DISSERTATIONES
LINGUISTICAE
UNIVERSITATIS
TARTUENSIS
37

ELERI AEDMAA

Detecting Compositionality of Estonian Particle Verbs with Statistical and Linguistic Methods

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS
37

DISSERTATIONES LINGUISTICAE UNIVERSITATIS TARTUENSIS
37

ELERI AEDMAA

Detecting Compositionality of Estonian Particle Verbs with Statistical and Linguistic Methods

UNIVERSITY OF TARTU
Press

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

Dissertation accepted for the commencement of the degree of Doctor of Philosophy on June 25^{th}, 2019 by the Committee of the Institute of Estonian and General Linguistics, Faculty of Arts and Humanities, University of Tartu

Supervisors: Associate Prof. Kadri Muischnek (University of Tartu)
Kristel Uiboaed, PhD

Opponent: Carlos Ramisch, PhD (Aix-Marseille Université)

Commencement: October 4^{th}, 2019 at 14:15, in room 139, University main building, Ülikooli 18, 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
ISBN 978-9949-03-162-7 (print)
ISBN 978-9949-03-163-4 (pdf)

University of Tartu Press
www.tyk.ee

ACKNOWLEDGEMENTS

I have received a great deal of support and assistance throughout my PhD studies. First and foremost, I am sincerely and deeply thankful to my supervisors Kadri Muischnek and Kristel Uiboaed for offering their knowledge and guidance, but also for giving me free rein to pursue my research interests. I appreciate all their contributions of time, belief and funding to make my PhD studies as productive and successful as possible. I am also grateful for the opportunity to work under their supervision for numerous scientific projects. I am extremely thankful to Kadri for addressing theoretical matters and for being always willing to take time to answer my questions. My heartfelt thanks are due to Kristel because she continued supervising me after she left academia. Her friendship and constant faith are invaluable to me.

Without the expertise of my outstanding reviewers Dr. Carlos Ramisch (Aix-Marseille Université) and Dr. Veronika Vincze (MTA-SZTE Research Group on Artificial Intelligence, University of Szeged), my thesis would be incomplete. They both went the extra mile in providing useful, constructive and encouraging feedback. I am proud and honoured that Carlos agreed to be my opponent at the defence. I would like to express appreciation to my annotators and all the reviewers of my papers for their hard work.

During my studies, I was fortunate to visit two great research groups abroad. I spent three fruitful months in 2015 with the Computational Linguistics group at the University of Uppsala. I am grateful to professor Joakim Nivre, dr. Mats Dahllöf and other colleagues for the interesting seminars and discussions. In 2017, I visited the Institute for Natural Language Processing at the Stuttgart University. It was one of the highlights of my studies when dr. Sabine Schulte im Walde agreed to welcome me in her group. I learned a lot from her and dr. Maximilian Köper; I thank both for their contributions to my research. Many special thanks to all the colleagues from Uppsala and Stuttgart for making me feel pleasant at their homes.

Appreciation is due to all the financial support I received during my studies. I acknowledge the contribution of the Estonian Students Fund in USA, the ESSLLI 2014 Student Grant, the IC1207 COST action PARSEME, the Archimedes Foundation, Carlson Wagonlit Travel, the Tartu University Foundation, the Graduate School of Linguistics, Philosophy and Semiotics and the Association for Computational Linguistics. Thanks for helping to keep my head above water during my studies and travels. My research has been supported by the European Regional Development Fund (University of Tartu ASTRA Project PER ASPERA), the Centre of Excellence in Estonian Studies (TK145) and by the following projects: "Tools for Multi-layered Annotation of Text (applied to Estonian Reference Corpus)" (EKT7), "Computational models for Estonian" (IUT20-56), "Digital Resources and Spatial Data in Linguistics" (EKKM16-425) and "Estonian Universal Syntax: Resources and Applications" (EKTB7).

Although I have not been around lately, I got a kick out of working with Kristel, Maarja-Liisa, Ann, Mari, Triin and other colleagues. I am indebted to Liina Lindström for always being supportive and for involving me in various activities. In addition to the people at the Institute of Estonian and General

Linguistics, I praise excellent colleagues from the Institute of Computer Science for helping me to solve various (technical) issues and for offering me numerous chances to collaborate. Special thanks go to Mark Fišel, Heiki-Jaan Kaalep, Heili Orav and Haldur Õim.

The care from my family and friends has always reminded me that every cloud has a silver lining. My sincere gratitude goes to my best friend Teele, who has stayed by my side through thick and thin. Words would not be enough to thank my parents Ene and Erlend, my sister Einike and brother Ets for their endless support. Cheers to Katriine, Adeele, Elias and Lisanna for being such sources of joy. I am also grateful to my grandfather Heiki, who has always been interested in my research. *Didi madloba* to Lasha for his constant love, great care, useful advice and encouraging words. I am forever thankful to him for going the whole nine yards for me.

On the 20th of August 2019 in Groningen

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

Abbreviations used in the dissertation:

AM	lexical association measure
CBOW	continuous bag-of-words
CS	cosine similarity
DSM	distributional semantic model
EED	Estonian Explanatory Dictionary
ERC	Estonian Reference Corpus
MWE	multiword expression
MWV	multiword verb
NLP	natural language processing
POS	part-of-speech
PV	particle verb
VPC	verb-particle construction
WSD	word sense disambiguation

Abbreviations used in the glosses:

1, 2, 3	person	IMP	imperative
ABL	ablative	IMPS	impersonal
ADE	adessive	INE	inessive
ALL	allative	INF	infinitive
CL	clitic	NEG	negation
COM	comitative	PRT	partitive
COND	conditional	PL	plural
COMP	comparative	PST	past
CONNEG	connegative	PTCP	participle
ELA	elative	QUOT	quotative
GEN	genitive	SG	singular
GER	gerundive	SUP	supine
ILL	illative	TRL	translative

1 INTRODUCTION

1.1 Motivation

Every once in a while, I find my social media friends sharing a screen capture of how 'inaccurately' the well-known Google Translate tool[1] translates sentences or phrases in Estonian, often with absurd results. Machine translation does not usually fail when translating single words, but has difficulty with longer word-units such as phrases and sentences. Expressions that are not translatable word for word are particularly problematic. For example, one can attempt to translate the sentence *Ta läks lepinguga alt* into English. The Google translation is 'He went under the contract'. Although the machine ignored the fact that the word *leping* 'contract' is in the comitative case, the direct translation of the words is correct. However, the translation does not express the real meaning of the sentence – 'He was deceived by the contract/He was deceived when he signed the contract'. As the meaning of the sentence is not a sum of the meanings of its component words, the machine is not able to express the meaning of the sentence. Why is Google not able to understand the correct meaning of the sentence? What additional information is required for the correct translation?

Despite this provocative example, analysing and improving the quality of the machine translation is not the topic of this thesis. Nevertheless, the example illustrates why it is important to improve on the quality of processing (including translating) compounds of words that behave like one word. These kinds of word-units that contain more than one word are called multiword expressions (MWEs) in natural language processing (NLP). MWE processing is important for machine translation; other known NLP tasks, such as named-entity recognition, information retrieval and question answering, also benefit from it. Successful MWE processing creates a situation in which the information that MWEs convey does not become lost.

The aim of MWE processing is clear, but it is recognised as one of the most complex tasks for NLP applications because MWEs are a highly diverse class of constructions; thus, they cannot be treated the same way. Therefore, MWE processing has been divided into subtasks (such as MWE discovery, MWE identification, and so on), and many more or less successful methods have been introduced to improve the quality of solving these problems. For example, lexical association measures (AMs), distributional methods and machine learning appear in many studies concerning MWE modelling and processing. While the work on English (and other larger languages) has developed rapidly, comparatively few experiments on Estonian MWEs have been conducted. Because Estonian is a morphologically rich language, the treatment of Estonian MWEs has proved to be even more challenging than has the treatment of English MWEs, for example.

In summary, the main motivation of this work is to test well known methods of MWE processing on Estonian and to study how they work on a morphologically rich but relatively under-resourced language. In addition, the successful processing of Estonian MWEs is important in order to improve the quality of Estonian NLP tools as well.

[1] `https://translate.google.ee` (accessed 07.07.2018).

1.2 Research object

MWEs are a highly diverse class of linguistic constructions. For example, Constant et al. (2017) identify the following MWE categories that are commonly seen in the literature – idioms, light-verb and verb-particle constructions (VPCs), nominal/noun/verb compounds, complex function words, multiword name entities and multiword terms. This list reveals that it is challenging to find one common feature among all MWEs. Why MWEs cannot be treated uniformly was explained long ago (see e.g. Sag et al. 2002); thus, investigating different MWE types separately has been revealed to be an effective and well-recognised method.

Based on this widely adopted approach, the thesis does not study the entire range of MWEs, but focuses specifically on the exploration of the automatic processing of Estonian particle verbs (PVs). Similarly to other MWEs, PVs are a frequent phenomenon in Estonian, but the automatic processing thereof has not been studied in detail. Moreover, the compositionality of PVs has received minimal attention. Estonian PVs are similar to the English VPCs that Sag et al. (2002) categorised as syntactically flexible expressions. In addition, as Estonian has adopted numerous PVs from the German language (Hasselblatt 1990, as cited in Erelt 2013; Erelt et al. 2017), there is a significant amount of research on German PVs. This provides the opportunity to compare the behaviour of two similar linguistic phenomena in two languages from different language families.

1.3 Selection of methods

MWE processing involves many subtasks, and numerous methods have been proposed over the years (see the overview of computational research on MWEs in Section 2.3). Constant et al. (2017) mentioned two main subtasks, namely MWE discovery and MWE identification[2]. In general, MWE discovery focuses on finding new MWEs in text corpora, and MWE identification addresses automatic annotation of MWEs. From a semantic point of view, MWEs are considered to have some degree of compositionality; thus, the detection of compositionality has become central to semantic research on MWEs. The main methods used in MWE discovery are AMs, substitution and insertion, semantic similarity, and supervised learning. MWE identification has primarily been done using rule-based methods, classifiers and sequence tagging methods. (Constant et al. 2017) Some of these methods have also been employed for Estonian data. For example, AMs have been applied successfully to the automatic discovery of Estonian PVs (see Aedmaa 2014) and tested for classifying PVs (see Aedmaa 2017); Aedmaa (2016) demonstrated the possibility of using distributional methods to detect the compositionality of PVs. The first supervised classifier to detect the (non-)literalness of Estonian PVs was proposed by Aedmaa et al. (2018).

The discovery of Estonian PVs (that is, the detection of the compositionality of PVs, see Chapter 5) was conducted using distributional semantic models

[2]The literature often follows Baldwin and Kim (2010) in considering MWE identification to be a token-level task for determining individual occurrences of MWEs in running text, and MWE discovery to be a type-level lexicon induction task. We follow the distinction introduced by Constant et al. (2017).

(DSMs); a supervised classifier was developed for the identification of PVs – the classification of the literal and non-literal usage of PVs (see Chapter 6). The selection of the applied methods is based on previous research on particle-verb constructions. These methods have proven to be successful in the discovery and identification of MWEs in other languages. The choice of methods also follows the current trends in NLP. More precisely, the use of DSMs is motivated by the rise of word embeddings (that is, word or phrase representations that are mapped onto vectors of real numbers, see Section 3.2) that are widely adopted for performing a wide range of NLP tasks, including MWE processing. In addition, the feasibility of adopting these methods for research on the Estonian language was an important factor in selecting these methods, as it includes the potential for evaluating the applied methods. The evaluation of MWE discovery is considered to be complex (Constant et al. 2017) because human judgements, dictionaries and special datasets are often needed, and the creation of these kinds of datasets is unquestionably expensive.

MWEs have been studied in the context of many frameworks within the different sub-fields of (computational) linguistics (see Section 2). As a result, the methods applied in this study originated in different approaches to MWE processing. Nonetheless, this thesis does not follow any particular theoretical framework; instead of fitting the study into a particular approach, the results are described through definitions adopted from related research. Following common practice in computational linguistics, the study is corpus-based.

In summary, the applied methods were chosen based on their success in previous studies, their applicability to Estonian data and the feasibility of the assessment.

1.4 Research purposes and questions

This dissertation has multiple purposes.

The first aim is the automatic detection of the compositionality of Estonian PVs. Although some research on differentiating between compositional and non-compositional PVs was done previously, the goal of this dissertation is to determine which information is necessary in order to distinguish among the different levels of compositionality of PVs. The focus is on the (linguistic) information that can be obtained from electronic resources.

The second purpose is to introduce and apply methods that have been adopted widely and successfully for computational research on different kinds of MWEs in various languages, but which have not yet been used extensively for linguistic research on the Estonian language. While the ultimate goal is to create a model that detects the compositionality of PVs successfully and automatically, the thesis further investigates the advantages and drawbacks of the applied methods that are crucial with regard to the investigation of the compositionality of PVs.

The third goal is to provide a study that is influential for further research in this field. The analyses of methods and models presented in this thesis could be used as guidelines for further investigation of the semantic compositionality of other MWEs, or of other linguistic phenomena. The methods and models applied, the datasets that were created, and the trained word and multi-sense embeddings are

useful not only for various topics concerning NLP but also for other areas, such as lexicography.

In order to accomplish these goals, the research addresses the following questions:

1. **To what extent do human annotators agree with each other when evaluating the compositionality of PVs? What are the main reasons for disagreement?**

   The evaluation of the degree of compositionality might be challenging not only for computers but also for humans. Computational models predicting compositionality are evaluated against datasets containing human annotations (gold standards). Therefore, novel datasets including human judgements of PV compositionality were created for this research. The quality of the human-annotated datasets is evaluated by seeking agreement among annotators. This assessment not only indicates the value of the dataset but also presents cases of disagreement. The reasons for variance in annotations can thus be analysed and discussed.

2. **How well do DSMs predict the compositionality of Estonian PVs? Which training parameters and other aspects influence the quality of word and multi-sense embeddings for detecting the degree of compositionality?**

   DSMs have become widely used and successful methods in NLP. Among other tasks, a variety of distributional techniques has been applied to predict MWE compositionality in various languages (e.g. Baldwin et al. 2003; Reddy et al. 2011b; Cordeiro et al. 2016a). Although the models are applicable to all languages that have large text corpora, good human-annotated datasets are necessary for the evaluation of the DSMs. The compositionality predictions of DSMs are compared to the created datasets containing human judgements of compositionality. In addition, based on the research, there are general discussions concerning how model training parameters affect the quality of embeddings. Consequently, which parameter settings help to train embeddings that achieve the best possible results in terms of predicting the degree of compositionality of Estonian PVs can be determined. However, applying the embeddings in linguistic research to Estonian is a relatively new approach, and little work has been undertaken. Hence, the impact of some vital parameters on word and multi-sense embeddings to detect the compositionality of Estonian PVs is explored.

3. **Which (linguistic) features predict the use of compositional versus non-compositional PVs? How well are the values of these features automatically acquirable?**

   In DSMs, the meaning of the targeted linguistic unit (word, phrase, clause, sentence and so on) emerges via its context. Other than the lemmas of the surrounding words, there might be other contextually hidden features that help to predict the meaning of the target linguistic unit. Work on other languages has ascertained some standard features that predict literal versus

non-literal language usage. PV-specific features are unigrams (contextual lemmas), affective ratings (such as the abstractness of surrounding words), and the subject and the object of the PV. In addition to these features, Estonian language-specific features are suggested, and are applied to predict the literal versus the non-literal usage of Estonian PVs. In order to train a supervised classifier, the data must be labelled with information regarding the studied features. This allows for the analysis of the cost of annotation of these features.

4. **How does frequency affect the compositionality of Estonian PVs? How are human judgements of PV compositionality and automatic compositionality predictions associated with frequency?**

   The co-occurrence frequency of the components of PVs has been demonstrated to work well for the automatic processing of PVs, Krenn and Evert (see, for example, 2001); Uiboaed (see, for example, 2010); Aedmaa (see, for example, 2014). In addition, the role of frequency in compositionality (judgements) has been investigated (McCarthy et al. 2003; Bott and Schulte im Walde 2014) because the effect of frequency on language systems has long been discussed (see, for example, Bybee et al. 2007; Gries and Divjak 2012). Furthermore, the results of DSMs are affected by the frequency of the words (Sahlgren and Lenci 2016). Moreover, frequent words tend to be more polysemous than are infrequent ones, which poses a challenge not only while detecting compositionality automatically but also for humans when evaluating the degree of compositionality of PVs. Hence, it is important to study the association between frequency and the results of the models, while also evaluating the human-annotated datasets.

5. **Are widely adopted and successful computational methods suitable for detecting the compositionality of Estonian PVs? What are the drawbacks and benefits of these methods?**

   The automatic detection of MWEs has been studied for decades. As a result, numerous more or less successful approaches have been suggested for this task. The success of the method often depends on the availability and universality of the required resources. Hence, methods using unlabelled corpora are preferred to approaches in which an annotated dataset or large databases are needed. The applied methods were thus selected by considering the results reported earlier, and the applicability to Estonian data. Therefore, the quality of the results and the cost and convenience of the suggested methods to Estonian are also analysed.

1.5 Contributions

Word embeddings and machine learning are methods that are used commonly and successfully in NLP for numerous different tasks, such as parsing, sentiment analysis, machine translation, text classification and so on. However, these methods have not been widely adopted to study Estonian MWEs or other linguistic

phenomena. Therefore, in addition to exploiting and describing these methods for the automatic processing of Estonian PVs, one of the extensive contributions of this dissertation is to describe the methods in such a way that encourages their application within wider linguistic research.

The creation, evaluation and analysis of three novel datasets for the Estonian language are an original contribution of this research. While the aim of building these datasets was to support the development and assessment of the models, the description, assessment and investigation of these datasets provide valuable knowledge regarding how datasets containing compositionality information should be built. For example, the problems with automatically generated or crowdsourced datasets are discussed, the differences between collecting the compositionality ratings of PVs and PV meanings are outlined, and the reasons that compositionality is difficult for both computers and humans to evaluate are highlighted.

The compositionality of PVs is predicted utilising two approaches, namely DSMs and supervised classifiers. While word embeddings have been adopted for compositionality predictions previously, context-dependent compositionality predictions have been explored less often. Thus, in addition to word embeddings, multi-sense embedding models are introduced. Both models were trained with a variety of parameter settings. Thus, this research provides an overview of the influence of the parameter values on the quality of embeddings. As there is scant existing research on the impact of DSM parameters on embeddings for Estonian, this is an original contribution that will be available for future research. In addition, pre-trained word and multi-sense embedding models are made publicly available, and can therefore prove useful for a variety of future studies.

In addition to standard (language-independent) features, the study of the distinction between literal versus non-literal PV usage introduces a set of linguistic features that have not been previously investigated thoroughly as indicators of compositionality. In addition, the challenge of acquiring labelled data that are crucial for the development of the supervised models is analysed.

The frequency of the studied PVs and their components receives special attention throughout the thesis. How frequency is associated with the compositionality annotations and the compositionality predictions of word and multi-sense embedding models is investigated. In addition, frequency, as an indicator of PV literalness, is described. Therefore, the study provides a comprehensive overview of the relationship between frequency and PV compositionality.

The evaluation of the employed models shows how the applied methods contribute to the automatic processing of Estonian PVs' compositionality. From a theoretical point of view, the results of corpus-based experiments contribute to a wider and deeper understanding of PVs and their compositionality than what has been provided in the existing literature thus far (see Section 2.4 for details). Such a comparison provides evidence that complements the current treatment of the compositionality of Estonian PVs.

1.6 Chapters outline

The remainder of this thesis is structured as follows: Chapter 2 provides a background to MWEs and Estonian PVs, in order to explain the general properties

of MWEs, and discusses the compositionality of Estonian PVs. The treatment of MWEs and Estonian PVs in theoretical and computational linguistics is also discussed. Chapter 3 introduces the theoretical background to the chosen methods of research. Machine learning as a statistical method is discussed first, followed by an overview of distributional methods and supervised classification. The text corpora used and the three created datasets are introduced in Chapter 4. This chapter contains an exhaustive analysis of score distributions, an evaluation of annotations and an analysis of the effect of frequency on human compositionality judgements. Chapter 5 details the experiments and results of the out-of-context discovery of Estonian PVs – the detection of the degree of compositionality of PVs using unsupervised distributional semantic models. The impact of the training parameters on models' predictions is explored. In addition, the results of word and multi-sense embedding models are compared. The context-dependent identification of the literal versus the non-literal usage of Estonian PVs is provided in Chapter 6. This chapter presents further details regarding how the studied context features, including frequency, influence the results of a suggested classifier.

2 MULTIWORD EXPRESSIONS AND PARTICLE VERBS

This chapter provides background pertaining to MWEs and PVs that is fundamental to the thesis. In Section 2.1, MWEs are defined and the main properties of MWEs are described. A brief overview of discussions on MWEs in theoretical linguistics is given in Section 2.2. The computational research on MWEs, more specifically the discovery of MWEs and their compositionality, is reported in Section 2.3. Estonian PVs and their compositionality are introduced in Section 2.4. The same section includes analyses of the existing theoretical and computational research on Estonian PVs.

2.1 Multiword expressions and their properties

MWEs have many different definitions. For example, Sag et al. (2002) defined MWEs as "idiosyncratic interpretations that cross word boundaries (or spaces)". Van De Cruys and Moirón (2007) described them as "expressions whose linguistic behaviour is not predictable from the linguistic behaviour of their component words". Alternate definitions have suggested that "an MWE is a combination of two or more simplex word, covering compounds as well as collocations", as suggested by Kühner and Schulte im Walde (2010). According to the definition proposed by Tsvetkov and Wintner (2014), MWEs are "lexical items that consist of multiple orthographic words". The most exhaustive research on Estonian MWEs, which took their computational treatment into account, was conducted by Muischnek (2006), who suggested that "an MWE consists of two or more words that occur together to express some meaning". This research adopts the definition by Baldwin and Kim (2010): "MWEs are lexical items that: (a) can be decomposed into multiple lexemes; and (b) display lexical, syntactic, semantic, pragmatic and/or statistical idiomaticity". Several studies contain an overview of the existing definitions of MWEs (e.g. Ramisch 2014; Constant et al. 2017).

Due to the dissimilarities among MWEs, multiple and conflicting definitions of MWEs have been suggested. These definitions emphasise different features of MWEs because each category of MWE is distinct. Within the existing literature, the following phenomena have been defined as MWEs: compounds (for example, noun-noun, noun-verb, and verb-particle constructions), idioms, collocations, multiword name entities and multiword terms (Constant et al. 2017). This research focuses on Estonian PVs, which are comparable to English VPCs.

Sag et al.'s (2002) contribution is notorious within the field due to the argument that MWEs are a key problem hindering the development of large-scale, linguistically sound NLP technology. Since Sag et al. called MWEs "a pain in the neck of NLP", many studies have been conducted on the discovery of MWEs. Existing work is introduced in more detail in Section 2.3.1.

One must take multiple properties of MWEs into account while attempting to process them automatically. For example, Ramisch (2014) identified arbitrariness, institutionalisation, limited semantic and syntactic variability and heterogeneity, while Constant et al. (2017) mentioned arbitrarily prominent co-occurrence (collocation), discontiguity, non-compositionality, ambiguity and variability. Subsequently, we present an overview of the main characteristics of MWEs focusing on those that are challenging for the discovery and identification of Estonian PVs.

2.1.1 Co-occurrence

Collocation, that is, arbitrarily prominent co-occurrence, is one of the properties of MWEs (Constant et al. 2017). The common distinction between co-occurrence and collocations is that 'co-occurrence' is used as a term when describing general statistical understandings, while 'collocation' corresponds to a linguistically grounded approach (Seretan 2008).

The co-occurrence of words is defined as occurrences within the same linguistic unit; that is, a window (clause, sentence, paragraph, article and so forth). In MWE research, co-occurrence commonly describes the relationship between two words (Evert 2005). For example, the window needs to be defined in order to find all the adverb and verb co-occurrences in example (1). As the components of the PV appear in the same clause, it is reasonable to define the clause as a window. Thus, one needs to note that there are three clauses in the sentence and process these clauses separately. The adverb *lahti* 'open'[3] and the verb *tegema* 'to do' occur in the first clause *Ma tegin silmad lahti ja*, the verb *mõtlema* 'to think' in the second clause *mõtlesin*, and the verb *olema* 'to be' in the third clause *et on esmaspäev*. Therefore, there is only one adverb-verb co-occurrence (*lahti tegema*) in example (1).

