together based on shared features and the monotonic change of each feature (i.e., increase or decrease) along the level of the salience of the dimension.

The results suggest that strongly or moderately affective, subjective, spontaneous, and interactive texts are characterized by increasing importance of the centrality of the author and the addressee. This can be observed in the Text Sample 1 through the increased use of pronouns (blue text), verb-related features such as first- and second-person verb forms, and finite verb forms in general (**bold** text), and the use of interjections and discourse particles (red text). Also, the more affective, subjective, spontaneous and interactive the texts, the more negative polarity is used, such as *ära*, *ei* 'no' (**underlined bold text**). Furthermore, statistically significant features associated with these dimensions exemplify the characteristics of written-spoken language where linguistic efficiency plays an essential role, e.g., using more core verbs (such as 'to go' in *läksin* 'went', and 'to be' in *on* 'is'), which have general meanings and can be used in a wide range of contexts.

Affective, subjective, spontaneous, and interactive language tends to use fewer common nouns, nominals, genitives, and nominal modifiers that provide additional information about the noun phrases. Also, words tend to be shorter on average, i.e., 4.8 characters per word compared to 5.8 characters per word for texts where affectivity, subjectivity, spontaneity and interactivity were considered not present (see Appendix C).

Text Sample 1.
(Original)
Kui ma aga **kõndisin** & **kõndisin** , siis **mõtlesin** ma ikka ja jälle Kristoferi sõnade peale , mida ta mulle Joonasest **rääkis**. **Jõudes** siis Kätu palatisse , **nägin** ma temajuures ....ANDREST . Ma **läksin** närvi ja ma **mõtlesin** et mida teha. Kuid ma siiski siis **küsisin** vihaselt : „ Midahe*kki sa siin **teed** mees , ? " . Kätu **ütles** mulle suhteliselt tasa häälega temaeest : „ **Ära ärritu** , kõik **on** korras. ta **tuli** mind ainult **vaatama** ja **tahtis** ka sind näha. " . Mu silmad **läksid** suureks ja ma **küsisin** : „ Misasja ? , pmst ta **on** Janaga mestis ja siis ta **tuli** sind **vaatama** v ? ja oi imet , **tahtis** mind ka veel näha. omg , see **on** kuidagi kahtlane , kas sa **ei arva**
(source: *www_mul je_ee.ela*_177932)

(Translation)
As I continued to walk & walk, I kept pondering Christopher's words about Joonas. Upon arriving in Kätu's room, I saw ...ANDRES with her. I became nervous and wondered what to do. Nevertheless, I asked angrily, "What the heck are you doing here, man?" Kätu, speaking softly, reassured me, "Don't get upset, everything is fine. He just came to see me and wanted to see you too." My eyes widened, and I asked, "What? Basically, he's with Jana, and then he came to see you and wanted to see me as well? Oh my god, this is somehow suspicious, don't you think?"

However, it is important to bear in mind that the dataset includes texts from different websites. This raises the possibility that the frequent occurrence of imperative mood could also be partly due to preprocessing errors. For example, within the texts showing strong or moderate subjectivity, one can observe a considerable number of texts collected from chat rooms where the 'REPLY' button is placed at the end of each user's comment (see *Vasta* 'Reply' in Text Sample 2). Pre-processing errors could inflate the perceived association between imperative mood and subjectivity in this dataset.

Text Sample 2.
(Original)
??? 18.12.2010 13:21 - Mis siin keerulist on pange nupuga varustatud valgustusriba ülekäiguraja märgi külge , kui inimene tahab üle minna vajutab nuppu, märk hakkab vilkuma sõidukijuhile on paremini märgatav inimese soov üle tee minna, olukord on üheselt mõistetav ja jalakäija saab turvalisemalt üle tee **Vasta**
.:.: 18.12.2010 13:25 - astub, kõnnib, seisab need on kõik erinevad asjad. juht ei tea kunagi mida jalakäija plaanib teha **Vasta**
(source: *www_delfi_ee.ela_*361703)

(Translation)
??? 18.12.2010 13:21 - What's complicated about this is attach a light strip with a button to the pedestrian crossing sign , when a person wants to cross, they press the button, the sign starts flashing, making the pedestrian's intention to cross more noticeable to drivers, the situation is clear and the pedestrian can cross the road more safely Reply
.:.: 18.12.2010 13:25 - steps, walks, stands these are all different things. the driver never knows what the pedestrian plans to do Reply

---

When interactivity is excluded from the unplanned-verbal dimensions, new linguistic features emerge that contribute to texts being more affective, subjective, and spontaneous. These include the use of more adverbs and copula sentences and fewer proper nouns. Affective, spontaneous, and subjective language together evoke emotions and feelings that are created without extensive planning and reflect the author's personal opinions, perspectives, and biases (see Text Sample 3). Thus, affective, spontaneous, and subjective language seems to favour simpler syntactic constructs, such as copulas linking subjects to predicative expressions, typically *olema* 'to be' (underlined text). The use of adverbs (**bold text**) also increases as the language becomes more affective, spontaneous, and subjective, and since different verb categories were considered significant for the unplanned-verbal dimensions and adverbs modify verbs, it is natural to have a higher rate of adverbs present in a text.

Text Sample 3.
(Original)
Lõpuks kui me hakkasime **eraldi** vahetama (kolmanda klassi algul)poistega(**algul** tüdrukud ja **siis** poisid) hakkasid nad tütrukute vahetuse ajal mind igat **moodi** solvama,minu ees ülbitsema.Kui ma ütlesin ,et räägin õpetajale ütlesid nad ,et kes mind usub kui nemad väidavad vastu pidist.**Nii** see käib **siiamaani** kuigi mul on **nüüd** kaks sõpra(tüdrukud) üks teine põlaalune ja mu pinginaaber.Nad on normaalsed.**Vahepeal** proovisin **ka** selle klassiõe meele järgi olla ,aga see ei tulnud mul **hästi** välja ja ta oli mind **juba tõsiselt** vihkama hakkand. Palun aidake!Ma ei tea mida teha! Tänudega Katri!
(source: *www_lapsemure_ee.ela_*140896)

(Translation)
Finally, when we started changing separately (at the beginning of third grade) with boys (girls first and then boys), they started insulting me in every way and being arrogant in front of me during the girls' rotation. When I said I would tell the teacher, they said, "Who will believe me when they argue against me?" This still continues, although I now have two friends (girls), one other acquaintance, and my desk neighbor. They are normal. At one point, I also tried to please this classmate, but it didn't work out well, and she had already started seriously hating me. Please help! I don't know what to do! Thank you Katri!

Interactivity and spontaneity are associated, among other features, with the use of more second-person plural pronouns (**bold text** in Text Sample 4). In Estonian, second-person plural pronouns are often used as a form of politeness, contributing to a more inclusive and engaging tone. This suggests that spontaneous and interactive texts may prioritize politeness and inclusion.

Text Sample 4.
(Original)
au.Basil 09.06.2010 10:15 - 'king of herbs' tahab saada päikest. Siis kasvab mühinal. Aga jah lõikuda üldse ei tohiks isegi kääridega kuna parem kraam jääb metallile ja osa maitsest saamata. Kuivatada ka pole mõtet. Vasta
09.06.2010 11:46 - Tahtsin just sama teadjatelt küsida, et mitme kiloga peaks piirduma, et raskemeelsusest vabaneda. Vasta
xxx 27.03.2012 00:06 - Oleneb, mida **te** raskemeelsuse all mõtlete. Võib olla aitaks õhtune tee basiilikust. Oleneb inimesest, mõnel juba öösel, mõnel järgmisel päeval on tihe jooks kempsu vahet. Kõik üleliigne või ka vajalik vesi tahab teist välja tulla.
(source: *www_aialeht_ee.ela_*457123)

(Translation)
au.Basil 09.06.2010 10:15 - The 'king of herbs' wants to get sunlight. Then it grows rapidly. But yes, it should not be cut at all, not even with scissors, as the better stuff stays on the metal, and part of the flavor is lost. There's no point in drying it either. Reply
09.06.2010 11:46 - I was just about to ask the same from the experts, with how many kilos one should limit themselves to get rid of the heaviness. Reply
xxx 27.03.2012 00:06 - It depends on what you mean by heaviness. Maybe an evening tea with basil would help. It varies from person to person, some may have a frequent need to visit the restroom at night, while for others, it's the next day. All excess or necessary water wants to come out from you.

Interestingly, a line can be drawn between the affectivity-interactivity and spontaneity-subjectivity pairs. For example, spontaneous and subjective texts used more nominal copular subjects (**bold text**) and less oblique nominals (as seen in Text Sample 5). Copular sentences, known for their simpler sentence structure, are used to describe or rename the subject, such as in 'The cat is black' and 'I think she is great'. This suggests that under time constraints, simpler sentence constructions may be strategically chosen to convey subjective experiences. As oblique nominals are noun phrases that usually function as arguments or adjuncts to more subjective and spontaneous discourse, they could provide additional details about concepts, relationships or qualities.

Text Sample 5.
(Original)
Takistuseks on aga peamiselt siiski meie VÕLGADEL (mitte VÄÄRTUSTEL) nn " **majandussüsteem**" mis ning sunnib MÕLEMAT vanemat rügama hommikust õhtuni jättes lapsed võõraste (mitte)kasvatada. Peamiseks probleemiks on **see**,

et juhul kui kasvatab **riik** tegelikku kasvatust EI TOIMU. Lapsed lihtsalt viiakse kohta kus **nad** saavad olla. Öeldakse, et "riik **see** oleme meie" - kui asjaga tegeleb korraga palju siis asjaga sisuliselt ei tegeleta. Oma äri võid paari tunniga netis registreerida,aga katsu Sa äri pidada.Riigi- ja Brüsseli bürokraatia on asjad nõnda seadnud,et väikese äriga läbi ei löö.
(source: *kalah_zzz_ee.ela_*73709)

(Translation)
However, the main obstacle is primarily our DEBTS (not VALUES) in the so-called "economic system," which compels BOTH parents to toil from morning till night leaving the children to be raised by strangers. The main problem is that if the state raises children real upbringing DOES NOT OCCUR. Children are simply taken to a place where they can exist. Like it is said "the state it is us" - when many deal with something at once then essentially nothing is done about it. You can register your business online in a couple of hours, but try to run a business. State and Brussels bureaucracy have arranged things in such a way that small businesses cannot break through.

Affectivity and subjectivity share the largest number of features. An example of affective and subjective language is in Text Sample 6. Language that is strongly or moderately affective and subjective tends to have a greater number of subordinate clauses (red text) which serve multiple functions, such as providing background details, expressing causal relationships, indicating temporal sequences, or elaborating on ideas, thus expanding the narrative and offering a more accurate portrayal of thoughts and emotions. Furthermore, affective and subjective language strives to be present and personal, where the role of personal pronouns, especially first- and third-person plural pronouns (blue text), personal voice, and indicative mood (**bold text**) can play a role in embodying personal or third-person perspectives and experiences, involving an interactive exchange between individuals. The first- and third-person plural pronouns not only acknowledge the presence of multiple individuals but also create a collective identity of 'we' that binds speakers and listeners in a shared narrative.

A noteworthy aspect is the presence of a demonstrative pronoun, such as *see* 'it' as a pronoun (**bold underlined text**), which denotes a specific situation, attribute, or action that is evident either from the speech or from the context. In Estonian, this pronoun is usually used to refer to inanimate objects or events, but it can also be used as a derogatory expression. Supine constructions (*ma*-suffixes, as *vihkama* 'to hate' in Text Sample 6) also have a higher frequency in affective and subjective discourse. Since core and modal verbs are common to all unplanned-verbal dimensions, the use of supine in affective and subjective texts might point to the higher use of complex verbs (*verb* + *verb* constructions), such as *ta peab tegema* 'he has to do', *ma hakkan tegema* (lit.) 'I'm going to do'.

Text Sample 6.
(Original)
Mul **on** selline mure ,et mul **on** üks klassiõde alguses **tundus** ta normaalne **olime** parimad sõbrad **olime** kogu aeg koos.2klassis **juhtus** selline asi ,et ta **varastas** minu poolt ühe asja pärast kui mu ema seda ta ema käest tagasi **küsis** ei **uskunud** ta ema seda ja mu klassiõde **rääkis** mu emaga päris ebaviisakalt . Pärast seda **hakkas** klassiõde minuga **ülbitsema** . Pärast **läks** asi veel hullemaks ja ta **hakkas** mind **solvama** ,**aias** mu sõbrad minu juurest ära(ma **sain** enne kõigiga hästi läbi) ja siis **hakkasid** need teised klassiõed mind ka juba **solvama** ja ta **rääkis** iga-suguseid asju minu kohta.Lõpuks kui me **hakkasime** eraldi **vahetama** (kolmanda klassi algul)poistega(algul tüdrukud ja siis poisid)**hakkasid** nad tütrukute vahetuse ajal mind igat moodi **solvama**,minu ees **ülbitsema**.Kui ma **ütlesin** ,et **räägin** õpetajale **ütlesid** nad ,et kes mind **usub** kui nemad **väidavad** vastu pidist.Nii see **käib** siiamaani kuigi mul **on** nüüd kaks sõpra(tüdrukud) üks teine põlaalune ja mu pinginaaber.Nad **on** normaalsed.Vahepeal **proovisin** ka selle klassiõe meele järgi olla ,aga see ei **tulnud** mul hästi välja ja ta **oli** mind juba tõsiselt **vihkama hakkand**. **Palun aidake**!Ma ei **tea** mida teha! Tänudega Katri!
(source: *www_lapsemure_ee.ela*_140896.*txt*)

(Translation)
I have a problem with one of my classmates. At first, she seemed normal, and we were best friends. We were always together. In second grade, she stole something from me. When my mom asked her mom to get it back, her mom didn't believe it and my classmate was really rude to my mom. After that, my classmate started acting bossy towards me. Things got worse, and she started calling me names. She turned my friends against me (I used to get along with everyone). Then, the other girls in the class started calling me names too, and she would say all sorts of bad things about me. Finally, when we started changing with the boys separately (girls first, then boys), they would insult me in every way possible during girls' changing time and act bossy in front of me. When I said I would tell the teacher, they said, "Who would believe you if they say the opposite?" This has been going on until now, even though I have two friends now (girls), one from another class and my seatmate. They are normal. I tried to be nice to that classmate, but it didn't work out well, and she really started to hate me. Please help! I don't know what to do! Thanks, Katri

---

Interestingly, affective and subjective language tends to use fewer adjectives, less impersonal voice and have fewer instances of adessive. The reduced use of adjectives, although seemingly counterintuitive given their subjective nature, may

correlate with a reduced use of nouns. This could be due to the modifying function of adjectives accompanying noun phrases. Since an impersonal voice hides the identity of the agent or downplays individual involvement, affective and subjective texts may be more inclined to express personal involvement. In general, the functions of these features are not so straightforward and need to be explored further.

Interactivity, spontaneity, and subjectivity can be distinguished from affectivity by having a lower type/token ratio (the median for spontaneity, subjectivity, and interactivity was 0.53, and 0.63 for affectivity). This ratio is a key measure of lexical diversity, reflecting the richness and complexity of a text's vocabulary. The higher the ratio, the more diverse the vocabulary and potentially the more complex the writing. These results suggest that one underlying communicative function of spontaneity, subjectivity, and interactivity is to connect with the audience, which is expressed through language economy and by conveying the message through repetition without being vague or ambiguous.

Comparing affectivity with spontaneity, subjectivity and interactivity, affectivity (see Table 7 and Text Sample 7) can be uniquely characterized by the use of more singular third-person pronouns (**bold text**). In addition, affectivity uses more determiners (**underlined bold text**) and open clausal complements (red text). The diverse functional roles of open clausal complements complicate the task of providing generalizations for their increased usage. The increased use of determiners in affective texts is noteworthy, as determiners serve to link noun phrases with their contextual referents and require further research.

Text Sample 7.
(Original)
Tedagi häirib, kui **ta** töölt koju jõudes avastab, et nõud on pesemata ja saapad esikus pilla-palla. «Aga siis küsin endalt: kumb on olulisem, kas see, et panen **igal** õhtul saapaid ritta, või see, et lähen hoopis jooksma ja tunnen end hästi?» On hetki, mida tahaks nii väga vanematega jagada. Näiteks siis, kui seisad ülikooli lõpuaktusel ja laulad Gaudeamust. Või siis, kui võtad kätele oma vastsündinud lapse. Kõige enam kurvastab Annelid aga see, et **tal** jäi vanematele ütlemata, kui väga **ta** neid armastab. **Sama** viga teistega **ta** korrata ei kavatse ja soovitab kõigile: öelge oma lähedastele, kui kallid nad teile on! Iial ei tea, mida toob homne päev.
(source: *www_naisteleht_ee.ela*_355886)

(Translation)
It bothers her when she comes home from work and finds that the dishes are unwashed and boots are scattered in the hallway. "But then I ask myself: what is more important, lining up the boots every evening or going for a run and feeling good instead?" There are moments that she so dearly wishes to share with her parents. For instance, when standing at the university graduation ceremony and

singing Gaudeamus. Or when holding her newborn baby in her arms. What saddens Annelid the most is that she never told her parents how much she loves them. She does not intend to repeat the same mistake with others and advises everyone: tell your loved ones how dear they are to you! You never know what tomorrow will bring.

Interactivity can also be distinguished from affectivity, spontaneity, and subjectivity, and all other dimensions by its two unique features (see Text Sample 8), namely, increased use of the vocative (red text) and fewer *hapax legomena*. As interactivity is about participants actively engaging with each other, the use of more vocatives is predictable as it directly addresses a specific person or group. A reduced frequency of *hapax legomena*, i.e., words that occur only once in a given text or a collection, suggests that (written) interactive communication is subject to temporal and memory-related constraints, leading authors to rely on familiar and consistent language. Less interactive texts have more words occurring once, as opposed to more interactive texts (5 vs. 4.5 words appearing once, respectively, see Appendix C). Frequent repetition of words suggests a degree of predictability and uniformity in the language, possibly indicating a comfortable and straightforward writing style.

Text Sample 8.
(Original)
Esimees Ene Ergma
- Suur tänu, härra peaminister! Lõpetan selle küsimuse käsitlemise, kuna aeg on läbi. Ma tahaksin proua kultuuriministrilt vabandust paluda! Me ei jõudnudki teieni, kuna teie meeskolleegid kasutasid väga täpselt oma aja ära. Aitäh, head kolleegid saalis, küsimuste esitamise eest!
(source: *www_riigikogu_ee.ela*_224866)

(Translation)
Chairperson Ene Ergma:
Thank you very much, Mr. Prime Minister! I will conclude by addressing this question as time has run out. I would like to apologize to the Minister of Culture! We didn't even reach you because your male colleagues used their time very precisely. Thank you, dear colleagues in the hall, for asking questions!

Subjectivity stands out with the highest level of agreement between the annotators, reaching 0.76 (see Table 4). An example of subjective language is shown in the Text Sample 9. Texts with high subjectivity use more punctuation (underlined text), clausal copular subjects (**bold text**) and fewer adjectival modifiers. Punctu-

ation suggests the need to structure sentences while providing rhythm, emphasis, and tonality, allowing for more expressive communication. The relationship between these features and subjectivity requires further research.

Text Sample 9.
(Original)
Maailmas, kus operatsioonisüsteemist kasumlikum on **toota** riistvara ning rakendusi, tundub platvormi nii selge esiletoomine pisut liialdatud. Oleks veider **mõeldagi**, et Apple võtaks logona kasutusele oma iPhone'i (või iOS'i!) ikooni, mis üksiku tootena toodab neile ometi rohkem raha kui kõik Microsofti tooted kokku. See ei ole logo, mida saaks tugevalt vihata või armastada ja see vist mind häiribki. Näib, et Microsoft **püüab** meelega suruda end sinna kasti, kuhu ta näiteks Apple'i poolt ammu mängitud on. Igavuse kasti. Sõna otseses mõttes. Maailmas, kui kõik püüavad meeleheitlikult "kastist väljaspool" mõelda, näitab Microsofti uus logo justnimelt kastisisest mõtlemist, kujutades end kastidena.
(source: www_reklaamitrikk_ee.ela_315713)

(Translation)
In a world where it is more profitable to produce hardware and applications than an operating system, emphasizing the platform so clearly seems a bit exaggerated. It would be strange to even think that Apple would use its iPhone (or iOS!) icon as a logo when, as a single product, it generates more revenue for them than all of Microsoft's products combined. This is not a logo that can be strongly hated or loved, and that is what bothers me. It seems that Microsoft is deliberately trying to push itself into the box where, for example, Apple has long played. The box of boredom. Literally. In a world where everyone is desperately trying to think "outside the box," Microsoft's new logo precisely represents thinking inside the box, depicting itself as boxes.

The linguistic profile of spontaneity was distinct across all dimensions, and no statistically significant features were identified when compared to other unplanned-verbal dimensions (see Table 7). Nevertheless, an agreement of 0.6 was achieved among the annotators, indicating that spontaneity may be challenging to distinguish based solely on the features associated with the unplanned verbal dimensions.

6.3.2. Planned-nominal dimensions

The planned-nominal dimensions represent compact and more regulated language use that adheres to prescribed grammar and vocabulary, such as formality, impersonality, and information density. This section examines the linguistic variation between these dimensions.

Table 8: The distribution of features between the planned-nominal dimensions.

FORM	IMP	INFO	INCREASING USAGE	DECREASING USAGE
form	imp	info	- common noun - avg word length (longer)	- pron - pron/noun ratio (lower) - 1st sg/pl pron - 1st prs verb
form	imp		- adjective - nominal - imp voice - genitive case - nominal mod.	- adverb - determiner - core verb - personal voice - verb type ratio (higher) - finite verb - indicative mood
form		info		- 2nd prs verb
	imp	info		- open clausal comp.
UNIQUE FEATURES				
form			- proper noun - avg sentence length (longer)	- punctuation - imperative - nominal cop. subject - copula
	imp		- abbreviation	- interjection - 3rd prs pl pron - supine - discourse

The planned-nominal dimensions employ specific linguistic features to convey complex ideas concisely and precisely, often expressing nuances that are difficult to articulate with general language. In contrast to the unplanned-verbal dimensions, impersonality shifts the focus from the individual to the action or recipient of the action. By minimizing details about the doer (agent) of an action, impersonal writing emphasizes conveying information rather than the personal attributes or experiences of the writer. Information density is concerned with providing more information with fewer words and clauses while being packaged economically. The general results are shown in Table 8, which summarizes how the features were distributed across the planned-nominal dimensions. It shows how the dimensions cluster together based on the shared features and the monotonic change of each feature (i.e. increased or decreased usage) along the level of the

salience of the dimension.