(1) Ma teg-i-n silma-d lahti ja mõtle-si-n, et on esmaspäev.
I do-PST-1SG eye-PL open and think-PST-1SG that be.3SG Monday
'I opened my eyes and thought it was Monday.'

The frequent co-occurrence of words might indicate that the words form a collocation. For example, the adverb *ette* 'in advance/ahead/forward' and the verb *heitma* 'to throw' occur together more frequently than do the adverb *ette* and the verb *virutama* 'to whack'[4]. The components of the first combination occur together much more often than do the components of the second combination. *Ette heitma* is a PV with a non-compositional meaning 'to reproach/blame', but *ette* and *virutama* do not occur frequently in the same clause. Although simple co-occurrence information alone is not sufficient for the automatic discovery of MWEs (Manning and Schütze 1999), the feature has been a basis for many more sophisticated methods, such as AMs (see Section 2.3), developed for the MWE discovery task.

2.1.2 Non-compositionality

The compositionality of MWEs is a consequential relationship between the whole and its parts (Bannard et al. 2003), expressed by the degree to which the semantics of the parts of an MWE contribute to the meaning of the whole (Li and Sporleder 2009). Van De Cruys and Moirón (2007) claimed that, of all the properties,

[3]Note that the English translations are provided for better readability for non-Estonian speakers; they do not include an exhaustive list of interpretations.

[4]According to the Google search outputs (18.07.2018) for the queries 'ette heitma' and 'ette virutama', the lemmas *ette* and *heitma* are adjacent to each other 5,360 times, but *ette* and *virutama* only once.

semantic non-compositionality was the most influential property of MWEs. A generally adopted idea is that MWEs form a continuum from fully compositional to fully non-compositional expressions (Moon 1998). The meaning of compositional MWEs is transparent, and is the sum of the meanings of the components of the MWE. The meaning of non-compositional MWEs is not derivable from the meanings of its components. For example, the meaning 'to admit defeat' is not obtainable from the meanings of the components of the phrasal verb 'to give up'.

Addressing with the compositionality of MWEs is an important task not only for NLP applications but also for lexicographers in order to decide which expressions should be treated as lexicon entries (Kühner and Schulte im Walde 2010: 47). Due to their idiosyncratic behaviour, Sag et al. (2002) suggested that MWEs needed to be described in lexicons. However, this is not a realistic approach because it is prohibitively expensive to build large lexical resources and to keep them updated. Therefore, automated methods for the discovery of MWEs in large text corpora are more beneficial for NLP.

Non-compositional MWEs are considered to pose a special challenge for NLP applications (Lin 1999: 317) because their meaning is not easily ascertainable. This is also why word-to-word translations of non-compositional expressions are prone to generating absurd outcomes. Thus, it is necessary to identify non-compositional expressions automatically. Instead of classifying MWEs according to binary classes, which is a difficult and somewhat outdated approach, an increasing number of empirical studies concentrate on the degree of compositionality of MWEs. Some of the earliest studies on MWE compositionality were carried out more than 15 years ago (e.g. McCarthy et al. 2003). The compositionality of Estonian PVs is explained in Section 2.4.1.

Because of MWEs' non-compositionality, synonyms, equivalent words or constructions cannot be substitutes for the components of MWEs. For example, the verb *heitma* 'to throw' cannot be substituted by its synonyms (for example, *viskama*, *pilduma* and *paiskama*) and maintain the meaning of 'to reproach/blame' in the PV *ette heitma*. Methods based on substitution and insertion are used for automatic discovery of MWEs (see Section 2.3.1).

Besides the notion of compositionality other terms are used to designate the formation and usage of MWEs. A non-compositional expression, a combination of words, meaning or usage is frequently defined as 'non-literal' or 'idiomatic'. Similarly, instead of the word 'compositional', the terms 'literal' and 'non-idiomatic' are often employed. The (non-)compositionality of one kind of MWE – idioms – has been discussed thoroughly in the existing literature, with some authors arguing that non-compositionality is a defining property of idioms, while others have claimed that both non-compositional and compositional idioms exist (Villada Moirón 2005). The notions 'compositionality', 'literality', and 'literalness' are used as synonyms in this thesis[5], but the terms 'idiom' and 'idiomaticity' are avoided for the sake of clarity. Therefore, PVs are MWEs with various degrees of compositionality (or literalness).

[5]Note that 'compositionality' has not always viewed as 'literality/literalness' (Reddy et al. 2011b).

2.1.3 Ambiguity

Although MWEs tend to be less polysemous than single-words (Finlayson and Kulkarni 2011: 20), the fact that many words have multiple meanings complicates the task of MWE discovery. From a computational linguistics point of view, it is important to discriminate among the different senses of a word within a given context (Iomdin et al. 2016: 214). According to Constant et al. (2017), the most influential type of ambiguity for MWE processing is the choice between a compositional and a non-compositional reading of a sequence of words.

For example, the following sentence (see example (2)) can have two readings. The first, interpretation would be that a friend used physical force – 'My friend pushed me forward'. Hence, the reading is more compositional. The second interpretation would suggest that this friend was acting in an encouraging manner – 'My friend pushed me (to do something)'. This reading is less compositional.

(2) Mu sõber **tõuka-s** min-d **tagant**.
I-GEN friend push-PST.3SG I-PRT from behind
Lit. 'My friend pushed me from behind.'
'My friend pushed me forward.'
'My friend encouraged me.'

In some cases, morphological and syntactic analyses can aid in determining whether the sequence of words should be recognised as an MWE (Constant et al. 2017). For example, the meaning of the PV *järele vaatama* 'to watch someone' in the first sentence (see example (3)) is more compositional, but less compositional – 'to look up' – in the second sentence (see example (4)).

(3) Ta **vaata-s** mehe-le **järele**.
s/he look-PST.3SG man-ALL after
Lit. 'She/he looked after the man.'
'Her/his eyes followed the man.'

(4) Ta **vaata-s** interneti-st **järele**.
s/he look-PST.3SG Internet-ELA after
Lit. 'She/he looked after from the Internet.'
'She/he looked up on the Internet.'

The difference in the meanings is detectable with the help of the case of the argument – the compositional meaning requires an argument in the allative case, while the non-compositional meaning is detectable with the help of an argument in the elative case.

2.1.4 Other properties

Among the properties of MWEs mentioned previously, there are further considerations for automatic MWE processing. Two of them – discontiguity and variability – are discussed briefly below.

In the text, the words not belonging to an MWE can appear in between the core elements of discontinuous MWEs (Kaalep and Muischnek 2008). For example, Estonian word order is heterogeneous, which means that the components of the PV (or other MWEs) are not necessarily adjacent or in a particular order (particle + verb). In fact, there can be several intervening words between the components. The problem of discontiguity could be solved with the help of syntactic analysis, or parsing (Constant et al. 2017). Numerous parsing tools are available for Estonian (see Muischnek et al. 2016) which can be applied to resolve the discontiguity of Estonian MWEs (including PVs).

Variability is also one of the properties of MWEs that might create obstacles to finding all possible forms of the same MWE (Constant et al. 2017). As Estonian is a morphologically rich language, the verbal components of an Estonian PV can occur in a variety of surface forms within a text. Nevertheless, the quality of the existing Estonian morphological analyser (Kaalep 1997) is sufficient so that variability does not pose a challenge for processing Estonian PVs.

2.2 Multiword expressions in theoretical linguistics

Theoretical discussions about MWEs usually adopt one of two approaches – the first takes a previously developed theory and adapts it to MWEs. The second observes the properties of MWEs and modifies the theory accordingly. Hence, MWEs are modelled inside frameworks, or new insights into the properties of words and grammatical processes are collected. (Sailer and Markantonatou 2018)

MWEs (and idioms, collocations, and phraseologisms) have been discussed by many researchers in multiple sub-fields of linguistics using numerous theoretical frameworks. Based on the distinction between lexicon and grammar, approaches are generally divided into dual-system theories and single-system theories (Snider and Arnon 2012). For example, in recent decades, different approaches to generative grammar have discussed the treatment of idioms (e.g. Chomsky 1965; Nunberg et al. 1994; Chafe 1968; Weinreich 1969; Fraser 1970; Jackendoff 1975; Wasow et al. 1983; Jackendoff 1997; Culicover 1999), including Lexical Functional Grammar (e.g. Attia 2006) and Head-driven Phrase Structure Grammar (e.g. Krenn and Erbach 1994; Sailer 2003; Richter and Sailer 2009; Webelhuth et al. 2018). Snider and Arnon (2012) concluded that, as generative approach is a dual-system theory, it differentiate between compositional and non-compositional phrases. Compositional phrases are generated by a grammar, while non-compositional ones originate from the lexicon where they are stored, together with their idiosyncratic, syntactic and semantic features. Compositional phrases are derived in a predictable way, and there is no need to keep them in the lexicon. (Snider and Arnon 2012)

Single-system models do not divide compositional and non-compositional linguistic units in such a way. Instead, compositional and non-compositional phrases are processed similarly, and like any other linguistic pattern. This approach is characteristic of construction grammar (Fillmore et al. 1988) and usage-based approaches in which grammatical knowledge emerges from linguistic experience. (Snider and Arnon 2012)

Many other approaches have anticipated views on the treatment of MWEs, although the term 'MWE' has not been always used. To name only a few, lexicon-

grammar (e.g. Giry-Schneider 1978; Freckleton 1985; Gross 1986; Laporte 2018), meaning-text theory (e.g. Mel'čuk and Polguere 1987; Mel'čuk 1998) and frame semantics (e.g. Fontenelle 2001). In addition, collocations have been studied in connection with semantic prosody (Louw 1993; Hoey 1997). An increasing amount of (theoretical and experimental) psycholinguistic and cognitive research on MWEs has been conducted (e.g. Dahlmann and Adolphs 2007; Eskildsen and Cadierno 2007; Lavagnino and Park 2010; Nematzadeh et al. 2013; Siyanova-Chanturia 2013). For example, Dahlmann and Adolphs (2007) explored whether the analysis of pauses might be useful in the validation of automatically discovered MWE candidates as MWEs. The authors' aim was to test the assumption that MWEs are stored in the mental lexicon, and are therefore produced without pauses. Nematzadeh et al. (2013) studied how children learn to identify and interpret different types of multiword lexemes.

This study does not concentrate on developing any of these theories; thus, these approaches are not discussed further. Moreover, several thorough overviews of the MWE treatment describing some of these frameworks have recently been proposed (e.g. Gries 2008; Seretan 2008; Sailer and Markantonatou 2018). As the scope of this thesis is focused narrowly on Estonian PVs and their acquisition, an overview of the theoretical treatment of Estonian PVs is provided in Section 2.4.2.

2.3 Multiword expressions in computational linguistics

The study of MWEs in computational linguistics is connected firmly to the availability of large text corpora and computers that are sufficiently powerful to analyse them. The first papers in the field described methods of collocation discovery. Choueka (1988) proposed a method based on n-gram statistics, while Smadja (1993) suggested a tool called Xtract that uses part-of-speech (POS) filters and some statistical measures. Church and Hanks (1990) introduced mutual information an AM, and Justeson and Katz (1995) combined POS information and frequency for the discovery of technical terms from the text. As a criticism of the use of AMs thus far, Dunning (1993) represented a more sound theory of his approach – he suggested the log-likelihood measure, which is a popular method for the automatic discovery of MWEs. One of the most well-known works on MWEs in NLP is by Sag et al. (2002), who analysed MWEs and explained why they pose such a challenge for NLP.

Since the aforementioned seminal papers, much work on different areas of MWE processing has been conducted. With regard to the goals of this thesis, the following sections introduce previous research that has focused solely on the automatic discovery and identification of MWEs and their compositionality by focusing mainly on studies of VPCs and PVs. As the following overview describes the work on other languages, the related computational research on Estonian PVs is introduced in Section 2.4.3. Other tasks in MWE processing are beyond the scope of this study.

2.3.1 Discovery and identification of multiword expressions

The previous work on the discovery and identification of MWEs, focusing on their compositionality, is described in this section. According to the explanation by Constant et al. (2017), MWE processing consists of MWE discovery and MWE identification. Following this distinction, the automatic detection of MWEs falls under the umbrella of MWE discovery, namely finding new MWEs in corpora. As one of the main features of MWEs, semantic compositionality has attracted much attention. The purpose of the research is to decide whether the semantics of the sequence of words is compositional. This includes tasks such as detecting the degree of compositionality of MWEs and the classification of MWEs according to their compositionality. Large corpora are usually used, and the studies are focused on a certain type of MWE at a time (Ramisch 2014). As most existing work has been undertaken in English and German, the overview focuses on these languages. However, some notable studies on other languages are also mentioned.

As noted earlier (see Section 1.3), the automatic discovery of MWEs began using different AMs – a method to estimate the strength of the association between two words based on their frequency of co-occurrence. In addition to pointwise mutual information (PMI) (Church and Hanks 1990) and log-likelihood measure (Dunning 1993), other more or less sophisticated AMs have been utilised for candidate discovery. Some of them (for example, t-score) are based on hypothesis testing. Others (such as χ^2) use contingency tables and are considered to be more sophisticated. For example, Pedersen (1996) suggested Fisher's test for automatic MWE discovery. While the initial studies concentrated on one AM at a time, multiple AMs were later explored and compared (e.g. Krenn 2000; Weeber et al. 2000; Schone and Jurafsky 2001; Pearce 2002; Bartsch 2004; Evert 2005; Ramisch et al. 2008; Hoang et al. 2009; Ramisch et al. 2012). For example, Pecina (2008) provided an exhaustive overview of 54 AMs in his study. AMs and their combinations have been applied to discover MWEs such as German PP-verb constructions (e.g. Krenn and Evert 2001; Wermter and Hahn 2004), German adjective-noun pairs and preposition-noun-verb triples (e.g. Evert and Krenn 2001), and noun-verb constructions in Japanese and English (e.g. Pereira et al. 2014). Some studies have concentrated on specific MWEs (e.g. Evert and Kermes 2003), while others have addressed a vast number of different MWEs (in different languages) (e.g. Seretan 2008). Constant et al. (2017) stated that, although much (comparative) work on AMs has been published, no single best measure has been identified.

In addition to and in combination with AMs, other methods have been applied to MWE discovery. For example, Pearce (2001) used WordNet to discover collocations while substituting synonyms of candidate components. Baldwin and Villavicencio (2002) combined syntactic evidence using automatic POS taggers and statistical chunkers, and added evidence from a number of tokens to a memory-based learner for the discovery of verb-particle expressions. While still using AMs such as mutual information (MI) and χ^2, Ramisch et al. (2008) proposed a new unsupervised measure that used syntactic permutations obtained via reordered words inside MWEs. As another example of unsupervised learning, Blaheta and Johnson (2001) discovered English VPCs using log-linear models.

They demonstrated that, instead of estimating which particles belonged together, and with which verbs, this information should be received from the parsed input.

One of the first papers to explore the discovery of idiomatic expressions was by Lin (1999), who presented a method that used the statistical properties of non-compositional expressions in a text corpus. He compared the MI values of non-compositional phrases and phrases that were similar to their literal meaning. Their experiment showed that non-compositional phrases have significantly different MI values than have compositional ones.

Models based on semantics have been studied widely due to MWEs' compositionality. DSMs are gaining popularity for the discovery of MWEs. In these models, senses are represented as vectors of co-occurring context words (see e.g. Baldwin et al. 2003). For example, McCarthy et al. (2003) and Bannard et al. (2003) focused on VPCs. While Bannard et al. (2003) studied distributional techniques to classify verb-particles as compositional and non-compositional, McCarthy et al. (2003) judged the compositionality of verb-particles using an automatically acquired thesaurus, and examined various measures using the nearest neighbours of the PV, thus creating a list of 116 VPCs with compositionality judgements from four annotators. They detected a significant relationship between human compositionality judgements and the measures taking into account the semantic of the particle. Venkatapathy and Joshi (2005) measured the relative compositionality of verb-noun collocations using a Support Vector Machine-based ranking function. They concluded that the correlation between the ranks computed by SVM-based ranking function and the human ranking is significantly better than the correlation between the ranking of individual features (such as frequency, distributed frequency of the object, dissimilarity of the collocation with its constituent verb using the LSA model and so on) and human ranking. Fazly and Stevenson (2006) adopted an automatically generated thesaurus from Lin (1999), and used lexical and syntactic fixedness as partial indicators of non-compositionality. They demonstrated that the measures they introduced outperformed a widely used PMI. Schulte im Walde et al. (2013) predicted the compositionality of German noun-noun compounds and found that window-based features outperformed syntax-based features. Bott and Schulte im Walde (2014) succeeded in evaluating the degree of compositionality of German PVs via a DSM that relied on word window information without syntactic information. The ranking of PVs according to the distributional distance they showed from their corresponding base verbs correlated to human judgement. The same authors consequently presented a successful distributional approach to the same task by modelling changes in syntactic argument structure (Bott and Schulte im Walde 2015). Schulte im Walde et al. (2016) predicted the degree of compositionality of the English and German noun-noun compounds within a vector space model, and demonstrated that the empirical and semantic properties of the compounds and the head nouns played a significant role.

Classification methods of MWEs often use contextual features. For example, the approach introduced by Katz and Giesbrecht (2006) is based on latent semantic analysis (LSA) to examine the compositionality of MWEs. In order to classify German verb-noun MWEs based on their compositionality, they proposed that vector similarity between distribution vectors associated with an MWE

as a whole, and those associated with its parts, would be a good measure of the degree to which an MWE was compositional. Cook et al. (2007) developed techniques for the semantic classification of potential English verb-noun expressions. They successfully used information about the syntactic behaviour of an expression type to determine whether a specific expression was used idiomatically or literally. Their unsupervised approach reached an accuracy of 72.8%. Boukobza and Rappoport (2009) proposed a supervised learning method that used surface features of sentences based on the canonical forms of expressions. They showed that their, more sophisticated model than others introduced previously for MWE identification, was able to improve the performance of identifying (any kind of) English MWEs. (Bhatia et al. 2017) explored the extent to which the English VPCs could be treated compositionally as opposed to idiomatically, relying on the WordNet hierarchy. In order to compositionally compute the meaning of a wide range of VPCs, they identified core senses of particles that have broad application across verb classes and used this information to build computational lexicons. In addition to English and German MWEs, models for other languages have been developed. For example, Uchiyama et al. (2005) identified Japanese verb compounds using statistical and rule-based methods. While the rule-based method outperformed the statistical method, they suggested that fine-grained semantic analysis is important for the Japanese compound verb disambiguation. Hashimoto and Kawahara (2008) explored the behaviour of Japanese idioms and improved upon the previous state-of-the-art method on transitive verb disambiguation task and on a compositionality detection task using a method for jointly learning compositional and non-compositional phrase embeddings by weighting both types of embeddings using compositionality scoring function.

From unsupervised methods, clustering has been applied to identify MWEs. For example, Birke and Sarkar (2006) introduced a system that differentiated between the literal and non-literal usage of English verbs with an accuracy that was 24.4% higher than the baseline. They suggested that their system is applicable to all sorts of non-literal language and that it could be adapted to other POS and other languages. Kühner and Schulte im Walde (2010) used unsupervised clustering to determine the degree of compositionality of German PVs. Their simpler cluster approach predicted the degree of compositionality for 59% of the particle verbs, while the correlation with the gold standard was 0.43. They concluded these results as reasonable because they worked with very simple distributional features. In order to improve the prediction of compound-constituent compositionality, Bott and Schulte im Walde (2017) suggested soft clustering to separate the different senses of a word type and predicted the degrees of compositionality of German noun-noun compounds and PVs. They showed that both types of MWEs benefited from the use of clustering in distributional modelling.

Translation has been helpful for discovering MWEs. For example, Pichotta and DeNero (2013) discovered English phrasal verbs by using a token-based method, which was applied on parallel corpora for 50 languages. They demonstrated that combining statistical evidence from many parallel corpora using the ranking-oriented boosting algorithm resulted to a list of MWEs that was comparable to human-curated set. Salehi and Cook (2013) proposed a type-based approach, utilising translation data from multiple languages, and string similarity

between an MWE and each constituent component. They found their method to be competitive with others addressing English noun compounds and VPCs. Salehi et al. (2014) reimplemented their method using the longest common substring, and subsequently combined it with a distributional similarity-based method that measured the distributional similarity between each component word and the overall expressions. They showed that using translation and multiple target languages enhances compositionality modelling.

Salehi et al. (2015) presented the first attempt to use word embeddings to predict the compositionality of MWEs. They combined their approach with information from string similarity, and achieved satisfying results for three compositionality datasets. For the prediction of compositionality of German noun-noun and PVs, Köper and Schulte im Walde (2017b) compared a neural network DSM, relying on textual co-occurrences, with a multi-modal model extension that integrates visual information. They showed that visual features contribute differently for verbs than for nouns and images complement textual information if the textual modality is poor and appropriate image subsets are used or the textual modality is rich and large images are added.

Verbal MWE identification was the goal for the two editions of PARSEME Shared Tasks in 2017 and 2018. The first edition of the PARSEME Shared Task in 2017 (Savary et al. 2017) provided annotated datasets for 18 languages, where the goal was to identify verbal MWEs in context. The main outcome of the shared task was a multilingual 5-million-word annotated corpus. However, most of the seven systems submitted for the task used techniques originally developed for parsing, one system exploited neural networks. The best F-scores were obtained for Farsi, Romanian, Czech and Polish, while modest performance resulted for Swedish, Hebrew, Lithuanian and Maltese. For the 1.1. edition of the PARSEME Shared Task in 2018 (Ramisch et al. 2018), corpora were created for 20 languages. Most of the submitted systems exploited neural networks (such as the best system TRAPACC that was ranked first for 8 languages), but syntactic trees and parsing methods (such as TRAVERSAL that was ranked first for 7 languages), tree-structured CRF, statistical methods, association measures and Naïve Bayes classifier were also employed in some systems. The highest F-scores (90.31 and 85.28) were obtained for Hungarian and Romanian, the lowest scores (23.28, 32.88 and 32.17) for Hebrew, English and Lithuanian. For both shared tasks, it was suggested that the results depend on the amount of annotated training data.

To summarise, many methods for the discovery and identification of MWEs have been introduced in the literature. For MWE processing, a wide range of results has reported as state-of-the-art depending on the language and datasets (Taslimipoor et al. 2018). However, the work on MWE discovery and identification continues because large-scale discovery evaluation and broad-coverage MWE identification are still open issues (Constant et al. 2017). The results of some above-mentioned studies are described in detail and compared to the current research in Section 7.2.

2.3.2 Compositionality datasets of particle-verb constructions

One of the problems with compositionality studies is the lack of representative datasets of human judgements. These datasets are necessary for training and evaluating models of discovery and for the identification of MWEs. Some of the datasets that have been created for other languages are described in this section. Given that the main aim of this research is to explore the compositionality of PVs, the primary focus is on datasets containing compositionality information of similar phenomena from other languages.

The dataset introduced by Bannard et al. (2003) contains English VPCs. They defined compositionality as an entailment relationship between a whole and its component parts. Twenty-six annotators evaluated 40 sets of five sentences, in which each of the five sentences contained one particular VPC. Two questions were asked for each set – 1) whether the VPC implied the verb and 2) whether the VPC implied the particle. Annotators evaluated the VPCs by answering the questions with 'yes', 'no' or 'don't know'. Hence, the dataset contains binary classifications of VPCs with regard to the base verb and the particle.

Another dataset with human judgements is available for 389 English VPCs, with the particle 'up' balanced across three different frequency bands (Cook and Stevenson 2006). This dataset was created for the prediction of particle senses. Two annotators evaluated all the VPCs, and the final dataset includes only VPCs upon which both annotators agreed. They focused on the classification of meanings of the particle 'up'; thus, the dataset is not specifically a compositionality dataset.

Hartmann (2008), who collected compositional ratings for 99 German PVs across 11 different particles and eight frequency bands, compiled the first compositionality dataset for German PVs. Three annotators were asked to use their own words to indicate ambiguities and to describe differences. These compositionality ratings do not distinguish among different word senses.