The results demonstrate that, for strongly or moderately formal, information-dense and impersonal texts, it is more characteristic to use common nouns (**bold text** in Text Sample 10) and longer words on average (6.1 characters per word versus 4.8 characters per word, see Appendix C). Both of these features carry specific meanings, add layers of complexity, and enhance the formal tone of a text. Longer words are less common and typically consist of compound words and loan words. The use of longer words can create distance between participants, as they can create a more academic or professional atmosphere by prioritizing precision over colloquialism. As the focus for formality, impersonality, and information density shifts more often from the individual to the action, it is natural that the use of pronouns decreases due to their personalizing functions. The results show that first-person verb forms are used much less in more impersonal, formal, and information-dense texts. Taken together, these dimensions seem to indicate the importance of structured and coherent communication that conveys precise meanings, factual information, and a more formal tone that enhances credibility and authority while minimizing ambiguity and personal bias.

Text Sample 10.
(Original)
Täidesaatev **võim** Eesti **seisukohti** ELi **algatuste** ning **eelnõude** kohta koostavad **ministeeriumid** lähtuvalt oma **valitsemisalast**. Sealjuures on **ministeeriumide ülesandeks** korraldada **seisukohtade kaitsmine** ELi **institutsioonides** ning jõustunud ELi **õigusaktide rakendamine**. **Ministeeriumide seisukohtade sidusus** ning ELi **õigusloome** korrektne **rakendamine** eeldab tugevat **koordineerimist**. ELi **küsimuste** siseriiklikku **koordineerimist** juhib **peaminister**, kes ühtlasi kannab poliitilist **vastutust protsessi** tõhusa ja eesmärgipärase **kulgemise** eest. **Valitsust** ja **peaministrit** nõustab Riigikantselei EL **sekretariaat** (ELS), kes lisaks teenindab ametkondadevahelist **koordinatsioonikogu**, korraldab Euroopa Liidu **dokumendihaldust**, osaleb **valitsuse** Euroopa Liidu alaste **seisukohtade kujundamises** ning seirab ELi **õigusloome rakendamist**.
(source: *valitsus_ee.ela*_604239)

(Translation)
The executive power prepares Estonia's positions on EU initiatives and draft laws based on their respective areas of governance. It is the ministries' responsibility to organize the defense of these positions within EU institutions and to implement enacted EU legislation. The coherence of ministries' positions and the correct implementation of EU legislative processes require strong coordination. The domestic coordination of EU matters is led by the Prime Minister, who also bears political responsibility for the efficient and purposeful progress of the process. The Government and the Prime Minister are advised by the State Chancellery EU Secretariat (EUS), which also serves as the secretariat for inter-agency coordi-

nation, manages European Union document handling, participates in shaping the government's positions on EU matters, and monitors the implementation of EU legislative processes.

The results show that a moderate or high salience of formality and impersonality in texts (see Text Sample 11) contributes to the higher use of features that modify nouns, such as adjectives, genitive case forms (underlined text), and nominal modifiers (**underlined bold text**), nominal word class in general (**bold text**), and impersonal voice (red text). Formality and impersonality together show the highest proportion of shared linguistic features among the planned-nominal dimensions. In general, the results suggest that impersonality and formality are associated with texts that have been carefully curated and may include sentences that pack multiple events into one, often presented in noun phrases.

Text Sample 11.
(Original)
Eelnõus sätestatakse **kasvuhoonegaaside saastekvootidega kauplemise süsteemi rajamist** vastavalt **Kyoto protokollile**. Samuti täpsustatakse **nõudeid**, **mis** sätestavad **bensiini** ja **diislikütuse kvaliteeti**; **biokütuste** ja muude **taastuvkütuste kasutamise edandamist transpordisektoris** ja **osoonikihti** kahandavate **ainete vähendamist**. **Seaduseelnõuga** kehtestatavad **piirangud** on mõeldud **selleks**, et tagada keskkonnasõbralikumate **toodete kasutamine**. **See** omakorda võimaldab säästa **elanikkonda** kasvavast **välisõhu saastest**.
(source: *valitsus_ee.ela*_132104)

(Translation)
The draft law also provides for the establishment of a greenhouse gas emissions trading system in accordance with the Kyoto Protocol. It also specifies requirements that govern the quality of gasoline and diesel fuel; promotes the use of biofuels and other renewable fuels in the transport sector and reduces substances that deplete the ozone layer. The restrictions imposed by the draft law are intended to ensure the use of more environmentally friendly products. Thereby helping to protect the population from increasing air pollution.

Two further dimension pairs were identified: formality & information density versus impersonality & information density. For formality and information density, fewer second-person verb forms are used. This feature directly addresses the reader or audience, introducing a sense of informality or personal involvement that may not be consistent with the desired tone of formality or focus on dense, factual information. This shows that in formal and information-dense texts, there

is no direct communication between the author and the reader or addressee. For impersonality and information density, less open clausal complements are used. Due to the polyfunctional nature of open clausal complements, interpreting their role in this context is challenging and needs further investigation.

Impersonality has distinct characteristics compared to information density and formality. In particular, it is characterized by a higher frequency of abbreviations and a lower presence of interjections, discourse particles, third-person plural pronouns, and supine forms (see Text Sample 12). Abbreviations (**bold text**) inherently compress language, which is consistent in conveying information but may not always be easily understood by everyone. In addition, the use of fewer interjections, discourse particles, third-person plural pronouns, and supine forms in impersonality further emphasizes the shift towards a more formal style of communication. Interjections and discourse particles, often used to express emotion, attitude, or conversational nuance, are minimized in impersonal writing to maintain a sense of professionalism and neutrality.

Text Sample 12.
(Original)
Esimene **CDROM** (ja üldine internetis levitatav) versioon oli **FreeBSD** 1.0, mis lasti välja 1993. aasta detsembris. See põhines **4.3BSD-Lite** (**"Net/2"**) lindil **U.C.** Berkeleyst, koos mitmete muude komponentidega **386BSD** ja Free Software Foundationi projektidest. Esimest versiooni saatis esimese väljaande kohta üsna hea edu ja tema järglaseks sai äärmiselt edukas FreeBSD 1.1 väljaanne 1994. aasta märtsis. Umbes sellel ajal moodustusid silmapiirile üsna ootamatud tormipilved, kui Novell ja **U.C.** Berkeley lahendasid oma pikaajalise kohtuprotsessi Berkeley **Net/2** lindi seaduslikkuse üle. Selle protsessi üheks tingimuseks oli **U.C.** Berkely mööndus, et suur osa **Net/2** koodist on "piiratud" kood ja kuulub Novellile, kes on selle omakorda omandanud **AT&T** käest. Vastutasuks andis Novell 4.4BSD-Lite'ile "õnnistuse" ja lubaduse, et kui **4.4BSD-Lite** välja tuleb, saab ta olema ilma litsentsipiiranguteta ja kõigil **Net/2** kasutajatel soovitatakse tungivalt sellele üle minna.
(source: *www_bsd_ee.ela*_638246)

(Translation)
The first CDROM (and generally distributed on the internet) version was FreeBSD 1.0, released in December 1993. It was based on the 4.3BSD-Lite ("Net/2") tape from U.C. Berkeley, along with several other components from the 386BSD and Free Software Foundation projects. The first version saw fairly good success with its initial release, and its successor became the highly successful FreeBSD 1.1 release in March 1994. Around that time, unexpected storm clouds appeared on the horizon as Novell and U.C. Berkeley settled their long-standing court case over the legality of the Berkeley Net/2 tape. One condition of this process was a concession by U.C. Berkeley that a significant portion of the Net/2 code is "en-

cumbered" code and belongs to Novell, who had acquired it from AT&T. In return, Novell granted "blessing" to 4.4BSD-Lite and a promise that when 4.4BSD-Lite is released, it would be unrestricted, strongly recommending all Net/2 users to switch to it.

Formality has clear characteristics when compared to information density and impersonality. Formal texts are characterized by a higher frequency of proper nouns and longer sentences (see the Text Sample 13). The increased use of proper nouns (**bold text**) in formality serves a dual purpose. It adds specificity and clarity to content by naming particular entities or concepts, while also contributing to a sense of formality in communication. Proper nouns can provide concrete references that help to establish a clear context for the reader. Furthermore, the results show that compared to other dimensions, formal texts have longer sentences on average, i.e., 25 words per sentence compared to 16 words per sentence in texts where formality was not considered present (see Appendix C).

Formal texts use less punctuation, less imperative mood, less nominal copular subjects, and fewer copulas. Punctuation may be related to longer sentences where the strategic plan is to pack as much information within a single clause, resulting in longer sentences with less punctuation. Nominal copular subjects and copulas are often associated with more conversational or informal language, whereas formal writing typically relies on more concise and condensed structures, which may create more complex sentences. The use of imperative mood is not characteristic in formal texts because it seems to adopt a more suggestive and persuasive tone, instead of giving direct commands.

Text Sample 13.
(Original)
23. aprill 2010 **Tallinnas** jätkub täna **NATO** välisministrite mitteametlik kohtumine, mis keskendub teisel päeval **NATO-Vene** suhetele ning toimub **Afganistani** rahvusvahelistesse julgeoleku abijõududesse panustavate riikide välisministrite kohtumine. **NATO** välisministrid otsustasid eile hilisõhtul anda **Bosnia** ja **Hertsegoviinale** liikmelisuse tegevuskava, samuti kinnitasid välisministrid veelkord vajadust rajada **NATO** raketikaitse-süsteem, vahendasid **ERRi** teleuudised. Kuigi välisministrid **Tallinnas** muid otsuseid vastu ei võta, hindas välisministeeriumi poliitika asekantsler **Harri Tiido** juba **Tallinna** kohtumise esimest päeva väga produktiivseks.
(source: uuseesti_ee.ela_112397)

(Translation)
On April 23, 2010, in Tallinn, today continues the informal meeting of NATO foreign ministers, focusing on the second day on NATO-Russia relations, along with

a meeting of foreign ministers of countries contributing to international security assistance in Afghanistan. NATO foreign ministers decided late yesterday evening to grant Bosnia and Herzegovina a Membership Action Plan, and also reaffirmed the need to establish a NATO missile defense system, as reported by ERR television news. Although the foreign ministers in Tallinn will not be making any other decisions, the Deputy Secretary General for Political Affairs of the Ministry of Foreign Affairs, Harri Tiido, already assessed the first day of the Tallinn meeting as very productive.

As shown in 4.4 and Table 8, information density was related to the planned-nominal dimensions by sharing many features with formality and impersonality, despite the lowest inter-annotator agreement ($\alpha = 0.25$, see Table 4). The information density lacked statistically significant distinctive features, and the majority of the texts (59%, as detailed in Section 4.2.3) were considered to represent strong or moderate information density. This finding suggests two possibilities: either information density requires a more refined operationalization to capture the subtleties within this dimension, or the exploration of alternative dimensions may be more effective in differentiating the collective communicative functions of the planned-nominal dimensions.

6.3.3. Complementary dimensions

The third group of dimensions includes abstractness, temporality, instructability, and complexity. These dimensions can be categorized as complementary dimensions that do not fit into the macro dimensions. The complementary dimensions demonstrate a degree of independence within the unplanned-verbal and planned-nominal dimensions, and within the group itself. Further, inter-annotator agreement was lower for these dimensions compared to the unplanned-verbal and planned-nominal dimensions.

As outlined in Section 3.2, complexity and abstractness share the characteristic of requiring a certain level of cognitive engagement from the reader. Complexity focuses on difficulty or intricacy in structure and language, requiring extra effort to comprehend. Abstractness emphasizes generalization and conceptualization, moving from specific details to broader concepts, and may likewise require a higher level of cognitive processing to grasp the ideas. Instructability, on the other hand, aims to give clear instructions or outline processes. Temporality involves arranging events or ideas in a coherent and sequential order, emphasizing the chronological or logical progression of the discourse. Argumentativeness differs from the other complementary dimensions in that it is specifically concerned with the act of presenting one's point of view on an issue without overtly presenting subjective arguments. In contrast, abstractness and complexity focus on cognitive engagement, while instructability and temporality relate to the clarity of

instructions or processes sequentially. Argumentativeness focuses on the expression of a point of view without cognitive processing or temporal organization as an underlying parameter.

Instructability appears to differ from other complementary dimensions. For example, it did not share any linguistic features with the planned-nominal or unplanned-verbal dimensions. Moreover, the sole statistically significant feature was the use of the present tense (as illustrated in **bold text** within the Text Sample 14). Contrary to expectations, the imperative mood did not emerge as a statistically significant feature of instructability, despite its intuitive association with the communicative function of instructability (i.e., providing the reader with instructions on how to perform certain activities or describing the stages of a process). It is, therefore, necessary to conduct further research to reach any definitive conclusions regarding the role of instructability within the Dimensional Text Model.

Text Sample 14.
(Original)
Lõika alõtša-ploomid pooleks ja **eemalda** kivid. **Pane** koos veega potti ja **kuumuta** keemiseni. **Keeda** tasasel tulel kaanega potis 15 minutit, kuni ploomid **on** pehmed. **Tambi** uhmris koriandriseemned, apteegitilliseemned, hakitud küüslauk, Cayenne ja sool ühtlaseks pastaks. Kui ploomid **on** pehmed, **aja** need läbi hakkmasina ja **pane** puhta poti sisse. **Kuumuta** keemiseni ning **keeda** mõõdukal tulel pidevalt segades umbes 3 minutit. **Lisa** vürtsipasta ja **keeda** veel umbes 5 minutit, kuni pasta pakseneb kergelt. **Lisa** hakitud münt ja koriander ja **tõsta** pott tulelt. **Kalla** tkemali kuumalt purkidesse. **Jahuta** toatemperatuurini ja siis **säilita** külmkapis. Kui **soovid** tkemalit kauem säilitada, siis **kaaneta** purgid õhukindlalt.
(source: www_nami-nami_ee.ela_173054)

(Translation)
Cut the plums in half and remove the pits. Put them in a pot with water and heat to boiling. Simmer covered for 15 minutes until the plums are soft. Crush coriander seeds, fennel seeds, chopped garlic, cayenne, and salt in a mortar to make a smooth paste. Once the plums are soft, pass them through a meat grinder and place them in a clean pot. Bring to a boil and simmer, stirring constantly, for about 3 minutes. Add the spice paste and simmer for another 5 minutes until the paste thickens slightly. Add chopped mint and coriander, then remove the pot from the heat. Pour the tkemali into jars while hot. Cool to room temperature and then store in the refrigerator. If you want to preserve tkemali for a longer period, seal the jars airtight.

Similar to instructability, temporality did not share features with the planned-nominal or unplanned-verbal dimensions. Texts with a moderate or high salience of temporality use more past tense (**bold text** in the Text Sample 15) and numerical modifiers (underlined text). The latter feature emerges as a unique feature that is characteristic of texts where temporality was considered moderate or strong. This indicates the importance of specific numerical data in clearly communicating temporal information. Where temporality is considered strong or moderate, present tense with infinitive verb forms and *da*-infinitives show a reduced usage.

Text Sample 15.

(Original)

Sellest ajast on meie lähikonnas teada näiteks Kolga olemasolu. Edasised teated pärinevad juba lääniürikutest. Kui taanlased **vallutasid** Põhja-Eesti, **läänistati** Kolga ümbruse maad Ojamaa tsistertslaste kloostrile ning aastal 1290 on mungad üles **lugenud** rea külasid, nende hulgas ka Vanaküla. Kuskilt olen ma **lugenud**, et praegune Valgejõe kaldaid palistav Vanaküla olevat algul hoopis Valgejõe nime **kandnud**. Kui küla suures tulekahjus **hävis**, **ehitati** uued majad paar kilomeetrit ülesvoolu, praeguse Valgejõe küla (või keskuse, nagu uhkelt viitab maantesilt) kohale. Mõne aja pärast ehitati aga taas üles ka vana külakoht ja sellest siis ka see nimi. Aastal 1990, kui külad **said** 700-aastaseks, **oli** kolme küla peale kokku 77 talu, neist Valgejõel 23, Vanakülas 37 ja Parksis 17. Püsivalt **elati** neist 53-s ning elanikke **oli** 101. Praegust seisu ei tea, kuid arvan, et viimased numbrid on väiksemaks **jäänud**: külad surevad aegamisi välja. Ja talud muutuvad suvituskohtadeks. Mitmed neist on "uutele" inimestele **müüdud**. Ja isegi põliselanike nooremat põlvkonda enam ei tunta: minu lapsi vaevalt et osatakse kaugema külaotsa rahva poolt "kodusse ajada" ehk siis ära määrata, kes ja mis talust nad on.
(source: kaja_ekstreem_ee.ela_71641)

(Translation)

Since then, the existence of Kolga in our vicinity has been known. Further reports come from historical records. When the Danes conquered Northern Estonia, the lands around Kolga were granted to the Ojamaa Cistercian monastery, and in 1290, the monks listed a series of villages, including Vanaküla. I have read somewhere that the current Vanaküla, which lines the banks of the Valgejõe River, was originally named Valgejõe. When the village was destroyed in a major fire, new houses were built a couple of kilometers upstream, at the present site of Valgejõe village (or center, as proudly indicated by the road sign). After some time, the old village site was rebuilt, hence the name. In 1990, when the villages turned 700 years old, there were a total of 77 farms among the three villages, with 23 in Valgejõe, 37 in Vanaküla, and 17 in Parks. 53 of them were permanently inhabited, with a population of 101. I do not know the current situation, but I believe the latest figures have decreased: villages are slowly dying out. Farms are turning into vacation spots. Several of them have been sold to "new" people. Even the

younger generation of native inhabitants is no longer recognized: my children are hardly known by the people from the other end of the village - in other words, it is difficult to determine who and from which farm they are.

Compared to instructability and temporality, argumentativity exhibits similarities with both the unplanned-verbal and planned-nominal dimensions. An example of an argumentative text is seen in the Text Sample 16. Argumentativity exhibits similarities with affectivity, interactivity, spontaneity, and subjectivity in using negative polarity (**bold text**). Furthermore, it shares similarities with the unplanned-verbal dimensions, characterized by the use of fewer proper nouns. Argumentativity has an inverse relation with formality: the more formal the text, the more proper nouns are used, while the more argumentative (and affective, spontaneous, and subjective) the text, the fewer proper nouns are used.

Text Sample 16.
(Original)
Üks põhjus, miks lapsed koolis edasi **ei** jõua, on, et nad **ei** oska teksti mõttega lugeda **ei** kirjanduses, ajaloos, geograafias **ega** ka matemaatika tekstülesandes, rääkimata definitsioonidest ja reeglitest. Kõige rohkem peab tekstiga tööd tegema eesti keele õpetaja, et kellelgi poleks põhjust öelda: mis te seal teinud olete, lapsed **ei** oska lugeda **ega** kirjutada. Mille vastu võideldakse? Õppekavaga **ei** ole 2003. a eksamis vastuolus midagi, pealegi on õppekava üldosa nii üldsõnaline, et hea tahtmise korral võib neid mõtteid piiritult avardada või soovi kohaselt koomale tõmmata, kuidas parasjagu tahetakse. Me võiksime skolastilist vaidlust arendada lõputult.
(source: arhiiv_koolielu_ee.ela_481254)

(Translation)
One reason why children do not progress in school is that they do not know how to read text with meaning, not in literature, history, geography, or even in a math assignment, let alone definitions and rules. The most work with text must be done by the Estonian language teacher, so that no one has a reason to say: what have you been doing there, the children cannot read or write. What is being fought against? There is nothing contradictory in the 2003 curriculum's exam, moreover, the general part of the curriculum is so general that with good intentions, these ideas can be infinitely expanded or reduced as desired, how one wants at the moment. We could develop scholastic dispute endlessly.

Furthermore, argumentativity is linked with other dimensions from the complementary dimensions, such as temporality and instructability. For instance, highly

argumentative texts prefer to use present tense over past tense (red text in Text Sample 16). This aligns with the emphasis on present-focused actions observed in instructability. However, temporality presents an intriguing contrast. When temporality is strong or moderate, the use of present tense decreases, with past tense becoming more prominent. Argumentativity and temporality also have an inverse relation to the use of *da*-infinitives. Argumentative texts have a higher frequency of *da*-infinitives (underlined text), while their use diminishes when expressing temporal relations becomes more important in a text. While *da*-infinitives in Estonian are multi-functional, their specific role within argumentative contexts merits further investigation through qualitative research.

Complexity shows a different pattern compared to the other complementary dimensions. Its statistically significant features are mainly aligned with the planned-nominal dimensions, specifically formality and impersonality. For example, complexity shows similarities with impersonality in the use of more abbreviations (**bold text** in Text Sample 17), with formality in the use of less punctuation, and with formality and impersonality in the use of more nominal modifiers (underlined text) and less finite verb forms. However, complexity cannot be considered as one of the dimensions of the planned-nominal dimensions because it has only a moderate positive correlation with information density (r=.41). However, complexity and information density do not share any statistically significant features.

Text Sample 17.
(Original)
Käesoleval aastal reguleerib küsimusi sotsiaalministri 22. jaanuari 1998. **a** määrus **nr** 8 "Tervisekaardi ja saatekirja kehtestamine" (**RTL** 1998,58/59,262) , kus punkt 2 sätestab järmist: Ambulatoorset arstiabi andvatel tervishoiuasutustel ja perearstidel kasutada tervisekaarti nii laste kui täiskasvanute kohta vormistatava dokumendina, mida peetakse paber- ja/või elektroonilisel kandjal. Elektroonilisel kandjal peetava tervisekaardi kasutamise programm peab tagama andmete säilimise muutumatuna nende sisestamise kuupäeva seisuga.
(source: www_hambaarst_ee.ela_423040)

(Translation)
This year, the issues are regulated by the Regulation No. 8 of January 22, 1998, of the Minister of Social Affairs, titled "Establishment of the Health Card and Referral" (RTL 1998,58/59,262), where point 2 stipulates the following: Healthcare facilities providing outpatient medical care and family doctors are to use the health card as a document issued for both children and adults, which is maintained in paper and/or electronic form. The program for maintaining the health card in electronic form must ensure the preservation of data unchanged as of the date of entry.