Ghost-PV (Bott et al. 2016) is another gold standard for German PVs, and consists of 400 randomly selected PVs. The resource is balanced across several PVs and three frequency bands. The annotations for this dataset were crowdsourced. Annotators were asked to evaluate the extent to which the meaning of a PV was related to the meaning of its base verb. Rating was done on a scale from 1 to 6. The final dataset consisted of 400 PVs and information about the frequency bands of PVs, the number of human ratings for each PV, the standard deviation of ratings among rates and so on.

For the classification of the literal versus the non-literal usage of German PVs, Köper and Schulte im Walde (2016b) created a dataset containing sentences with 159 PVs and information about the degree of compositionality of the PV for each sentence. Three annotators evaluated 8,128 sentences on a six-point scale, and determined the degree of literalness of usage of the PV for each sentence. The dataset was used to train the classifier for the binary classification of the literal versus the non-literal usage of German PVs.

The PARSEME shared task 1.1 on automatic identification of verbal MWEs included numerous languages that contain VPCs, such as Arabic, English, German, Greek, Hebrew, Hungarian and Italian. They annotated two types of VPCs –

fully (in which the particle totally changes the meaning of the verb) and semi (in which the particle adds a partly predictable but non-spatial meaning to the verb) non-compositional VPCs. Although the annotation did not include compositionality ratings, the context-based manual binary classification of VPCs was provided. (Ramisch et al. 2018)

In conclusion, there are multiple datasets available for English VPC and German PVs. However, the fact that English and German are large languages allows them to create very specific datasets (e.g. Cook and Stevenson 2006), collect human judgements via crowdsourcing (e.g. Bott et al. 2016) and annotate large number of sentences (e.g. Köper and Schulte im Walde 2016b).

2.4 Estonian particle verbs

An Estonian PV is a productive MWE that consists of a verb and a verbal particle. For example, in example (5), the verbal component of the PV *tõusma* 'to wake up' is *tõusma* 'to wake' and the particle is *üles* 'up'. In example (6), *maha* 'off' is the particle and *müüma* 'to sell' is the verbal component. According to Estonian grammar (Erelt et al. 1993), a verb is the semantic core of the PV and the verbal particle specifies the connotation which expresses direction, perfectivity, state or modality.

Particles such as *alla* 'down', *eemale* 'away/off', *ette* 'in advance/ahead/forward', *juurde* 'by/up', *kaasa* 'along', *maha* 'down/off', *peale* 'on', *sisse* 'in', *taha* 'behind', *üles* 'up', *all* 'below/under', *alt* 'from under', *läbi* 'through', *mööda* 'along', *ringi* 'round', *üle* 'over' and so forth express direction (like the particle *üles* in example (5)). Perfectivity is expressed by particles such as *ära* 'away/out/off', *läbi* 'through', *maha* 'off', *otsa* 'out', *täis* 'up', *valmis* 'ready', *välja* 'out' and so on. (like the particle *maha* in example (6)). Particles *kinni* 'up/to', *lahti* 'open/loose', *kokku* 'together/up', *viltu* 'wrong' may indicate the state (such as particle *kokku* in example (7)). Particles *vaja* 'need' and *tarvis* 'need' express modality (like the particle *vaja* in example (8)).

(5) Andres **tõus-is** kell kaheksa **üles**.
Andres **rise-PST.3SG** o'clock eight **up**
Lit. 'Andres rose up at eight o'clock.'
'Andres woke up at eight o'clock.'

(6) Mari **müü-s** kõik raamatu-d **maha**.
Mari **sell-PST.3SG** all book-PL **off**
'Mari sold off all the books.'

(7) Korsten **kukku-s** eile **kokku**.
chimney **fall-PST.3SG** yesterday **together**
Lit. 'The chimney fell together yesterday.'
'The chimney collapsed yesterday.'

(8) Mu-l **on** pliiatsi-t **vaja**.
I-ADE **be.3SG** pen-PRT **need**
'I need a pen.'

PVs are challenging not only for NLP tasks but also for theoretical linguistics. For example, Veismann and Sahkai (2016: 270) highlighted four issues regarding the determination of PVs: 1) The distinction between noun-verb constructions and adverb-verb constructions is not clear, 2) the category of verbal particles as a separate word class is not explicit, 3) the category of PVs is not sound, and 4) the syntactic status of PVs is not definite. As these problems are not associated directly with the compositionality of PVs, these problems are not discussed further in this thesis. However, a brief overview of the theoretical and computational research on Estonian PVs with slightly more attention to semantic work is presented. The description begins with a synopsis of how semantic compositionality of PVs has been treated within existing research.

2.4.1 Compositionality of Estonian particle verbs

According to theoretical approaches (Rätsep 1978; Erelt et al. 1993, 2017), Estonian PVs can be classified as compositional or idiomatic, depending on whether the constituents retain their meanings or not. The components of a compositional PV take a literal meaning. For example, in example (9), the meaning of the PV *maha võtma* 'to take down' is a composition of the meanings of its components, *maha* 'down' and *võtma* 'to take'. The meaning of a non-compositional PV is idiosyncratic and cannot be inferred from the literal meanings of the verb and the verbal particle; for example, the meaning of the PV *maha võtma* is 'to lose weight' in example (10).

(9) Lava **võe-t-i** kohe pärast kontserti-∅ **maha**.
stage **take-IMPS-PST** immediately after concert-PRT **down**
'The stage was taken down right after the concert.'

(10) Ta **on** kümme kilo-∅ **maha** **võt-nud**.
s/he **be.3SG** ten kilogram-PRT **down** **take-PST.PTCP**
Lit. 'She/he has taken down 10 kilograms.'
'She/he has lost 10 kilograms.'

Some observations about the division of PVs have been stressed: a) the binary division of PVs is most obvious within verbal particles that express orientation (Erelt et al. 1993: 21), b) adverbs expressing state appear as verbal particles only as components of non-compositional PVs. (Erelt et al. 1993: 21–22), and c) the classification into compositional and non-compositional PVs is most difficult for the PVs that contain verbal particles that express perfectivity (Veismann and Sahkai 2016: 272). The aforementioned, widely adopted view that MWEs form a continuum between entirely compositional and entirely non-compositional expressions (see Section 2.1.2) was mentioned by Muischnek (2006: 12), but has not

been broadly adopted by Estonian theoretical linguists. The next section describes how PVs and their compositionality have been described in previous research in more detail.

2.4.2 Theoretical research on Estonian particle verbs

The notion *ühendverb* 'particle verb' was first used by Muuk (1938), who stated that a PV was a verb construction that has its own figurative meaning. This definition included verb-particle and noun-verb compounds, such as the constructions shown in examples (11), (12) and (13).

(11) jalga-∅ laskma
foot-PRT let
Lit. 'to shoot the foot'
'to leave'

(12) osa-∅ võtma
part-PRT take
'to take part'

(13) keha-∅ kinnitama
body-PRT assure
Lit. 'to assure the body'
'to eat'

Mihkla (1964) suggested that verbs are the head of phraseological units, and that complements of these kinds of expressions can be nouns (for example, *aega viitma* 'to dawdle', lit. 'to spend time', *jagu saama* 'to overcome', lit. 'to receive a part', *nõuks võtma* 'to decide', lit. 'to take for advice' and so forth), adpositional phrases (for example, *südame peale panema* 'to urge/to recommend strongly', lit. 'to put on the heart'), adjectives (or participles) in partitive or translative cases (*haiget tegema* 'to hurt', lit. 'to do/make sick', *pahaks minema* 'to become spoiled', lit. 'to go bad'), prefixal adverbs (such as *alt vedama* 'to let down', lit. 'to pull from under') and sometimes even a pronoun. Hence, he categorised non-compositional PVs as phraseological units, while compositional ones were called 'particle verbs'. Later (see Mihkla et al. 1974), a subgroup of phraseological units was defined as 'phraseological PVs', which included expressions formed by a verb as the head of a phrase, and a nominal as the complement of a phrase. While the nominal is usually in the partitive or illative case, the expression is only used figuratively, as in *jalga laskma* 'to leave', lit. 'to shoot the foot', *nõuks võtma* 'to decide', lit. 'to take for advice' and *ühte hoidma* 'to stick together', lit. 'to keep one'.

In the first part of his grammar, Tauli (1973) described PVs as expressions that contain verbs and nouns/adverbs in which the noun or adverb can occur before or after the predicate. In the second part of the grammar, Tauli (1983) mentioned an MWE that consisted of a verb and an object. He suggested that some of these kinds of expressions are semantic compounds, most of which are so-called 'PVs', such

as *habet ajama* 'to shave', *kätt andma* 'to give hand for handshake', *tööd tegema* 'to work', *välku lööma* 'to lighten', and so on. There are no PVs (according to the definition followed in the present thesis, see Section 2.4) among his examples.

Rätsep (1978) differentiated between verb-particle and verb-noun compounds, and only considered only verb-particle compounds to be PVs. Hence, his distinction is based on the POS of the complement. Before his approach, the specifications of PVs were not clear. He suggested four – orthographical, morphological, syntactical and semantic – criteria that have been the basis of the differentiation of PVs in other verb compounds. According to the orthographic criterion, PVs are sometimes written as one unit and sometimes separately. The morphological principle indicates that a PV consists of a verb and an adverb. The syntactical criterion states that a particle-verb compound occurs as one POS, for example, as a predicate. The semantic criterion proposes that adverbs appear as components of PVs only if they complement the verb meaning, or add new dimensions to the PV meaning. Rätsep (1978: 28) presented the division of PVs as compositional (*korrapärased*) and non-compositional (*ainukordsed*) based on their productivity, as introduced earlier (see Section 2.4.1). He offered also a brief description of how the word compounds formed a range – free compounds of words, compositional PVs, non-compositional PVs and verb-noun compounds (which are defined as idiomatic). Rätsep's approach to PVs has been followed in different descriptions of Estonian syntax ever since (e.g. Erelt et al. 1993, 2017).

Muischnek (2006) illustrated the division of verb+noun constructions based on the transparency of their meanings. She differentiated between idioms (with opaque meanings and with transparent meanings) and collocations (half idioms and support verb constructions). Although she did not comment on the semantic compositionality of PVs in her dissertation, her approach followed the idea that MWEs form a continuum from fully non-compositional to fully compositional expressions. According to Kaalep and Muischnek (2009), who described the automatic processing of Estonian MWEs, idiomatic expressions can be opaque or transparent. Veismann and Sahkai (2016), who explored verb-particle combinations and their status as PVs as opposed to syntactic phrases, discussed this claim. They emphasised that, as semantic non-compositionality is a scalar feature, expressions can have different degrees of compositionality, then a binary division is insufficient because there are expressions that do not have clearly compositional or non-compositional meanings (Veismann and Sahkai 2016: 272).

Veismann and Sahkai (2016: 272) restated the special status of PVs that contain an adverb expressing perfectivity. While Rätsep (1978) did not analyse these along with non-compositional PVs, Veismann and Sahkai (2016) suggested that these kinds of PVs cannot be treated as non-compositional because they did not gain new, fully non-compositional meaning when the components occurred together. Nonetheless, there are PVs that do not have fully transparent meanings, such as *ära keelama* 'to forbid' and *üles kirjutama* 'to write up/down' (Veismann and Sahkai 2016). This discussion demonstrates clearly that the division of PVs should not be binary, nor based on the role of the adverb. Each PV should be placed on the continuum from fully compositional to non-compositional based on its degree of compositionality.

To conclude, since Rätsep (1978), Estonian PVs have been divided into two

classes – compositional and non-compositional. Although it has been discussed that not all expressions can be determined as being one or the other, exhaustive studies on the compositionality of PVs have not been conducted. One of the contributions of this thesis is to study the compositionality of PVs by running experiments following the binary classification of PVs, as well as an approach suggesting that they form a continuum.

2.4.3 Computational studies of Estonian particle verbs

This section provides an overview of previous computational studies of Estonian MWEs, with a particular focus on Estonian PVs. The main aim of the studies has been to discover the PVs in text corpora automatically using different methods. However, some resources that contain PVs are also described.

Kaalep and Muischnek (2002) used the language- and task-specific software tool SENVA to discover Estonian multiword verbs (MWVs) from text corpora. The outcome of this work was a comprehensive list of 16,000 MWVs. An overview of the tool, the manual post-editing principles, and an evaluation of the dataset was provided in the article. The tool itself uses a mutual expectation measure to calculate the degree of cohesiveness between n-grams and the GenLocalMass algorithm to filter out the candidate MWVs. Kaalep and Muischnek found that approximately 15% of the frequently occurring MWVs remained undiscovered by SENVA.

Kaalep and Muischnek (2006) focused on verbal MWEs such as PVs and on combinations of a verb and nominal phrases; that is, idiomatic expressions, support verb constructions and collocations. The primary finding concerning Estonian PVs was that, due to the free order, the context for the automatic identification of PVs should be limited to a clause. Therefore, the detection of clause boundaries is important; otherwise, the output is inaccurate. In addition, the authors emphasised that for Estonian, which is a morphologically rich and flective language, morphological analysis and disambiguation prior to the identification of MWEs are crucial. They highlighted two main problems they encountered while performing the automatic identification of PVs – the order of the components vary, and the components of PVs are not adjacent to each other within a clause. Moreover, most particles are homonymous with adpositions, which causes problems for disambiguation.

Kaalep and Muischnek (2008) introduced a database[6] of 13,000 Estonian MWVs and a 300,000-word corpus with annotated MWVs. They described the types of Estonian MWVs, and discussed how the word order and inflection also influenced work with MWVs. Among MWVs such as noun-verb constructions (idiomatic expressions and collocations), support verb constructions and catenative verb constructions, PVs are isolated as one type of MWV in the database and in the corpus. The database of Estonian verbal MWEs contains PVs, noun(phrase)-verb expressions (idiomatic expressions and collocations), support verb and catenative verb constructions.

[6] `http://www.cl.ut.ee/ressursid/pysiyhendid` (accessed 01.10.2017).

Uiboaed (2010) applied different AMs to discover PVs from the Corpus of Estonian Dialects. The log-likelihood was the best measure for dialect data in general, but MI and χ^2 worked well for low-frequency data. The role of PVs in Estonian syntax was also analysed in within the framework of Constraint Grammar (Muischnek et al. 2013). In an applied two-fold approach – non-compositional PVs were listed in a lexicon and compositional ones were composed by the rules – the authors achieved high levels of recall and precision.

This study is a continuation and extension of the research previously conducted and published by the author earlier. Aedmaa (2014, 2015) concentrated on the comparison of lexical AMs for the automatic discovery of Estonian PVs from text corpora. AMs distinguish between random co-occurrences and true statistical associations by computing the association score for each word pair. The outcome can then be used for ranking or selection by setting a cut-off threshold (Evert 2005). The co-occurrence of the components of the MWEs introduced previously (see Section 2.1.1) has been the basis of these widely used methods. Over time, many different AMs have been suggested for the discovery of MWEs (and for VPCs, see Section 2.3.1), but the most common AMs are (P)MI, t-score, log-likelihood, dice and χ^2.

For the study of discovering Estonian PVs in text corpora, seven AMs (t-test, log-likelihood, χ^2, MI, minimum sensitivity, ΔP and conditional probability) were compared to the co-occurrence frequency of a verb and a particle. The primary conclusions of these studies showed that the t-test performed better than did the other AMs, and that the corpus size had an impact on the results. (Aedmaa 2014, 2015)

After determining that AMs could be used for the automatic discovery of Estonian PVs, based on the results of previous research (particularly that by Evert and Krenn 2001; Evert and Kermes 2003), it was hypothesised that fully compositional and fully non-compositional PVs could be differentiated via the use of AMs (Aedmaa 2016). Tests were conducted to determine whether the AMs (t-score, MI, χ^2, log-likelihood function, minimum sensitivity and co-occurrence frequency) discovered fully compositional and/or fully non-compositional PVs successfully (with high recall). However, it was not assumed that AMs divided all PVs into two fixed classes. As a first step, the association scores for all PVs were computed. A higher score indicated a stronger association between the adverb and the verb. As a second stage of analysis, the PVs were ranked according to their association score values. As five different AMs were compared, a distinct ranking for each AM was created. These rankings were compared to the binary division according to human judgements (see Section 4.2), and the percentages expressing the amount of compositional and non-compositional PVs that each AM discovered were calculated. It was ascertained that none of the AMs discovered fully compositional or fully non-compositional PVs particularly well. Compositional or non-compositional Estonian PVs are thus not automatically identifiable using AMs. In addition, it was demonstrated that there were no linear relationships among the rankings. (Aedmaa 2016)

Aedmaa (2016, 2017); Aedmaa et al. (2018) explored the automatic discovery and identification of PVs based on their compositionality. Aedmaa (2016) showed that the degree of compositionality of PVs could be predicted with the help

of cosine similarity (CS). The results were not evaluated because appropriate resources were not available. Aedmaa (2017) presented a novel dataset and demonstrated that two rankings – one based on human judgement and another on CS – correlated poorly, and that a more sophisticated model needed to be developed for the task. The follow-up to these studies is proposed as part of the current thesis (see Section 5). Aedmaa et al. (2018) studied the compositionality of PVs by classifying sentences based on the usage of the PVs. While focusing on the language-specific features, a random forest classifier, which predicted the literal versus the non-literal usage of the PVs, was introduced. The work on this topic is continued and complemented in Section 6.

In summary, the automatic processing of MWEs, particularly MWVs, has been on-going work in Estonian for more than 15 years. While different, more or less successful methods have been proposed for the discovery and identification of PVs, there is little existing work in the direction of detecting the compositionality as one of the main characteristics of MWEs, including PVs. However, several studies of MWE compositionality have been published recently, and the current thesis continues this trend.

3 METHODS

This chapter provides a theoretical survey of the methods employed in this thesis. DSMs are implemented to predict the compositionality of PVs, and the random forest classifier to distinguish between the literal and the non-literal usages of PVs was developed. The methods were chosen based on their previous success in solving similar problems in other languages, as well as for their applicability to Estonian data.

Recent trends in computational MWE research show that the focus has changed from using the co-occurrence of MWEs to investigating the compositionality of MWEs (see Section 2.3). Widely used AMs that are based on the co-occurrence frequencies have been used several times to discover Estonian MWEs (see Section 2.4.3). Therefore, this study focuses on exploring the compositionality of PVs using methods that have been applied successfully to many NLP tasks, including MWE processing, but which have not been investigated thoroughly with regard to Estonian data.

While many of the proposed methods for compositionality research have used previously created resources, the selection of methods for this study had to also take the availability of necessary resources in Estonian into account. Accordingly, DSMs were selected, as the training of these models only requires large amounts of text and lemmatisation, both of which are available for research in Estonian. For the second method, we chose supervised learning that requires labelled data, but is presumably more accurate approach. Although it demands longer and more expensive preparation period, it was thus possible to provide an analysis of the collection of (automatically) annotated data using the available resources and tools for Estonian. Furthermore, two methods used for investigating the compositionality of Estonian PVs could be compared.

This chapter is organised as follows: In Section 3.1, the general introduction to machine learning is presented in order to provide a general background to the applied models. Section 3.2 provides an overview of distributional semantics, as well as the models and toolkits developed to predict the degree of compositionality of PVs. The third part of the chapter, Section 3.3, explains the means of automatic identification of the literalness of PVs, and introduces the random forest classifier, feature selection and cross-validation. Section 3.4 concerns evaluation measures that are used for evaluating the agreement of the annotators of datasets used to evaluate models and assess the work of the suggested models in both studies.

3.1 Machine learning

Machine learning is generally defined as a set of computational methods using past information available to the learner to improve the performance of predictions (Mohri et al. 2012). Machine learning is an intersection of statistics, artificial intelligence and computer science (Müller and Guido 2016). Electronic data are indispensable for machine learning, as efficient and accurate prediction algorithms are designed based on the data (Mohri et al. 2012).

Learning algorithms can be applied to different tasks; examples include text and document classification (spam detection), NLP (morphological analysis, POS

tagging, statistical parsing and named-entity recognition), speech recognition, speaker verification and so forth. These applications correspond to a variety of learning problems, such as classification, regression, ranking, clustering, dimensionality reduction and so on. (Mohri et al. 2012)

Machine learning techniques are typically classified according to two broad categories, namely supervised and unsupervised learning. Supervised learning is used most commonly for classification, regression and ranking problems. In supervised learning, the user provides a set of labelled examples as training data, and the learner makes predictions for all unseen points. (Mohri et al. 2012) Hence, an algorithm can create an output for an input it has never seen without any human help. Unsupervised learning is a machine learning technique in which there is no known output and no teacher to instruct the learning algorithm. (Müller and Guido 2016) The training data are unlabelled, and the learner makes predictions for all unseen points. The quantitative evaluation of unsupervised models can be challenging because labelled datasets are often not available. Examples of unsupervised learning are clustering and dimensionality reduction. (Mohri et al. 2012)

Aside from these broad categories, there are more complex and intermediate learning scenarios. For example, among other techniques, Mohri et al. (2012: 7–8) presented among other techniques semi-supervised learning. The learner in the semi-supervised approach receives labelled and unlabelled data, and makes predictions for all unseen points. This technique is often used when unlabelled data are easily accessible, but labels are difficult to obtain. For example, one of the semi-supervised learning techniques is pre-training, in which the unsupervised model is trained using unlabelled data and the weights that the model has learnt are then applied to supervised models and trained on the labelled data.

Following the broad classification of machine learning methods, both approaches are employed in this thesis. The word and multi-sense embedding models of compositionality predictions were trained in an unsupervised manner (as described in Section 3.2). The supervised random forest classifier was developed to detect the literal versus the non-literal usage of PVs (as described in Section 3.3).

3.2 Distributional semantic modelling

The statistical distribution of words in context is key for characterising their semantic behaviour (Lenci 2008). The methods for the computational analysis of word distributional properties have been developed over many years, but they all rely on a distributional hypothesis. In this section, the theoretical background to modelling distributional semantics is provided – overviews of distributional semantics and DSMs are given first, followed by a description of the specific tools used for learning word and multi-sense representations in this study.

3.2.1 Distributional semantics

Every study of meaning involves semantic similarity – a case in which two words have exactly the same meaning and are thus semantically similar. Semantic sim-

ilarity can be described through linguistic distributions. Distributional semantics, which has recently attracted much attention in computational linguistics, has its roots in work within the field of linguistics undertaken by structuralists such as Charles Hockett, Martin Joos, George Trager and Zellig Harris (Lenci 2008). It has been highlighted (e.g. Lenci 2008; Sahlgren 2008) that Harris' distributional procedure (Harris 1951) is a starting point for distributional semantics in general, but it is important to note that distributional procedure was not proposed first in the context of semantics, but in the phonemic analysis (Goldsmith 2005).

Harris transferred the distributional methodology from mathematics to linguistics – he defined the distribution of element as "the sum of all its environments" (Harris 1951). The environment of an element A was understood as an existing array of its co-occurrents; that is, the other elements, each in a particular position, with which A occurs to yield an utterance. According to Harris' distributional methodology, the classes of the basic entities of a language behave distributionally similarly, and can be grouped according to their distributional behaviour. Harris believed that a language can be described in terms of distributional structure. According to Harris' view of meaning, differences in meaning can be expressed by differences in distribution. (Harris 1970)

To conclude, the Distributional Hypothesis states that there is a correlation between distributional similarity and meaning similarity, which allows one to utilise distributional semantics in order to estimate the meaning similarity (Sahlgren 2008). The hypothesis is the basis of models of distributional semantics, also known as vector-space methods, which are introduced in the following section.

3.2.2 Distributional semantic models

Unlike theoretical linguistics[7], corpus linguistics and lexicography have adopted the idea of studying word meaning with the help of the distributional analysis of linguistic contexts. For example, distributional methods bring corpora and statistical methods closer to the lexicographers who can examine the behaviour of words in different contexts.

Lenci (2008) mentioned some terms that are commonly used when describing semantic approaches based on the distributional hypothesis, such as distributional, corpus-based, statistical, vector semantics, and so on. Mathematical and computational techniques help to convert the informal notion of contextual representation into empirically testable semantic models. Corpora are connected to distributional semantics because they provide thousands of examples of language usage; thus, they are a primary source of information to identify the distributional properties of a word. Many tools and techniques for building distributional semantic representations from huge text corpora have been developed over the years. (Lenci 2008)

"You shall know a word by the company it keeps" (Firth 1957: 11) is the motto behind vector-based models. These models have to take the core features of semantic representations into account – namely, that they are context-based (the

[7]See an overview of the criticism of the structuralist distributional methodology by Lenci (2008: 5–6).

composition of the word is derived from the context), distributed (the semantic content of a word is the result of its global distributional history), quantitative and gradual (the meaning is represented in terms of its statistical distribution in various linguistic contexts). (Lenci 2008)

The popularity of DSMs has been increased by their ability to represent word meanings using only distributional statistics. They operate as follows: The semantic properties of words are captured in a multi-dimensional space by vectors. Vectors are constructed from a large amount of text by adopting the distributional patterns of co-occurrences with their neighbouring words. Information about co-occurrences is collected in a frequency matrix in which rows correspond to target words and columns show the given linguistic context. The DSMs of word co-occurrence (also known as vector-space or word-space models) have been applied to various NLP tasks, such as word sense discrimination (e.g. Schütze 1998) and ranking (e.g. McCarthy et al. 2004), text segmentation (e.g. Choi et al. 2001), and automatic thesaurus extraction (e.g. Grefenstette 1994). These models have also been used in cognitive science; for example, in studies of simulated human behaviour (e.g. Landauer and Dumais 1997; McDonald 2000; McDonald and Brew 2004).

It has been suggested repeatedly that using just raw co-occurrence is often insufficient. In addition to the count-based models (in which the semantic similarity between words is learned by counting co-occurrence frequency), another type of DSM, prediction-based representations (often called 'embeddings'), has been proposed. These models produce word representations that are using neural network language models (Bengio et al. 2003), and they learn a function between words and their contexts using co-occurrence information (Mikolov et al. 2013b). Numerous studies have described and compared these two different types of DSMs (e.g. Baroni et al. 2014; Levy et al. 2015); therefore, the differences will not be discussed further at this point. However, in this study, embeddings are employed to predict the compositionality of PVs.

The semantic similarities of the vector representations of words are quantified using some type of distance measure (Padó and Lapata 2007). For example, cosine similarity (CS) is one of the similarity measures used to determine the degree of similarity between two objects. It measures the angle between two vectors and normalises the vector length. A value of CS close to 1 indicates high similarity, while a score close to 0 shows that the two objects are not related. In addition to CS, other similarity functions are used to measure the distance between the vectors. For example, Chakraborty et al. (2011) used the value of Euclidean distance to measure the semantic similarity of the components of noun-noun bigrams to identify MWEs in Bengali.

The creation of embeddings in distributional models has several parameters that affect the performance of these models when solving different NLP tasks. For example, the level of corpus processing, the number of vector dimensions and the type of context window can affect the results of the DSM in MWE discovery (Cordeiro et al. 2016a). The description of parameter settings within the models used is provided in Section 3.2.2.1, and the impact of the parameters on the compositionality predictions is analysed in Chapter 5. The evaluation of methods based on distributional similarity can employ dedicated test sets or can

use hand-built resources such as WordNet (Constant et al. 2017).

From an MWE discovery point of view, semantic similarity methods have been evaluated successfully evaluated on small samples of VPCs (e.g. Baldwin et al. 2003; Bannard 2005) and noun compounds (e.g. Reddy et al. 2011b; Yazdani et al. 2015; Cordeiro et al. 2016a), for example. For MWE discovery, the distributional vectors have been combined and compared in several ways. The most common approach is to measure the CS between the vector of an MWE and the member word vectors; for example, Baldwin et al. (2003) and McCarthy et al. (2003) used this approach.

Most DSMs represent each word via a single vector; thus, the different meanings of one word are combined. Nevertheless, several approaches that produce multiple representations of the same words and distinguish between different senses have been proposed in the literature (e.g. Huang et al. 2012; Tian et al. 2014; Li and Jurafsky 2015; Neelakantan et al. 2015). These kinds of models are mainly applied to WSD (e.g. Iacobacci et al. 2015; Pelevina et al. 2016), but some studies have utilised multi-sense embeddings for compositionality studies (e.g. Reddy et al. 2011a; Cheng and Kartsaklis 2015; Kober et al. 2017). Köper and Schulte im Walde (2017a) used multi-sense embeddings to detect the compositionality of German PVs. The compositionality of Estonian PVs was predicted using word and multi-sense embedding models. The tools used for computing the representations are introduced in the following sections.

3.2.2.1 word2vec

Mikolov et al. (2013a) proposed word2vec for learning the vector representations of words. It has consequently become a widely applied toolkit in NLP. The reason for its popularity might be derived from the fact that it has a simple and available implementation.

Word2vec has two model architectures for learning the distributed representations of words. These models are based on the Feedforward Neural Net Language Model (NNLM) introduced by Bengio et al. (2003). The NNLM consists of input, projection, hidden and output layers. At the input layer, previous words are encoded using 1-of-V coding, where V is the size of the vocabulary. The input layer is then projected onto a projection layer using a shared projection matrix. As only a limited number of previous words (input) are active at any given time, the composition of the projection layer is a relatively inexpensive operation. The NNLM architecture becomes complex for computation between the projection and the hidden layer, as values in the projection layer are dense. The hidden layer is used to compute the probability distribution over all the words in the vocabulary, resulting in an output layer.

The Continuous Bag-of-Words Model (CBOW) and the Continuous Skip-gram Model (Skip-gram) attempt to minimise computational complexity. These models do not have the non-linear hidden layer that NNLMs have – therefore, the data are not represented as precisely, but more data are trained efficiently as opposed to within recurrent neural networks. Figure 1 illustrates the architectures of the CBOW and Skip-gram models. The CBOW is similar to the NNLM, but without the non-linear hidden layer. The projection layer is shared for all words – thus,

all words are projected onto the same position. The order of words in the history does not influence the projection. Moreover, this model uses words from the future. Unlike the standard bag-of-words model, it uses a continuous, distributed representation of context. In other words, the current word is predicted based on the context. The Skip-gram is similar to the CBOW, but it predicts words within a certain range before and after the current word. Thus, each current word is used as input for a log-linear classifier, with a continuous projection layer. Increasing the range improves the quality of the resulting word vectors, but the computational complexity also increases. As the more distant words are usually less related to the current word than are closer words, the distant words are less weighted by taking fewer samples from words in the training examples. (Mikolov et al. 2013a)

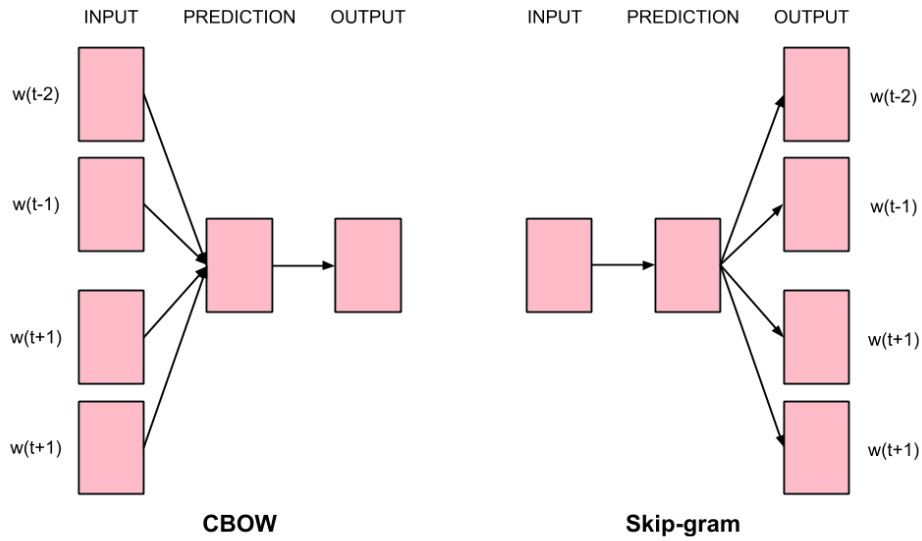

Figure 1: Architectures of the CBOW and Skip-gram models (adapted from Mikolov et al. (2013a)).

Regardless of the model type, there are parameters that could be modified in order to improve the quality of the produced embeddings. The parameters to be studied later are as follows:

- The number of dimensions – the size of the vector; this indicates how many dimensions the vector space contains. It has been suggested that more is better, but this is not always the case. Increased dimensionality should provide more fine-grained patterns of co-occurrence. (Google 2013)
- The size of the context window – this determines how many words before and after a given word are included as the context words of a given word. The authors of word2vec suggested around 10 for the Skip-gram model and around five for the CBOW model (Google 2013).

- A minimum-count threshold – this compels the model to exclude words that occur less often than the number dictates; the threshold is usually set to a value between 1 and 5.

- Training iterations – this is a count of training iterations, indicating the number of times the algorithm's parameters are updated. When the iteration number is small, the word embeddings exhibit a lack of learning but, when it is large, the model is prone to overfitting (Lai et al. 2016).

Although there are some general recommendations regarding how parameters should be set, they depend largely on both the data and the task. Therefore, in order to determine the best parameter settings for predicting the compositionality of Estonian PVs, several models were trained and analysed, as described in Chapter 5.

In addition to these parameters, there are others, such as sampling methods and sub-sampling, but the impact of these parameters on the compositionality predictions is not explored in the current study. However, some research on the influence of the parameters for the compositionality predictions of MWEs in other languages has been conducted previously (see, for example, Lai et al. 2016; Cordeiro 2017; Caselles-Dupré et al. 2018).

The output of word2vec is a file containing word vectors. In order to determine the similarities among these vectors, the Gensim tool (Řehůřek and Sojka 2010) is used. Gensim[8] is a Python library for topic modelling, document indexing and similarity retrieval with large corpora. Gensim provides the CS value among requested word vectors. For example, after training on Estonian word embeddings using word2vec, it is possible to ascertain that the CS value between the Estonian synonyms *telekas* and *televiisor* 'television' is 0.87 (indicates relatively high similarity), but is 0.13 (low similarity) for unrelated words such as *sülearvuti* 'laptop' and *kaalikas* 'swede'. In the same way, it is possible to request the CS value between verbs such as *jooksma* 'to run' and *sörkima* 'to jog' (CS = 0.69) or between *jooksma* and *magama* 'to sleep' (CS = 0.34).

3.2.2.2 SenseGram

SenseGram is an approach to learn word-sense (multi-sense) embeddings that was proposed by Pelevina et al. (2016). Figure 2 illustrates the four main stages of their method – a) learning word embeddings, b) building a graph of nearest neighbours based on vector similarities, c) inducting word senses using ego-network clustering, and d) aggregating word vectors with regard to the induced senses. Hence, the model uses existing word embeddings and word similarity graphs.

SenseGram learns word vectors using the word2vec toolkit (see Section 3.2.2.1); one can choose between CBOW and Skip-gram, and modify other parameters (see the previous section) as well. Word vectors are saved separately and are given in order to calculate word similarity graphs. The 200 nearest neighbours

[8]Gensim and its documentation can be accessed at `https://radimrehurek.com/gensim/` (accessed 12.05.2017).

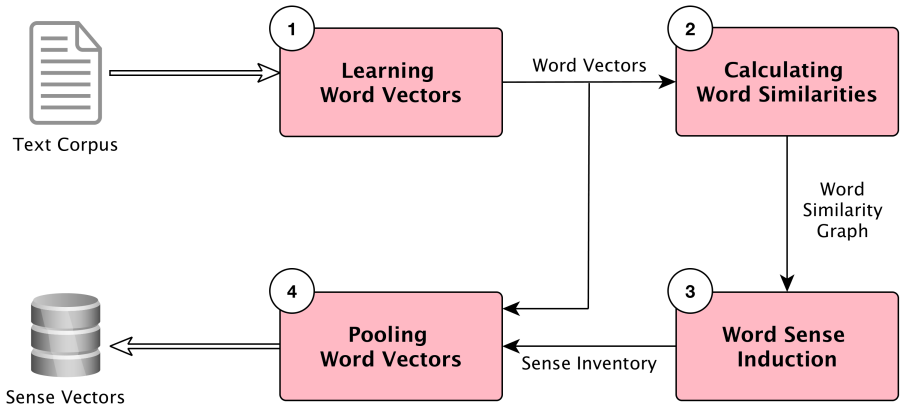

Figure 2: Schema of the multi-sense embeddings learning method (adapted from Pelevina et al. (2016)).

are retrieved for each word. The nearest neighbours are ascertained by calculating the CS among word vectors – words with higher scores are their nearest neighbours. The similarity graphs are the input for the word-sense induction. The sense of the word is represented by a word cluster. For the derivation of senses, an ego-network of a word is constructed, and the graph clustering of this network is performed. The graph clustering uses the Chinese Whispers algorithm suggested by Biemann (2006). The multi-sense embeddings are calculated for each sense induced previously. A more detailed explanation of the stages was described by Pelevina et al. (2016).

The difference between this approach and other methods that learn multi-sense embeddings is that, instead of learning multi-sense embeddings directly from the corpus, it learns the embeddings from the existing word embeddings. Hence, it encourages the reuse of resources. In addition, SenseGram does not rely on any knowledge base (such as WordNet), which means it does not require the existence of other resources. Moreover, Köper and Schulte im Walde (2017a: 537–538), who tested multiple models for the prediction of German PV compositionality, showed that, using this approach, (ChinWhisp) performed significantly better than the baseline and achieved the best result.

There are two other reasons that SenseGram was used for the learning senses for the automatic discovery of Estonian PVs. Firstly, the pilot studies (Aedmaa 2016, 2017) demonstrated that word2vec could be used for the detection of the degree of compositionality of Estonian PVs. Thus, besides expanding the previous studies by running more experiments with word2vec, the use of SenseGram guarantees a valid comparison of how the results of word embeddings impact on the work of multi-sense embeddings. Secondly, SenseGram is a freely available tool[9] with relatively easy implementation, and has obtained results that are comparable

[9]The implementation of the method with several pre-trained models is available at `https://github.com/tudarmstadt-lt/sensegram` (accessed 03.04.2018).

to state-of-the-art instruments.

Gensim is also used for retrieving the CS scores of sense vectors, inquiring into the number of senses each word obtained and investigating the meaning of each sense by requesting the nearest neighbours (the most similar words) for each sense. For example, the polysemous word *klaas* 'glass' has two meanings. Since SenseGram also calculates the probabilities for each sense, this information is included when requesting the number of senses. The probabilities of the senses of the word *klaas* are 0.64 and 0.36, respectively. Hence, the first meaning of the word *klaas* is more probable in the text than is the second meaning. It is possible to study the meanings of the different senses by examining their nearest neighbours. The nearest neighbours for the first meaning of *klaas* are *mattpind* 'matte surface', *klaasvahesein* 'glass partition wall', *plastliist* 'plastic slat', *metallkest* 'metal sheath', and *pimendav* 'darkening'. The second meaning of *klaas* is similar to words such as *topka* 'shot glass', *napsiklaas* 'shot glass', *pits* 'shot glass', *tass* 'cup', and *vahumüts* 'cap of foam'. Therefore, the first meaning represents 'a transparent solid material' and the second represents 'a drinking glass'. The ability of SenseGram to predict the compositionality of Estonian PVs is discussed in detail in Section 5.5.

3.3 Supervised learning

Supervised learning, whereby decisions are made by generalising from known examples, is considered to be the most successful of the machine learning algorithms. In supervised learning, machine learns a mapping from the input to an output whose correct values are provided by a supervisor (Alpaydin 2009). Thus, the user (a supervisor) gives the machine (a learner) an input and a desired output, and the machine finds a way to produce this desired output from the input provided to it. Compared to unsupervised learning algorithms, the supervised ones are understood well, and their performance is easy to measure. Therefore, when one is able to create a dataset containing the desired output, a supervised learning algorithm could solve the problem. (Müller and Guido 2016: 2–3)

In this section, an introduction to the method employed to detect the literal versus the non-literal usages of Estonian PVs – a random forest classifier – is provided. An explanation of the importance of the feature selection process for the creation of such a model and the evaluation technique are also presented.

3.3.1 Random forests

The random forest algorithm is one of many supervised learning methods that are applied to solve classification and regression problems. The main idea behind the algorithm is that it constructs a number of randomised decision trees during the training phase and makes predictions by averaging the results. The random forest algorithm was proposed by Breiman (2001), and has become widely used as a data analysis tool (Scornet et al. 2015).

The random forest algorithm combines a set of decision trees that differ slightly from each other (Müller and Guido 2016: 83). A decision tree is a decision-making device whereby the probability of each of the possible choices is

calculated based on the context of a given decision (Magerman 1995). A decision tree is a data structure in which: a) each leaf is labelled with the name of a class, b) the root and each internal node (called a decision node) are labelled with the name of an attribute, and c) every internal node has a set of at least two children, the branches of which are labelled with disjointed values or sets of values of that node's attribute. (Mahmood et al. 2011)

A decision tree is applied by starting at the top node, and finding the answer to the test question for this node. Depending on the test result, it branches to the subnode, and repeats this process until a terminal node is reached, the label of which is returned as the result class. (Clark et al. 2013) In other words, the idea of the decision tree algorithm is to select the attribute with the value that best separates the training set into subsets as the root of the tree and to repeat it recursively for each child node until a stopping criterion is met (Mahmood et al. 2011).

While each tree might work well for prediction, combining them reduces overfitting and leads to a more reliable classifier. Building a random forest model begins with a decision concerning how many trees the models contains. Each tree is built independently from the other and is derived from bootstrapped samples of the original training set. A bootstrapped sample has the same size as the original dataset, and is created by drawing a random sample with a replacement from the original set. Hence, some data points are missing from the new dataset, and some of them are repeated. A decision tree is built based on the new dataset, while the algorithm randomly selects a subset of features and looks for the best test in each node. The maximum number of features is controlled by the user. The selection of subset features is repeated in each node. Hence, the decision in each node could be taken based on a different subsets of features. (Müller and Guido 2016)

The classifications of the different trees are combined by choosing the most frequently predicted class (Clark et al. 2013). A so-called soft voting strategy is used – each algorithm provides a probability for each possible class, and the probabilities predicted by all the trees are averaged. Therefore, the class with the highest probability is predicted. (Müller and Guido 2016)

A description of developing a random forest classifier for predicting the literal versus the non-literal usage of Estonian PVs is provided in Chapter 6.

3.3.2 Feature selection

In machine learning, particularly when dealing with multivariate data, dimensionality reduction is a prerequisite to decrease the number of random variables under consideration (Roweis and Saul 2000). One way to decrease the number of variables is to conduct a 'feature selection' (Pudil and Novovičová 1998), which is a process that selects a subset of features occurring in the training set and uses these features to solve a classification task (Manning et al. 2008). The purposes of the feature selection are to increase the classifier accuracy, provide faster and more cost-effective classifiers, and to provide a better understanding of the data generation process (Guyon and Elisseeff 2003).

For better accuracy, the elimination of noise features, which are features that increase the classification errors in new data, is conducted. Noise features may

cause overfitting, which is an incorrect generalisation from an accidental property of the training set (Manning et al. 2008). For example, a feature, such as the infrequent word *animatsioon* 'animation', does not convey any information about a class (for example, the compositionality of Estonian PV), but all instances of *animatsioon* happen to occur in sentences in which one particular PV is used with a non-compositional meaning in the training set. In this case, the learning method might produce a classifier that incorrectly predicts the class of the sentences that contain the word *animatsioon* as having a non-compositional meaning. In statistics, 'overfitting' is defined as 'the production of an analysis which corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or reliably predict future observations'[10].

Overall, the reduction of feature dimensionality is important to decrease the computational complexity and to improve the generalisation ability of the classifier – fewer features require less run time, and the low-dimensional representation reduces the risk of overfitting (Liu and Zheng 2006). To conclude, feature selection is a method to replace a complex classifier that uses all the features with a simpler one that uses a subset of features (Manning et al. 2008).

There are different methods for feature selection. Some methods focus on the construction and selection of subsets of features that are useful in order to build a good predictor, while others aim to find or rank all potentially relevant features. Useful feature is not necessarily relevant (that is a feature that provides some information about the target label), and vice a versa, meaning that a subset of useful features may exclude redundant yet relevant features. (Guyon and Elisseeff 2003) Feature selection techniques have been applied in many different applications, such as text categorisation and data visualisation (Liu and Zheng 2006).

In general, feature selection methods are grouped according to three classes – filter, wrapper and embedded methods (Guyon and Elisseeff 2003). The filter methods are independent of the classifier and evaluate the performance via some indirect assessments. The principal criteria for the feature selection of filter methods are simple and successful for practical applications – ranking (Chandrashekar and Sahin 2014). The features are ranked according to the score and to a threshold that is either selected or removed from the dataset. The features are usually considered independently or with regard to the dependent variable. These kinds of methods include information gain, correlation coefficients, and chi-squared test and so on. (Guyon and Elisseeff 2003)

Wrapper methods assess all possible subsets of features according to their relative usefulness, they are classifier-dependent and are based on classification accuracy. Therefore, it is necessary to define how the space of all possible subsets is searched (such as best-first, branch-and-bound, genetic algorithms), how to assess the predictions performance (usually done using a validation set or by cross-validation), and which predictor to use (popular ones include decision trees, Naïve Bayes and support vector machines). A predictive model evaluates a subset of features and assigns a score based on the model's accuracy. (Guyon and Elisseeff 2003)

[10] `https://en.oxforddictionaries.com/definition/overfitting` (accessed 10.05.2018).

Embedded methods select features in the process of training and are usually specific to particular learning machines (Guyon and Elisseeff 2003). These methods aim to decrease the computation time of reclassifying different subsets; thus, feature selection is performed as part of the training process (Chandrashekar and Sahin 2014). One of the embedded feature selection methods is regularisation (or penalisation), a method that introduces additional constraints into the optimisation of a predictive algorithm that biases the model towards lower complexity.

For feature selection, two different filter methods (correlation coefficient and information gain), and a wrapper method described by Kohavi and John (1997) are implemented. Detailed descriptions of the methods and their results for the dataset are presented in Section 6.3.1.

3.3.3 Cross-validation

The evaluation of a machine learning model shows how accurate it is for data it has never seen previously. Cross-validation is a technique that is used to evaluate models when data are limited. The main idea in this approach is that some data are removed before training; after the training is complete, the removed part of the data is used to evaluate the performance of the model. Cross-validation can be carried out by applying different methods (such as the holdout method, k-fold cross-validation, leave-one-out cross-validation, and so on). The classifier presented in this thesis was evaluated using k-fold cross-validation.

K-fold cross-validation is a widely adopted method that is used when the labelled data are too small to reserve a validation sample, since that would result in an insufficient amount of training data. Therefore, the labelled data are used for both model selection and for training (Mohri et al. 2012); instead of making a single partition, the full set of available data is partitioned into k subsets, or folds (Clark et al. 2013). The approach is used widely because it is a less biased estimate of the model's skill than are other methods, such as the simple train/test split (the holdout method).

The evaluation is conducted as follows: Each k time, one of the k subsets is used as the test set and the other $k-1$ subsets are combined to form a training set. The error estimation is averaged over all k trials to ascertain the total effectiveness of the model. Every item (data point) is used for training as well as for testing – once in the test set and $k-1$ times in the training set. Thus, the results do not reflect a particularly good or particularly bad choice of test set. It also reduces the bias because most of the data are used for fitting. (Clark et al. 2013) In machine learning applications, k is typically chosen to be 5 or 10 (Mohri et al. 2012).

The output of the evaluation model usually contains values of several evaluation metrics. The classification metrics used in this thesis are described in Section 3.4.3.

3.4 Evaluation measures

In order to determine how well the suggested models work, evaluation measures needed to be applied. In addition, the quality of the datasets containing human

judgements was assessed. This section introduces the inter-rater agreement measures that are used to evaluate the datasets, and the correlation and classification metrics used for the evaluation of the machine learning algorithms.

3.4.1 Inter-rater agreement measures

In order to evaluate human annotations, multiple inter-rater reliability measures were developed. The aim of such measures is to determine the degree of agreement among rater. The simplest way is to calculate the percentage of annotations in which all the annotators agree.