Similarly, abstractness is associated with subjectivity and spontaneity from the unplanned-verbal dimensions. The more abstract (and also subjective and spontaneous) the text, the fewer oblique nominals were used. No other statistically significant features were identified for abstractness, and no significant positive correlations were observed with other dimensions. The Text Sample 18 illustrates that abstractness is expressed through lexical features, e.g., *jumal* 'God', *Vaim* 'spirit', and *külvama* 'to sow'. However, the results do not give us enough to establish a definitive connection between abstractness and the unplanned-verbal dimensions, nor to draw any definitive conclusions about the dimension itself.

Text Sample 18.
(Original)
Jumala sõna pidi avalikuks saama. Meie oleme Jumala viinamäe töölised. Jeesus on meie kohtagi ju sellise käsu andnud, et kui me Jumalasse usume, siis me peame seda ka avalikult üles tunnistama. Aga ometi ei pea me väga pead vaevama ja muretsema selle üle, kuidas siis minu elus see kuulutamine välja näeb või mida siis mina ära tegema pean hakkama, et Jumal oleks minu kuulutusega rahul. Sest Jumala sõna, mis meisse on külvatud, saadab ise korda selle, et ma mulle seatud ajal ütlen seda, mida Vaim mulle öelda annab. See tegelikult ka ei tähenda muud kui minemist oma igapäevaelusse julgusega, teades, et see, mida tuleb teha, on üksnes vastata, kui minult küsitakse.
(source: www_eestikirik_ee.ela_530609)

(Translation)
The word of God was meant to be made public. We are the laborers in God's vineyard. Jesus has also given us such a command that if we believe in God, we must also confess it publicly. Yet we do not need to worry too much or be anxious about how this proclamation looks in our lives or what I must do to ensure that God is pleased with my proclamation. Because the word of God that has been sown in us will itself bring about the situation where, at the appointed time, I will speak what the Spirit gives me to say. This actually means nothing more than stepping into our daily lives with courage, knowing that what needs to be done is simply to respond when asked.

6.4. Summary

This chapter addresses RQ2 by determining which linguistic features, if any, are significantly associated with the level of salience of the twelve dimensions proposed by the Dimensional Text Model. The objective was to determine the distinctiveness of the dimensions by examining their linguistic profiles, which reflect communicative functions. To determine the linguistic profiles of each dimension,

it was necessary to assess the statistical relevance of each feature in predicting the dimensional salience. To this end, a one-way non-parametric analysis of variance (ANOVA) was performed. This analysis compared the variances of the medians of the relative frequencies between the three groups representing different levels of dimensional salience, i.e., *strong/moderate*, *weak*, or *non-existent*. *Post hoc* tests were subsequently conducted to determine which groups were significantly different.

The results presented in Section 6.2 and Section 6.3 demonstrated that all dimensions exhibited a unique linguistic profile. Most linguistic features were not exclusive to one particular dimension but rather appeared to be shared across multiple dimensions. Although the linguistic profiles were distinct for all dimensions, the objective was to identify those dimensions that exhibited features that were exclusive to that particular dimension. The analysis demonstrated that five dimensions – subjectivity, affectivity, interactivity, formality, and temporality – had distinct linguistic features that set them apart from the other dimensions.

The objective of RQ2 was to determine how the linguistic profiles, if at all, between the dimensions differ. While the dimensions are assumed to be distinct, the analysis in Section 4.4 revealed that the dimensions do not exist independently. The results presented in this chapter provide further empirical support for this claim. The results demonstrated that some of the dimensions could be categorized as macro dimensions, namely unplanned-verbal, planned-nominal dimensions. The unplanned-verbal dimensions include *subjectivity*, *affectivity*, *spontaneity*, and *interactivity*, and are characteristic of texts that are personal, opinionated, and interactive. The planned-nominal dimensions include *formality*, *impersonality*, and *information density*, and are found in texts that are more formal and focused on conveying factual information. Most linguistic features identified as statistically significant by the ANOVA and *post hoc* tests are common to the unplanned-verbal and planned-nominal dimensions. The linguistic features tend to have an inverse relationship between the unplanned-verbal and planned-nominal dimensions. If the median of the relative frequency of a particular linguistic feature increased in association with unplanned-verbal dimensions, the median of a relative frequency conversely decreased in planned-nominal dimensions, and vice versa.

The results also indicate the existence of a third group of dimensions, which are referred to as complementary dimensions. This complementary group includes *argumentativity*, *abstractness*, *complexity*, and *instructability*. The term 'complementary' is employed to refer to dimensions which do not fit into the macro dimensions. The complementary dimensions demonstrate a degree of independence within the unplanned-verbal and planned-nominal dimensions, and within the group itself. Complexity correlates positively with information density, suggesting that complex texts may also have a higher concentration of information. In addition, complexity shows a particularly strong relationship with formality and impersonality. They are associated with the use of less punctuation and more finite

verb forms, abbreviations, and nominal modifiers. Thus, complexity can generally be associated with the planned-nominal dimensions. Similarly, abstractness shows a negative correlation with temporality, suggesting that more abstract texts tend to be less temporally bound. Abstractness can likewise be linked to spontaneity and subjectivity from the unplanned-verbal dimensions through a reduced use of oblique nominals. Argumentativity presents a more complex picture. It can function as a complementary dimension for both the planned-nominal and unplanned-verbal dimensions and even co-occur with other complementary dimensions. Compared to the other complementary dimensions, temporality shows more relative independence, sharing no significant features with the planned-nominal or unplanned-verbal dimensions, and having an inverse relationship in terms of features with argumentativity and instructability. Due to the lack of data on instructability, further research is needed before any conclusions can be drawn.

7. SUMMARY

Web corpora have transformed the way languages are studied. Their size, real-world nature, and ability to reflect language change offer unique advantages for linguistic research. As highlighted in Chapter 2, Web corpora are a valuable resource for many research fields. However, their unknown composition poses a major obstacle to linguistic analysis, as researchers may not have a clear picture of the types of text represented in the corpora. The main issues include the dynamic and evolving nature of Web content, a lack of universally agreed-upon terminology, the abundance of different register classifications, and the level of granularity in the already existing taxonomies. Even if there were a widely accepted scheme, achieving a high level of agreement between annotators remains a challenge. These issues were the main motivation behind this dissertation.

This dissertation introduces a language- and corpus-independent framework, the Dimensional Text Model, to identify the core characteristics that differentiate various texts based on their communicative function. The Dimensional Text Model is built on the strengths of two established and influential frameworks, the Multidimensional analysis (Biber 1988, MDA) and Functional Text Dimensions (Sharoff 2018, FTD).

The MDA (Biber 1988) is a methodological framework that identifies and compares the underlying dimensions of linguistic variation from a wide range of registers of English. This approach is predicated on a fundamental assumption that linguistic features co-occur because they share similar communicative functions. The framework utilizes factor analysis to identify the dimensions by examining the communicative purposes realized by the co-occurring linguistic patterns. Similar to Biber's approach to studying language variation, the FTD framework (Sharoff 2018) follows more traditional methods in automatic genre classification by developing a taxonomy where each label is motivated by a function. Large Web corpora are then classified into registers by those functional dimensions. Each dimension represents a functional category that describes a text's communicative purpose (e.g., *to what extent is the text concerned with expressing feelings or emotions?*). The FTD framework enables multi-labelling, which describes texts through a combination of primary (mandatory) and secondary (optional) functional dimensions. The FTD also introduces the multidimensional space, where each text can be represented as a vector of functions. These vectors allow mapping functionally similar texts across corpora.

While the MDA and FTD have been pivotal methodological approaches in contemporary corpus-based research, their constraints motivated the development of the theoretical framework proposed in this dissertation - the Dimensional Text Model (DTM). Since the DTM aimed to identify the universal core characteristics that differentiate various texts based on their communicative functions, different MDA studies had a central role in generating the set of dimensions proposed by the DTM. The MDA dimensions are inherently functionally complex factors in-

corporating many communicative functions at once. These were broken down into smaller elementary dimensions, each of which realizing a single communicative function. Based on the decomposition of the MDA dimensions into smaller dimensions motivated by a singular function, 12 dimensions for the DTM were defined: (1) *abstractness*, (2) *affectivity*, (3) *argumentativity*, (4) *impersonality*, (5) *interactivity*, (6) *instructability*, (7) *formality*, (8) *complexity*, (9) *spontaneity*, (10) *information density*, (11) *temporality* and (12) *subjectivity*. From FTD, the concept of *multidimensional space* was expanded to represent all texts within a comprehensive dimensional space. Unlike the FTD, where the multidimensional space described the texts through primary (mandatory) dimensions and secondary (optional) dimensions, the DTM expanded the multidimensional space to view the texts as all dimensions are salient simultaneously. In other words, some dimensions may be present to a limited extent or may gravitate towards being more strongly present. By modelling the texts in a unified 12-dimensional space, the model enables the measurement of the spatial proximity of all texts, regardless of whether the communicative function is more or less salient. This may indicate that texts that are closer together in the multidimensional space share similar register characteristics.

The Dimensional Text Model employs a hierarchical structure to characterize texts at three levels (see Chapter 3 for a more detailed description of the framework). The lowest level is represented by quantifiable linguistic features, such as average sentence length and the frequency of specific word classes. The middle layer consists of twelve latent dimensions, each representing a single communicative function, such as abstractness, affectivity, and argumentativity. These dimensions manifest through the co-occurrence of statistically significant linguistic features. Finally, the highest layer identifies text registers, based on the level of salience of the co-occurring dimensions present in a given text. Building on this framework, the following research questions were developed:

(RQ1) To what extent is it feasible to annotate the data using the dimensions outlined in the Dimensional Text Model?

(RQ2) Do the dimensions defined by the proposed model exhibit statistically significant differences in their associated linguistic features (e.g., vocabulary choices, grammatical structures)?

These research questions are discussed in Section 7.1. The limitations along with suggestions for further research are outlined in Section 7.2 and Section 7.3 briefly summarizes the whole dissertation.

7.1. Addressing the Research Questions

7.1.1. RQ1: Validating the framework

The goal of RQ1 was to assess whether the theoretically proposed twelve dimensions are recognizable and distinguishable to the annotators. To measure whether

the dimensions were distinguishable, an annotation study was conducted in which the annotators were asked to rate the level of salience of the proposed dimensions on a four-point Likert scale (*strong*, *moderate*, *weak* or *non-existent*) in a collection of 120 texts gathered from the EtTenTen Web corpus (Koppel & Kallas 2022). While the annotation study was designed to measure the non-existence of dimensions, the underlying framework posits that all dimensions are inherently salient in every text, albeit with varying levels.

The inter-annotator agreement was measured using Krippendorff's α coefficient (2004). Since the $\alpha \geq 0.667$ (Krippendorff 2004: 241) was considered to be too conservative, the following conventions proposed by Landis & Koch (1977) were adopted:

- $\alpha \geq 0.61$ as substantial,
- $0.41 \leq \alpha < 0.6$ as moderate,
- $0.21 \leq \alpha < 0.4$ as fair.

The results demonstrated substantial agreement on the dimensions of subjectivity, affectivity, formality, and spontaneity, with inter-annotator agreements ranging from 0.6 to 0.76. Moderate agreement was achieved on the dimensions of instructability, interactivity, impersonality and temporality, with inter-annotator agreement ranging between 0.40 and 0.47. Fair agreement was observed on the dimensions of complexity, argumentativity, abstractness and information density, with inter-annotator agreement ranging between 0.25 and 0.38.

A lower inter-annotator agreement reflects the issues in the quality and consistency of annotations. One of the contributing factors could be the lack of well-defined guidelines. For instance, the complexity dimension received an agreement score of 0.38. Complexity is defined as a phenomenon that refers to the level of difficulty or intricacy in its structure and language while demanding additional effort from the reader. Following this definition, it could be challenging for the annotators to decide whether complexity stems from the writer's style or their interpretation based on their experience and knowledge. Similarly, the argumentativity dimensions received a α score of 0.375. As defined in Section 3.2, argumentativity is a dimension which presents the point of view on a topic or phenomenon. However, difficulty arises in clearly differentiating between bias and neutrality, a task that can be challenging even with additional context. Similar issues can be applied to the abstractness dimension ($\alpha = 0.33$) since the level of abstractness is open to interpretation and varies between individuals - what one person considers a highly abstract concept, another might find relatively concrete. The lowest agreement ($\alpha = 0.25$) was observed for the information density dimension. Its definition of conveying more information while being economic is susceptible to various interpretations, such as all content should qualify as information or only texts which resemble entries from Wikipedia or another encyclopaedia. In particular, 59% of all the evaluated texts were classified as containing information strongly or moderately (discussed in Section 4.2.3). The term 'information' is

inherently vague and lacks a precise definition, which results in annotators depending on personal interpretations, leading to discrepancies. Tasks like these require refined definitions and typically rely on personal conceptions, which could complicate achieving a relatively high agreement.

Another contributing element to a low inter-annotator agreement might be the lack of suitable texts for the annotation task. When designing the annotation task, the etTenTen (*etTenTen: Corpus of the Estonian Web* 2021) corpus was the only Estonian Web corpus available, and its labelling had not been validated. The corpus was pre-classified into categories of *government*, *forum*, *religion*, *blog*, *periodicals*, *informative*, and *unknown*. In the absence of essential metadata, the objective was to ensure that the proportions of the existing categories in the corpus were maintained when selecting texts for the annotation study. For instance, the category *unknown* constituted the largest proportion in the etTenTen corpus, indicating that approximately 36% of the texts (43 texts) had to be selected from this category. This selection method was less than optimal, as it resulted in some more homogeneous categories (e.g. *religion*) being overrepresented, while more heterogeneous categories (e.g. *informative*, *periodicals*) were underrepresented.

Considering the limitations of the annotation task, it can be concluded that the generally satisfactory inter-annotator agreements indicate that texts can be annotated by the dimensions proposed by the Dimensional Text Model. Agreements ranging from moderate to substantial show that the guidelines were generally well-defined, demonstrating that certain dimensions had clear characteristics across various texts. This is particularly evident for subjectivity, affectivity, formality, spontaneity, instructability, interactivity, impersonality, and temporality. However, dimensions that achieved fair agreement, such as abstractness, complexity, argumentativity, and information density, require further refinement, and additional training sessions for annotators could be beneficial in achieving more consistent annotations.

7.1.2. RQ2: Dimensional variation

Although RQ1 demonstrated that the twelve dimensions are more or less distinguishable, it was unclear how the linguistic profiles between these dimensions differ. Thus, the objective of RQ2 was to identify how the proposed dimensions, if at all, differ in terms of their statistically significant features. Given the limited research in Estonia on the linguistic features contributing to the discrimination of registers, the choice of linguistic features was based on the output of Stanford's Stanza (Qi et al. 2020) parser. The objective was to automatically extract a diverse range of features, including lexical and textual features (e.g. type/token ratio, average sentence length, a list of core verbs) and various grammatical features (e.g. verb categories, parts of speech). All feature frequencies were normalized to the length of the text. The list of the extracted features and their potential functional roles were discussed in Chapter 5.

To assess the statistical significance of the role of each feature in dimensional saliency, a one-way non-parametric analysis of variance (ANOVA) with Dunn's multiple comparison tests was conducted. The ANOVA compared the variances of the medians of the relative frequencies between the three groups representing different levels of dimensional salience: *strong/moderate*, *weak*, or *non-existent*. A statistically significant ANOVA result suggests that there is a difference between the groups in question, but it is typically followed by post hoc tests to tell which specific groups are significantly different from each other. In the case of non-parametric data, Dunn's multiple comparison test was used to identify which specific pairwise comparisons of the groups demonstrated statistically significant differences.

The analysis demonstrated that all dimensions exhibited a unique linguistic profile, indicating that each dimension has a distinctive communicative function within the framework. While the linguistic profiles of all dimensions were distinct, the objective was to identify those that exhibited a unique feature exclusive to that particular dimension. The analysis uncovered that out of the twelve dimensions, the linguistic profiles of five dimensions, such as subjectivity, affectivity, interactivity, formality, and temporality, included unique features specific only to them. For example, interactivity can be characterized by distinctive features of vocatives and a smaller vocabulary. Formal texts, on the other hand, prioritize clarity, precision, and objectivity. This is reflected in the longer sentence structures found in formal texts. Longer sentences allow for more subtle details and the inclusion of compressed information. The temporality dimension, emphasizing the coherence and sequential order of events, is showcased using less infinitive verb forms and more numerical modifiers to quantify entities or specify the timing of various events. Most linguistic features were not unique to a single dimension but rather shared across multiple dimensions. This observation suggests that linguistic features are not exclusive, but rather capable of communicating a range of functions when used in conjunction with other features. This exploration only scratches the surface of this framework, but the results show that there is variation among the dimensions proposed by the Dimensional Text Model, and it should be further investigated.

Besides identifying the possible variational aspects of dimensions, Section 4.4 together with Chapter 6 revealed a noteworthy observation that the dimensions can be categorized into two macro (the planned-nominal and unplanned-verbal), and complementary dimensions. The unplanned-verbal dimensions are linked to subjectivity, affectivity, spontaneity, and interactivity. The planned-nominal dimensions are expressed through dimensions of formality, impersonality, and information density. The complementary dimensions include argumentativity, abstractness, complexity, and instructability which demonstrate a degree of independence within the unplanned-verbal and planned-nominal dimensions, and within the group of complementary dimensions itself.

The planned-nominal and unplanned-verbal distinction aligns with the established recognition that spoken and written language are distinct modes. These parallels in the spoken-written spectrum have been observed consistently in different languages and cultures, e.g., Biber (1988), Besnier (1988), Biber & Hared (1992), Kim & Biber (1994), Crystal (2001), Sardinha et al. (2014), Cvrček et al. (2020), Biber et al. (2020), etc. The planned-nominal and unplanned-verbal dimensions exhibited an inverse pattern of characteristic linguistic features. In essence, if a particular linguistic feature showed a significant increase in association with unplanned-verbal dimensions, it conversely displayed a significant decrease in planned-nominal dimensions, and vice versa. For example, when an unplanned-verbal dimension (e.g., interactivity) uses more pronouns, core verbs, and interjections, then conversely, a planned-nominal dimension (e.g., impersonality) uses fewer pronouns, core verbs, and interjections. Most features appear to be delineated between the planned-nominal and unplanned-verbal dimensions, i.e. if its increased usage is statistically significant for the unplanned-verbal dimensions, then its decreased usage is statistically significant for the planned-nominal dimensions, and vice versa. To conclude, a qualitative analysis is necessary to fully understand the nature of these linguistic patterns.

As highlighted in Section 7.1.1, the absence of precise and unambiguous guidelines for dimensions such as complexity, argumentativity, abstractness and information density contributed to lower inter-annotator agreement, indicating the need for further refinement of the annotation process. For example, the information density was demonstrated to be more confusing for the annotators, with the lowest agreement ($\alpha = 0.25$). The low inter-annotator agreement for information density indicates that this dimension is inherently complex to define and detect. However, despite the low agreement, factor analysis and the shared linguistic features identified a strong connection between information density, formality, and imperativity. This implies that even if the essence of information density is tricky, it likely plays a significant role alongside other planned-nominal dimensions.

Although some dimensions presented considerable difficulties in terms of low inter-annotator agreement, achieving a moderate inter-rater agreement does not necessarily indicate the absence of potential issues or complications. To illustrate, instructability achieved moderate agreement among the annotators (α= 0.47), yet exhibited no statistically significant correlations with other dimensions. Furthermore, the ANOVA and *post hoc* tests revealed only a single statistically significant feature: where instructability was considered to be present more strongly, the more present tense is used. This finding suggests two possible explanations. Firstly, the definition provided to the annotators may have been overly specific, limiting their ability to identify instructability in a broader sense. Secondly, instructability may not be a fundamental characteristic within this framework.

7.2. Future work

This section discusses the limitations of the research presented in this dissertation and introduces some suggestions for future work.

First, some dimensions achieved a low inter-annotator agreement. In general, a high inter-annotator suggests that the annotators share a common understanding of the task and the criteria for making judgements. This ensures more reliable annotations and a well-annotated dataset leads to better performing models. Conversely, low inter-annotator agreement introduces noise and negatively affects model performance. Similarly, in this dissertation, high inter-annotator agreement was used to assess the reliability of the dimensional text model and identify areas for improvement. The exact reasons some dimensions achieve low inter-annotator agreement are difficult to determine. One reason may be that some dimensions proposed in the Dimensional Text Model need to be fine-tuned or re-calibrated. For example, information density and argumentativity could have been too ambiguous for consistent annotation. In addition, some dimensions may be strongly associated with their literal name. For example, texts which were considered highly instructive were mostly recipes and manuals, neglecting those texts where instructions are conveyed more subtly (e.g. someone giving instructions in a forum-type conversation). The other reason may have been the lack of clear guidelines provided for the annotators, which could have led to inconsistent interpretations of the task and, consequently, a lower inter-annotator agreement. There are many options for the inter-annotator agreement score. For example, the whole annotation study could benefit from providing annotators with examples to help them in their decision-making process. The annotation study presented in this dissertation only included short definitions. No text examples were provided to avoid bias. Another option would be to include training sessions for the annotators to develop a common understanding of the study. In addition, it may be beneficial to ask annotators to explain the reasoning behind the choices. This kind of introspection not only helps to identify potential problems in the design of the annotation study but may also lead to more consistent annotations.

Secondly, the insufficient amount of appropriate data has been a bottleneck for many models in natural language processing. The lack of data often limits the capacity of the models and overfits the data where the model becomes too specialized for the training data, etc. In this dissertation, the annotation study used only 120 texts, which proved sufficient to address the research questions, but future research would benefit from a larger, dimensionally annotated dataset. For example, instead of manually collecting dimensionally annotated data, Large Language Models (LLMs) could be used. There is no denying that LLMs can be a promising solution for annotating the data: scalability for handling large amounts of data, cost-effectiveness compared to manual annotation, and the potential for higher agreement scores when appropriately guided, see Gilardi et al. (2023), Karjus (2023), Ziems et al. (2024). However, LLMs are not without limitations. For

example, biases inherent in their training data may be reflected in their annotations, and LLMs may struggle to grasp the context, leading to misinterpretations. One option would be to use the annotators' reasoning to guide the LLMs into automatically annotating new unseen texts. Combining LLMs and human expertise could provide insights into cases where they disagree, which could be used to identify areas where the framework needs improvement or where specific annotation guidelines need clarification. In addition, the design of the annotation study set limits on text length. It would be interesting to see if the shorter texts are more pronounced in terms of their dimensional salience, while the longer texts have less dimensional salience.

This dissertation has provided an initial validation of the Dimensional Text Model through a case study. The annotation study produced a small corpus where each text has dimension-specific annotations using a four-point Likert scale, which could be extended using the power of the LLMs. The extended dataset is then used to fine-tune the pre-trained models, such as BERT-like models (Devlin et al., 2019, Liu et al., 2019) to classify and compare the results with pre-labelled corpora, such as CORE for English (Egbert et al. 2015) and Finnish (Laippala et al. 2019). The DTM could provide a complementary view of register variation in already established annotated corpora, and ultimately classify corpora whose composition is unknown.

7.3. Conclusions

This dissertation proposed the Dimensional Text Model (DTM), a hierarchical framework, to address the challenges associated with the existing taxonomies for classifying Web corpora. The theoretical foundations of DTM are a synthesis of the Multidimensional analysis of Biber (1988) and the Functional Text Dimensions of Sharoff (2018) which represent pivotal methodological approaches that have significantly advanced corpus-based research on register variation. The DTM proposed a set of 12 latent dimensions comprising *abstractness*, *affectivity*, *argumentativity*, *impersonality*, *interactivity*, *instructability*, *formality*, *complexity*, *spontaneity*, *information density*, *temporality* and *subjectivity*. These dimensions manifest through co-occurring linguistic features (linguistic profiles) that serve distinct communicative functions. A single dimension is not sufficient to define a register. Instead, it describes a particular characteristic of a register, e.g., characterizing academic language by its complexity, informativeness, and abstractness.

The primary objectives of the dissertation were to determine whether the theoretically proposed twelve dimensions could be reliably recognized and distinguished by annotators when applied to real-world data and to investigate the differences between the linguistic profiles of these dimensions. To measure whether the dimensions were distinguishable, an annotation study was conducted in which the annotators were asked to rate the level of salience of the proposed dimensions

on a four-point Likert scale in a collection of 120 texts gathered from the EtTenTen Web corpus. Inter-annotator agreement ranged from fair to substantial, with more than half of the dimensions achieving at least moderate agreement. These results suggest that the Dimensional Text Model can differentiate texts based on the identified dimensions. The analysis showed that all dimensions had a unique linguistic profile, indicating that each dimension has a distinct communicative function within the framework. While the linguistic profiles of all 12 dimensions were distinct, the aim was also to identify profiles that had a unique feature exclusive to that particular dimension. The results uncovered that out of the twelve dimensions, the profiles of five dimensions included distinct features. Thus, most of the linguistic profiles of the dimensions had a high degree of overlap in the distribution of features. In addition, the correlation analysis showed that many of the dimensions were strongly positively correlated with each other, and the factor analysis showed that the dimensions were not isolated, but rather interacted and influenced each other in more ways than expected. The Dimensional Text Model is not intended to replace existing taxonomies for automatic register classification, but rather to provide a complementary approach to the study of register variation.

BIBLIOGRAPHY

Aava, K. (2004). Narratiiv müüdiloome teenistuses: arhetüübid meediatekstides, *in* R. Kasik (ed.), *Tekstid ja taustad III. Lingvistiline tekstianalüüs*, Vol. 28 of *Tartu Ülikooli eesti keele õppetooli toimetised*, Tartu Ülikooli Kirjastus, Tartu, pp. 10–32.

Akoglu, H. (2018). User's guide to correlation coefficients, *Turkish Journal of Emergency Medicine* **18**: 91–93.

Artstein, R. & Poesio, M. (2008). Inter-coder agreement for computational linguistics, *Computational Linguistics* **34**(4): 555–596.
URL: *https://doi.org/10.1162/coli.07-034-R2*

Asheghi, N. R., Sharoff, S. & Markert, K. (2016). Crowdsourcing for Web genre annotation, *Language Resources and Evaluation* **50**(3): 603–641.
URL: *https://doi.org/10.1007/s10579-015-9331-6*

Beaudouin, V., Fleury, S. & Pasquier, M. (2002). Analyzing Internet Uses Through a Description of Web Pages Visited. Personal and Business Websites, *Réseaux* **116**(6): 9–51.
URL: *https://www.cairn-int.info/journal-reseaux-2002-6-page-19.htm*

Berninger, V. F., Kim, Y. & Ross, S. (2008). Building a document genre corpus: a profile of the KRYS I corpus.

Besnier, N. (1988). The Linguistic Relationships of Spoken and Written Nukulaelae Registers, *Language* **64**(4): 707–736.
URL: *http://www.jstor.org/stable/414565*

Bhatia, V. K. (2002). Applied genre analysis: A multiperspective model, *Ibérica* **4**: 3–19.

Biber, D. (1986). Spoken and written textual dimensions in English: resolving the contradictory findings, *Language* **2**: 384–414.

Biber, D. (1988). Variation across Speech and Writing, Cambridge University Press.

Biber, D. (1995). Dimensions of Register Variation: A Cross-Linguistic Comparison, Cambridge University Press.

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style*, Cambridge Textbooks in Linguistics, Cambridge University Press.

Biber, D., Davies, M., Jones, J. K. & Tracy-Ventura, N. (2006). Spoken and written register variation in Spanish: A multi-dimensional analysis, *Corpora* **1**(1): 1–37.
URL: *https://doi.org/10.3366/cor.2006.1.1.1*

Biber, D. & Egbert, J. (2018). *Register Variation Online*, Cambridge University Press.

Biber, D., Egbert, J. & Keller, D. (2020). Reconceptualizing register in a continuous situational space, *Corpus Linguistics and Linguistic Theory* **16**(3): 581–616.
URL: *https://doi.org/10.1515/cllt-2018-0086*

Biber, D., Egbert, J., Keller, D. & Wizner, S. (2021). *Extending text-linguistic studies of register variation to a continuous situational space: Case studies from the web and natural conversation*, Studies in Corpus Linguistics, John Benjamins Publishing Company, Netherlands, pp. 19–49. Publisher Copyright: © 2021 John Benjamins Publishing Company.

Biber, D. & Hared, M. (1992). Dimensions of register variation in Somali, *Language Variation and Change* **4**(1): 41–75.

Biggs, J. & Madnani, N. (2022). Factoranalyzer documentation, `https://factor-analyzer.readthedocs.io/en/latest/index.html`. Accessed: 2024-24-06.

Brown, P. & Fraser, C. (1979). Speech as a marker of situation, *Social markers in speech* pp. 33–62.

Bulygin, M. V. & Sharoff, S. (2018). Using machine translation for automatic genre classification in arabic, *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018"*.

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature, *Spoken and written language: Exploring orality and literacy* .

Clarke, I. & Grieve, J. (2019). Stylistic variation on the Donald Trump Twitter account: A linguistic analysis of tweets posted between 2009 and 2018, *PLOS ONE* **14**(9): 1–27.
URL: *https://doi.org/10.1371/journal.pone.0222062*

Cohen, J. (1960). A coefficient of agreement for nominal scales, *Educational and Psychological Measurement* **20**(1): 37–46.
URL: *https://doi.org/10.1177/001316446002000104*

Crossley, S. A. & Louwerse, M. M. (2007). Multi-dimensional register classification using bigrams, *International Journal of Corpus Linguistics* **12**(4): 453–478.

Crowston, K., Kwasnik, B. H. & Rubleske, J. (2010). Problems in the use-centered development of a taxonomy of web genres, *Genres on the Web*, Text, Speech and Language Technology, Springer Netherlands.

Crowston, K., Kwaśnik, B. & Rubleski, J. (2011). Problems in the use-centered development of a taxonomy of web genres, *in* A. Mehler, S. Sharoff & M. Santini (eds), *Genres on the Web: Computational Models and Empirical Studies*, Vol. 42 of *Text, Speech and Language Technology*, Springer Publishing Company, Dordrecht, pp. 69–84.

Crystal, D. (2001). *Language and the Internet*, Cambridge University Press.

Cvrček, V., Komrsková, Z., Lukeš, D., Poukarová, P., Řehořková, A., Zasina, A. J. & Benko, V. (2020). Comparing web-crawled and traditional corpora, *Lang. Resour. Eval.* **54**(3): 713–745.
URL: *https://doi.org/10.1007/s10579-020-09487-4*

de Beaugrande, R. & Dressler, W. (1981). *Introduction to Text Linguistics*, London & New York: Longman.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.
URL: *https://arxiv.org/abs/1810.04805*

Dewe, J., Karlgren, J. & Bretan, I. (1998). Assembling a balanced corpus from the internet, *11th Nordic Computational Linguistics Conference*, Copenhagen, Denmark.

Dunn, O. J. (1961). Multiple comparisons among means, *Journal of the American Statistical Association* **56**: 52–64.

Egbert, J., Biber, D. & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification, *Journal of the Association for Information Science and Technology* **66**(9): 1817–1831.
URL: *https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23308*

Erelt, M. (2017). *Eesti keele süntaks ['The syntax of Estonian']*, Tartu ülikooli Kirjastus, chapter Sekundaartarindiga laused ['Sentences with secondary constructions'], pp. 756–840.

etTenTen: Corpus of the Estonian Web (2021). `https://www.sketchengine.eu/ettenten-estonian-corpus/#toggle-id-3`. Accessed: 2024-16-06.

Fang, A. & Cao, J. (2010). Use of terms and term-related units as feature sets for automatic text classification, *CEUR Workshop Proceedings* **673**.

Finn, A. & Kushmerick, N. (2006). Learning to classify documents according to genre, *Journal of the American Society for Information Science and Technology* **57**(11): 1506–1518.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.20427*

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters, *Psychological Bulletin* **76**(5): 378–382.
URL: *https://doi.org/10.1037/h0031619*

Friginal, E. (2008). Linguistic variation in the discourse of outsourced call centers, *Discourse Studies* **10**(6): 715–736.
URL: *http://www.jstor.org/stable/24049379*

Gailit, K. G. (2023). *Leksikonide ja kaalude lisamine veebitekstide formaalsuse ja spontaansuse dimensioonide hindamise mudeli arendamiseks ['Lexicons and feature weights as an addition to improve the evaluation of the dimensions of formality and spontaneity of online texts']*, Master's thesis, University of Tartu.

Gilardi, F., Alizadeh, M. & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks, *Proceedings of the National Academy of Sciences* **120**(30).
URL: *http://dx.doi.org/10.1073/pnas.2305016120*

Gries, S., Newman, J. & Shaoul, C. (2011). N-grams and the clustering of registers, *The ELR Journal* **5**.

Grieve, J. (2014). *A Multi-Dimensional analysis of regional variation in American English*, pp. 3–34.

Grieve, J., Biber, D., Friginal, E. & Nekrasova, T. (2011). *Variation Among Blogs: A Multi-dimensional Analysis*, Springer Netherlands, pp. 303–322.
URL: *https://doi.org/10.1007/978-90-481-9178-9_14*

Halliday, M. (1985). *Spoken and Written Language*, Oxford University Press.

Hennoste, T., Metsalng, H., Habicht, K., Jürine, A., Laanesoo, K. & Ogren, D. (2015). Üldküsimuse vorm ja funktsioonid läbi nelja sajandi ja kuue tekstiliigi ['Forms and functions of polar questions across four centuries and six text types'], *Emakeele Seltsi aastaraamat* **61**: 80–109.

Hennoste, T., Metslang, H., Habicht, K. & Prillop, K. (2021). Kuue (inter)subjektiivsuspartikli kasutus eesti keele registries ['The use of six (inter)subjectivity particles in Estonian registers'], *Emakeele Seltsi aastaraamat* **66**: 91–123.

Hennoste, T., Prillop, K., Habicht, K., Metslang, H., Laanesoo, K., Pärismaa, L., Pärt, E., Rumm, A., Rääbis, A. & Simmul, C. E. (2022). Komplementlausega predikaatidel põhinevate diskursusemarkerite kasutus eri registries ['Complement-taking predicate markers in different registers in Estonian'], *Keel ja Kirjandus* **1–2**: 130–150.

Heylighen, F. & Dewaele, J.-M. (2002). Variation in the contextuality of language: An empirical measure, *Foundations of Science* **7**(3): 293–340.

Holm, S. (1979). A simple sequentially rejective multiple test procedure, *Scand. J. Statist.* **6**(2): 65–70.

Hymes, D. (1974). *Foundations in Sociolinguistics: An Ethnographic Approach*, Communications, Linguistics, Anthropology, University of Pennsylvania Press, Incorporated.

Irvine, J. T. (1979). Formality and informality in communicative events, *American Anthropologist* **81**(4): 773–790.
URL: *https://doi.org/10.1525/aa.1979.81.4.02a00020*

Jakubíček, M., Kovář, V., Rychlý, P. & Suchomel, V. (2020). Current challenges in web corpus building, *in* A. Barbaresi, F. Bildhauer, R. Schäfer & E. Stemle (eds), *Proceedings of the 12th Web as Corpus Workshop*, European Language Resources Association, Marseille, France, pp. 1–4.
URL: *https://aclanthology.org/2020.wac-1.1*

Kanaris, I. & Stamatatos, E. (2009). Learning to recognize webpage genres, *Information Processing and Management* **45**(5): 499–512.
URL: *https://www.sciencedirect.com/science/article/pii/S0306457309000491*

Karjus, A. (2023). Machine-assisted mixed methods: augmenting humanities and social sciences with artificial intelligence.

Kasik, R. (2000). Reklaamikeel tekstiliigina, *in* T. Hennoste (ed.), *Eesti keele allkeeled*, Vol. 16 of *Tartu Ülikooli eesti keele õppetooli toimetised*, Tartu Ülikool, Tartu, pp. 111–127.

Kasik, R. (2004). Interpersonaalse tähenduse konstrueerimine. Ühe intervjuu analüüs, *in* R. Kasik (ed.), *Tekstid ja taustad III. Lingvistiline tekstianalüüs*, Vol. 28 of *Tartu Ülikooli eesti keele õppetooli toimetised*, Tartu Ülikooli Kirjastus, Tartu, pp. 33–50.

Kasik, R. (2007). *Sissejuhatus tekstiõpetusse ['Introduction to Text Studies']*, Tartu Ülikooli Kirjastus, Tartu.

Kasik, R. (2008). Uudistekstide struktuur ja keelekasutus, *in* R. Kasik (ed.), *Tekstid ja taustad V. Meediatekstide keelekasutus ja selle sotsiokultuurilised taustad*, Tartu Ülikooli Kirjastus, Tartu, pp. 44–64.

Katinskaya, A. & Sharoff, S. (2015). Applying multi-dimensional analysis to a Russian webcorpus: Searching for evidence of genres, *in* J. Piskorski, L. Pivovarova, J. Šnajder, H. Tanev & R. Yangarber (eds), *The 5th Workshop on Balto-Slavic Natural Language Processing*, INCOMA Ltd. Shoumen, BULGARIA, pp. 65–74.
URL: *https://aclanthology.org/W15-5311*

Kerge, K. (2000). Kirjakeel ja igapäevakeel, *in* T. Hennoste (ed.), *Eesti keele allkeeled : [9.-10.12.1999 Tartu Ülikoolis toimunud seminari kogumik]*, Vol. 16 of *Tartu Ülikooli eesti keele õppetooli toimetised*, Tartu Ülikool, Tartu, pp. 75–110.

Kerge, K. (2003a). Autori stiil ja allkeele tekst, *in* R. Kasik (ed.), *Tekstid ja taustad II. Tekstianalüüsi vaatepunkte*, Vol. 26 of *Tartu Ülikooli eesti keele õppetooli toimetised*, Tartu Ülikooli Kirjastus, Tartu, p. 59–89.

Kerge, K. (2003b). *Keele variatiivsus ja mine-tuletus allkeelte süntaktilise keerukuse tegurina*, Ph.D. disseration, Tallinna Pedagoogikaülikool.

Kerge, K. (2004). Veebikommentaariumi mitmetahuline maailm, *in* R. Kasik (ed.), *Tekstid ja taustad III. Lingvistiline tekstianalüüs*, Vol. 28 of *Tartu Ülikooli eesti keele õppetooli toimetised*, Tartu Ülikooli Kirjastus, Tartu, pp. 51–73.

Kerge, K. (2009). Kirjažanrite keeleparameetrid mitme tekstiliigi taustal ['Linguistic parameters of letter genres with regard to oral and written language'], *Emakeele Seltsi aastaraamat* **55**: 32–62.

Kerge, K. & Pajupuu, H. (2010). Text-types in speech technology and language teaching, *in* J. L. Bueno Alonso et al. (eds), *Analizar datos > Describir variación / Analysing data > Describing variation*, Universidade de Vigo, Servizo de Publicacións, Vigo, pp. 380–390.

Kerge, K., Pajupuu, H. & Altrov, R. (2007). Tekst, kontekstuaalsus ja kultuur ['Text, contextuality and culture'], *Keel ja Kirjandus* **nr 8**: 624–637.

Kerge, K., Pajupuu, H., Tamuri, K. & Meier, H. (2008). Kõnetehnoloogia vajab žanrilist lähenemist ['Speech technology needs a genre-based approach'], *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics* **4**: 53–65.

Kilgarriff, A. & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus, *Computational Linguistics* **29**(3): 333–348.
URL: *https://aclanthology.org/J03-3001*

Kim, Y.-J. & Biber, D. (1994). A Corpus-Based Analysis Of Register Variation In Korean, *Sociolinguistic Perspectives On Register*, Oxford University Press.
URL: *https://doi.org/10.1093/oso/9780195083644.003.0008*

Kongot, L. (2013). Kui x otsib y-it eesmärgil z ehk mehed ja naised tutvumiskuulutustes, *Oma Keel* **2**: 32–39.

Koppel, K. & Kallas, J. (2022). Estonian National Corpus 2013–2021: The largest collection of Estonian language data ['Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu'], *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics* **18**: 207–228.

Krippendorff, K. (2004). *Content Analysis: An Introduction to Its Methodology*, Content Analysis: An Introduction to Its Methodology, Sage.
URL: *https://books.google.ee/books?id=q657o3M3C8cC*

Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association* **47**(260): 583–621.
URL: *https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441*

Kuldnokk, K. (2011). *Militaarne retoorika. Argumentatsioon ja keeleline mõjutamine Eesti kaitsepoliitilises diskursuses ['Military rhetoric. The influence of language used by the media in the discourse on Estonian Defence policy.']*, Dissertationes philologiae Estonicae Universitatis Tartuensis, Tartu Ülikooli Kirjastus, Tartu.

Kulkarni, V., Al-Rfou, R., Perozzi, B. & Skiena, S. (2015). Statistically significant detection of linguistic change, *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, p. 625–635.
URL: *https://doi.org/10.1145/2736277.2741627*

Kutuzov, A., Øvrelid, L., Szymanski, T. & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: A survey, *in* E. M. Bender, L. Derczynski & P. Isabelle (eds), *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, pp. 1384–1397.
URL: *https://aclanthology.org/C18-1117*

Kuzman, T., Rupnik, P. & Ljubešić, N. (2022). The GINCO training dataset for web genre identification of documents out in the wild, *in* N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk & S. Piperidis (eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 1584–1594.
URL: *https://aclanthology.org/2022.lrec-1.170*

Laippala, V., Kyllönen, R., Egbert, J., Biber, D. & Pyysalo, S. (2019). Toward multilingual identification of online registers, *in* M. Hartmann & B. Plank (eds), *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, Linköping University Electronic Press, pp. 292–297.
URL: *https://aclanthology.org/W19-6130*

Laippala, V., Salmela, A., Rönnqvist, S., Aji, A. F., Chang, L.-H., Dhifallah, A., Goulart, L., Kortelainen, H., Pàmies, M., Prina Dutra, D., Skantsi, V., Sutawika, L. & Pyysalo, S. (2022). Towards better structured and less noisy web data: Oscar with register annotations, *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, Association for Computational Linguistics, pp. 215–221.
URL: *https://aclanthology.org/2022.wnut-1.23*

Lamb, W. (2002). *Speaking and Writing in Scottish Gaelic: A Study of Register Variation in an Endangered Langauge*, Ph.D. disseration, University of Edinburgh.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data, *Biometrics* **33**(1): 159–174.
URL: *http://www.jstor.org/stable/2529310*

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

Lepajõe, K. (2011). *Kirjand kui tekstiliik. Riigieksamikirjandite tekstuaalsed, retoorilised ja diskursiivsed omadused ['The Estonian high school matriculation examination essay as a text genre: textual, rhetorical and discursive features']*, Dissertationes philologiae Estonicae Universitatis Tartuensis, Tartu Ülikooli Kirjastus, Tartu.

Liimatta, A. (2019). Exploring register variation on Reddit: A multidimensional study of language use on a social media website, *Register studies* **1**(2): 269–295.

Lindström, L. (2005). *Finiitverbi asend lauses. Sõnajärg ja seda mõjutavad tegurid suulises eesti keeles ['The position of the finite verb in a clause: word order and the factors affecting it in spoken Estonian.']*, Dissertationes philologiae Estonicae Universitatis Tartuensis, Tartu Ülikooli Kirjastus, Tartu.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach.
URL: *https://arxiv.org/abs/1907.11692*

Madjarov, G., Vidulin, V., Dimitrovski, I. & Kocev, D. (2019). Web genre classification with methods for structured output prediction, *Information Sciences* **503**: 551–573.
URL: *https://www.sciencedirect.com/science/article/pii/S0020025516318965*

Mandra, K. (2009). Haldusteksti struktuurist Tartu linnavalitsuse korralduste näitel ['The structure of the administrative texts on the example of Tartu city government'], *in* H. Metslang, M. Langemets, M.-M. Sepper & R. Argus (eds), *Eesti Rakenduslingvistika Ühingu aastaraamat 5*, Eesti Keele Sihtasutus, Tallinn, pp. 131–141.