However, more statistics are considered to be better than the simple percentage agreement calculation because they take the possibility of the agreement occurring by chance into account; for example, kappa coefficients. Cohen's kappa (Cohen 1960) and Fleiss' kappa (κ) (Fleiss 1971) coefficients are the most frequently used measures to evaluate datasets created by several annotators. These measures presume that all the annotators have evaluated the same items. Cohen's kappa can only be used when measuring the agreement between two annotators, while Fleiss' kappa is suitable for evaluating an agreement among a larger number of evaluators. Fleiss's kappa score takes the following features into account – the number of evaluators, the number of items and the number of categories; it also measures pairwise agreement (Artstein and Poesio 2008). Value $\kappa \leq 0$ indicates that there is no agreement among the rates. When the annotators are in complete agreement, $\kappa = 1$.

One of the limitations of κ is that all disagreements are treated equally, but for some features (especially semantic and pragmatic ones) disagreements are not all alike. Therefore, several weighted agreement coefficients have been proposed. One of them, Krippendorff's α (Krippendorff 2004), applies to multiple coders, incomplete data, and it allows for different magnitude of disagreement. It is suitable for nominal, interval, ordinal, and ratio scales. (Artstein and Poesio 2008) As suggested by Krippendorff (2004), α = 0.8 is a threshold of good reliability, but tentative conclusions can be made also when $\alpha \geq 0.667$.

In sum, for the assessment of inter-rate agreement among annotators of (non-)literalness ratings (see Section 4.3), Fleiss's kappa and Krippendorff's α are calculated.

3.4.2 Correlation measures

The comparison of two datasets with numerical values is feasible using correlation measures. In the case of compositionality detection for Estonian PVs, the system's predictions can be compared to a human-annotated dataset to determine the quality of the performance of the model. Hence, having human judgements and system predictions for the same PVs allows one to evaluate how good the predictions of the respective systems are.

As it is suggested that MWEs form a continuum ranging from fully compositional to fully non-compositional expressions, system-based and human-annotated ratings of degrees of compositionality of PVs can be viewed as rankings. Accordingly, the datasets were compared using Spearman's rank correlation coefficient

(Spearman's ρ). Spearman's ρ is a special case of Pearson's correlation coefficient (r) that is applied to ranked variables; thus, it relies on the rank order of values. Pearson's r between two datasets (X and Y) is calculated using the expected values (E), means (μ_X, μ_Y), and the standard deviation (σ_X, σ_Y), as shown in the following formula:

$$r_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

The calculation of Spearman's ρ is similar to that of Pearson's r, but it uses ranks instead of raw scores. Therefore, it allows exploring how similarly two sets of scores rank the same items. The value of Spearman's ρ can range from -1 to +1. When one variable increases, another variable increases by a consistent amount; thus, the value of Spearman's ρ is +1. When the amount of increase is inconsistent, the value of ρ is still positive, but is less than +1. In the event of having a random or non-existent association between variables, the ρ value is close to 0. A negative value occurs in a situation in which one variable increases and another decreases. When the amount of the decrease is consistent, Spearman's ρ value is -1.

In addition to Pearson's and Spearman's correlation coefficients, Kendall rank correlation coefficient (Kendall's τ, Kendall 1938) is another widely applied measure to express the association between two measured quantities. In this study, Kendall's τ is not presented because ρ has been shown to be more appropriate for small datasets (Xu et al. 2013) and in significance testing these two coefficients usually produce similar results (Colwell and Gillett 1982).

3.4.3 Classification metrics

Classification performance can be measured using many different metrics such as accuracy, precision and recall, f-measures, a Matthews correlation coefficient (MCC), a receiver operating characteristic curve (ROC curve), a precision-recall curve (PRC curve) and so forth.

The classification of the literal versus the non-literal usage of Estonian PVs is evaluated using classification accuracy and an f-measure. The accuracy reflects the number of correct predictions among all the predictions made. As the target variable classes (literal and non-literal sentences) in the data are not balanced, the accuracy is used to show the general performance of the model. For a more precise evaluation of the performance of the models, the f-measure (F_1, f-score) is calculated and reported.

The f-measure is a harmonic mean of precision and recall. The formula for the calculation of the f-measure is the following:

$$F_1 = 2 \cdot \frac{precison \cdot recall}{precision + recall}$$

Precision itself is defined as the proportion of true positives (TPs) among the

sum of TPs and false positives (FPs). TPs are data points classified as positive by the model that are actually positive. FPs are that cases that the model labels incorrectly as positive that are actually negative. Recall is the proportion of TPs among the sum of TPs and false negatives (FNs). FNs are data points the model identifies as negative that actually are positive.

In the event of classifying non-literal versus literal sentences, the f-measure values are calculated and presented for non-literal and literal sentences. For non-literal sentences, TPs are sentences that are classified as non-literal and are non-literal, FPs are classified as non-literal but are literal, and FNs are predicted to be literal but are actually non-literal. For literal sentences, TPs are literal sentences that are predicted to be literal, FPs are literal sentences classified as non-literal and FNs are literal sentences that were predicted to be non-literal.

4 MATERIAL AND DATASETS

This chapter introduces the data used in the present study. Firstly, Section 4.1 describes two corpora that were used for the development of the datasets and for training the DSMs. The outline of the corpora also explains the selection of the corpora for this study. Detailed overviews of three datasets are presented in the following sections. The dataset of compositionality ratings is represented in Section 4.2. Among other information, crowdsourcing as a data sourcing model is discussed, the evaluation of the annotation is conducted and the content of the dataset is described. The (non-)literalness ratings for Estonian PVs are described in Section 4.3. The annotation is evaluated and a detailed overview of the content of this dataset is also presented. In Section 4.5, the automatically created dataset of abstractness/concreteness ratings for Estonian lemmas is presented.

4.1 Corpora

The datasets introduced in this chapter are based on the newspaper subcorpora of the Estonian Reference Corpus[11] (ERC), containing 170 million words. The corpus only includes full texts that represent written Estonian. The largest volume of texts is from the daily newspapers Eesti Päevaleht (88 million words), Õhtuleht (45.5 million words) and Postimees (33 million words); several magazines and local newspapers are also represented. The texts sourced were from 1995–2008. The corpus was analysed and disambiguated morphologically, and the clause boundaries were detected prior to building the datasets. As the current study is a continuation of the author's previous research on discovering Estonian PVs in different corpora (see Aedmaa (2015)), the selection of PVs relied largely on the results presented previously. Moreover, as newspaper texts are considered to reflect common language usage better than are fiction or scientific texts, continuing with newspaper texts was a legitimate choice. Furthermore, this corpus had already been processed which enabled the automatic extraction of sentences for the datasets.

Embeddings for this study were trained on the Estonian Web Corpus[12] (etTenTen) text corpus, which is a web corpus containing approximately 270 million words. The etTenTen is part of the TenTen Corpus Family[13] that contains text corpora created from the web and which now contains corpora in more than 30 languages. The etTenTen has a diverse selection of text types – approximately a quarter of the corpus consists of periodicals, 20% are forums and 10% are blogs. Informative, religious and government texts are also classified. Approximately 32% of the texts are unclassified. The corpus contains texts from 686,000 Estonian web pages, and was the largest corpus of Estonian until the Estonian National Corpus 2017[14] was created in 2017. The etTenTen corpus was chosen for this study because it was the biggest morphologically analysed and disambiguated text

[11] `http://www.cl.ut.ee/korpused/segakorpus/` (accessed 01.11.2017).

[12] `http://www.keeleveeb.ee/dict/corpus/ettenten/about.html` (accessed 14.09.2018).

[13] `https://www.sketchengine.eu/documentation/tenten-corpora/` (accessed 14.09.2018).

[14] `http://doi.org/10.15155/3-00-0000-0000-0000-071E7L` (accessed 14.09.2018).

corpus at the time that the embeddings needed to be trained[15]. Having much of the data covering different types of texts made the etTenTen suitable for the current study, but the experiments should definitely also be conducted on a larger number of texts.

4.2 Compositionality ratings for Estonian particle verbs

The dataset of compositionality ratings for Estonian PVs[16] is described in this section. The purpose of the dataset and the choice of the scale of measurement are explained first. The target PVs and example sentences are then described, and crowdsourcing as a data collection method is outlined and discussed. The annotation is evaluated and the content of the dataset is also presented. In addition, the PVs that were challenging to evaluate are analysed.

4.2.1 Purpose and the choice of scale

The development of the dataset containing compositionality ratings was required for the evaluation of the methods of distributional semantics described in Section 3.2. These methods are applied to rank Estonian PVs automatically according to their degree of compositionality (see Chapter 5). It was thus necessary to collect a set of ratings expressing the degree of compositionality of PVs from human annotators.

The output of the DSM model helps to determine the semantic similarity of words. The CS score indicates how similar the meanings of a PV and a verb are. Based on the CS scores, the PVs can be ranked according to their degree of compositionality. In order to evaluate such an output, the human-annotated dataset also had to reflect the degrees of compositionality of PVs. Therefore, the annotators were asked to evaluate the compositionality of the PVs using a scale.

The decision to use an interval scale (Stevens 1946) was based on the presumption that intervals between the points were equal. For example, the difference between 1 and 2 is the same as the difference between 4 and 5. Moreover, the numbers are comparable in terms of greater-than and lesser-than. For example, 5 is greater than 3, which is greater than 1; therefore, a PV with a rating of 5 is more compositional than is a PV with a rating of 3, which is more compositional than is a PV with a rating of 1. In fact, the value *1* was chosen arbitrarily; hence, it does not signify that the PV is lacking compositionality. Using an interval scale ensures that different statistical methods for the analysis of the data can be applied. For example, standard deviations and rank-order correlations can be calculated. (Stevens 1946)

The adapted scale had an odd number of options. This kind of Likert scale is commonly used in linguistic research because it has a precise middle point (Podesva and Sharma 2014). Furthermore, the odd-numbered scale provides the

[15]The analysed corpus is downloadable from META-SHARE repository at `http://doi.org/10.15155/1-00-0000-0000-0000-00158L` (accessed 14.09.2018).

[16]The final version of the dataset is accessible from `https://github.com/eleriaedmaa/compositionality` (accessed 10.11.2018).

opportunity to consider the PVs that are difficult to evaluate at a binary level – the middle point would be used for the PVs that the annotators found challenging to evaluate as being more compositional than non-compositional and vice versa. In addition, the aim was not to create a binary division of Estonian PVs; therefore, there was no need to use an even-numbered scale. Furthermore, the meanings of some PVs can be compositional or non-compositional depending on the context. It is therefore likely that these kinds of PVs will be given the rating of 3 most often. Moreover, the sixth option, 'I don't know', was added for very difficult cases. This option was required to be used only when an evaluation of the PV was impossible. However, the annotators were not forced to submit a judgement when to do so was challenging for them.

To conclude, the annotators were asked to rate the compositionality of the PVs on a scale from 1 to 5; 1 meant that the meaning of the PV agreed fully with the meanings of the verb and particle – thus, the PV was compositional. However, 5 meant that the meaning of the PV did not agree at all with the meanings of the verb and particle; thus, the PV was non-compositional.

4.2.2 Target particle verbs and example sentences

The selection of the target PVs relied on the list of 1,676[17] PVs that were previously extracted automatically from the newspaper subcorpora of the ERC (Aedmaa 2015). The decision to select the automatically extracted PVs instead of the 1,737 PVs presented in the Explanatory Dictionary of Estonian[18] (EED henceforth) was motivated by the fact that the automatically detected PVs really occurred in the corpora.

The remainder of the PVs were ranked according to their frequency, and infrequent ones (that is, PVs that occurred less than nine times in the corpus) were removed from the selection. Note that the frequency of the PV expresses the co-occurrence frequency of a verb and an adverb in the same clause, not the real frequency of a PV. This is due to by the fact that PVs are not annotated in the corpus. In order to discover automatically PVs, all of the co-occurrences of a verb and an adverb were counted. As an adverb and a verb can occur in the same sentence as independent units, the frequencies of the PVs are approximate.

The number of target PVs was chosen based on a rough estimation of how many people could possibly participate in the data collection. The aim was to collect at least 10 assessments for each PV and to keep the number of the PVs each annotator evaluated to a relatively reasonable level. It was thus decided to collect ratings for approximately 200 PVs. Moreover, 200 PVs formed a representative sample of all the PVs. After establishing the number of target PVs, a random sample of 193 PVs with different frequencies was generated. In addition, in order to study the effect of the frequency of the PV on compositionality, all 20 most

[17]Note that most PVs formed with the particle *ära* 'away/out/off' were excluded from the dataset because the meaning of *ära* is not transparent (Veismann and Sahkai 2016) and needs to be studied as a special case. However, two PVs containing *ära – ära tegema* 'to finish doing' and *ära võtma* 'to take away' – were included in the dataset in order to analyse whether these PVs posed a challenge for annotators. The annotation of these PVs is studied in Section 4.2.6.

[18]`http://www.eki.ee/dict/ekss` (accessed 14.01.2015).

frequent PVs were added to the selection. Therefore, 18 more PVs were included in the dataset, and 211 PVs were selected for the evaluation.

The third stage of dataset creation was to select sentences that would help the annotators to evaluate the degree of compositionality of the PVs. To achieve this, nine sentences for each PV were extracted automatically from the corpus. Three example sentences were then selected manually as example sentences to be displayed to the annotators when they rated the degree of compositionality of the PVs. This manual work was inevitable because, although the corpus was analysed and disambiguated morphologically, it contained some noise. For example, adverbs and verbs can often appear in the same clause as independent units without forming a PV. In order to remove sentences without a PV, the selection was reviewed manually.

Example sentences were selected in such a way that their different meanings as represented in the EED would be reflected. However, when automatic extraction did not provide the meaning of the PV, it was included in the selection of example sentences. Therefore, the most frequent (and prototypical) meanings were presented in the example sentences. For example, the PV *välja paiskama*[19] has two meanings in the EED – 'to throw something out from somewhere' and 'to blurt out something'. Thus, both meanings were reflected in the example sentences. However, the PV *peale tungima*[20] has four meanings, and not all of them were reflected in the example sentences. For example, the meaning 'to be insistent' was not presented as an example because the meaning did not appear in the automatically extracted sentences. Thus, it might not be as frequent as are the other meanings. Some PVs have only one meaning – in these cases, all three sentences expressed the same meaning. It was thus anticipated that these PVs would be easier to evaluate as being fully compositional or fully non-compositional. For example, all the example sentences containing the PV *alla tingima* expressed one meaning, 'to bargain/beat down', because only one meaning is represented in the EED.

In summary, the selection of the target PVs was random, but the aim was to cover different frequency ranges equally. In addition, all 20 of the most frequent PVs were included. The example sentences provided for the annotators were selected randomly from the corpora. Therefore, the more frequent meanings of the PVs were represented.

4.2.3 Crowdsourcing annotations

The compositionality ratings for Estonian PVs were crowdsourced. Human annotators were asked 'to what extent the meaning of the PV agreed with the meanings of its components'. The number 1 on one side of the scale represented the answer 'not at all', while the number 5 on the other side of the scale reflected the answer 'fully'. Hence, the PVs with a higher score are more compositional than are PVs with a lower score.

The annotations were collected via the Qualtrics platform[21]. The decision to

[19] `http://www.eki.ee/dict/ekss/index.cgi?Q=välja+paiskama` (accessed 02.08.2018).

[20] `http://www.eki.ee/dict/ekss/index.cgi?Q=peale+tungima` (accessed 02.08.2018).

[21] `https://www.qualtrics.com` (accessed 13.03.2016).

crowdsource the ratings was motivated by the possibility of collecting as many annotations as possible. For the same reason, the backgrounds of the annotators were not surveyed. Since it is not possible to have many people answer these kinds of questionnaires, it was decided not to ask annotators for any personal details. It was assumed that, by not asking about their backgrounds, people without any background in linguistics would be more likely to participate. Furthermore, as the annotators were not paid for their contributions, it was reasonable not to ask them to spend more of their time answering secondary questions, as the aim was to collect compositionality ratings.

The questionnaire was distributed among the author's friends, family and colleagues via social media and mailing lists. Only native speakers of Estonian were asked to take part in the survey, and the annotators were encouraged to use their intuition, not linguistic knowledge. It was emphasised that there were no correct or incorrect answers. Each annotator was asked to annotate 21 PVs. The annotators could not return to the PVs that they had evaluated previously. Around 130 fully completed surveys were collected within a few months. The annotators who completed the questionnaire very quickly (in less than two minutes) or who evaluated all of the PVs in the same way (for example, all 21 PVs were evaluated as being fully non-compositional) were excluded from the analysis.

The most problematic aspect of collecting compositionality ratings was the polysemy of PVs and their components. For example, the verb *saama*[22] has 12, the adverb *kokku*[23] has 11 and the PV *kokku saama*[24] has four different meanings in the EED. Thus, it was clear that it would be difficult for annotators to evaluate all of the meanings of the components and PVs using the same rating. Nevertheless, despite considering other options, this was the chosen procedure.

More precisely, it was not reasonable to ask annotators to evaluate the specific meanings of the PVs and their components because the dataset was created for the evaluation of the DSMs that do not differentiate among possible senses of the same word. The collection of compositionality scores for different meanings of the PVs would have complicated the assessment process. Another aspect was the choice of meanings to submit for the assessment. As it was clear that the annotators were not able to evaluate all the meanings of all the PVs and their components, a selection had to be made. Furthermore, many meanings are fixed in the dictionary, but this does not ensure that all the possible readings of all the words and PVs are included. On the contrary, the differences in meanings that dictionaries distinguish are not always clear to everyone, and can be subjective. Therefore, the choice of meanings can be problematic.

Instead of specifying the meanings of the PVs and their components, some example sentences were provided. While there was an option not to give example sentences in order to not prevent bias among the annotators, it was instead decided to supply some information. This decision was mostly motivated as a means to make the task more engaging and clear for the annotators who were lacking a background in linguistics. While dropping all possible cognitive aspects of deciding which meanings are predominant, it was assumed that frequent meanings

[22] `http://www.eki.ee/dict/ekss/index.cgi?Q=saama` (Accessed 12.03.2016)

[23] `http://www.eki.ee/dict/ekss/index.cgi?Q=kokku` (Accessed 12.03.2016)

[24] `http://www.eki.ee/dict/ekss/index.cgi?Q=kokku+saama` (Accessed 12.03.2016)

are dominant. The most frequently occurring meanings of a random selection of sentences were thus picked as example sentences displayed to annotators. Yet, the annotators were not expected to make their decisions based on these sentences.

From the data collection point of view, the selected method – crowdsourcing – was challenging but also compelling at the same time. As described earlier, the main motivation to crowdsource the ratings was to collect as many evaluations as possible while saving time and money. The annotators could choose a time and place when to answer the questionnaire and stay anonymous. At the same time, there were circumstances that might have influenced the quality of the annotations, such as the clarity and complexity of the task, the annotators' motivation and background, the author's lack of experience, and so on. These issues are discussed further in Section 4.4, in which the compositionality ratings of two different datasets are compared, and in Chapter 5, in which the DSMs are evaluated using these compositionality ratings.

To conclude, the dataset does not reflect the compositionality of certain meanings of the PVs. Instead, the rating of the PV suggests its overall compositionality thereof. However, as the example sentences shown to annotators expressed only the most frequent meanings, it can be assumed that the ratings reflect the predominant meaning of each PV. It was also expected that the annotators would select the 'I don't know' option when rating the PV was difficult to rate the PV.

4.2.4 Evaluation of the annotations

One hundred and ten satisfactory replies were collected as a result of the annotation process, each containing annotations for 21 PVs; each PV received at least 10 ratings, up to a maximum of 20. Fifty-four PVs received at least one 'I don't know' response. These PVs are not included in the dataset, but they are discussed in Section 4.2.6 as PVs that are difficult to evaluate. The evaluation of the dataset was thus conducted based on the 157 PVs forming the final dataset of the compositionality ratings of Estonian PVs.

To measure the difficulty of annotations, we followed Cordeiro et al. (2016b) and the standard deviations among the scores assigned by the annotators were calculated. The high value of the standard deviations indicates low agreement among the annotators, while low values suggest that the human judgements were similar. The standard deviations were calculated for the 157 PVs. Figure 3 illustrates the distribution of the standard deviation values across these PVs.

The average standard deviation per rating was 1.13, while the lowest was 0.5 and the highest 1.85. The PVs *ette jõudma* 'to get ahead' or 'to outstrip' and *läbi tungima* 'to penetrate/go right through' had the lowest standard deviation (0.50). The PV *läbi tungima*[25] has one meaning in the EED, and it can therefore be assumed that the annotators assigned the same meaning, and that it is non-compositional. *Ette jõudma*[26] has two meanings in the EED; thus, as the annotators agreed, it can be assumed that both meanings are non-compositional.

[25] `http://www.eki.ee/dict/ekss/index.cgi?Q=läbi+tungima` (accessed 10.11.2018).

[26] `http://www.eki.ee/dict/ekss/index.cgi?Q=ette+jõudma` (accessed 10.11.2018).

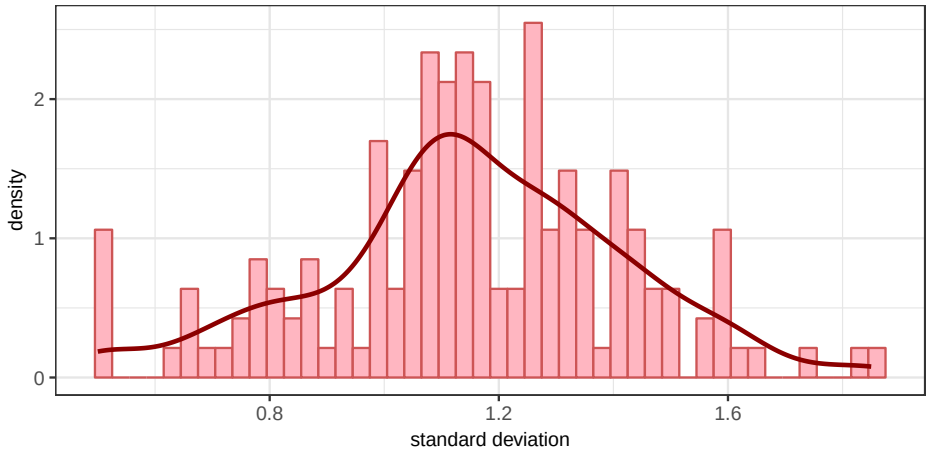

Figure 3: Distribution and density of standard deviation values across PVs.

The PVs *ette heitma* 'to reproach/blame' and *ümber riietama* 'to change somebody's clothes' had the greatest deviation (1.85 and 1.83, respectively), and they both had average compositionality ratings of around 3.0. While both PVs[27],[28] have one meaning in the EED, the annotators did not agree about the compositionality of these PVs.

It is a well-known fact that frequent words tend to be more polysemous than infrequent ones (see, for example, Zipf (1945); Hamilton et al. (2016)). It can thus be suggested that, compared to the infrequent PVs, frequent PVs with frequent components are more difficult to evaluate, and that this is reflected in the standard deviations scores. In order to test this claim, the associations between frequency and standard deviations were studied by considering the standard deviation scores of three frequency groups of PVs, particles and verbs.

Frequent PVs occurred 1,000–35,929 times, infrequent PVs 9–96 times, and PVs with medium frequency 102–999 times. The sizes of the PV frequency groups were 52, 45 and 59, respectively. The group of PVs with frequent particles consisted of five different particles that appeared 108,758–322,547 times. This group contained 63 PVs. The group of PVs with infrequent particles also consisted of five different particles that occurred in six different PVs. These particles occurred 2,304–9,258 times. Eighty-seven PVs consisted of one to 20 particles with medium frequency (11,589–99,135). The 20 most frequent verbs appeared 111,459–756,544 times in 41 different PVs. Sixty-five PVs consisted of one to 61 verbs with low frequency (117–9,586). The remainder of the PVs (50) had a verb with moderate frequency – these 38 PVs occurred 10,482–82,453 times. Figure 4 illustrates the distributions of standard deviation values across three frequency groups of the PVs, adverbs and verbs.

[27] `http://www.eki.ee/dict/ekss/index.cgi?Q=ette+heitma` (accessed 10.11.2018).

[28] `http://www.eki.ee/dict/ekss/index.cgi?Q=ümber+riietama` (accessed 10.11.2018).

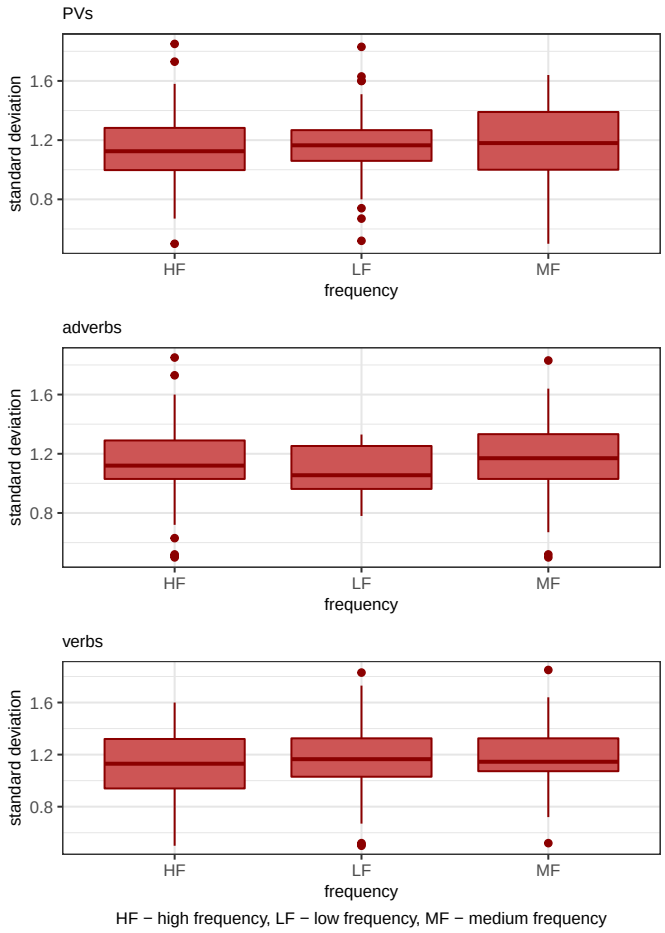

Figure 4: Distribution of standard deviation values for compositionality across the frequency bands of PVs adverb and verbs.

The medians of the standard deviations of all frequency groups of PVs were similar. The figures were insignificantly lower for PVs with high frequency than for those with low frequency. As both groups included PVs that had relatively high standard deviations, the data do not confirm that the highly frequent PVs created more disagreement among annotators than did infrequent PVs. The compositionality of frequent PVs was not more complicated to evaluate than was the compositionality of infrequent PVs.

The frequencies of the components of the PVs can also have an impact on the compositionality. Figure 4 illustrates that the standard deviation values of the PVs with infrequent particles were lower than they were for the PVs with frequent particles. Thus, it is evident that the annotators disagreed more when evaluating the PVs with particles with high and medium frequency than they did when evaluating the PVs with infrequent particles. However, it is important

to note that there were only five infrequent particles in the dataset, and that the association between the particle frequency and the compositionality requires further exploration.

The distribution of standard deviation values across the frequency bands of verbs suggested that the disagreement among annotators was slightly lower when evaluating PVs with frequent verbs than it was when evaluating PVs with infrequent verbs. Some PVs with infrequent verbs gave rise to the disagreement, but some did not. In general, it seems that the frequency of the verb did not affect the degree of disagreement among annotators.

To conclude, the observational study suggested that the frequency did not have any significant impact on how similar the annotators' judgements were. The only finding that might suggest some association was that there was less disagreement in the evaluation of PVs with infrequent particles. The effect of frequency on the compositionality ratings is studied in the next section.

4.2.5 Analysis of the compositionality ratings

The final dataset consisted of 157 PVs. The following information was provided for each PV in the dataset – the adverb of the PV, the verb of the PV, the frequency of the adverb, verb and PV, the number of human ratings, the average compositionality rating and the standard deviation of the ratings. The content of the dataset is analysed in this section.

The 157 PVs in the dataset consisted of 30 different adverbs and 119 different verbs. Figure 5 shows the distribution of the average compositionality ratings across adverbs. The adverbs are ranked based on their frequency.

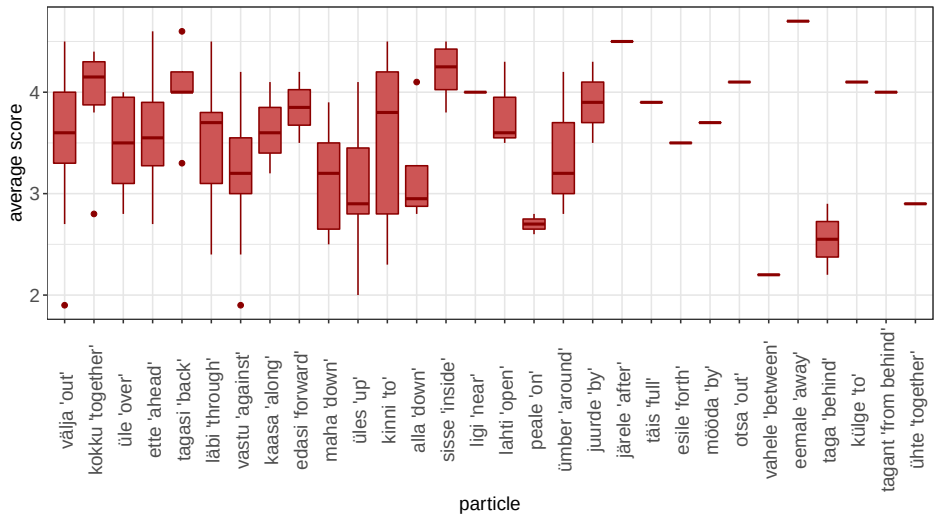

Figure 5: Distribution of the average compositionality ratings across adverbs.

The particle *välja* 'out' formed 34 PVs, which was more than any other adverb. Most of the PVs with this adverb were evaluated as being more compositional

than non-compositional, but there was one PV (*välja nägema*) 'to appear to your eyes/see outside' that was evaluated as having a compositionality degree of less than 2.0. The adverb *kokku* 'together' appeared in 13 PVs, and most of them had a compositionality degree of more than 4.0 – therefore, they are highly compositional. The PV *kokku leppima* 'to agree' was least compositional PV among the PVs with the particle *kokku*. The adverbs *peale* 'on' and *taga* 'behind', which belonged to three and two PVs, respectively, appeared to be the only PVs that had an average compositionality rating of less than 3.0 – thus, they are more non-compositional than compositional. Ten particles, such as *ligi* 'near', *juurde* 'by', *vahele* 'between' and so on, were part of only one PV. These adverbs were generally less frequent than were the adverbs that formed more than one PV.

The PVs in the dataset consisted of 118 different verbs. The verb *käima* 'to go' was the most frequent, and appeared in five different PVs. The verb *vaatama* 'to watch' occurred in four different PVs. The verbs *andma* 'to give', *jääma* 'to stay/remain', *minema* 'to go', *pääsema* 'to escape' and *tõmbama* 'to pull' were components of three different PVs. The other 112 verbs appeared in 1–2 PVs. The least compositional PV with the verb *käima* was *maha käima* 'to go down/run down/go/degenerate' (with an average rating of 2.5), and the most compositional was *kinni käima* 'to close/be able to be closed' (with an average rating of 3.8). The least compositional PV with the verb *vaatama* 'to watch' was the PV *ette vaatama* 'to foresee/look ahead' (with a rating of 2.7) and the most compositional was the PV *sisse vaatama* 'to look inside/visit something for a moment' (with a rating of 4.4).

Table 1 presents the most compositional PVs (that is, PVs with an average compositionality rating of ≥ 4.5) according to the compositionality ratings assigned by human annotators. The co-occurrence frequency of adverbs and verbs in the same clause, the numbers of the senses in the EED and the standard deviations (σ) are also provided. The most compositional PV was *eemale tõukama* 'to push away/scare off/repel' with an average rating of 4.7. The value of the standard deviation shows that the level of agreement among the annotators was relatively high. Of these 10 PVs, only two were among the PVs with high frequency (>1,000). The numbers of EED senses indicate that the most compositional PVs are not highly ambiguous. In addition, the standard deviation values suggested that the annotators agreed about the compositionality of these PVs. Therefore, several observations can be made: 1) The most frequent PVs are not highly compositional, 2) highly compositional PVs are not highly polysemous, and 3) polysemous PVs have one prominent meaning, or several meanings have similar compositionality.

Muru (2018) studied the effect of frequency on the compositionality of Estonian PVs in her bachelor's thesis and concluded, based on the six most compositional PVs in this dataset, that most compositional PVs were relatively monosemous and that annotators agreed about their degree of compositionality. However, the most compositional PVs, *eemale tõukama*[29] and *ette jõudma*[30], have two meanings in the EED ('to push away' and 'to repel'), and both were provided in the example sentences. It can therefore be assumed that these meanings are all compositional, or that the compositional ones are prominent.

[29] `http://www.eki.ee/dict/ekss/index.cgi?Q=eemale+tõukama` (accessed 10.11.2018).

[30] `http://www.eki.ee/dict/ekss/index.cgi?Q=ette+jõudma` (accessed 10.11.2018).

Table 1: The most compositional PVs according to human judgement, HJ – human judgement, σ – standard deviation among the scores assigned by the annotators.

HJ	PV	in English	frequency	EED senses	σ
4.7	*eemale* 'away' *tõukama* 'to push'	to push away	221	2	0.78
4.6	*ette* 'ahead' *jõudma* 'to reach'	to get ahead	1,671	2	0.50
4.6	*tagasi* 'back' *minema* 'to go'	to go back	5,023	3	0.79
4.5	*järele* 'after' *kihutama* 'to race'	to chase	57	2	0.52
4.5	*järele* 'after' *vahtima* 'to stare'	to stare after somebody	23	1	0.82
4.5	*kinni* 'up' *traageldama* 'to tack'	to baste up	10	1	1.17
4.5	*sisse* 'in' *kutsuma* 'to invite'	to invite in	186	1	1.13
4.5	*välja* 'out' *rändama* 'to migrate'	to emigrate	215	1	0.53

Table 2 shows the least compositional PVs (that is, PVs with an average compositionality rating of ≤ 2.4) according to the compositionality ratings assigned by the human annotators. The frequency indicates the co-occurrence of a verb and an adverb in the same clause. The numbers of the senses in the EED and the values of the standard deviations (σ) are also shown. Similarly to the most compositional PVs, there were two PVs with very high frequency amongst the least compositional PVs – *välja nägema* 'to appear to your eyes/see outside' and *läbi viima* 'to conduct/pass through'. Hence, very frequent PVs occurred amongst compositional and non-compositional PVs. In addition, the average compositionality scores indicated that the non-compositional meanings of ambiguous PVs such as *välja nägema*, *taga kihutama* 'to encourage/chase', *vahele kukkuma* 'to get caught' and so forth were predominant.

The least compositional PVs – *vastu põrutama* 'to snap back at somebody/shoot back' and *välja nägema* 'to appear to your eyes/see outside' – both have two meanings, but the example sentences expressed only one (non-compositional) meaning for each PV. Therefore, the annotators assigned the following meanings: 'to snap back at somebody' (*vastu põrutama*) and 'to appear to your eyes' (*välja nägema*). Nevertheless, this was not the case for all non-compositional PVs.

Muru (2018) concluded that the correlation between frequency and compositionality was low and statistically not significant. As she used the same dataset for the statistical analysis, a further statistical analysis of the effect of frequency

Table 2: The least compositional PVs according to human judgement, HJ – human judgement, σ – standard deviation among the scores assigned by the annotators.

HJ	PV	in English	frequency	EED senses	σ
1.9	*vastu* 'back' *põrutama* 'to knock'	to snap back at somebody	107	2	1.00
1.9	*välja* 'out' *nägema* 'to see'	to appear, to see outside	11,986	2	0.74
2.0	*üles* 'up' *kloppima* 'to fluff'	to fluff up	50	3	0.74
2.2	*taga* 'back' *kihutama* 'to race'	to encourage	23	2	1.08
2.2	*vahele* 'between' *kukkuma* 'to fall'	to get caught	16	1	1.17
2.3	*maha* 'down' *käima* 'to go'	to degenerate	339	3	0.90
2.4	*läbi* 'through' *viima* 'to carry'	to conduct, to pass through	14,144	2	1.28
2.4	*vastu* 'against' *raiuma* 'to chop'	to raise objections	55	1	1.50

is not provided in the following overview. However, some additional insights into the association between frequency and compositionality are presented.

Figure 6 shows the distribution of the averaged rating scores across the frequency bands of PVs, adverbs and verbs. The frequency groups of PVs, adverbs and verbs were the same as reported in Section 4.2.4. The frequent PVs were slightly less compositional than were the infrequent ones, but the difference is not substantial. The compositionality scores for infrequent PVs were more evenly distributed on the scale from 1–5. PVs with medium frequency tended to be more non-compositional than were other PVs, but there were PVs with high and low degrees of compositionality that had a moderate frequency. Overall, there was no clear difference in the degree of compositionality between frequent and infrequent PVs, and the PV frequency did not seem to correlate with the compositionality ratings.

PVs with frequent adverbs had a higher compositionality rating than did PVs with infrequent adverbs. The range of the average compositionality score was wider for the PVs with infrequent adverbs than it was for PVs with frequent ones. Except for the PV *välja nägema* 'to appear to your eyes/see outside' the least compositional PVs contained adverbs with medium frequency. While most of the PVs with frequent particles had a compositionality degree greater than 3.5 and most of the PVs with infrequent adverbs had a compositionality degree less than 3.5, the frequency of the adverb did not seem to correlate with the degree of

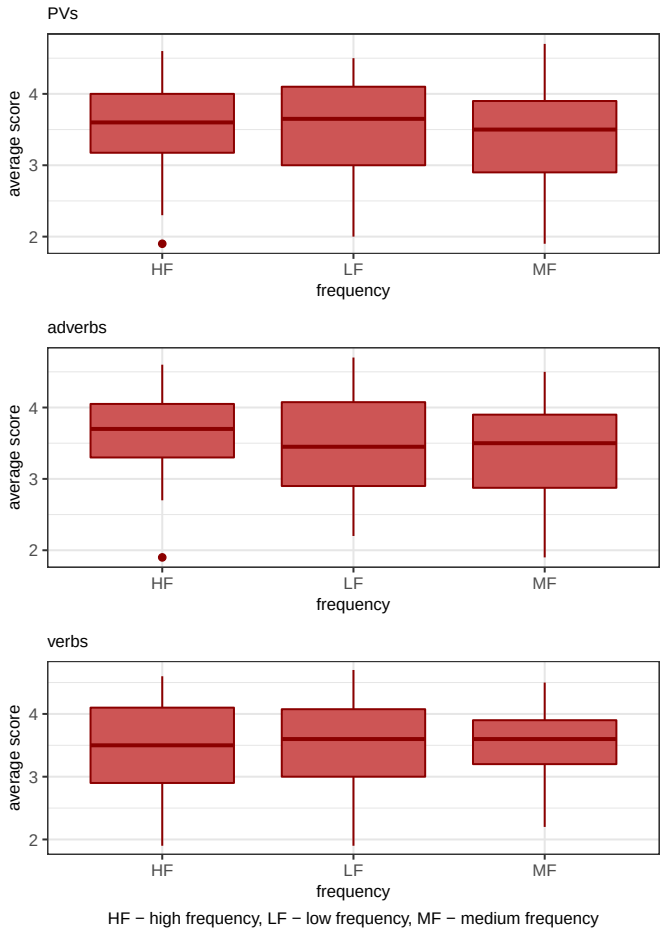

Figure 6: Distribution of averaged compositionality scores across frequency bands of PVs, adverbs and verbs.

compositionality of the PVs.

The PVs with infrequent verbs were rated as being slightly more compositional than were the PVs with frequent verbs. Nevertheless, there were highly non-compositional and highly compositional PVs among the PVs with frequent and infrequent verbs. The PVs with verbs of medium frequency tended to have medium compositionality. Depending on the frequency of the verb, the compositionality degrees varied slightly, but it cannot be claimed that the frequency of the verb had a considerable impact on the compositionality ratings of the PVs.

In summary, the dataset of compositionality ratings for the PVs contained 157 PVs with different frequencies. As demonstrated by Muru (2018) and evidenced here, the frequencies of the PV and its components did not have a substantial impact on the compositionality ratings.

4.2.6 Particle verbs that are difficult to evaluate

The PVs that received at least one 'I don't know' annotation are discussed in this section. Of 211 PVs, 40 received this rating once, 11 PVs twice, and three PVs three times. These PVs were not included in the final dataset because the assessment of their compositionality posed a challenge for the annotators. Some possible reasons that the evaluation of these PVs was complicated for the annotators are described here.

As the annotators were not asked to evaluate a particular meaning of the PVs, polysemous PVs were definitely problematic to evaluate because the target meaning was not unequivocal. For example, three annotators found the PV *maha ajama* 'to drive/push/shave off/remove' difficult to evaluate. In the EED, it[31] has four meanings, of which the example sentences expressed three. It is likely that the annotators could not select one meaning to evaluate. Furthermore, the PV *üles võtma* 'to take something up/start something (song, conversation)/record' has multiple meanings that were represented in the example sentences, and two annotators found it impossible to evaluate the compositionality of this PV. In addition to these two PVs, the compositionality of PVs such as *kõrvale tõrjuma* 'to displace/push aside', *sisse virutama* 'to push inside/break something by throwing something else at it', *vahele pistma* 'to interlard a conversation with/stick between something', *üle pingutama* 'to strain/overdo' and *üles keerama* 'to wind up/provoke' was probably difficult to evaluate for the same reasons.

Muru (2018) analysed the effect of frequency on PVs that were difficult to evaluate. She concluded that the frequency of the problematic PVs did not influence the results, although the frequent components of the PVs and their polysemy might have posed a challenge for the annotators (Muru 2018). For example, six PVs contained the adverb *järele* 'after', and four of them received at least one 'I don't know' rating. The adverb *järele* has 11 meanings according to the EED. Thus, it can be suggested that the polysemous adverb made it difficult to evaluate the compositionality of PVs such as *järele kiitma* 'to chime in', *järele kuulama* 'to inquire/expiscate', *järele laskma* 'to loosen' and *järele vaatama* 'to watch someone/check or investigate'. As the adverb in the PV *ringi sõitma* 'to take a detour/drive around' has eight meanings in the EED, the PV might have been too complex to evaluate due to the ambiguity of the adverbial component.

The role of a particle might also complicate the evaluation process. For example, the function of the particle *ära* 'away/out/off' is to express perfectivity. The difference when using the PV *ära võtma* 'to take away' instead of the verb *võtma* 'to take' alone emphasises the completion of the activity of taking. Thus, as the meaning of the PV can be understood almost entirely from the meaning of the verb, it might be difficult to decide how great an impact the adverb in the formation of the meaning of the PV has. As both PVs with the particle *ära* were difficult to evaluate, it can be speculated that the reason was the role of the particle. The same factor made it difficult to evaluate the compositionality of the PVs *maha saagima* 'to saw off', *maha salgama* 'to deny', *maha tantsima* 'to dance off' and *üles vuntsima* 'to soup up'.

In addition, the polysemy of the verbs could lead to difficulties in the as-

[31] `http://www.eki.ee/dict/ekss/index.cgi?Q=maha+ajama` (accessed 10.11.2018).

sessment of the compositionality of the PVs. This was particularly the case for frequent verbs. For example, the relatively frequent verb *tegema* 'to do' has a total of 18 meanings in the EED. All three PVs with the verb *tegema* were marked as difficult to evaluate at least once. Therefore, it may be the case that the ambiguity of a verb can cause difficulties for the annotators with the evaluation of PVs such as *ette tegema* 'to do beforehand', *sisse tegema* 'to conserve' and *ära tegema* 'to win somebody/finish doing something'. The evaluation of PVs such as *kokku ajama* 'to herd together/gather', *kokku saama* 'to meet, *läbi tulema* 'to come through', *otsa panema* 'to add', *sisse laskma* 'to let somebody in', *sisse taguma* 'to beat in(to)', *sisse õnnistama*, *välja kurnama* 'to wear out/filter', *välja minema* 'to go out', *välja tulema*, *välja võtma* 'to take out', *üle käima* 'to go/walk over', *üle mängima* 'to overplay/outplay', *üle võtma* and *ümber tõmbama* 'to put something around somebody or something/encircle' might also have been difficult because of the polysemy of the verb. According to the EED, many of these PVs have more than 10 readings, such as *minema* 'to go', *saama* 'to get/receive/have', *tulema* 'to come' and so on.

4.2.7 Summary of the compositionality ratings

The compositionality ratings for Estonian PVs were collected in order to evaluate the DSMs. The ratings were crowdsourced by asking annotators to evaluate the compositionality of PVs without specifying the meaning they should evaluate. Therefore, there is one compositionality rating per PV in the dataset, although the meaning that it represents is not determined. However, the dataset is suitable for evaluating DSM models because they produce one representation per word without discriminating among meanings.

The annotators evaluated 210 PVs, although the final dataset contained 157 PVs. The remainder of the PVs were excluded because the annotators found them difficult to evaluate. The omitted PVs were often polysemous or had ambiguous components; thus, the annotators could not decide which meaning to evaluate. Nonetheless, the hypothesis that frequent PVs are polysemous and thus more difficult to evaluate was not confirmed. However, the PVs with infrequent particles occasioned marginally less disagreement.

Based on the compositionality ratings of the Estonian PVs, no statistically significant correlation between frequency and the compositionality of Estonian PVs was found.

4.3 Literalness ratings for Estonian particle verbs

In this section, the creation of the dataset of the literal and non-literal usage of Estonian PVs (Aedmaa 2018) is described. The development of such a dataset was necessary for the evaluation of the machine learning model for the classification of the literal versus the non-literal usages of PVs, described in Section 6. The compositionality ratings from this dataset were also used for the evaluation of DSMs to detect the degree of compositionality of PVs (see Section 5). The dataset was first introduced by Aedmaa et al. (2018), but has been modified slightly for the

current, more exhaustive, research. The differences are explained in the following sections.

4.3.1 Purpose and choice of measurement

The aim of the machine learning model was to distinguish between literal and non-literal PV usage. In order to train the model for the task, a labelled dataset that included information about the (non-)literalness of Estonian PVs was required. Therefore, a set of sentences containing PVs needed to be evaluated based on the meaning of the PVs. Accordingly, human annotators were asked to evaluate a PV in each sentence based on the given context (that is, the sentence). The precise question was the following: 'To what extent does the meaning of the PV in the sentence agree with the meanings of its components?' The annotators were asked to answer to the question using a scale from 0 to 5, on which 0 indicated that the meaning of the PV agreed fully with the meaning of its components; hence, the meaning of the PV was compositional. A rating of 5 meant that the PV was non-compositional; in other words, the meaning of the PV did not agree at all with the meaning of its components. Following the assessment, it was possible to place the PVs were possible to place the PVs along a continuum based on their degree of (non-)literalness (or compositionality)[32].

When choosing a scale for the (non-)literalness ratings, it was decided to use a similar scale to that which was used for the collection of the compositionality ratings. However, following Köper and Schulte im Walde (2016b), a scale with even numbers was selected. The reason was to force the annotators to decide whether the PV was more literal than non-literal, or vice versa. The main purpose of this dataset was to apply it to the classification task described in Section 6, for which binary classification was necessary. For the final dataset, all those sentences that caused disagreement among annotators at the binary level were discarded (see Section 4.3.3). Using an even-numbered scale provided an opportunity to detect problematic sentences that needed special treatment.

There were two main reasons for asking the annotators to evaluate the degree of (non-)literalness of PVs on the scale instead of asking them to group the PVs into two strict classes. Firstly, as mentioned earlier (see Section 2.1.2), it has become common practice not to classify MWEs based on their compositionality, but to place them along a continuum from compositional to non-compositional expressions. Secondly, this dataset should be comparable to the dataset of the compositionality ratings, as both are rankings of Estonian PVs based on their

[32]For the sake of a transparent distinction between the two datasets, the ratings of the datasets were differentiated as follows – the dataset detailed in Section 4.2 contained compositionality ratings for PVs and the current dataset presents the (non-)literalness ratings for PVs. However, as mentioned earlier (see Section 2.1.2), the terms *compositionality* and *literalness* are treated as synonyms in this thesis; thus, the datasets express the same thing. Moreover, the annotators evaluated the same thing, namely the extent of the meanings of the components in the meanings of the PVs. Nonetheless, the annotators for the compositionality ratings evaluated the overall degree of compositionality of the PVs, while the annotators for the literalness ratings evaluated the meanings of PVs in a given context (that is, a sentence). Therefore, the latter is used to distinguish between the literal versus the non-literal usages of PVs.

compositionality. While the compositionality of the PVs was evaluated similarly for both datasets, there were differences in the data collection methods for these two datasets. Based on the obstacles encountered while crowdsourcing the compositionality ratings for PVs (see Section 4.2.3, an alternative but more costly method was used for the data collection for the (non-)literalness ratings of PVs. More specifically, three paid annotators with linguistic backgrounds evaluated the (non-)literalness of the sentences. The evaluation of the distributional semantic methods (see Section 5) also included a discussion of how differences in data collection methods can influence the results.