Matthiessen, C. M. (1993). Register in the round: Diversity in a unified theory of register analysis, *in* M. Ghadessy (ed.), *Register analysis: theory and practice*, Pinter, London, pp. 221–292.

Mehler, A., Sharoff, S. & Santini, M. (2010). *Genres on the Web: Computational Models and Empirical Studies*, Vol. 42.

Meier, H. (2002). Olulisi aspekte tekstitüübi võrdluses, *in* R. Kasik (ed.), *Tekstid ja taustad: artikleid tekstianalüüsist*, Vol. 23 of *Tartu Ülikooli eesti keele õppetooli toimetised*, Tartu, pp. 101–114.

Meier, H. (2003). Essee allkeelte võrdluses, *in* R. Kasik (ed.), *Tekstid ja taustad II. Tekstianalüüsi vaatepunkte*, Vol. 26 of *Tartu Ülikooli eesti keele õppetooli toimetised*, Tartu Ülikooli Kirjastus, Tartu, pp. 116–135.

Meyer zu Eissen, S. & Stein, B. (2004). Genre classification of web pages, *in* S. Biundo, T. Frühwirth & G. Palm (eds), *KI 2004: Advances in Artificial Intelligence*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 256–269.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space.

Nivre, J., Zeman, D., Ginter, F. & Tyers, F. (2017). Universal Dependencies, *in* A. Klementiev & L. Specia (eds), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Association for Computational Linguistics, Valencia, Spain.
URL: *https://aclanthology.org/E17-5001*

Pajupuu, H., Altrov, R. & Pajupuu, J. (2016). Identifying polarity in different text types, *Folklore. Electronic Journal of Folklore* (64): 25–42.
URL: *https://doi.org/10.7592/FEJF2016.64.polarity*

Pajupuu, H., Kerge, K. & Altrov, R. (2012). Lexicon-based detection of emotion in different types of texts: Preliminary remarks, *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics* (8): 171–184.
URL: *https://doi.org/10.5128/ERYa8.11*

Pajusalu, R. (2017). *Eesti keele süntaks ['The syntax of Estonian']*, Tartu ülikooli Kirjastus, chapter Määratlejad. ['Determiners'], pp. 382–385.

Parodi, G. (2007). Variation across registers in Spanish: exploring the El Grial PUCV corpus, *Working with Spanish corpora* pp. 11–53.

Passonneau, R. J., Ide, N., Su, S. & Stuart, J. (2014). Biber Redux: Reconsidering Dimensions of Variation in American English, *in* J. Tsujii & J. Hajic (eds), *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pp. 565–576.
URL: *https://aclanthology.org/C14-1054*

Pritsos, D. & Stamatatos, E. (2018). Open set evaluation of web genre identification, *Language Resources and Evaluation* **52**.

Puksand, H. & Kerge, K. (2011). Õpiteksti analüüs kirjaoskuse omandamise kontekstis ['Analysis of learning texts in the context of literacy acquisition'], *Emakeele Seltsi aastaraamat* **57**: 162–217.

Purvis, T. M. (2008). *A Linguistic and Discursive Analysis of Register Variation in Dagbani*, Ph.D. disseration, Indiana University.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. (2020). Stanza: A Python Natural Language Processing toolkit for many human languages.

Rehm, G. (2002). Towards automatic Web genre identification: A corpus-based approach in the domain of academia by example of the Academic's Personal Homepage, *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, pp. 1143–1152.

Rehm, G., Santini, M., Mehler, A., Braslavski, P., Gleim, R., Stubbe, A., Symonenko, S., Tavosanis, M. & Vidulin, V. (2008). Towards a reference corpus of Web genres for the evaluation of genre identification systems, *in* N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco.
URL: *http://www.lrec-conf.org/proceedings/lrec2008/pdf/94_paper.pdf*

Reid, T. B. W. (1956). Linguistics, structuralism, and philology, *Archivum Linguisticum* **8**: 28–37.

Reinsalu, R. (2019). *Juhendavad haldustekstid žanriteoreetilises raamistikus ['Instructive Administrative Texts in a Genre Theory Framework']*, Dissertationes philologiae Estonicae Universitatis Tartuensis, Tartu Ülikooli Kirjastus, Tartu.

Repo, L., Skantsi, V., Rönnqvist, S., Hellström, S., Oinonen, M., Salmela, A., Douglas, B., Egbert, J., Pyysalo, S. & Laippala, V. (2021). Beyond the English Web: Zero-shot cross-lingual and lightweight monolingual classification of registers, *Proceedings of the 16th Conference of the European Chapter of the Associationfor Computational Linguistics: Student Research Workshop*, pp. 183–191.
URL: *https://aclanthology.org/2021.eacl-srw.24/*

Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J. & Liu, X. (2001). Genre based navigation on the Web, *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, pp. 10 pp.–.

Rääbis, A. (2009). *Eesti telefonivestluste sissejuhatus: struktuur ja suhtlusfunktsioonid ['Openings in Estonian Telephone Conversations: Structure and Interactional Functions']*, Dissertationes philologiae Estonicae Universitatis Tartuensis, Tartu Ülikooli Kirjastus, Tartu.

Rönnqvist, S., Skantsi, V., Oinonen, M. & Laippala, V. (2021). Analyzing the unrestricted Web: The Finnish corpus of online registers, *Nordic Conference on Computational Linguistics*, p. 157–165.

Salla, S. (2002). Jututuba kui võrgusuhtlusvorm, *in* R. Kasik (ed.), *Tekstid ja taustad I*, Vol. 23 of *Tartu Ülikooli eesti keele õppetooli toimetised*, Tartu Ülikooli Kirjastus, Tartu, pp. 128–156.

Santini, M. (2007). *Automatic Identification of Genre in the Web Pages*, Ph.D. dissertation, University of Brighton, Brighton.

Santini, M. (2010). Cross-testing a genre classification model for the web, *Genres on the Web*, Springer, pp. 87–128.

Santini, M., Mehler, A. & Sharoff, S. (2010). *Riding the Rough Waves of Genre on the Web*, Springer Netherlands, pp. 3–30.
URL: *https://doi.org/10.1007/978-90-481-9178-9_1*

Sarapuu, K. (2008). Suhtlustasandi tähendused ajalehtede juhtkirjades, *in* R. Kasik (ed.), *Tekstid ja taustad V. Meediatekstide keelekasutus ja selle sotsiokultuurilised taustad*, Tartu Ülikooli Kirjastus, Tartu, pp. 100–192.

Sardinha, T. B. (2022). A text typology of social media, *Register Studies* **4**(2): 138–170.
URL: *https://www.jbe-platform.com/content/journals/10.1075/rs.22008.ber*

Sardinha, T. B., Kauffmann, C. & Acunzo, C. M. (2014). A multi-dimensional analysis of register variation in Brazilian Portuguese, *Corpora* **9**(2): 239–271.
URL: *https://doi.org/10.3366/cor.2014.0059*

Schäfer, R. & Bildhauer, F. (2013). Web corpus construction, *Synthesis Lectures on Human Language Technologies* **6**: 1–145.

Shakir, M. & Deuber, D. (2019). A multidimensional analysis of Pakistani and U.S. English blogs and columns, *English World-Wide* **40**(1): 1–24.
URL: *https://www.jbe-platform.com/content/journals/10.1075/eww.00020.sha*

Sharoff, S. (2010). In the garden and in the jungle, *in* A. Mehler, S. Sharoff & M. Santini (eds), *Genres on the Web: Computational Models and Empirical Studies*, Springer Publishing Company, Dordrecht, pp. 149–166.

Sharoff, S. (2018). Functional Text Dimensions for the annotation of web corpora, *Corpora* **13**: 65–95.

Sharoff, S. (2021). Genre Annotation for the Web: text-external and text-internal perspectives, *Register Studies* **3**(1): 1–32.

Sharoff, S., Wu, Z. & Markert, K. (2010). The Web Library of Babel: Evaluating genre collections, *in* N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (eds), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, European Language Resources Association (ELRA), Valletta, Malta.
URL: *http://www.lrec-conf.org/proceedings/lrec2010/pdf/28_Paper.pdf*

Sirel, R. (2013). Meetodeid tekstide leksikaalsete ja grammatiliste erinevuste tuvastamiseks meditsiiniliste tarbetekstide näitel ['Methods for identifying lexical and grammatical differences in medical applied texts'], *Eesti Rakenduslingvistika Ühingu aastaraamat* **9**: 265–278.
URL: *http://dx.doi.org/10.5128/ERYa9.17*

Skantsi, V. & Laippala, V. (2023). Analyzing the unrestricted web: The Finnish corpus of online registers, *Nordic Journal of Linguistics* p. 1–31.

Song, J., Qu, Y., Zhu, X., Wang, X. & Zhang, Y. (2021). A multi-dimensional approach to register variations in Mandarin Chinese, *Glottometrics* **51**: 39–69.

Sorokin, A., Katinskaia, A. & Sharoff, S. (2014). Associating symptoms with syndromes: Reliable genre annotation for a large Russian webcorpus, *Proc. Dialogue, Russian International Conference on Computational Linguistics*, pp. 646–659.

Stubbe, A., Ringlstetter, C. & Schulz, K. U. (2007). Genre as noise – noise in genre, *Proceedings of the IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text Data*.

Suchomel, V. (2020). *Better Web Corpora For Corpus Linguistics And NLP*, Ph.D. disseration, Masaryk University, Faculty of Informatics Brno.
URL: *https://theses.cz/id/moc9kj/*

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*, Cambridge University Press, Cambridge.

Tragel, I. (2003). *Eesti keele tuumverbid. ['Estonian core verbs']*, Ph.D. disseration, University of Tartu.

Vaik, K. & Muischnek, K. (2018). Eestikeelsete veebitekstide automaatne liigitamine ['Classifying Estonian Web texts'], *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics* **14**: 215–229.

Vaik, K., Sirts, K. & Muischnek, K. (2020). Dimensionaalne tekstimudel. Teoreetiline ülevaade ['Dimensional Text Model. A Theoretical Overview.'], *Keel ja Kirjandus* **10**: 875–898.

Vidulin, V., Lustrek, M. & Gams, M. (2009). Multi-label approaches to Web genre identification, *JLCL* **24**: 97–114.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nature Methods* **17**: 261–272.

Werlich, E. (1983). *A Text Grammar of English*, 2nd edn, Quelle & Meyer, Heidelberg.

Wu, Z., Markert, K. & Sharoff, S. (2010). Fine-grained genre classification using structural learning algorithms, *in* J. Hajič, S. Carberry, S. Clark & J. Nivre (eds), *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Uppsala, Sweden, pp. 749–759.
URL: *https://aclanthology.org/P10-1077*

Zhang, M. (2016). A multidimensional analysis of metadiscourse markers across written registers, *Discourse Studies* **18**(2): 204–222.

Ziems, C., Held, W., Shaikh, O., Chen, J., Zhang, Z. & Yang, D. (2024). Can Large Language Models transform computational social science?

Appendix A. GUIDELINES FOR THE ANNOTATORS

Dear Participant,
The aim of this experiment is to construct a corpus that can be used for exploratory work in order to study the relationship between linguistic features and different texts.

This is session number X, so there will be four sessions. In this session, we will look at attributes such as DIM_y, DIM_z, and DIM_w. Their definitions*:

- DIM_y = definition for DIM_y
- DIM_z = definition for DIM_z
- DIM_w = definition for DIM_w

*No need to remember these; they will appear on every page.

Process:
Each session has 60 questions. In each question, two different texts are displayed simultaneously, TEXT A and TEXT B. Your task is to judge whether DIM_y, DIM_z, and DIM_w are more characteristic of TEXT A or TEXT B (or not characteristic of either of those). There are two evaluation panels below the text panel:

- in the **left panel**, you have to choose, one by one, whether the dimension is more characteristic of TEXT A or TEXT B (or NONE).
- in the **right panel**, you have to choose, one by one, the "strength" of a dimension on the scale of 'weak', 'moderate', 'strong' or 'non-existent' (choose the latter if you chose 'None' in the left panel).

There are no right or wrong answers. The only aim is to collect judgements based on your intuition. Do not get stuck on the definitions, they are laconic but as informative as possible.

Appendix B. DIMENSION DEFINITIONS

SPONTAANSUS - see on iseloomulik sellistele tekstidele, kus esitatud tekst on vahetu ja toimetamata. Selle vastandiks võib pidada viimistletud teksti.
SPONTANEITY - is characteristic of texts that are immediate and unedited. Its opposite can be considered a polished text.

IMPERSONAALSUS - see on iseloomulik sellistele tekstidele, kus aktiivse tegija asemel on fookuses tegevus/sündmus/olukord vmt või see kellele/millele too tegevus/sündmus on suunatud.
IMPERSONALITY - is characteristic of texts where, instead of an active actor, the focus is on the action/event/situation or on whom/what this action/event is directed.

ARGUMENTATIIVSUS - see on iseloomulik sellistele tekstidele, kus esitatakse põhjendatud seisukohti, soovides sellega midagi väita, tõestada vmt.
ARGUMENTATIVITY - is characteristic of texts that present reasoned positions, seeking to assert, prove, or demonstrate something.

ABSTRAKTSUS - see omadus on iseloomulik sellistele tekstidele, kus kirjutatakse/räägitakse nähtustest ja ideedest, mida pole võimalik meeltega vahetult kogeda või mis on üldised ja mittekonkreetsed.
ABSTRACTNESS - is characteristic of texts that write/speak about phenomena and ideas that cannot be directly perceived by the senses or are general and non-specific.

INFORMATSIOONITIHEDUS - see omadus on iseloomulik sellistele tekstidele, mis sisaldavad tihedalt kokkupakitud infot.
INFORMATION DENSITY is characteristic of texts that contain tightly packed information.

AJALISUSE OLULISUS - see omadus on iseloomulik sellistele tekstidele, kus ajaline mõõde on oluline ning tekstis toimuvaid sündmusi saab paigutada ajatelljele.
TEMPORALITY - is characteristic of texts where the temporal dimension is important and the events in the text can be placed on a timeline.

AFEKTIIVSUS - see omadus on iseloomulik sellistele tekstidele, kus väljendatakse tundeid ja emotsioone (võib olla positiivne, negatiivne, sarkastiline jm). Selle omaduse vastandiks on neutraalsus.
AFFECTIVITY is characteristic of texts that express feelings and emotions (can be positive, negative, sarcastic, etc.). Its opposite is neutrality.

FORMAALSUS - see omadus on iseloomulik sellistele tekstidele, mis on oma laadilt ametlikud.
FORMALITY is characteristic of texts that are official.

INTERAKTIIVSUS - see omadus on iseloomulik sellistele tekstidele, kus on tüüpiliselt rohkem kui üks osaleja ning mida iseloomustab vastastikune suhtlus.
INTERACTIVITY is characteristic of texts that typically involve more than one participant and are characterized by mutual communication.

INSTRUEERIVUS - see omadus on iseloomulik sellistele tekstidele, kus lugejale antakse juhiseid teatud toimingute tegemiseks või kirjeldab mingisuguse protsessi toimumise etappe.
INSTRUCTIVITY - is characteristic of texts that provide the reader with instructions for performing certain actions or describe the stages of a process.

KEERUKUS - see omadus on iseloomulik sellistele tekstidele, millest arusaamine nõuab lisapingutust. Keerukus võib avalduda nii vormi kui sisu tasandil.
COMPLEXITY is characteristic of texts that require additional effort to understand. Complexity can be expressed in both form and content.

SUBJEKTIIVSUS - see omadus on iseloomulik sellistele tekstidele, kus on tuntav autori enda arvamus või hinnang.
SUBJECTIVITY is characteristic of texts where the author's own opinion or assessment is evident.

Appendix C. ANOVA AND *POST HOC* TEST RESULTS (RAW DATA)

This table only includes those linguistic features which were considered to be statistically significant (51 out of 85). The primary goal was to identify monotonic changes in the medians of relative frequencies of features across the dimensions, meaning a consistent increase or decrease from NE to W to S/M. Each feature row comprises the pairwise comparisons (S/M - W, S/M - NE, and W - NE), and the colour (light grey or dark grey) depends on comparing the relative medians of the S/M, W and NE groups separately. For example, Kruskal-Wallis and Dunn's test identified that a 'noun' has a monotonic change across pairwise comparisons. The direction of the monotonic change was determined based on the median of the relative frequency of a feature (e.g., regarding nouns, the median relative frequency increases towards the S/M group - 0.321 vs. 0.216). Thus, the overall reasoning based on these results was that the use of nouns increases as texts become more formal, impersonal, and information-dense (light grey), and the use of nouns decreases as texts become more affective, interactive, spontaneous and subjective (dark grey).

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
noun	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.335; 0.347	0.321; 0.328	0.302; 0.3	0.22; 0.209	0.212; 0.205	0.212; 0.198	0.224; 0.217
	W median; avg						0.299; 0.302	0.262; 0.263	0.236; 0.247	0.241; 0.246	0.227; 0.222	0.231; 0.237	0.263; 0.263
	NE median; avg						0.245; 0.242	0.216; 0.214	0.219; 0.23	0.308; 0.316	0.286; 0.289	0.296; 0.296	0.306; 0.311
adj	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.102; 0.111	0.096; 0.096		0.059; 0.064			0.061; 0.064
	W median; avg						0.076; 0.077	0.072; 0.072		0.072; 0.074			0.072; 0.074
	NE median; avg						0.067; 0.07	0.061; 0.067		0.084; 0.091			0.084; 0.091
propn	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg					0.02; 0.032	0.071; 0.088			0.014; 0.024		0.009; 0.013	0.014; 0.024
	W median; avg					0.038; 0.047	.079; 0.085			0.038; 0.06		0.008; 0.024	0.032; 0.041
	NE median; avg					0.076; 0.09	0.022; 0.038			0.079; 0.09		0.071; 0.077	0.087; 0.093
adv	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.054; 0.054	0.06; 0.066		0.126; 0.13		0.12; 0.127	0.118; 0.126
	W median; avg						0.058; 0.068	0.087; 0.093		0.113; 0.108		0.13; 0.132	0.113; 0.113
	NE median; avg						0.114; 0.114	0.114; 0.119		0.06; 0.065		0.074; 0.077	0.058; 0.061
intj	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg							0.0; 0.0		0.0; 0.002	0.0; 0.002	0.0; 0.003	0.0; 0.001
	W median; avg							0.0; 0.0		0.0; 0.0002	0.0; 0.001	0.002; 0.003	0.0; 0.001
	NE median; avg							0.0; 0.001		0.0; 0.0	0.0; 0.0	0.0; 0.0	0.0; 0.0

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
sconj	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg									0.029; 0.031			0.029; 0.03
	W median; avg									0.032; 0.034			0.023; 0.025
	NE median; avg									0.014; 0.015			0.017; 0.018
det	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.005; 0.005	0.01; 0.014		0.02; 0.021			
	W median; avg						0.015; 0.019	0.013; 0.016		0.019; 0.022			
	NE median; avg						0.019; 0.02	0.024; 0.025		0.009; 0.013			
punct	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg		0.118; 0.127				0.112; 0.11						0.151; 0.155
	W median; avg		0.125; 0.126				0.134; 0.14						0.153; 0.159
	NE median; avg		0.148; 0.148				0.151; 0.155						0.128; 0.133
pron	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.032; 0.033	0.026; 0.028	0.027; 0.032	0.089; 0.088	0.065; 0.075	0.065; 0.082	0.081; 0.085
	W median; avg						0.028; 0.036	0.048; 0.048	0.064; 0.064	0.06; 0.067	0.089; 0.093	0.082; 0.082	0.052; 0.05
	NE median; avg						0.068; 0.071	0.076; 0.08	0.051; 0.064	0.026; 0.033	0.037; 0.042	0.037; 0.041	0.026; 0.034
abbr	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg		0.054; 0.045					0.012; 0.023					
	W median; avg		0.008; 0.017					0.004; 0.007					
	NE median; avg		0.0; 0.006					0.002; 0.005					

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
TTR	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg										0.561; 0.539	0.555; 0.549	0.544; 0.553
	W median; avg										0.543; 0.549	0.547; 0.543	0.593; 0.587
	NE median; avg										0.633; 0.627	0.633; 0.629	0.642; 0.634
nominal	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.538; 0.524	0.48; 0.487		0.384; 0.379	0.385; 0.376	0.392; 0.383	0.383; 0.385
	W median; avg						0.457; 0.455	0.413; 0.414		0.395; 0.401	0.39; 0.395	0.403; 0.391	0.407; 0.41
	NE median; avg						0.405; 0.399	0.372; 0.379		0.471; 0.474	0.437; 0.448	0.452; 0.453	0.467; 0.468
avg word len	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						6.568; 6.602	5.925; 6.062	5.82; 5.908	4.957; 4.885	4.725; 4.82	4.63; 4.689	4.957; 4.92
	W median; avg						5.942; 5.819	5.331; 5.451	5.134; 5.317	5.416; 5.509	4.957; 5.123	5.043; 4.998	5.364; 5.319
	NE median; avg						5.107; 5.148	4.993; 5.005	4.645; 4.83	5.992; 6.069	5.624; 5.757	5.754; 5.835	5.986; 6.074
avg sent len	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						25.08; 23.573						
	W median; avg						18.835; 21.56						
	NE median; avg						15.59; 16.436						
hapax legomena	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg										0.451; 0.423		
	W median; avg										0.416; 0.421		
	NE median; avg										0.503; 0.512		

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
pron/ noun ratio	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.086; 0.102	0.082; 0.089	0.091; 0.116	0.386; 0.499	0.259; 0.461	0.29; 0.548	0.383; 0.468
	W median; avg						0.095; 0.129	0.167; 0.189	0.28; 0.318	0.246; 0.315	0.371; 0.508	0.347; 0.344	0.196; 0.251
	NE median; avg						0.274; 0.364	0.331; 0.467	0.219; 0.298	0.087; 0.109	0.122; 0.166	0.122; 0.153	0.089; 0.115
see as pron	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg									0.014; 0.014			0.017; 0.017
	W median; avg									0.017; 0.016			0.012; 0.013
	NE median; avg									0.0; 0.007			0.0; 0.008
1st pron sg	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.0; 0.0	0.0; 0.0	0.0; 0.0	0.004; 0.014	0.006; 0.013	0.005; 0.021	0.008; 0.016
	W median; avg						0.0; 0.001	0.0; 0.003	0.0; 0.008	0.0; 0.012	0.013; 0.018	0.002; 0.006	0.0; 0.004
	NE median; avg						0.0; 0.01	0.004; 0.012	0.004; 0.005	0.0; 0.0	0.0; 0.001	0.0; 0.002	0.0; 0.0
1st pron pl	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.0; 0.0	0.0; 0.001	0.0; 0.001	0.004; 0.005			0.003; 0.005
	W median; avg						0.0; 0.002	0.0; 0.001	0.004; 0.006	0.0; 0.006			0.004; 0.005
	NE median; avg						0.0; 0.004	0.004; 0.007	0.004; 0.005	0.0; 0.001			0.0; 0.0
2nd pron sg	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg									0.0; 0.004	0.0; 0.003	0.0; 0.004	0.0; 0.004
	W median; avg									0.0; 0.001	0.0; 0.003	0.002; 0.006	0.0; 0.001
	NE median; avg									0.0; 0.0	0.0; 0.0	0.0; 0.0	0.0; 0.0

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
2nd pron pl	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg										0.003; 0.003	0.0; 0.002	
	W median; avg										0.0; 0.0	0.0; 0.0	
	NE median; avg										0.0; 0.0	0.0; 0.0	
3rd pron sg	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg									0.009; 0.013			
	W median; avg									0.0; 0.004			
	NE median; avg									0.0; 0.005			
3rd pron pl	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg							0.0; 0.0		0.001; 0.004			0.0; 0.004
	W median; avg							0.0; 0.002		0.0; 0.003			0.0; 0.002
	NE median; avg							0.0; 0.004		0.0; 0.001			0.0; 0.001
core verb	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.014; 0.016	0.011; 0.013		0.032; 0.033	0.032; 0.031	0.036; 0.034	0.031; 0.033
	W median; avg						0.013; 0.017	0.026; 0.028		0.028; 0.027	0.032; 0.035	0.031; 0.034	0.021; 0.025
	NE median; avg						0.027; 0.029	0.032; 0.032		0.012; 0.017	0.017; 0.02	0.017; 0.02	0.014; 0.017
per voice	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.111; 0.114	0.111; 0.115		0.15; 0.152			0.148; 0.148
	W median; avg						0.116; 0.119	0.123; 0.127		0.133; 0.137			0.132; 0.132
	NE median; avg						0.144; 0.145	0.152; 0.152		0.112; 0.119			0.116; 0.121

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
imp voice	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.027; 0.029	0.018; 0.025		0.007; 0.009			0.009; 0.011
	W median; avg						0.015; 0.021	0.019; 0.021		0.012; 0.016			0.014; 0.017
	NE median; avg						0.011; 0.012	0.008; 0.011		0.018; 0.023			0.024; 0.024
1st prs verb	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.0; 0.0	0.0; 0.0		0.012; 0.017	0.015; 0.016	0.011; 0.017	0.015; 0.019
	W median; avg						0.0; 0.002	0.0; 0.005	0.006; 0.012	0.017; 0.016	0.017; 0.02	0.003; 0.009	0.006; 0.021
	NE median; avg						0.007; 0.013	0.02; 0.027	0.018; 0.016	0.0; 0.005	0.0; 0.007	0.0; 0.005	0.0; 0.0
2nd prs verb	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.0; 0.0		0.0; 0.004	0.006; 0.01	0.01; 0.015	0.012; 0.014	0.004; 0.009
	W median; avg						0.0; 0.0		0.0; 0.005	0.0; 0.001	0.003; 0.007	0.006; 0.007	0.0; 0.001
	NE median; avg						0.0; 0.013		0.014; 0.019	0.0; 0.007	0.0; 0.003	0.0; 0.00	0.0; 0.006
verbtype ratio	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.933; 1.225	1.375; 1.718					
	W median; avg						1.944; 2.226	2.421; 3.032					
	NE median; avg						2.0; 2.27	2.244; 2.299					
da-inf	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg				0.009; 0.00	0.025; 0.026							
	W median; avg				0.017; 0.02	0.021; 0.025							
	NE median; avg				0.026; 0.02	0.014; 0.012							

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
fin verb	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg		0.074; 0.081				0.081; 0.084	0.098; 0.093		0.126; 0.128	0.127; 0.129	0.133; 0.131	0.126; 0.125
	W median; avg		0.106; 0.103				0.105; 0.102	0.113; 0.117		0.116; 0.116	0.126; 0.125	0.123; 0.125	0.114; 0.116
	NE median; avg		0.119; 0.117				0.121; 0.121	0.127; 0.126		0.098; 0.101	0.105; 0.106	0.105; 0.105	0.102; 0.102
inf verb	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg				0.041; 0.04								
	W median; avg				0.058; 0.06								
	NE median; avg				0.069; 0.07								
supine	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg							0.0; 0.005		0.012; 0.014			0.01; 0.012
	W median; avg							0.007; 0.007		0.006; 0.007			0.005; 0.009
	NE median; avg							0.013; 0.013		0.004; 0.005			0.005; 0.006
pres tense	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg	0.1156; 0.11			0.031; 0.04	0.112; 0.107							
	W median; avg	0.1084; 0.1			0.088; 0.08	0.098; 0.098							
	NE median; avg	0.0707; 0.0			0.11; 0.105	0.05; 0.063							
past tense	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg				0.101; 0.09	0.022; 0.038							
	W median; avg				0.06; 0.057	0.037; 0.044							
	NE median; avg				0.03; 0.036	0.079; 0.073							

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
ind mood	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.081; 0.08	0.092; 0.09		0.111; 0.115			0.11; 0.113
	W median; avg						0.104; 0.098	0.107; 0.105		0.112; 0.111			0.111; 0.111
	NE median; avg						0.106; 0.105	0.117; 0.116		0.096; 0.09			0.097; 0.093
imp mood	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.0; 0.0			0.003; 0.005	0.009; 0.012	0.005; 0.01	0.003; 0.005
	W median; avg						0.0; 0.0			0.0; 0.001	0.0; 0.004	0.0; 0.002	0.0; 0.0
	NE median; avg						0.0; 0.01			0.0; 0.007	0.0; 0.002	0.0; 0.005	0.0; 0.006
neg polarity	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg					0.017; 0.019				0.019; 0.021	0.019; 0.022	0.021; 0.025	0.019; 0.02
	W median; avg					0.014; 0.015				0.02; 0.019	0.016; 0.019	0.015; 0.017	0.015; 0.017
	NE median; avg					0.004; 0.006				0.0; 0.005	0.008; 0.011	0.007; 0.01	0.0; 0.007
gen case	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.2; 0.205	0.166; 0.156		0.074; 0.083	0.063; 0.062	0.063; 0.057	0.081; 0.09
	W median; avg						0.15; 0.15	0.124; 0.122		0.118; 0.124	0.097; 0.105	0.086; 0.099	0.103; 0.113
	NE median; avg						0.084; 0.103	0.092; 0.1		0.16; 0.159	0.146; 0.142	0.15; 0.146	0.174; 0.16
ade case	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg									0.019; 0.021			0.016; 0.019
	W median; avg									0.018; 0.028			0.019; 0.023
	NE median; avg									0.036; 0.042			0.038; 0.043

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
nsubj cop	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.007; 0.008					0.028; 0.03	0.024; 0.027
	W median; avg						0.017; 0.02					0.035; 0.03	0.022; 0.024
	NE median; avg						0.02; 0.023					0.016; 0.017	0.011; 0.015
modal	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg									0.027; 0.028	0.03; 0.029	0.031; 0.03	0.026; 0.025
	W median; avg									0.019; 0.02	0.02; 0.025	0.017; 0.021	0.019; 0.021
	NE median; avg									0.007; 0.01	0.012; 0.015	0.012; 0.015	0.008; 0.012
csubj cop	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg												0.0; 0.002
	W median; avg												0.0; 0.002
	NE median; avg												0.0; 0.0
xcomp	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg							0.006; 0.007	0.008; 0.01	0.02; 0.02			
	W median; avg							0.013; 0.015	0.018; 0.019	0.016; 0.014			
	NE median; avg							0.019; 0.018	0.018; 0.021	0.008; 0.01			
obl	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg			0.0538; 0.05								0.069; 0.072	0.083; 0.084
	W median; avg			0.0898; 0.09								0.091; 0.088	0.09; 0.094
	NE median; avg			0.1048; 0.10								0.108; 0.111	0.11; 0.114

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
nmod	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg		0.168; 0.155				0.15; 0.165	0.121; 0.127		0.05; 0.055	0.049; 0.049	0.048; 0.045	0.053; 0.059
	W median; avg		0.118; 0.115				0.11; 0.118	0.083; 0.082		0.086; 0.089	0.071; 0.068	0.058; 0.064	0.071; 0.079
	NE median; avg		0.08; 0.08				0.054; 0.069	0.052; 0.061		0.12; 0.124	0.107; 0.106	0.107; 0.11	0.121; 0.123
nummod	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg				0.029; 0.032								
	W median; avg				0.01; 0.017								
	NE median; avg				0.007; 0.009								
amod	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg												0.033; 0.034
	W median; avg												0.042; 0.044
	NE median; avg												0.051; 0.053
voc	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg										0.0; 0.001		
	W median; avg										0.0; 0.0		
	NE median; avg										0.0; 0.0		
cop	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg						0.009; 0.011			0.025; 0.028		0.033; 0.034	0.027; 0.03
	W median; avg						0.018; 0.021			0.02; 0.025		0.035; 0.03	0.024; 0.026
	NE median; avg						0.022; 0.025			0.016; 0.017		0.016; 0.019	0.015; 0.016

FEATURES	GROUPS	INST	COMP	ABS	TEMP	ARG	FORM	IMP	INFO	AFF	INTER	SPONT	SUBJ
discourse	S/M - W												
	S/M - NE												
	W - NE												
	S/M median; avg							0.0; 0.0		0.0; 0.002	0.0; 0.002	0.0; 0.003	0.0; 0.001
	W median; avg							0.0; 0.0		0.0; 0.0	0.0; 0.001	0.002; 0.003	0.0; 0.001
	NE median; avg							0.0; 0.001		0.0; 0.0	0.0; 0.0	0.0; 0.0	0.0; 0.0

SISUKOKKUVÕTE

Väljaspool žanre: Dimensionaalne tekstimudel tekstide klassifitseerimiseks

Tänapäeval kasutatakse korpuslingvistikas ja keeletehnoloogias andmete allikana peamiselt veebist automaatselt kogutud tekstide kogusid ehk veebikorpuseid, mis on oma mahu ja tekstilise variatiivsuse tõttu väärtuslikud keeleressursid ning mille koostamine on märgatavalt kulutõhusam korpuste käsitsi koostamisest (Schäfer & Bildhauer 2013; Jakubíček et al. 2020). Olgugi, et veebikorpused peegeldavad vaid osa veebis olevast sisust (Mehler et al. 2010: 16), võimaldab neis sisalduvate tekstide variatiivsus ja hulk teha suuremahulist kvantitatiivset analüüsi mitmetes teadusvaldkondades. Seega on igasugune tekstiline info muutunud kättesaadavamaks, kuid puuduvad head lahendused tekstide liigitamiseks ehk klassifitseerimiseks, mis takistab keelelise varieerumise uurimist ning sellega arvestamist ka keeletehnoloogilistes rakendustes. Erinevalt traditsiooniliste tekstikorpuste loomisest, kus tekstiallikaid on hoolikalt valitud, tuleb metaandmete, sh tekstiliigiline info, puudumise tõttu veebist kogutud tekste käsitleda teisiti. Lisaks on kasvanud ka tekstide variatiivsus, kus traditsiooniliste kirjalike tekstide kõrvale (nt uudisteartiklid, ilukirjandus jne) on tekkinud üha enam kasutaja ja arvuti poolt genereeritud sisu (nt erinevad blogid, sotsiaalmeedia postitused, sisuturundus, erinevad foorumid jne), mis võib varieeruda hästi kirjutatud ja informatiivsetest kuni mitteformaalsete ja grammatiliselt ebakorrektse sisuni. Seda suurt ja mitmepalgelist tekstihulka on vaja automaatselt klassifitseerida, kuid katsed teha seda traditsiooniliste tekstikorpuste taksonoomiate alusel pole andnud rahuldavaid tulemusi (Mehler et al. 2010; Schäfer & Bildhauer 2013).

Tekste saab mõistagi liigitada igasuguse äranägemise järgi, kuid tulemuslikum lahendus on anda klassifitseerimisalgoritmile sisendiks eelnevalt defineeritud taksonoomia ja sellesse taksonoomiasse kuuluvate kategooriate näidistekstid. Nii on loodud mitmesuguseid taksonoomiaid, mis erinevad üksteisest metodoloogiliselt ja ka struktuurilt. Näiteks saab eristada *ülalt-alla* või *alt-üles* meetodeid. Ülalt-alla meetodi puhul esmalt koostatakse esialgne taksonoomia, mille kategooriad pärinevad, kas juba olemasolevatest taksonoomiatest, traditsioonilistest tekstikorpustest (nt uudis, ilukirjandus, teadus) või veebilehedelt (nt e-pood, blogi, koduleht, KKK), või neid kõiki kombineerides, vt Rehm et al. 2008; Santini et al. 2010; Sharoff et al. 2010; Jakubíček et al. 2020; Kuzman et al. 2022. Ülalt-alla meetod hõlmab tüüpiliselt ka annoteerimiskatset, kuna taksonoomia kategooriate sobitumist tekstidega tuleb märgendajate abiga ka valideerida (Berninger et al. 2008; Vidulin et al. 2009). Alt-üles meetodi puhul palutakse märgendajatel ise tekstide kuuluvust määrata, kasutades selleks etteantud juhendeid või otsustuspuid (Dewe et al. 1998; Meyer zu Eissen & Stein 2004; Crowston et al. 2011; Egbert et al. 2015). Näiteks märgendajate käest võidakse küsida selliseid küsimusi nagu *kuidas te nimetaksite sellist tüüpi veebilehte?* (Crowston et al. 2011) või *kas peate seda*

žanrit vajalikuks/sobilikuks? (Meyer zu Eissen & Stein 2004). Mõlemad meetodid pakuvad erinevaid vaatenurki, kuid nende kombineerimine võib kaasa aidata usaldusväärsema ja üldistatavama taksonoomia loomisele. Näiteks Asheghi et al. (2016) kasutasid ülalt-alla lähenemisviisi esialgse taksonoomia loomiseks ja altüles meetodit pilootuurimuse läbiviimiseks, mille käigus märgendajatelt kogutud tagasidet kasutati uuesti esialgse taksonoomia täpsustamiseks.

Taksonoomiaid saab eristada ka vastavalt sellele, kas nende struktuur on lame või hierarhiline. Hierarhilised taksonoomiad sarnanevad raamatukogude kataloogisüsteemidega, mida esitatakse puulaadse struktuurina ja kus ülemkategoorial on alamkategooriad (Berninger et al. 2008, Kuzman et al., 2022). Mõlemad strktuurid võivad suurte ja mitmekesiste korpuste liigitamisel osutada probleemseks, muutes taksoomiat liiga detailseks või vastupidiselt, liiga üldiseks. Lisaks saab eristada taksonoomiaid vastavalt sellele, kas üks tekst saab kuuluda ainult ühte või korraga mitmesse kategooriasse (*single- vs multilabelling*). Tekste, mis kuuluvad samaaegselt mitmesse kategooriasse, nimetatakse hübriidseteks. Neid tekste iseloomustavad korraga mitme erineva kategooria tunnused, nt blogipostitus võib olla üheaegselt olla nii tootearvustus, uudis, arvamusartikkel jne. Taksonoomiad, kus tekstid saavad külge vaid ühe kategooria, sarnanevad lameda struktuuriga taksonoomiatele. Mõlemad struktuurid lihtsustavad klassifitseerimisprotsessi, kuid ei arvestada tekstide hübriidusega. Seevastu hierarhilised ja hübriidtekstidega arvestavad taksonoomiad võivad olla olemuselt keerulisemad, kuid samas on võimelised edasi andma tekstide mitmetahulisemat olemust.

Kuigi veebitekstide liigitamiseks on loodud mitmeid taksonoomiaid, ei ole leitud üht universaalset või vähemalt enamikule juhtudele sobivat lahendust. Olemasolevad taksonoomiad kipuvad olema sisemiselt ebaühtlased, sisaldades nii üldisemaid (nt ajakirjandus, ilukirjandus) kui spetsiifilisemaid (nt arvamuslood, spordireportaažid, reklaamid) kategooriaid, nii teksti kommunikatiivsest funktsioonist lähtuvaid (nt seletav, informatiivne, juhendav) kui ka arbitraarseid (nt *lingikogu*, *allalaadimisleht*, *veateade*) kategooriaid. Taksonoomiad sisaldavad tihti ka nö prügikastikategooriat, kuhu paigutatakse tekstid, mis ei sobi kuhugi mujale. Lisaks on mõned taksonoomiad loodud selliselt, et need liigitaks vaid teatud allkeeli, näiteks akadeemiline keel, veebikeel. Selline kategooriate mitmekesisus on põhjustanud selle, et olemasolevad taksonoomiad erinevad üksteisest märkimisväärselt, varieerudes näiteks seitsmest (Santini 2010), kahekümnest (Vidulin et al. 2009) kuni 292 (Crowston et al. 2011) kategooriani. Veebikorpuste taksonoomiate loomist on püütud ühtlustada, kehtestades veebitekstidele määratud kategooriatele teatud põhireeglid, kuid see initsiatiiv väga kaugele ei jõudnud, vt Rehm et al. (2008).

Nii jõuame järgmise probleemini, kus ka märgendajatel on raske saavutada rahuldav kooskõla loodud taksonoomia valideerimiseks (Meyer zu Eissen & Stein 2004; Crowston et al. 2010; Sharoff et al. 2010; Egbert et al. 2015; Suchomel 2020). See tõstatab küsimuse, kas veebitekstide klassifitseerimiseks on üldse võimalik luua katvat taksonoomiat. Veebikorpuste erinevate taksonoomiate rohkus

näitab, et kõiki rahuldavat lahendust ei pruugi üldse eksisteeridagi, mis omakorda viitab vajadusele uurida alternatiivseid lahendusi.

Töö lähtepunktid

Siinne väitekiri pakub välja ühe võimaliku teoreetilise mudeli eelnevalt kirjeldatud probleemidega tegelemiseks. Töös tutvustakse keelest ja korpusest sõltumatut raamistikku, Dimensionaalset tekstimudelit (*Dimensional Text Model*), mis tuginеb hüpoteesile, et tekste saab kategoriseerida tuginedes nende kommunikatiivsetele funktsioonidele, mis omakorda avalduvad koosesinevate keeleliste tunnuste kaudu.

Siinkohal tuleks lühidalt peatuda terminikasutusel. Automaatse tekstide klassifitseerimisega tegelevad autorid on eelistanud kasutada terminit *žanr*[20] viitamaks kategooriale, mis lähtub rohkem teksti kommunikatiivsete funktsioonide perspektiivist (Santini 2007; Wu et al. 2010; Crowston et al. 2011; Sharoff 2018; Madjarov et al. 2019). Mõistet *register* kasutavad need, kes lähtuvad teksti lingvistilisest perspektiivist ehk kuidas erinevad situatiivsed ja kommunikatiivsed eesmärgid leksikogrammatiliste tunnuste alusel avalduvad (Biber 1995; Biber & Conrad 2009; Egbert et al. 2015; Biber & Egbert 2018; Laippala et al. 2019). Arvestades nende terminite taga olevate mõistete kattuvust, ambivalentsust ning selge konsensuse puudumist nende defineerimises, sõltub terminikasutus sageli uurimusvaldkonna praktikast. Kuigi käesolevas väitekirjas kirjeldatakse vaid põgusalt nende terminite ja mõistete ümber toimuvat arutelu, siis järjepidevuse huvides kasutakse selles väitekirjas läbivalt terminit *register*.

Dimensionaalse tekstimudeli raamistik lähtub Biberi (1988) Multidimensionaalsel analüüsil (*Multidimensional analysis*) ja Sharoffi 2018 Funktsionaalsete Tekstidimensioonide (*Functional Text Dimensions*) raamistikul. Mõlemad raamistikud on andnud olulise panuse keelelise varieerumise korpuspõhisele uurimisele.

Multidimensionaalne analüüs (edaspidi **MDA**) on korpuspõhine metodoloogiline lähenemine, mille eesmärgiks on kirjeldada, kuidas koosesinevad keeleliste tunnuste komplektid ehk dimensioonid eristavad registreid, mida Biber mõistab kui kasutussituatsioonist ja kommunikatiivsest funktsioonist tulenevaid keelekasutusvariante. MDA põhirõhk oli algselt kirjaliku ja suulise keelekasutuse erinevuste kvantitatiivsel uurimisel, mille kohta Biberi (1986: 384-387; 1988: 52-53) arvates varasemad tööd ei suutnud pakkuda adekvaatseid üldistusi, kuna toetusid vaid vähestele tekstidele ning üksikute registrite (tüüpiliselt üks kirjaliku, teine suulise allkeele esindaja) võrdlusele. Biberi eesmärk oli vältida varasemate uurimuste kitsaskohti, kasutades kvantitatiivset analüüsi, heterogeensemat ja suuremahulisemat korpust ning suurendada keeleliste tunnuste hulka. Aegade jooksul on MDA-d keelteülese varieeruvuse uurimiseks rakendatud mitmetes eri keeltes, näiteks Nukulaelae tuvalu (Besnier 1988), somaali (Biber & Hared 1992), korea (Kim & Biber 1994), mandariini (Song et al. 2021), gaeli (Lamb 2002), hispaania

[20] Sellele viitab ka inglisekeelne termin *automatic genre classification*.