4.3.2 Overview of the target PVs and sentences

In order to compare the two datasets, the annotations for the same PVs represented in the dataset of compositionality ratings for Estonian PVs (described in Section 4.2) were collected. For each PV, 20 sentences were extracted automatically from the ERC and revised manually. All extracted sentences included a PV, and there were at least three sentences representing each PV. The following seven PVs appeared in less than three sentences, and were subsequently omitted from the dataset – *kinni traageldama* 'to baste up', *läbi kobama* 'to fumble through', *läbi kompama* 'to touch through', *üles käänama* 'to roll up', *üles tursuma* 'to swell up', *ümber reastuma* 'to change a lane' and *ümber riietama* 'to change somebody's clothes'. In order to still have a representative number of PVs in the dataset, sentences with six new PVs – *alt minema* 'to fail or to be deceived', *juurde lõikama* 'to cut/add (land)', *maha võtma* 'to take down', *peale käima* 'to be insistent/impose/enforce', *üle ajama* 'to flow over' and *üles lööma* 'to dress up/toss (upward)' – were added to the dataset. All these PVs can be used in literal and non-literal senses and were thus included in the dataset as interesting cases.

The number of sentences for each PV in the dataset was not equal. The number mainly depended mostly on how many of these 20 automatically extracted sentences included a real PV, and how many did not. Some sentences were omitted because they were too long or confusing. Naturally, frequent PVs tended to have a higher number of sentences. Those with the highest number of sentences represented by PVs included *üles keerama* 'to wind up/provoke' (17), *vastu hakkama* 'to resist/detest' (16), *vastu kajama* 'to sound like an echo' and *maha ajama* 'to drive/push/shave off/remove' (15). The PVs *kokku valguma* 'to join/melt together', *kokku kiskuma* 'to shrink', *maha tantsima* 'to dance off', *sisse taguma* 'to beat in(to)', *sisse virutama* 'to push inside/break something by throwing something else at it', *tagasi nimetama* 'to give back old name', *välja nutma* 'to cry yourself out', *üle uhtuma* 'to flush/wash' and *üles ehmatama* 'to startle' all appeared in three sentences. Most of the PVs that were represented by a low number of sentences were less frequent than were other PVs. For example, the frequency of *kokku valguma* was 18, while the frequency of *vastu hakkama* was 3,271. Overall, most PVs appeared in 7–12 sentences.

In the dataset, all the sentences with the same PV followed each other; six incorrect sentences were added in order to ensure that annotators were paying attention. In these sentences, homonymous adpositions were presented instead of verbal particles; thus, it was impossible to evaluate the compositionality of the

PVs. Examples (14)–(17) illustrate some of these incorrect sentences[33].

(14) Millise-d osa-d selle-st ei tohi-∅ avalikkuse-∅
which-PL part-PL this-ELA NEG can-CONNEG public-GEN
ette sattu-da?
front get into-INF
Lit. 'Which parts of it cannot get in front of the public?'
'Which parts of it cannot become public?'

(15) Jä-i-n üksi parve-∅ **külge**.
stay-PST-1SG alone raft-GEN to
'I stayed attached to the raft alone.'

(16) Külm tungi-b **läbi** luku-ga viltsaabas-te ja
cold invade-3SG through zipper-COM felt boot-PL.GEN and
sunni-b jala-lt jala-le tammu-ma.
force-3SG foot-ABL foot-ALL step-SUP
'Cold air passes through the zippered felt boots and forces one to step from foot to foot.'

(17) Prantsuse-∅ kriitiku-d ei vaevu Alain Deloni-∅
French-GEN critic-PL NEG trouble-CONNEG Alain Delon-GEN
viimase-i-d filme-∅ maa-∅ **sisse** tagu-ma-gi.
last-PL-PRT film-PL.PRT ground-GEN into slog-SUP-CL
Lit. 'The French critics hardly bother to slog Alain Delon's latest films into the ground.'
'The French critics hardly bother to criticise Alain Delon's latest films.'

For example, in example (15), the word *külge* 'to' is not a verbal particle, but a postposition – together with the word *parv*, it forms a postpositional phrase *parve külge* 'attached to the raft'. In example (16), the word *läbi* forms together with the word *viltsaabas* a prepositional phrase *läbi viltsaabaste* 'through felt boots', not a PV *läbi tungima* 'to penetrate/go right through' with the verb *tungima* 'to force'. These sentences were excluded from the final dataset before the evaluation described in the following section.

4.3.3 Results of the human annotation

All three annotators with linguistic backgrounds evaluated the same 1,838 sentences on a scale of 0 to 5. The evaluation of the dataset (1,832 sentences), based on Fleiss' kappa (κ) score, indicated that the agreement among the annotators was fair ($\kappa = 0.36$); while the agreement was substantial for two categories, as

[33]The sentences might have been shortened to improve the readability thereof.

$\kappa = 0.71$. In case of Krippendorff's α, corresponding values were $\alpha = 0.68$ and $\alpha = 0.71$.

For the classification task (see Section 6), the sentences needed to be divided into two groups – the sentences were categorised as literal and non-literal based on the degrees of (non-)literalness of the PVs. In order to accomplish this, the three scores given by the annotators were averaged – if the PV received an average degree of (non-)literalness equal to or lower than 2.4, the sentence was labelled as literal. The sentences with an average degree of (non-)literalness of 2.5 or more were classified as non-literal. The sentences with considerable disagreement were rejected – if one annotator rated a PV as literal at the binary level, while the other two rated it as non-literal, or vice versa, the sentence was excluded from the dataset. After removing problematic sentences, the agreement among annotators for six categories was moderate ($\kappa = 0.43$) and reliable ($\alpha = 0.85$). Altogether, 351 sentences with 73 different PVs were omitted from the final dataset. Two groups were formed for the following overview and analysis of the omitted sentences, namely the PVs that were completely excluded and the PVs that were partially excluded from the dataset.

Of the original 210 PVs, 26 were omitted completely from the dataset – all the sentences containing these PVs led to disagreement among the annotators. There are several explanations regarding why the annotators did not agree unanimously in their judgements. Except for the scale of measurement, the annotators were not restricted in any way. Therefore, their evaluations may have been based on information obtained from other resources, such as the EED. Furthermore, the annotators were not required to explain their ratings. The subsequent analysis suggested possible reasons why the annotators disagreed.

The following PVs were represented with one meaning in the sentence selection – *edasi müüma* 'to sell on', *kokku trehvama* 'to run across' *maha müüma* 'to sell off', *maha saagima* 'to saw off', *maha salgama* 'to deny', *maha tapma* 'to kill', *maha tantsima* 'to dance off', *tagasi nimetama* 'to give back old name', *välja loosima* 'to raffle off', *välja nutma* 'to cry yourself out', *välja pakkima* 'to unbox', *välja tahuma* 'to hew out', *üle küsima* 'to ask again', *üles harima* 'till the soil', *üles joonistama* 'to draw', *üles vuntsima* 'to soup up', *ümber ristima* 'to rename' and *ümber kohendama* 'to readjust'. As the annotators tended to evaluate the same meaning via the same ratings, all the sentences with these PVs caused disagreement among the annotators, and were thus excluded from the final dataset.

The verbs in most of these PVs have only one (prototypical) meaning. It is thus likely that the reason for disagreement was the polysemy of particles. For example, the PVs with the particle *maha* 'off/down' might have generated variance when one annotator based the evaluation on the directional meaning ('down') of the particle, while others assessed the meaning as expressing perfectivity.

Although the PV *välja mõõtma* 'to measure out' has one meaning in the EED, one annotator did not evaluate all of the sentences with this PV in the same way. Examples (18) (literalness score 2.67) and (19) (literalness score 3.00) illustrate how they differentiated between the measurement of land (which was the most common object to be measured in the dataset) and the measurement of other things, such as energy.

(18) Pendli-ga **on** ta **mõõt-nud** näiteks Tartu-s
pendulum-COM be.3SG s/he measure-PST.PTCP for example Tartu-INE
välja kõige positiivse-ma-∅ energia-ga koha-∅.
out most positive-COMP-GEN energy-COM place-GEN
'With a pendulum, she/he has measured out, for example, the place with the most positive energy in Tartu.'

(19) Krunti-∅ **ei** **ole-∅** veel **välja** **mõõde-tud**,
Plot-PRT NEG be-CONNEG yet out measure-PST.PTCP
kuid kõik õiguse-d maa-le on ole-mas.
but all right-PL land-ALL be-3SG be-PRS.SUP
'The land has not been measured out yet, but all rights to the land are available.'

The PVs that were represented as having multiple meanings were *ringi sõitma* 'to take a detour/drive around', *välja tõrjuma* 'to crowd off/displace/supersede', *üle pesema* 'to wash (again/over)' and *üles kündma* 'to plough up'. It is possible that the annotators compared the meanings of the same PV and made their decisions based on analogy. Furthermore, depending on how the annotators understood their task, one may have assessed the components as having the same meaning regardless of the meaning of the PV as a whole. Moreover, other annotators changed the meaning of the components based on the meaning of the PV. This may have been because the components are polysemous, and the particular meanings that were to be assessed were not specified. For example, *ringi sõitma* has two meanings – the first is to express taking a detour, and the other is to express driving around. Both the particle and the verb have at least two different meanings, and it cannot be expected that all the annotators evaluated the same meanings. The disagreement might also have been caused by polysemous verbs such as *koguma* 'to gather' and *kiiluma*. For example, *täis kiiluma* expresses a situation in which something is crowded. One possible meaning of *kiiluma* is 'to become stuck'. When an annotator had this meaning in mind, they probably assessed the PV differently from the others.

In addition, the subjective understanding of literalness plays a role in disagreement. Some PVs are very easy to evaluate subjectively at a binary level. For example, all of the sentences containing the PV *otsa sõitma* 'to run down' expressed hitting something or somebody by driving. Although two of the annotators determined that the PV was literal, one decided that this PV was more non-literal than literal. In order to understand the reasons for such decisions, a more exhaustive study needs to be conducted.

Forty-seven PVs appeared at least in one sentence which led to disagreement, and in at least one sentence, which did not cause disagreement among the annotators. It could be argued that the reasons for the differences were the same as discussed above, but there are some additional sources of disagreement that affect these types of PVs. One reason for the conflict amongst the annotators was the lack of sufficient context. For example, the sentence in example (20) does not provide sufficient context to determine clearly whether the meaning of the PV *edasi jõudma* expresses spatial movement or progress in something. Therefore,

each annotator added his or her own context and carried out the estimation based on that. Furthermore, the disagreement might indicate that this PV did not have one dominant meaning for all of the annotators.

(20) Niimoodi **edasi** **ei** **jõu-a**
like that forward NEG reach-CONNEG
Lit. 'Cannot reach forward like this.'
'Cannot get ahead like this.'
'Cannot succeed like this.'

Another reason for the difference might be an unusual usage of the PV. For example, while some sentences containing the PV *läbi sadama* expressed its main meaning of 'to leak when raining', the usage of this PV in the sentence in example (21) is unexpected and fuzzy. Based on the context, it is not clear whether the verb *sadama* expresses the meanings of 'falling' or of 'unexpected appearance'[34]. Thus, which meaning was required to be evaluated was not absolutely clear. This resulted in an outcome in which the annotators understood the context differently and did not agree about the literalness of this PV.

(21) Pundi-ga **või-b** **läbi** **sada-da**, hoiata-si-d korraldaja-d.
group-COM might-3SG through fall/rain-INF warn-PST-3PL organiser-PL
Lit. 'The organisers warned with a group might fall/rain.'
'The organisers warned that it may fall when there is a group on it.'
'The organisers warned that she/he might drop by with a group.'

When a PV has one (or more) clearly non-literal or literal meaning(s), it might be difficult to evaluate other meanings of the same PV. The PV *üles lööma* was presented as having different meanings in the initial selection of the sentences. While the highly non-literal meaning 'to dress up' (as in example (22)) and the extremely literal meaning 'to toss (upward)' (as in example (23)) did not spark any differences, the third meaning was difficult to assess at a binary level. As the meaning of the PV in the sentence in example (24) is neither fully literal nor fully non-literal, the annotators evaluated it as being more literal than the sentence in example (22), and less literal than the sentence in example (23). However, they did not agree whether the usage was literal or non-literal at the binary level.

(22) Et naine en-d moodsalt **üles** **lüü-a** saa-ks.
that woman own-PRT fashionably up beat-INF can-COND
Lit. 'That woman can beat up fashionably.'
'That woman can dress up fashionably.'

(23) Mõnikord **löö-vad** oma-∅ jalg-u kõrgele **üles**.
sometimes beat-3PL own-GEN foot-PL.PRT high up
'Sometimes they kick their legs up high.'

[34]The annotators suggested these meanings.

(24) Tänavanurka **lüü-a-kse** silt **üles**.
street corner beat-IMPS-PRS sign up
'The sign is put up on the corner of the street.'

The same also occurred for PVs with fewer meanings. For example, *ära tegema* has one clearly non-literal meaning – 'to win (or be better than) somebody' (as in example (25)). Adjacent to the sentence in which *ära tegema* was used with this meaning were sentences in which the meaning of the PV was more literal; that is, '(to finish) doing something' (as in example (26)). While the non-literal meaning was unanimously evaluated as being extremely non-literal, the annotators did not agree regarding the extent of the literalness of the literal meaning of the PV.

(25) **Teh-ke** sakslas-te-le **ära**, Inglismaa-ga saa-me
do-IMP.2PL German-PL-ALL off England-COM can-1PL
ise hakkama.
ourselves cope with-SUP
Lit. 'You do off the Germans, we can cope with England ourselves.'
'You beat the Germans, we will handle England ourselves.'

(26) Inimene saa-b midagi **ära** **teh-a** oma-∅
human can-3SG something off do-INF own-GEN
vererõhu-∅ heaks.
blood pressure-GEN for
'A person can do something about his blood pressure.'

Another reason that the PV *ära tegema* in example (26) might have caused disagreement is the indistinct role of the particle. For example, the particles *ära* 'away/out/off' and *maha* 'off' are the main particles used to express perfectivity. Eleven of the 16 PVs containing the particle *maha* led to conflict amongst the annotators. The reason might have been that the meaning of the particle is complex to determine because its role as a part of the component of the PV is difficult to evaluate. Moreover, the annotators might have been assessing different meanings of the adverb.

Even though the sentences with the same PV expressed the same meaning, and the annotators agreed on whether the usage of the PV was literal or non-literal, there were some sentences with the same PV upon which the annotators did not agree. For example, the PV *kinni minema* has two main meanings – 'to close' (as in example (27)) and 'to go to prison' (as in example (28)). All the annotators determined that the sentences with the first meaning were literal and the sentences with the second meaning were non-literal. However, one sentence (see example (29)) caused disagreement because one annotator evaluated it differently from the others.

(27) Ukse-d **lä-k-sid** **kinni**.
door-PL go-PST-3PL closed
Lit. 'The doors went closed.'
'The doors closed.'

(28) Kaido **läk-s** **kinni** seitsme-ks aasta-ks, Markel viie-ks.
Kaido go-PST.3SG closed seven-TRL year-TRL Markel five-TRL
Lit. 'Kaido went closed for seven years, Markel for five.'
'Kaido went to jail for seven years, Markel for five.'

(29) Samuti **läk-s** esmaspäeva-st **kinni** osa Tartu-∅ maantee-st.
also go-PST.3SG Monday-ELA closed part Tartu-GEN highway-ELA
Lit. 'Part of the Tartu highway went closed also on Monday.'
'Part of the Tartu highway closed also on Monday.'

The reason for this might have been the unusual subject (*osa Tartu maanteest* 'part of the Tartu highway') – in general, roads are closed by somebody, and do not close themselves. This might have influenced the annotator's decision, thus causing the annotators to disagree amongst themselves.

In some cases, the annotator adhered strictly to the EED and the meanings presented in it. For example, the EED presents one meaning for the PV *vastu küsima*[35] – 'to answer the question with a question'. There are sentences in which the PV has been used with this meaning (as in example (30)), but there are also sentences in which the meaning of this PV is 'to ask something in return' (as in example (31)). One annotator evaluated all the sentences in the same way, but two annotators differentiated between these two meanings. Therefore, the annotators did not agree about the sentences in which the meaning of the PV was 'to answer the question with a question'.

(30) Kui Joala-lt küsi-da, **küsi-b** endine laulja **vastu**.
when Joala-ABL ask-INF ask-3SG former singer back
'When Joala is asked, the former singer asks back.'

(31) NRG **küsi-b** **vastu** pool loodava-∅ ühisfirma-∅
NRG ask-3SG in return half to be created-GEN joint venture-GEN
aktsia-te-st.
stock-PL-ELA
'In return, NRG asks for half of the stocks of the created joint venture.'

Three sentences in which the PV *üle kaaluma* was used to express the meaning 'to outweigh' (as in example (32)) received special attention from the annotators. In fact, two annotators suggested that *üle kaaluma* was a typographic error. They noted that, instead of the particle *üle*, the particle *üles* should have been used because the PV *üle kaaluma* does not have the meaning 'to outweigh' that the PV *üles kaaluma* has. They annotated these sentences as if the particle had been *üles* 'up'. The third annotator determined that the PV was fully non-literal. While two annotators did not assess the meaning of the adverb *üle* 'over', these three sentences were excluded from the final dataset.

[35] `http://www.eki.ee/dict/ekss/index.cgi?Q=vastu+küsima` (accessed 10.11.2018).

(32) Kasu **kaalu-b** **üle** selle-∅ kahju-∅, mis looma-d
benefit weigh-3SG over this-GEN harm-GEN what animal-PL
loomaaia-st saa-vad.
ZOO-ELA get-3PL
Lit. 'The benefit weighs over the harm that the animals get from the zoo.'
'The benefit outweighs the harm that the animals experience as a result of living in the zoo.'

In addition to the PVs mentioned in the previous discussion, at least one sentence with each of the following PVs led to disagreement amongst the annotators – *eemale tõukama* 'to push away/scare off/repel', *ette jõudma* 'to get ahead/outstrip/outdistance', *ette tegema* 'to do beforehand', *järele kihutama* 'to chase', *järele vahtima* 'to stare after somebody', *kaasa tõmbama* 'to persuade to join/pull along', *kokku monteerima* 'to assemble/edit video', *kokku sättima* 'to set/put together', *lahti pääsema* 'to break loose/get free', *ligi pääsema* 'to access something/get close to somebody/be comparable', *läbi lendama* 'to fly through/fail', *läbi põletama* 'to fuse something/to burn something out', *läbi tulema* 'to come through', *läbi valgustama* 'to x-ray/dissert', *maha jahtuma* 'to cool down', *maha kustutama* 'to erase/wipe out', *maha kõmmutama* 'to shoot dead', *maha minema* 'to get off', *maha rahunema* 'to calm down', *maha võtma* 'to take down', *mööda käima* 'to bypass', *otsa panema* 'to add', *peale hakkama* 'to start out/in', *sisse torkama* 'to stick in', *tagasi vaatama* 'to look back', *vastu rääkima* 'to talk back/dispute', *vastu vahtima* 'to stare back at somebody', *vastu võtma* 'to accept/welcome/admit', *välja ilmuma* 'to debouch/emerge/appear unexpectedly', *välja kostma* 'to be heard', *välja minema* 'to go out', *välja puhastama* 'to clean out/restore', *välja saagima* 'to saw out', *välja venima* 'to stretch out', *välja võtma* 'to take out', *ära võtma* 'to take away', *üles keerama* 'to wind up/provoke', *üle lugema* 'to read over/recount', *üle pakkuma* 'to exaggerate', *üles lööma* 'to dress up/toss (upward)' and *ümber tõmbama* 'to put something around somebody or something/encircle'. Although the annotators were provided with a description of the task, it is not possible to provide the exact reasons and explanations concerning which meaning they evaluated, or why they assessed the PVs as they did.

The initial selection of the target PVs for the literalness rating dataset was very similar to the target PVs for the compositionality dataset (see Section 4.2.1). Therefore, the PVs that caused disagreement amongst the annotators of this dataset were compared to the PVs the annotators of the compositionality dataset referred to as being difficult to evaluate (see Section 4.2.6). The comparison revealed that 19 PVs were omitted from the compositionality dataset because they were difficult to evaluate, and appeared in at least one sentence that caused disagreement amongst the annotators of literalness. The PVs with the most problematic compositionality, and which were thus omitted from both datasets, were *kokku trehvama* 'to run across', *maha saagima* 'to saw off', *maha salgama* 'to deny', *maha tantsima* 'to dance off', *ringi sõitma* 'to take a detour/drive around', *tagasi nimetama* 'to give back old name' and *välja nutma* 'to cry yourself out'. The reasons that these PVs were difficult to evaluate were discussed previously – the role of the particle was not clear, or the ambiguous components caused disagreement (see Section 4.2.6).

Overall, the final dataset contained 1,481 sentences – 1,096 sentences were

labelled as non-literal and 385 as literal. This information and averaged literalness scores given by annotators were added to the dataset for each sentence. Compared to the first version of the dataset introduced by Aedmaa et al. (2018), which included 1,490 sentences, the current version of the dataset excluded six incorrect sentences and three sentences containing the PV *üle kaaluma* 'to weigh again'. The reasons for the exclusions are discussed above.

4.3.4 Analysis of (non-)literalness ratings

There were (non-)literalness ratings for 184 Estonian PVs in the dataset. These PVs consisted of 120 verbs and 32 particles. In this section, the average (non-)literalness ratings of PVs, adverbs and verbs in the dataset are analysed. In addition, the association between frequency and (non-)literalness ratings is examined.

The 32 particles in the PVs in the dataset had different frequencies; hence, the number of PVs they formed was not equal. this means that some particles were part of more than 10 PVs, and some particles only occurred in one PV. For example, *välja* 'out' was the most frequent particle in the dataset and the most frequent adverb in the corpus. It was also a constituent of the greatest number of different PVs (28). While PVs containing the particle *välja* form the biggest group of PVs represented in the EED (226 PVs), it can be suggested that the particle *välja* forms the highest number of PVs in Estonian; this was reflected in the dataset. Other particles that, according to the EED, are components of many PVs, were also present in the dataset, including *ära* 'away/out/off', *maha* 'down/off', *läbi* 'through', *sisse* 'in', *kokku* 'together' and *üles* 'up'. While the particle *ära* is the second most productive particle after the particle *välja*, it only appeared in four sentences formed with two different PVs in the dataset. The reason for the poor representation of PVs containing the particle *ära* is that most of them were excluded from the initial selection of PVs (see Section 4.2.1). Moreover, the annotators found the literalness of these PVs challenging to assess (see Section 4.3.3).

In general, the number including the sentences of a particle was in compliance with the number of different PVs to which the particle belonged to. However, there were some exceptions. For example, 13 PVs containing the particle *üle* 'over' appeared in more sentences than did the 14 PVs containing the particle *läbi*. Furthermore, six PVs containing the particle *tagasi* 'back' were represented in more sentences than were the nine PVs containing the particle *sisse* 'in'. These kinds of differences occurred for two main reasons. Firstly, some automatically extracted sentences (20 for each PV) did not contain an actual PV, and were omitted from the annotation. Secondly, the annotators did not agree about some of the sentences containing some of the PVs; these were also omitted from the final dataset (see Section 4.3.3).

The number of sentences that included particles did not correlate to the particles' frequencies in the corpus. As the selection of the PVs had already been based on the frequency of various individual components (see Section 4.3.2), the frequencies of the components were not a factor when selecting PVs for the dataset. Moreover, as the components of a PV can occur independently in a sentence,

the high frequency of a particle does not necessarily imply that it is a frequent component of a frequent PV. For example, *ära* 'away/out/off' is a frequent adverb in the corpus and, according to the EED, it is also a component in many PVs. Nevertheless, the annotators evaluated 20 sentences containing two PVs with the particle *ära*, and only four of these sentences did not cause disagreement among the annotators. Therefore, the number of sentences that includes the particle does not reflect the frequency of the particle in the corpus.