(Biber et al. 2006; Parodi 2007), dagbani (Purvis 2008), Ameerika inglise (Grieve et al. 2011; Passonneau et al. 2014; Grieve 2014), Brasiilia portugali (Sardinha et al. 2014), vene (Katinskaya & Sharoff 2015), urdu (Shakir & Deuber 2019) ja tšehhi keel (Cvrček et al. 2020). MDA-d on rakendatud ka eritüüpi valdkondade uurimiseks, nt kõnekeskuse kõned (Friginal 2008), Reddit (Liimatta 2019), sotsiaalmeedia (Sardinha 2022) ja Donald Trumpi Twitteri säutsud (Clarke & Grieve 2019).

MDA lähtub vormi ja funktsiooni duaalsusest: keeleline sisu avaldub keeleliste tunnuste koosesinemisena tekstides, kuid nende koosesinemine pole juhuslik, vaid tuleneb sellest, et nad täidavad ühist kommunikatiivset funktsiooni või funktsioone. Näiteks esineb interaktiivse sisuga tekstides rohkem ase- ja küsisõnu. MDAs arvutatakse korpuse igas tekstis keeleliste tunnuste esinemissagedused, mida kasutatakse sisendina faktoranalüüsis, et tuvastada korpuses esinevaid latenseid mustreid. Neid faktoreid nimetatakse dimensioonideks ja tõlgendatakse induktiivselt faktoriga seotud keeleliste tunnuste kommunikatiivseid eesmärke silmas pidades. Oluline on märkida, et kuigi tunnuste koosesinemise mustrid on saadud kvantitatiivselt, on dimensiooni tõlgendamine faktori põhjal siiski eksploratiivne. Algne MDA uuring (Biber 1988) leidis inglise keele kirjaliku (LOB) ja suulise (London-Lund) korpuse põhjal seitse dimensiooni: D1 informatiivne vs. kaasav; D2 narratiivne vs. mitte-narratiivne tekst; D3 kontekstist sõltumatu viitamine vs. mittespetsiifiline olukorrast sõltuv viitamine; D4 veenmine (avalik, tekstitasandil väljendatud); D5 abstraktsus; D6 reaalajas teabe edastamine; D7 akadeemiline argumenteerimine. Esimesed viis dimensiooni on olulisemad. MDA sõltub kasutatud korpusest, kuid rakendades seda erinevatele korpustele ja keeltele on saadud tõendeid universaalsemate, kõigis keeltes ja korpustes esilduvate dimensioonide olemasolu kohta, näiteks dimensioonid, mis eristavad kõnekeelt ja kirjakeelt, narratiivseid ja mitte-narratiivseid, ning abstrakteseid ja vähem abstraktseid tekste.

Funktsionaalsete tekstidimensioonide (Sharoff 2018, edaspidi **FTD**) raamistik lähtub traditsioonilisest veebikorpuste automaatsest liigitamisest, kus taksonoomia alusel käsitsi liigitatud korpust kasutatakse masinõppemudeli treenimiseks ning mis omakorda on võimeline liigitama seninägemata tekste ettenähtud taksonoomia ja sellesse kuuluvate kategooriate alusel. FTD taksonoomia koosneb dimensioonidest, kus iga dimensioon esindab teksti suhtluseesmärki ehk tekstifunktsiooni. Need funktsionaalsed tekstidimensioonid on korpuses käsitsi märgendatud testküsimuste abil, kus iga küsimus esindab üht suhtluseesmärki ning selle eesmärgi prototüüpseimaid registreid (nt *mil määral sisaldab tekst ilmset argumentatsiooni lugeja veenmiseks?* või *mil määral paistab tekstis osalevat mitu osapoolt?*). FTDs hinnatakse suhtluseesmärkide esilduvust 4-pallilisel Likerti skaalal[21]. FTD taksonoomia loomine on toimunud mitmes märgendamise etapis (peamiselt Sharoff et al. 2010, Sorokin et al. 2014, Katinskaya & Sharoff 2015), kuid lõplikuks dimensioonide arvuks kujunes 18, mis jagunevad primaarseteks

[21] Dimensioon ei esildu - 0, esildub vähesel määral - 0.5, esildub osaliselt - 1, esildub tugevalt - 2.

ehk kohustuslikeks ning sekundaarseteks ehk valikulisteks dimensioonideks[22]. Sekundaarsete dimensioonide eesmärk on esitada primaarsete dimensioonide sisemist variatiivsust.

Sharoffi (2018: 4) väitel on FTD universaalne ja üldistatav kõikidele tekstidele või tekstikorpustele, kuna erinevad märgendajad suudavad selle taksonoomia alusel tekste sarnaselt kategoriseerida ehk hindajatevaheline kooskõla on suur. Lisaks on FTD lihtne, kuna võrreldes varasemate taksonoomiatega iseloomustab FTD-d vähene parameetrite ja kategooriate hulk. Olulise uuendusena võtab FTD kasutusele *multidimensionaalse ruumi* mõiste ja *distantsi mõõdiku*: tekste esitatakse funktsionaalsete dimensioonide multidimensionaalses ruumis, mis võimaldab modelleerida ka tekstide hübriidsust, ja distantsimõõdiku abil on võimalik mõõta mis tahes teksti kaugust dimensiooni n-ö prototüüpseimast registrist.

Dimensionaalne tekstimudel

MDA oli üks esimesi teoreetilisi lähenemisi, mis võrdles registrite omavahelist ja sisemist varieeruvust keeleliste parameetrite kaudu. Erinevalt MDA-st, mis uurib juba eelnevalt liigitatud tekstide keelekasutuse varieerumist dimensioonide kaudu, sai FTD alguse vajadusest klassifitseerida veebikorpusi, mille kohta pole teada, missugusesse registrisse mingi tekst kuuluda võiks. Mõlemad esindavad olulisi metodoloogilisi lähenemisi tänapäevastes korpuspõhistes uurimisvaldkondades, kuid neile mõlemal on teatud piirangud, mis said ajendiks Dimensionaalse Tekstimudeli (edaspidi **DTM**) väljatöötamisel.

MDA lähtub vormi ja funktsiooni duaalsusest, kus keelelised tunnused ja kommunikatiivsed funktsioonid sõltuvad üksteisest ehk keelelised tunnused väljendavad kommunikatiivseid funktsioone ning neid kommunikatiivseid funktsioone väljendatakse omakorda keeleliste tunnuste abil. DTM lähtub samast ideest, võttes MDA-st üle mõiste *dimensioon* kui latentne kvantifitseeritav parameeter, mida mõõdetakse koosesinevate leksikaalsete, grammatiliste ja tekstiliste tunnuste abil. Kuid erinevalt MDA-st, mis käsitleb dimensiooni kui skaalat, millel on kaks erinevat funktsioonipoolust, käsitletakse DTM-is dimensiooni kui üht kommunikatiivset funktsiooni esindavat pidevat skaalat ehk kontinuumi. MDA-l on ka keskne roll DTMis esitatud dimensioonide formuleerimisel. Probleemkohana saab MDA dimensioonide osas mainida nende hägust ning hübriidsust, kus samaaegselt üks dimensioon hõlmab mitmeid kommunikatiivseid funktsioone korraga. Kuna DTM-i eesmärk on tuvastada universaalseid konstrukte, mis eristavad teks-

[22] Primaarsed dimensioonid on *argumentatsioon* (nt juhtkirjad, arvamusartiklid), *ilukirjanduslik* (nt romaanid, jutustused), *juhendavad* (nt õpetused, käsiraamatud), *uudis* (nt uudislehed), *õiguslik* (nt seadused, lepingud), *isiklik* (nt päevikud, isiklikud kirjad), *kaubanduslik esitlus* (nt, reklaamid), *ideoloogia* (nt manifestid, propaganda), *teaduslik-tehniline* (nt artiklid, esseed), *informatiivne või entsüklopeediline*, *hindav* (nt tooteülevaade), *luuleline* (tähelepanu esteetikale), *apellatiivne* (palved, kaebekirjad, reklaamid). Sekundaarsed dimensioonid on *emotsionaalsus* (tundeid, emotsioone väljendavad tekstid), *meelelahutuslikkus* (lihtsad ja meelelahutuslikud tekstid), *mitteformaalsus* (keele hälbimine standardsest keelekasutusest), *erialalisus* (teksti mõistmine nõuab erialalist taustainformatsiooni) ja dialoogilisus (mitme osaleja vaheline suhtlus).

te nende suhtlusolukorra ja kommunikatiivse funktsiooni alusel, siis tugineti selle dimensioonide formuleerimisel erinevate MDA-d rakendanud keelte tulemustele. Niisiis jaotati MDA dimensioonid väiksemateks põhikonstruktideks, millest igaüks potentsiaalselt realiseerib vaid üht situatiivset või kommunikatiivset funktsiooni. Näiteks on mitmete keele MDA faktorite tõlgendamisel eristatud suulist ja kirjalikku keelekasutust.

Erinevalt MDA-st jagavad DTM ja FTD eesmärki luua taksonoomia, mille alusel hiljem veebikorpuseid automaatselt registritesse klassifitseerida. Erinevalt MDA keelelistest tunnusest inspireeritud lähenemisest, on FTD pinnapealsem ja sarnasem traditsioonilistele klassifitseerimismeetoditele, milleks kasutatakse eelnevalt defineeritud taksonoomiaid. Kui MDA ja DTM defineerivad dimensiooni kui parameetrit, mis on mõõdetav läbi keeleliste tunnuste komplektide, siis FTD vaatleb dimensiooni kui registrit ja üht teksti võib kirjeldada läbi mitme dimensiooni esildumise (ehk tekst on registriliselt hübriidne). FTD ühe kitsaskohana võib välja tuua dimensioonide taksonoomia kategooriate eklektilisuse. Üheltpoolt eksisteerivad ülemkategooriad nagu *ilukirjandus*, *uudised*, *teadus*, kuid samasse hulka kuuluvad ka erinevad modaalsused (nagu *dialoog*), keelekorraldamise- (nt *luule*) ja normimiseviisid (*formaalne keelekasutus*). DTM laenab FTD-lt *mitmemõõtmelise ruumi* mõiste, kuid laiendab seda nii, et mitmemõõtmeline ruum esindaks kõiki tekste korraga. Kui FTD-s toimub 18-dimensionaalse ruumi ahendamine kohustuslike ja valikuliste dimensioonide kaudu, siis DTM vaatleb kõikidel tekstidel dimensioonide esildumist. Need tekstid, mis paiknevad mitmemõõtmelises ruumis üksteisele lähedal, jagavad ka sarnast suhtlusfunktsiooni ning loomissituatsiooni ja võivad kuuluda sama(de)sse registri(te)sse.

Dimensionaalse Tekstimudeli tuuma moodustavad dimensioonid, mis potentsiaalselt realiseerivad vaid üht situatiivset või kommunikatiivset funktsiooni. Neid dimensioone on kokku 12: *abstraktsus*, *afektiivsus*, *argumentatiivsus*, *impersonaalsus*, *interaktiivsus*, *instrueerivus*, *formaalsus*, *keerukus*, *spontaansus*, *informatsioonitihedus*, *aja olulisus* ja *subjektiivsus*.

DTM on oma olemuselt hierarhiline raamistik, millel on kolm põhikomponenti.

I Mudeli alumisel tasandil asuvad keelelised **tunnused**, mis on otseselt tekstist mõõdetavad, nt nimisõnade arv, sõnavara suurus, relatiivlausete hulk, abstraktsete sõnade hulk jm. Tunnused võivad olla keelteülesed või keeltele spetsiifilised.

II Mudeli keskmisel tasandil asuvad latentsed **dimensioonid**, mis on kirjeldatavad tunnuste komplektide kaudu. Kõik dimensioonid esilduvad igas tekstis, aga erinevus tuleneb tugevusest, st mõnes tekstis on dimensioon tugevalt väljendunud, teises keskmiselt ja kolmandas väga vähe. Dimensioonid ise on keelteülesed, kuigi tunnuste komplektid, mille kaudu dimensioonid väljenduvad, võivad olla keeltespetsiifilised.

III Kõige ülemisel tasandil asuvad **registrid**, mis avalduvad läbi koosesinevate dimensioonide. Tekstid, mille dimensioonivektorid on sarnased, jagavad sarnaseid funktsioone ning seega võiksid kuuluda ühte registrisse.

DTM-i dimensioonidele kehtivad järgmised tingimused:

- dimensiooni aluseks on teksti situatiivne või kommunikatiivne funktsioon ehk eesmärk, mida saab mõõta keeleliste tunnuste kaudu;
- dimensioon on pidev skaala;
- dimensioon on iseseisev latentne konstrukt. Dimensioon üksi ei suuda kirjeldada ühtegi registrit, küll aga on dimensioon võimeline kirjeldama registri üht omadust. Näiteks teadustekst räägib sageli nähtustest üldistatult, pürgides tegelikkust peegeldava abstraktse süsteemi poole ja nii iseloomustab neid tekste lisaks muudele tunnustele ka abstraktse sõnavara kasutamine.

Siinse väitekirja eesmärk on pakkuda alternatiivne lahendus (veebi)korpuste registrilistele taksonoomiatele, vastates järgmistele uurimisküsimustele:

- Uurimusküsimus 1 – kas Dimensionaalses Tekstimudelis esitatud dimensioonid on üksteisest eristatavad?
- Uurimusküsimus 2 – kas ja kuidas Dimensionaalse Tekstimudelis pakutud dimensioonid üksteisest keeleliselt eristuvad?

Tulemused

Esimese uurimusküsimuse eesmärk on hinnata, kas teoreetiliselt formuleeritud dimensioonid on eristatavad ka eestikeelsetes tekstides. Selleks viidi läbi annoteerimiskatse, kus märgendajatel paluti hinnata 120 etTenTen korpusest (Koppel & Kallas 2022) pärinevat teksti. Tekstid valiti katsesse juhuslikult, kuid lähtudes sellest, et tekstidele eelnevalt automaatselt määratud kategooriate osakaalud terves korpuses oleksid proportsioonis annoteerimiskatse tekstide hulgaga. Näiteks, kui tekste märgendiga X oli korpuses kokku 10%, siis annoteerimiskatsesse valiti sama märgendiga tekste kaksteist. Märgendajate ülesanne oli hinnata dimensiooni esildumise määra nendes veebitekstides neljapunktilisel Likerti skaalal (*tugev*, *mõõdukas*, *nõrk* või *mitteeksisteeriv*). Siinkohal on oluline mainida, et kuigi annoteerimiskatses kasutatakse Likerti skaalas hinnangupunktina märgendit *mitteeksisteeriv*, siis DTM eeldab vaikimisi kõikide dimensioonide avaldumist tekstis, kuid mõistagi dimensioonide esildumine tekstides on erinev. Näiteks on akadeemiline tekst üldjuhul abstraktsem kui luuletus ja sõpradevaheline vestlus, kuid luuletus on omakorda abstraktsem kui vestlus jne. Märgendajatevahelise kooskõla hindamiseks kasutati Krippendorfi α ning tulemuste tõlgendamiseks Landis ja Kochi 1977 adapteeritud konventsiooni, kus $\alpha \geq 0.40$ võib pidada keskmiseks (*moderate*) hindajatevaheliseks kooskõlaks.

Annoteerimiskatse tulemuse põhjal saab väita, et subjektiivsus, afektiivus, formaalsus ja spontaansus olid märgendajatele jaoks selgemalt äratuntavad (α varieerus vahemikus 0.6 kuni 0.76). See viitab sellele, et subjektiivsus, afektivsus,

formaalsus ja spontaasus on DTM-i raamistikus selgemalt defineeritud ja märgendajate jaoks eristatavad. Instrueerivus, interaktiivsus, impersonaalsus ja ajaolulisus olid märgendajate jaoks keskmise kooskõlaga (α varieerus vahemikus 0.4 kuni 0.47). Abstraktsuse, keerukuse, argumentatiivsuse ja infotiheduse vahel oli kooskõla mõõdukas (α varieerus vahemikus 0.25 kuni 0.38), mis näitab, et nende dimensioonide osas on mõningaid ebakõlasid.

Madala hindajatevahelise kooskõla üheks põhjuseks võib olla see, et dimensioonide definitsioonid polnud märgendajate jaoks piisavalt selged. Näiteks keerukuse dimensiooni (α = 0.38) on DTM-is defineeritud kui nähtust, millest arusaamine nõuab lisapingutust. Definitsiooni järgi võib märgendajal olla raske otsustada, kas keerukus tuleneb süntaktilisest või kognitiivsest keerukusest. Sarnaseid probleeme võib laiendada ka abstraktsuse dimensioonile (α = 0,33), kuna abstraktsuse taseme määramine on inimeseti erinev - mida üks inimene peab väga abstraktseks tekstiks, võib teine hinnata selle küllaltki konkreetseks. Madalaim oli hindajatevaheline kooskõla informatsioonitiheduse dimensiooni märgendamisel (α = 0,25). Sarnaselt teistele madala hindajatevahelise kooskõlaga dimensioonidele saab ka informatsioonitihedust tõlgendada mitmel, üksteisest veidi erineval moel ning seda, mis üldse on informatsioon, saab mõisteta erineval moel. Kas informatsioonitihedus esildub tugevalt vaid sellistes tekstides, mis sarnanevad Wikipeedia või teiste entsüklopeediliste tekstidega, või igasugune suhtlus klassifitseerub kui tihedaks informatsiooniks? Märgendajate hinnangul esildus informatsioonitihedus tugevalt või mõõdukalt 59% tekstidest. Need tulemused vihjavad, et informatsioon mõistena on olemuslikult ähmane ja kuna annoteerimiskatses teadlikult ei antud märgendajatele detailseimaid juhiseid, siis võis juhtuda, et märgendajad tuginesid isiklikele tõlgendustele.

Annoteerimiskatse disainimisel oli etTenTen (Koppel & Kallas 2022) ainus kättesaadav eestikeelne veebikorpus, kus tekste oli eelnevalt liigitatud kategooriatesse (*valitsus*, *foorum*, *religioon*, *blogi*, *perioodika*, *informatsioon*, *tundmatu*), kuid selle liigituse korrektsust pole valideeritud. Kuna tekstidel puudusid muud metaandmed, siis annoteerimiskatse jaoks tekstide valikul tuli hoida alles algupärase tekstiliigilise jaotuse proportsiooni terves korpuses. Näiteks *tundmatu* kategooriaga tekstid moodustasid terves korpuses 36%, seega annoteerimiskatsesse valiti sellest kategooriast kokku 43 teksti. Tekstivaliku meetod ei pruukinud olla kõige optimaalsem, sest liikide alusel valimine võis põhjustada homogeensemate (nt *religioon*) tekstiliikide üleesindatust ja heterogeensemate (nt *informatiivne*, *perioodika*) tekstiliikide tekstid alaesindatud.

Annoteerimiskatse ülesehitus seadis ülesandele omad piirangud, kuid tulemused näitavad, et suures plaanis on enamik DTM-i poolt pakutavaid dimensioone üksteisest selgelt eristatavad. Subjektiivsuse, afektiivsuse, formaalsuse, spontaansuse, instrueerivuse, interaktiivsuse, impersonaalsuse ja aja olulisuse dimensioon saavutasid vähemalt keskmise hindajatevahelise kooskõla, näidates, et märgendajatele antud juhised ja dimensioonid olid selgemalt defineeritud. Abstraktsuse, keerukuse, argumentatiivsuse ja informatsioonitiheduse dimensioon saavuta-

sid mõõduka hindajatevahelise kooskõla. See omakorda näitab, et kuigi need dimensioonid tekstides teatud määral esildusid, siis edasiste töösuundate planeerimine eeldab eelnevat täiendavat kvalitatiivset analüüsi.

Kui esimese uurimusküsimuse eesmärk oli välja selgitada, kas raamistiku väljatöötamisel valitud ja defineeritud dimensioonid on märgendajate jaoks eristatavad, siis sellele järgnev samm oli vaadelda, kas ja kuidas need dimensioonid üksteisest eristuvad koosesinevate keeleliste tunnuste komplektide ehk lingvistiliste profiilide kaudu. DTM eeldab, et igal dimensioonil on ainulaadne lingvistiline profiil, mille keeleliste tunnuste esinemine pole juhuslik, vaid nende koosesinemine on motiveeritud selgest eesmärgist ehk situatiivsest või kommunikatiivsest funktsioonist.

Loomuliku keele töötluses on tekstide klassifitseerimisel levinud nii järjestikuste sõnade esinemissageduste (*bag of words*) kui ka sõnavektorite (*word embeddings*) kasutamine. Mõlemad meetodid on olnud väga tõhusad suurtes tekstikorpustes statistiliste mustrite leidmisel, kuid teisele uurimisküsimusele vastamiseks on vaja lähenemist, mis ei vaata pelgalt leksikaalseid ja semantilisi tunnuseid. Seetõttu võetakse dimensioonide lingvistiliste profiilide väljaselgitamiseks aluseks leksikaal-grammatilised tunnused. Arvestades et eesti keeles on registritevahelist keelelist varieerumist kvantitatiivselt vähe uuritud, siis oli selles väitekirjas eesmärgiks kaasata võimalikult mitmekesine tunnuste hulk. Tunnuste valikut määras Stanfordi Stanza parseri (Qi et al. 2020) väljund, st tunnuseid sai valida ainult selle väljundis olemasolevate hulgast. Lisaks kaasati ka mitmeid tekstilisi tunnuseid (nt sõnade ja sõnede suhe). Kokku ekstraheeriti dimensioonide hinnangutega märgendatud tekstidest 85 erinevat leksikaal-grammatilist tunnust ja arvutati nende suhtelised sagedused, nt suhteline nimisõnade arv, suhteline finiitsete verbide hulk, keskmine lause pikkus jne.

Dimensioonide statistiliselt oluliste tunnuste väljaselgitamiseks viidi läbi ühesuunaline mitteparameetriline dispersioonianalüüs (*nonparametric one-way analysis of variance*). Dispersiooninalüüsis võrreldakse iga tunnuse suhtelist esinemissagedust dimensiooni esilduvuse kolmel tasandil - *tugev/mõõdukas* (T/M), *nõrk* (N) või *mitteeksisteeriv* (ME). Annoteerimiskatse tulemusena olid Likerti skaala hinnangud N ja ME hinnangute poole kaldu, seega T ja M liideti üheks hinnanguks. Dispersioonianalüüsi tulemus näitas, kas konkreetse tunnuse kolme tasandi hinnangute mediaanide vahel on erisusi või mitte. Kui kolme tasandi mediaanide vahel oli statistiliselt olulised erinevused, järgnes dispersioonanalüüsile järeltest selgitamaks välja, milliste konkreetsete tasandite võrdluses statistiliseslt oluline erinevus eksisteeris. Kuna selles analüüsis oli iga tunnuse osas vaatluse all dimensiooni esilduvuse graduaalsus, siis muutus oluliseks see, milliste tasandite vahel eksisteeris statistiline olulisus. Töö seisukohalt relevantseteks peetud võrdlused olid:

1. samaaegselt (T/M ja N), (T/M ja ME) ja (N ja ME) vahel,
2. samaaegselt (T/M ja ME) ning (T/M ja N) või (N ja ME) vahel,
3. ainult (T/M ja ME) vahel.

Seejärel tuli kindlaks määrata, kas keelelise tunnuse esinemissageduse ja dimensiooni esilduvuse vahel eksisteeris monotoonne suhe. Monotoonsus näitas, kas keelelise tunnuse kasvav või kahanev esinemissagedus muutub järjekindlalt dimensiooni esilduvuse tugevnemisel või nõrgenemisel. Näiteks on monotoonne suhe olemas siis, kui nimisõnade esinemissagedus järjekindlalt kasvab üle kõigi kolme tasandi. Kui aga tunnuse esinemissagedus suureneb dimensiooni mõlemas otsas (tugev/mõõdukas ja mitteeksisteeriv) ja väheneb nõrga esilduvuse tasemel, siis monotoonsus puudub.

Tulemused osutasid, et kõikidel dimensioonidel avaldusid ainulaadsed lingvistilised profiilid, mis võib viidata sellele, et igal dimensioonil on raamistiku piires eristuv kommunikatiivne või situatiivne funktsioon. Kuigi kõigi dimensioonide lingvistilised profiilid olid erinevad, oli eesmärk tuvastada need dimensioonid, mis eristusid teistest dimensioonidest ainulaadsete statistiliselt oluliste tunnuste poolest. Analüüsist selgus, et kaheteistkümne dimensiooni seast viiel – subjektiivsusel, afektiivsusel, interaktiivsusel, formaalsusel ja aja olulisusel – tuvastati lingvistilised profiilid, mis sisaldasid vaid neile dimensioonidele iseloomulikke ainulaadseid keelelisi tunnuseid. Näiteks subjektiivsetele tekstidele oli iseloomulik kasutada vähem adjektiivseid täiendeid ning afektiivsetes tekstides oli iseloomulik kasutada rohkem ainsuse kolmanda isiku pronoomeneid. Interaktiivseid tekste saab iseloomustada vaid neile iseloomuliku sagedasema ütete kasutuse ja väiksema sõnavara kaudu. Formaalsetes tekstides seevastu kasutati pikemaid lausekonstruktsioone. Aja olulisuse dimensiooni eesmärk on rõhutada sündmuste sidusust ja järjekorda, ja tekste, kus aja olulisust on tugevaks hinnatud, kasutatakse ka vähem infinitiivseid verbivorme ja rohkem arvsõnalisi laiendeid. Kuid enamik analüüsi kaasatud keelelisi tunnuseid polnud omased ainult ühele dimensioonile, vaid iseloomulikud mitmele dimensioonile korraga, ja dimensiooni defineeribki tunnuste kimp, mitte üksik tunnus. Keeleliste tunnuste dimensionaalne jaotumine ja ka nende kvalitatiivne analüüs on väga huvitav uurimisteema edaspidiseks.

Annoteeritud testkorpuse analüüsimisel vaadeldi dimensioonidevahelisi suhteid lisaks veel Spearmani korrelatsiooni ja eksploratiivse faktoranalüüsiga. Korrelatsioonianalüüsi eesmärk oli hinnata, kas ja kuidas dimensioonid omavahel korrelleeruvad, ja eksploratiivne faktoranalüüs võimaldas tuvastada dimensioonide vahel eksisteerivaid latentseid mustreid. Korrelatsiooni- ja faktoranalüüsi põhjal võib DTM-i dimensioone kategoriseerida kahe makrodimensiooni kaudu – nominaalne-planeeritud ja verbaalne-planeerimata. Verbaalne-planeerimata makrodimensiooni saab iseloomustada subjektiivsuse, afektiivsuse, spontaansuse ja interaktiivsuse dimensioonide kaudu. Nominaalne-planeeritud makrodimensioon omakorda on seotud formaalsuse, impersonaalsuse ja informatsioonitiheduse dimensoonidega. Pea kõikidele makdrodimensioonidele statistiliselt oluliste koosesi-

nevate keeleliste tunnuste vahel esines pöördvõrdeline seos. Kui keelelise tunnuse esinemissagedus suurenes verbaalse-planeerimata dimensioonide tugevamal esildumisel, siis nominaalse-planeeritud dimensioonide tugevama esilduvuse puhul oli keelelise tunnuse väiksem esinemissagedus statistiliselt oluline. Ehk kui tugevalt interaktiivseks hinnatud tekst (verbaalne-planeerimata makrodimensioonist) kasutab rohkem asesõnu, tuumverbe ja ütteid, siis vastupidiselt tugevalt impersonaalsetes tekstides (nominaalne-planeeritud makrodimensioonist) kasutatakse vähem asesõnu, tuumverbe ja ütteid. Selline tunnuste jaotus kehtis suurema osa analüüsi kaasatud keeleliste tunnustega. Ülejäänud dimensioone nagu argumentatiivsus, abstraktsus, keerukus ja instrueerivus saab iseloomustada kui komplementaarseid dimensioone, kuna need dimensioonid olid vähem seotud nii üksteise kui ka verbaalse-planeerimata ning nominaalse-planeeritud makrodimensiooniga.

Selles väitekirjas esitleti keelest ja korpusest sõltumatut Dimensionaalse tekstimudeli raamistikku, mille eesmärk on pakkuda üht võimalikku lahendust tegelemaks veebikorpuste klassifitseerimisega seotud probleemidega. Dimensionaalse Tekstimudeli abi saab kategoriseerida tekste, toetudes nendes avalduvatele suhtlussituatsioonile ja kommunikatiivsetele eesmärkidele, mis omakorda avalduvad koosesinevate keeleliste tunnuste kaudu. Dimensionaalne tekstimudel tegi eelduse, et iga dimensiooni situatiivne või kommunikatiivne eesmärk avaldub ainulaadsete lingvistiliste profiilide kaudu ja et neid lingvistilisi profiile iseloomustavad ka ainuomased keelelised tunnused. Tulemused näitasid, et dimensioonid olid omavahel arvatust rohkem seotud. See viitab sellele, et dimensioonid ei ole isoleeritud, vaid pigem omavad üksteise suhtes tugevat mõju.

Töötulemused on huvitavad ja uudsed, kuid sellest hoolimata on töös probleeme, mis mõjutavad ka tulemusi. Annoteerimiskatse näitas, et teatud dimensioonide osas oli kooskõla pigem madal. Kõrge hindajatevahelise kooskõla saavutamine polnud küll eesmärk omaette, kuid selle puudumine annab vajalikku informatsiooni raamistiku kitsaskohtadest. Madal kooskõla võib viidata sellele, et annoteerimisjuhistes defineeritud dimensioonid olid märgendajate jaoks ebaselged või mõni dimensioon vajaks raamistikus ümberkalibreerimist. Kitsaskohtadele lahenduste leidmiseks tuleks teha uus annoteerimiskatse, kuid selle erinevusega, et märgendajatelt võiks koguda nende poolt antud hinnangute tagamaid. Töö edasised suunad hõlmavad masinõppemudelite treenimist ja ka moodsamate keelemudelite peenhäälestamist selleks, et Dimensionaalse tekstimudeli dimensioonide kaudu liigitada juba eelnevalt registriliselt annoteeritud korpuseid. Kuid selleks, et üldse klassifitseerida või peenhäälestada, on tarvis suuremat treeningkorpust. Käsitsi tekstide kogumise asemel oleks mõistlik sellesse protsessi integreerida keelemudelid, kelle juhendamiseprotsessi saaks kaasata esmalt märgendajatelt saadud hinnagute põhjused ja lasta seejärel keelemudelil iseseisvalt tekstide dimensionaalsuse esilduvust määrata. Keelemudeli treenimine on mõistagi iteratiivne protsess, kus keelemudelite annotatsioone tuleb valideerida ja vajadusel neid tulemusi keelemudelile uuesti sisendina kaasa anda.

CURRICULUM VITAE

Name:	Kristiina Vaik
Citizenship:	Estonian
E-mail:	kristiina.vaik@ut.ee

Education

2017-2024	University of Tartu, PhD in Estonian and Finno-Ugric Linguistics
2014–2016	University of Tartu, MA in Estonian and Finno-Ugric Linguistics (computational linguistics)
2010–2014	University of Tartu, BA in Estonian and Finno-Ugric Linguistics

Employment

2021–2024	University of Tartu, Faculty of Arts and Humanities, Institute of Estonian and General Linguistics, Junior Research Fellow
2018–2020	TEXTA, Data scientist
2016–2017	University of Tartu, Faculty of Science and Technology, Institute of Computer Science, Programmer

Scientific work

List of publications:

- Kanger, Laur; Tinits, Peeter; Pahker, Anna-Kati; Orru, Kati; Velmet, Aro; Sillak, Silver; Šeļa, Artjoms; Mertelsmann, Olaf; Tammiksaar, Erki; Vaik, Kristiina; Penna, Caetano C.R.; Tiwari, Amaresh Kumar; Lauk, Kalmer (2023). Long-term country-level evidence of major but uneven ruptures in the landscape of industrial modernity. Environmental Innovation and Societal Transitions, 48. DOI: 10.1016/j.eist.2023.100765.
- Kanger, Laur; Tinits, Peeter; Pahker, Anna-Kati; Orru, Kati; Tiwari, Amaresh Kumar; Sillak, Silver; Šeļa, Artjoms; Vaik, Kristiina (2022). Deep Transitions: Towards a comprehensive framework for mapping major continuities and ruptures in industrial modernity. Global Environmental Change, 72, 102447. DOI: 10.1016/j.gloenvcha.2021.102447.
- Vaik, Kristiina; Sirts, Kairit; Muischnek, Kadri (2020). Dimensionaalne tekstimudel. Teoreetiline ülevaade [‘The Dimensional Text Model: A theoretical overview’]. Keel ja Kirjandus, 10, 875-898. DOI: 10.54013/kk755a4.

- Ulčar, Matej; Vaik, Kristiina; Lindström, Jessica; Dailidėnaite, Milda; Robnik-Šikonja, Marko (2020). Multilingual Culture-Independent Word Analogy Datasets. In: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis (Ed.). Proceedings of The 12th Language Resources and Evaluation Conference. (4074-4080). Marseille, France: European Language Resources Association (ELRA).
- Vaik, Kristiina; Asula, Marit; Sirel, Raul (2020). Hybrid Tagger – An Industry-driven Solution for Extreme Multi-label Text Classification. In: Proceedings of the LREC2020 Industry Track. (26-30). Language Resources and Evaluation Conference (LREC 2020), Marseille, 11-16 May 2020. European Language Resources Association (ELRA).
  DOI: 10.5281/zenodo.4306169.
- Vaik, Kristiina; Muischnek, Kadri (2018). Eestikeelsete veebitekstide automaatne liigitamine ['Automatic Classification of Estonian Web Texts']. Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics, 14, 215-229. DOI: 10.5128/ERYa14.13.
- Vaik, Kristiina (2018). Classifying Estonian Web Texts. In: Proceedings of the ESSLLI 2018 Student Session. 30th European Summer School in Logic, Language & Information. (42-54). Sofia, Bulgaria: Sofia University "St. Kl. Ohridski".
- Vaik, Kristiina; Vihman, Virve-Anneli (2017). Eesti lastekeele korpuse morfoloogilise märgendamise kitsaskohtadest. ['Issues in the morphological annotation of the Estonian child language corpus']. Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics, 13, 205-221. DOI: 10.5128/ERYa13.13.

ELULOOKIRJELDUS

Nimi:	Kristiina Vaik
Kodakondsus:	eesti
E-post:	kristiina.vaik@ut.ee

Haridus

2017–2024	Tartu Ülikool, doktoriõpe (PhD)
2014–2016	Tartu Ülikool, humanitaarteaduste magister (MA), arvutilingvistika
2010–2014	Tartu ülikool, humanitaarteaduste bakalaureus (BA), üldkeeleteadus

Teenistuskäik

2021–2024	Tartu Ülikool, Humanitaarteaduste ja kunstide valdkond, eesti ja üldkeeleteaduse instituut, nooremteadur
2018–2020	TEXTA, andmeteadlane
2016–2017	Tartu Ülikool, Loodus- ja täppisteaduste valdkond, arvuteaduse instituut, programmeerija

Teadustegevus

Publikatsioonid:

- Kanger, Laur; Tinits, Peeter; Pahker, Anna-Kati; Orru, Kati; Velmet, Aro; Sillak, Silver; Šeļa, Artjoms; Mertelsmann, Olaf; Tammiksaar, Erki; Vaik, Kristiina; Penna, Caetano C.R.; Tiwari, Amaresh Kumar; Lauk, Kalmer (2023). Long-term country-level evidence of major but uneven ruptures in the landscape of industrial modernity. Environmental Innovation and Societal Transitions, 48. DOI: 10.1016/j.eist.2023.100765.
- Kanger, Laur; Tinits, Peeter; Pahker, Anna-Kati; Orru, Kati; Tiwari, Amaresh Kumar; Sillak, Silver; Šeļa, Artjoms; Vaik, Kristiina (2022). Deep Transitions: Towards a comprehensive framework for mapping major continuities and ruptures in industrial modernity. Global Environmental Change, 72, 102447. DOI: 10.1016/j.gloenvcha.2021.102447.
- Vaik, Kristiina; Sirts, Kairit; Muischnek, Kadri (2020). Dimensionaalne tekstimudel. Teoreetiline ülevaade [‘The Dimensional Text Model: A theoretical overview’]. Keel ja Kirjandus, 10, 875-898. DOI: 10.54013/kk755a4.
- Ulčar, Matej; Vaik, Kristiina; Lindström, Jessica; Dailidėnaite, Milda; Robnik-Šikonja, Marko (2020). Multilingual Culture-Independent Word Analogy Datasets. In: Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi

Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis (Ed.). Proceedings of The 12th Language Resources and Evaluation Conference. (4074-4080). Marseille, France: European Language Resources Association (ELRA).

- Vaik, Kristiina; Asula, Marit; Sirel, Raul (2020). Hybrid Tagger – An Industry-driven Solution for Extreme Multi-label Text Classification. In: Proceedings of the LREC2020 Industry Track. (26-30). Language Resources and Evaluation Conference (LREC 2020), Marseille, 11-16 May 2020. European Language Resources Association (ELRA).
  DOI: 10.5281/zenodo.4306169.
- Vaik, Kristiina; Muischnek, Kadri (2018). Eestikeelsete veebitekstide automaatne liigitamine ['Automatic Classification of Estonian Web Texts']. Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics, 14, 215-229. DOI: 10.5128/ERYa14.13.
- Vaik, Kristiina (2018). Classifying Estonian Web Texts. In: Proceedings of the ESSLLI 2018 Student Session. 30th European Summer School in Logic, Language & Information. (42-54). Sofia, Bulgaria: Sofia University "St. Kl. Ohridski".
- Vaik, Kristiina; Vihman, Virve-Anneli (2017). Eesti lastekeele korpuse morfoloogilise märgendamise kitsaskohtadest. ['Issues in the morphological annotation of the Estonian child language corpus.']. Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics, 13, 205-221. DOI: 10.5128/ERYa13.13.

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS

1. **Anna Verschik.** Estonian Yiddish and its contacts with coterritorial languages. Tartu, 2000, 196 p.
2. **Silvi Tenjes.** Nonverbal means as regulators in communication: socio-cultural perspectives. Tartu, 2001, 214 p.
3. **Ilona Tragel.** Eesti keele tuumverbid. Tartu, 2003, 196 lk.
4. **Einar Meister.** Promoting Estonian speech technology: from resources to prototypes. Tartu, 2003, 217 p.
5. **Ene Vainik.** Lexical knowledge of emotions: the structure, variability and semantics of the Estonian emotion vocabulary. Tartu, 2004, 166 p.
6. **Heili Orav.** Isiksuseomaduste sõnavara semantika eesti keeles. Tartu, 2006, 175 lk.
7. **Larissa Degel.** Intellektuaalsfäär intellektuaalseid võimeid tähistavate sõnade kasutuse põhjal eesti ja vene keeles. Tartu, 2007, 225 lk.
8. **Meelis Mihkla.** Kõne ajalise struktuuri modelleerimine eestikeelsele tekst-kõne sünteesile. Modelling the temporal stucture of speech for the Estonian text-to-speech synthesis. Tartu, 2007, 176 lk.
9. **Mari Uusküla.** Basic colour terms in Finno-Ugric and Slavonic languages: myths and facts. Tartu, 2008, 207 p.
10. **Petar Kehayov.** An Areal-Typological Perspective to Evidentiality: the Cases of the Balkan and Baltic Linguistic Areas. Tartu, 2008, 201 p.
11. **Ann Veismann.** Eesti keele kaas- ja määrsõnade semantika võimalusi. Tartu, 2009, 145 lk.
12. **Erki Luuk.** The noun/verb and predicate/argument structures. Tartu, 2009, 99 p.
13. **Andriela Rääbis.** Eesti telefonivestluste sissejuhatus: struktuur ja suhtlus-funktsioonid. Tartu, 2009, 196 lk.
14. **Liivi Hollman.** Basic color terms in Estonian Sign Language. Tartu, 2010, 144 p.
15. **Jane Klavan.** Evidence in linguistics: corpus-linguistic and experimental methods for studying grammatical synonymy. Tartu, 2012, 285 p.
16. **Krista Mihkels.** Keel, keha ja kaardikepp: õpetaja algatatud parandus-sekventside multimodaalne analüüs. Tartu, 2013, 242 lk.
17. **Sirli Parm.** Eesti keele ajasõnade omandamine. Tartu, 2013, 190 lk.
18. **Rene Altrov.** The Creation of the Estonian Emotional Speech Corpus and the Perception of Emotions. Tartu, 2014, 145 p.
19. **Jingyi Gao.** Basic Color Terms in Chinese: Studies after the Evolutionary Theory of Basic Color Terms. Tartu, 2014, 248 p.
20. **Diana Maisla.** Eesti keele mineviku ajavormid vene emakeelega üliõpilaste kasutuses. Tartu, 2014, 149 lk.
21. **Kersten Lehismets.** Suomen kielen väylää ilmaisevien adpositioiden *yli*, *läpi*, *kautta* ja *pitkin* kognitiivista semantiikkaa. Tartu, 2014, 200 lk.

22. **Ingrid Rummo.** A Case Study of the Communicative Abilities of a Subject with Mosaic Patau Syndrome. Tartu, 2015, 270 p.
23. **Liisi Piits.** Sagedamate inimest tähistavate sõnade kollokatsioonid eesti keeles. Tartu, 2015, 164 lk.
24. **Marri Amon.** Initial and final detachments in spoken Estonian: a study in the framework of Information Structuring. Tartu, 2015, 216 p.
25. **Miina Norvik.** Future time reference devices in Livonian in a Finnic context. Tartu, 2015, 228 p.
26. **Reeli Torn-Leesik**. An investigation of voice constructions in Estonian. Tartu, 2015, 240 p.
27. **Siiri Pärkson.** Dialoogist dialoogsüsteemini: partneri algatatud parandused. Tartu, 2016, 314 lk.
28. **Djuddah A. J. Leijen.** Advancing writing research: an investigation of the effects of web-based peer review on second language writing. Tartu, 2016, 172 p.
29. **Piia Taremaa.** Attention meets language: a corpus study on the expression of motion in Estonian. Tartu, 2017, 333 p.
30. **Liina Tammekänd**. Narratological analysis of Võru-Estonian bilingualism. Tartu, 2017, 217 p.
31. **Eva Ingerpuu-Rümmel**. Teachers and learners constructing meaning in the foreign language classrooms: A study of multimodal communication in Estonian and French classes. Tartu, 2018, 218 p.
32. **Kaidi Rätsep.** Colour terms in Turkish, Estonian and Russian: How many basic blue terms are there? Tartu, 2018, 181 p.
33. **Kirsi Laanesoo**. Polüfunktsionaalsed küsilaused eesti argivestluses. Tartu, 2018, 176 lk.
34. **Maria Reile**. Estonian demonstratives in exophoric use: an experimental approach. Tartu, 2019, 240 p.
35. **Helen Türk**. Consonantal quantity systems in Estonian and Inari Saami. Tartu, 2019, 149 p.
36. **Andra Rumm**. Avatud küsimused ja nende vastused eesti suulises argivestluses. Tartu, 2019, 217 lk.
37. **Eleri Aedmaa**. Detecting Compositionality of Estonian Particle Verbs with Statistical and Linguistic Methods. Tartu, 2019, 271 p.
38. **Kristina Koppel**. Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele. Tartu, 2020, 249 lk.
39. **Ilze Tālberga**. On the equivalents of the Latvian verbal prefixes in Estonian. Tartu, 2020, 210 p.
40. **Roger M. A. Yallop.** The affect and effect of asynchronous written artefacts (cover letters, drafts, and feedback letters) within L2 English doctorate writing groups. Tartu, 2020, 312 p.
41. **Mariann Proos**. Meaning and usage of Estonian experience perception verbs. Tartu, 2021, 175 p.
42. **Helen Hint**. From full phrase to zero: a multifactorial, form-specific and crosslinguistic analysis of Estonian referential system. Tartu, 2021, 271 p.

43. **Anton Malmi**. The production of Estonian palatalization by Estonian and Russian speakers. Tartu, 2022, 180 p.
44. **Mari Aigro**. In any case? Estonian spatial cases as argument markers. Tartu, 2022, 223 p.
45. **Maria Tuulik**. Adjektiivide süstemaatiline polüseemia eesti keeles taju-adjektiivide näitel. Tartu, 2022, 182 p.
46. **Rodolfo Basile.** Invenitive-locational constructions in Finnish: A mixed methods approach. Tartu, 2024, 180 p.