The particle *ligi* 'near' was the least frequent in the dataset, while *ühte* 'together' occurred less often in the corpus than did any other adverb in the dataset. The annotators evaluated 11 sentences containing one PV that included the particle *ligi*. As eight sentences caused the annotators to disagree, the particle *ligi* was represented in three sentences. The particle *ühte* was also represented via one PV (*ühte hoidma* 'to stick together'), but all seven sentences expressed one meaning that was assessed similarly by all the annotators.

Particles that appear as components of the PVs are present in the EED, but these were not in the dataset; example include *ilma* 'without', *järel* 'after', *kaasas* 'with/along', *kallale* 'on/upon/at', *koos* 'together', *kõrvalt* 'from aside', *kätte* 'into hands', *lahku* 'apart', *laiali* 'around', *minema* 'to go', *peal* 'on', *pealt* 'from', *pihta* 'at', *pärale* 'across', *püsti* 'up(right)', *ringi* 'round', *taha* 'behind', *takka* 'behind', *tasa* 'even', *tulema* 'to come', *täis* 'full', *vahelt* 'between', *valla* 'open', *valmis* 'ready', *ühes* 'with', *ülal* 'above' and *üleval* 'above'. While some other particles were not included in the collection at all, *ringi* and *täis* were evaluated as components of the PVs *ringi sõitma* 'to take a detour/drive around' and *täis kiiluma* 'to become stuck', respectively. However, this led to disagreement amongst the annotators, and the particles were consequently omitted from the final dataset (see Section 4.3.3).

As the PVs in the dataset contained particles with various frequencies, it can be assumed that the PVs containing frequent particles had more meanings represented in the dataset than did those containing infrequent particles. The distribution of the average (non-)literalness ratings across the particles in the literalness dataset is illustrated in Figure 7. The order of particles reflects the number of PVs containing the particles. Accordingly, the particle *välja* 'out' is a component of more PVs than is the particle *ühte* 'together'. Most sentences containing these particles were evaluated as being more non-literal than literal, as the medians for these particles tended to be greater than 3. Although *sisse* 'in' and *lahti* 'open' received ratings from 0–5, at least half of the scores for these particles signified more literal than non-literal meanings. It can thus be suggested that these two particles tended to be components of the PVs expressing a literal meaning.

The most frequent particle, *välja* 'out', appeared in 28 different PVs, and these PVs were evaluated as being fully literal (compositional), fully non-literal (non-compositional) and as having scores between these two extremes. Most of the meanings of the particle *ette* 'in advance/ahead/forward' are non-literal. However, as there were some outliers, the scores were different from the majority of the given evaluations in some cases. In the current study, these scores indicated highly literal meanings for *ette andma* 'to put something in front of somebody/feed/specify', *ette vaatama* 'to foresee/look ahead' and *ette sattuma* expressing 'to run across or

Figure 7: Average (non-)literalness scores across particles.

meet somebody or something on the way'. At the binary level, most of the PVs appeared at least once in a literal sentence and once in a non-literal sentence.

Even though 11 particles only appeared once in one PV, most of the PVs had multiple meanings. Only the particles *ühte* 'together', *ligi* 'near' and *alt* 'from under' formed part of PVs that had only one meaning in the dataset. While *ühte hoidma*[36] 'to stick together' and *alt minema*[37] 'to fail or to be deceived' also have one meaning according to the EED, the PV *ligi pääsema*[38] has three – 'to access something', 'to get close to somebody' and 'to be comparable'. The second meaning of this PV was represented in the dataset. The first meaning was evaluated, but it created disagreement amongst the annotators and was therefore excluded from the dataset. The third meaning was not included in the initial selection of sentences. It can thus be assumed that two meanings of the PV *ligi pääsema* are more frequent than is the third one.

The particle *eemale* appeared in one PV – *eemale tõukama* – which has two different meanings. The first meaning 'to push away' (see example (33)) was rated as being fully literal, while the other meaning 'to scare off' (see example (34)) was given an average score of 3. Therefore, the second meaning is less literal than is the first meaning, but neither of them is fully non-literal. This example is a good illustration of how the same PVs can have several meanings with various degrees of compositionality.

(33) Jüri püüd-is tundmatu-t **eemale** **tõuga-ta**, kuid too
Jüri try-PST.3SG stranger-PRT away push-INF but that
lõ-i ta-lle noa-∅ südame-∅ piirkonda-∅.
hit-PST.3SG s/he-ALL knife-GEN heart-GEN area-ILL
'Jüri tried to push the stranger away, but he stabbed him around the heart area.'

(34) Kooli-s olevad pitsliniku-d ja toataime-d **tõuka-vat** poisi-d
school-INE being lace doily-PL and houseplant-PL push-QUOT boy-PL
eemale.
away
Lit. 'The lace doilies and houseplants in the school seem to push boys away.'
'The lace doilies and houseplants in the school seem to repel boys.'

In summary, the frequency of the particles varied from frequent to relatively infrequent, and frequent particles were included in more PVs than did infrequent ones. Despite the fact that most of the sentences in the dataset were evaluated as being more non-literal than literal, the average (non-)literalness scores across the particles suggest that it is likely that most of the particles can be components of PVs with various degrees of (non-)literalness, regardless of the frequency of the particle. At the binary level, most particles occurred at least once in a PV that was evaluated as being literal.

[36] http://www.eki.ee/dict/ekss/index.cgi?Q=ühte+hoidma (accessed 10.11.2018).
[37] http://www.eki.ee/dict/ekss/index.cgi?Q=alt+minema (accessed 10.11.2018).
[38] http://www.eki.ee/dict/ekss/index.cgi?Q=ligi+pääsema (accessed 10.11.2018).

The group of verbs included in 184 PVs in the dataset was more diverse than was the set of adverbs – 120 different verbs were represented. Therefore, most of the verbs only appeared in one PV. However, this does not imply that the average compositionality scores for these verbs were similar. The average (non-)literalness scores for the verbs forming the PVs in the dataset are explored in the following section.

The frequency of the verbs varied – the most frequent verb was *saama* 'to get/receive/have' with a frequency of 1,150,781 and the most infrequent one was *kõlksuma* 'to clatter', which only occurred 64 times. Other infrequent verbs, such as *sülgama* 'to spit', *voltima* 'to fold', *kõmmutama* 'to bang', *tükkima* 'to intrude', *trumpama* 'to trump/ruff', *kajama* 'to echo', *tuhnima* 'to dig/grub/rout', *kloppima* 'to fluff', *soojenema* 'to warm up', *uhtuma* 'to flush/wash', *kuhjama* 'to heap up' and *kehastuma* 'to be incarnated', appeared 100–1,000 times. The number of the sentences is not associated with the frequency – there were many verbs with different frequencies that formed part of one PV. For example, both *saama* and *kõmmutama* were part of only one PV.

The number of sentences was not based on the frequency of the verbs, but does correlate with it to some extent. For example, the frequency of the verbs that appeared in seven or fewer sentences was less than 81,000. However, there were some verbs that had a relatively low frequency (less than 10,000), but which appeared in more than 10 sentences, such as *tõukama* 'to push', *pistma* 'to stick', *kajama* 'to echo' and *peksma* 'to beat'. Of the 17 verbs that appeared in 20 or more sentences, 11 had a frequency greater than 100,000. The exceptions were the verbs *ajama* 'to drive/run', *tõmbama* 'to pull', *lugema* 'to read', *laskma* 'to let/have/allow/shoot', *heitma* 'to throw' and *kasvama* 'to grow'.

The verbs *võtma* 'to take' and *käima* 'to go/walk' formed part of seven PVs, while the first occurred in 59 sentences and second is in 51 sentences. The verbs *saagima* 'to saw', *ilmuma* 'to appear', *valguma* 'to flow', *venima* 'to stretch', *ehmatama* 'to frighten', *kiskuma* 'to tear', *kõmmutama* 'to bang', *kustutama* 'to erase', *torkama* 'to prick/sting', *uhtuma* 'to flush/wash' and *virutama* 'to whack' appeared in one PV, but in less than five sentences. The reason for so few sentences might have been the low frequency of the PV (for example, *kokku kiskuma*, *sisse virutama* 'to push inside/break something by throwing something else at it' and *üle uhtuma*) 'to flush/wash' or disagreement amongst annotators (for example, *maha kustutama* 'to erase/wipe out', *välja venima* and *välja ilmuma* 'to debouch/emerge/appear unexpectedly').

Figure 8 illustrates the average (non-)literalness scores for the verbs in the PVs. The verbs are ranked based on the number of different PVs to which they belonged to. The verbs that appear only in one PV with one meaning are not presented. According to the EED, the most frequent verbs in the dataset are also polysemous. Therefore, it can be assumed that different meanings of the verbs are represented in the dataset forming both literal and non-literal PVs.

The verbs *käima* 'to go/walk' and *võtma* 'to take' were both part of seven PVs. These PVs were evaluated as being more non-literal than literal, but there were some meanings that were literal. Examples are the meaning 'to walk (expressing perfectivity of the activity)' of the PV *maha käima* and the meaning 'to take something out from somewhere' of the PV *välja võtma*. The verbs *minema* 'to

Figure 8: (Non-)literalness ratings across the verbal components of PVs.

go', *vaatama* 'to look', *ajama* 'to drive/run' and *tõmbama* 'to pull' were part of four or five PVs with various (non-)literalness scores. All the (non-)literalness scores for these three PVs containing the verb *lugema* 'to read' suggest that this verb only formed part of PVs that were non-literal at the binary level.

Many verbs were components of only one PV. Verbs such as *lendama* 'to fly', *lööma* 'to hit', *rääkima* 'to talk', *riputama* 'to hang', *siduma* 'to tie/bind', *valguma* 'to flow', *valgustama* 'to lighten', *viima* 'to carry' seemed to form part of PVs that had several meanings. For example, *lendama* formed part of the PV *läbi lendama* 'to fly through/fail', which had at least three meanings with very different degrees of (non-)literalness. The example (35) expresses the most literal meaning of the PV, which is 'to visit many places by flying'. The meaning in example (36) is 'to visit something quickly', and was evaluated as being more non-literal than was the meaning of the PV in example (35), but not fully non-literal like the meaning 'to fail the exam' in example (37).

(35) Ta ol-i selle-ga pool maailma-∅ **läbi** **lenna-nud**.
s/he be-PST.3SG this-COM half world-PRT through fly-PST.PTCP
Lit. 'She/he had flown through half the world with it.'
'She/he had flown half the world with it.'

(36) Ja järgmine kord **lenda-b** klient juba **läbi**
and next time fly-3SG customer already through
kuskilt mujalt
somewhere elsewhere
Lit. 'And next time the client will fly through someplace else.'
'And next time the client will visit someplace else.'

(37) Karemäe ei lase-∅ tuju-l lange-da: "Noh,
Karemäe NEG let-CONNEG mood-ADE fall-INF well
läbi **lenda-si-n**!".
through fly-PST-1SG
Lit. 'Karemäe does not let the mood fall: "Well, I flew through!"'
'Karemäe does not mourn: "Well, I failed!"'

Even though some verbs appeared in several PVs, the (non-)literalness degrees of these PVs were similar. For example, PVs containing the verbs *tegema* 'to do', *hakkama* 'to start', *heitma* 'to throw' and *hoidma* 'to keep' were all considered more non-literal than literal, and none of the PVs received an average (non-)literalness rating lower than 3. This observation indicates that some verbs were more likely to appear in non-literal PVs than in literal PVs.

In conclusion, the frequency of the 120 verbs in the PVs varied from 64 to 1,150,781, but this does not reflect upon the number of sentences or PVs. Frequent verbs tended to form part of more PVs, and thus appeared in a greater number of sentences. The average (non-)literalness scores reveal that few verbs only appeared as components of fully literal or non-literal PVs. However, the

sample of the verbal components of the PVs is not sufficiently representative to draw exhaustive conclusions about the compositionality/literalness of the verbal components of PVs.

The final dataset consisted of 184 PVs. The number of sentences containing PVs varied due to several reasons. Firstly, not all extracted sentences contained a PV and some PVs were so infrequent that it was impossible to collect 20 sentences for each PV (see from Section 4.3.2). In the following section, the PVs from the dataset of (non-)literalness ratings are analysed. As an overview, the PVs were grouped according to the degree to which the average (non-)literalness rating for the sentences containing the same PV varied. The variation was calculated based on the difference between the maximum and minimum average degrees of (non-)literalness.

The first group of PVs are considered to have a very high degree of variation in the average (non-)literalness ratings. The difference between the maximum and minimum ratings for these 49 PVs was 3.33–5. The second group consisted of 22 PVs with a high degree of variation (2.33–3) in their average (non-)literalness ratings. Thirty-two PVs had a moderate degree of variation (1–2) and 25 PVs had a low degree of variation in their average (non-)literalness ratings (0.33–0.67). The sentences with the 56 PVs from the last group were rated using the same scores and there was no variation.

Figure 9 illustrates the average (non-)literalness ratings for the PVs with very high variations. These PVs were the most polysemous PVs in the dataset, and they all had multiple meanings with different degrees of (non-)literalness. There were seven PVs that formed part of sentences that were evaluated as having a score of 0 in some cases, and a score of 5 in others – these were *läbi lendama* 'to fly through/fail', *tagant tõukama* 'to push from behind/boost', *üles lööma* 'to dress up/toss (upward)', *üles võtma* 'to take something up/start something (song, conversation)/record', *vahele kukkuma* 'to fall between/get caught', *vahele pistma* 'to interlard a conversation with/stick between something' and *välja pistma* 'to stick out'. Hence, they had at least two usages, with one being fully literal and the other fully non-literal. One of these PVs – *läbi lendama* 'to fly through/fail' – was discussed earlier in this section (see examples (35)–(37)) – three different meanings were represented in five sentences. The annotators identified different meanings for the verb *lendama* 'to fly' with varying degrees of (non-)literalness. *Üles lööma* and *üles võtma* were assessed similarly – there were sentences that included a fully literal meaning and a fully non-literal meaning, and sentences that had meanings with differing degrees of (non-)literalness in between. *Vahele pistma* 'to interlard a conversation with/stick between something' and *välja pistma* 'to stick out' were also used in sentences with different degrees of (non-)literalness – most sentences containing *vahele pistma* were more non-literal than literal, and most sentences containing *välja pistma* had more literal than non-literal meanings. The sentences containing the PVs *tagant tõukama* 'to push from behind/boost' and *vahele kukkuma* 'to fall between/get caught' had two different meanings – most of the sentences were non-literal and one sentence was literal. Therefore, the non-literal meanings of these PVs were more frequent (and predominant) than were the literal ones.

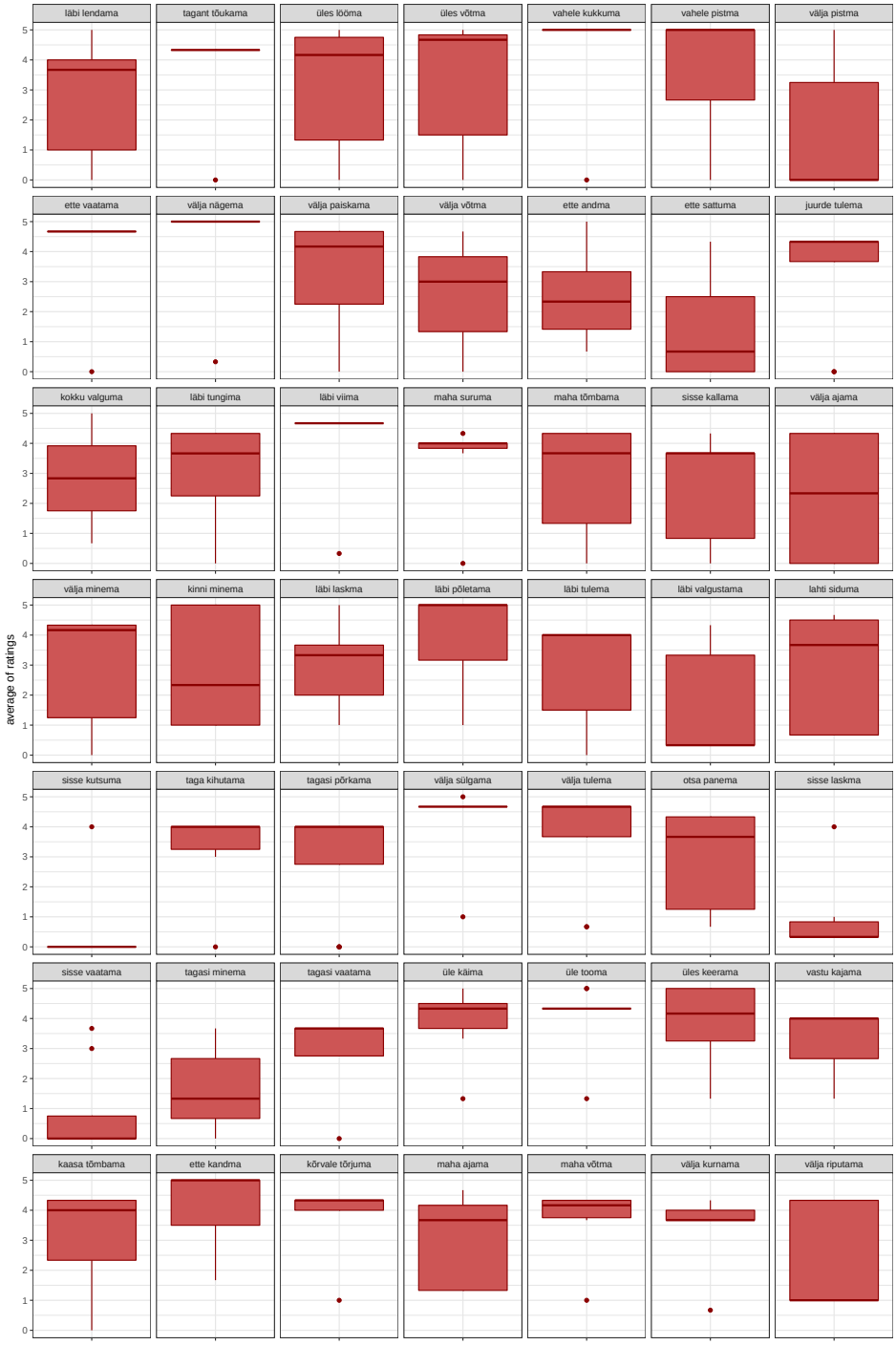

Figure 9: (Non-)literalness scores across PVs with very high variations in ratings.

The difference in the minimum and maximum ratings of four PVs – *ette vaatama* 'to foresee/look ahead', *välja nägema* 'to appear to your eyes/see outside', *välja paiskama* 'to throw something from somewhere/blurt out' and *välja võtma* 'to take out' – was 4.67. *Ette vaatama*[39] and *välja nägema*[40] both have two meanings – the more frequent one is non-literal and the infrequent one is literal. For example, *välja nägema* occurred in 14 sentences – 13 sentences were evaluated as having a fully non-literal meaning expressing *how something or someone appears to your eyes* and one sentence had the non-literal meaning 'to be able to see outside', with an average score of 0.33. These meanings corresponded to the meanings in the EED. According to the EED, *välja paiskama*[41] has two meanings – 'to throw something from somewhere' and 'to blurt out something'. The first meaning was assessed as being fully literal with a (non-)literalness rating of 0.0 (three sentences). The second meaning appeared in six sentences and had a (non-)literalness degree of 4.67. In addition, there was one sentence with a rating of 3.0 and two sentences with a rating of 3.67 – these were evaluated in a different way because of the abstractness of the subjects and objects. The most literal meaning of *välja võtma* is 'to take something out from somewhere' and the least literal is 'to tire out'. The degree of literalness of other meanings fits in between these two.

The difference between the minimum and maximum ratings was 4.33 for the following 11 PVs – *ette andma* 'to put something in front of somebody/feed/specify', *ette sattuma* 'to run across or meet somebody or something on the way', *juurde tulema* 'to approach/accrue', *kokku valguma* 'to join/melt together', *läbi tungima* 'to penetrate/go right through', *läbi viima* 'to conduct/pass through', *maha suruma* 'to suppress/bottle up/allay', *maha tõmbama* 'to cross off/out/pull down', *sisse kallama* 'to pour in/drink up', *välja ajama* 'to send off/out', and *välja minema* 'to go out'. Of these PVs, *läbi viima* has two meanings, which are also given in the EED[42] – 'to go through something' and 'to conduct'. *Maha suruma* appeared in 11 sentences that had four meanings; three of them had very similar degrees of (non-)literalness. The PV *juurde tulema* 'to approach/accrue' also has one meaning that is fully literal, while other two are non-literal and are probably very similar to each other. For other PVs, the scores were distributed more evenly. For example, *ette sattuma* appeared in seven sentences, half of which had ratings lower than 0.67. Twelve sentences containing the PV *maha tõmbama* were evaluated as having various degrees of (non-)literalness. The most literal meaning, 'to pull on the ground', had an average score of 0.0 and occurred in one sentence. The meaning 'to pull down' appeared in four sentences – when the object was a building, the average (non-)literalness degree was 1.67, and when the object was something else, it was 0.33. Three sentences with the meaning 'to reduce' received a rating of 3.67, and four sentences with the meaning 'to cross something out or delete' received a rating of 4.33. Half of the eight sentences containing *välja ajama* received an average score that was greater than 2.33, and there were sentences with a score of 0.0–4.33 amongst them.

[39] `http://www.eki.ee/dict/ekss/index.cgi?Q=ette+vaatama` (accessed 10.11.2018).

[40] `http://www.eki.ee/dict/ekss/index.cgi?Q=välja+nägema` (accessed 10.11.2018).

[41] `http://www.eki.ee/dict/ekss/index.cgi?Q=välja+paiskama` (accessed 10.11.2018).

[42] `http://www.eki.ee/dict/ekss/index.cgi?Q=läbi+viima` (accessed 10.11.2018).

Eleven PVs, namely *kinni minema* 'to close/go to prison', *läbi laskma* 'to let through/pretermit', *läbi põletama* 'to fuse something/to burn something out', *läbi tulema* 'to come through', *läbi valgustama* 'to x-ray/dissert', *lahti siduma* 'to untie/unbind', *sisse kutsuma* 'to invite in', *taga kihutama* 'to encourage/chase', *tagasi põrkama* 'to bounce back', *välja sülgama* 'to spit out' and *välja tulema* had a difference between the average minimum and maximum ratings for (non-)literalness of 4.0. While most of the PVs have multiple meanings according to the EED and these meanings were also represented in the dataset, there were some exceptions. For example, *sisse kutsuma* has one meaning in the EED[43] – 'to invite somebody in'. Most of the sentences containing this PV were also evaluated as being fully literal. Nevertheless, there was one sentence (see example (38)) in which the subject was inanimate; thus, the annotators determined the degree of the (non-)literalness of this meaning to be 4.0. The PV *lahti siduma*[44] also has one meaning – 'to unbind (tie or knot)' – and the sentences expressing 'unbinding a tie or knot or something similar' were evaluated as being literal. At the same time, there were sentences (see example (39)) that were assessed as being non-literal because the object of the PV was not a tie or knot but a currency – the sentences described currencies being 'tied' to each other.

(38) Tartu-∅ jaamahoone **ei** **kutsu-∅** mitte kuidagi **sisse**.
Tartu station building NEG invite-CONNEG not someway inside
Lit. 'Tartu station does not invite in in any way.'
'Tartu station does not look inviting.'

(39) Seetõttu suurene-ks surve **sidu-da** kroon saksa-∅
therefore increase-COND pressure bind-INF kroon Deutsche-GEN
marga-st **lahti**.
mark-ELA loose
'Therefore, the pressure to bind loose the Estonian kroon from the Deutsche mark would increase.'
'Therefore, the pressure to unpeg the Estonian kroon from the Deutsche mark would increase.'

PVs such as *taga kihutama* 'to encourage/chase', *tagasi põrkama* 'to bounce back' and *välja tulema* 'to get out (of)/turn up/come up with' had multiple meanings with similar degrees of (non-)literalness and one meaning that was much more literal. For example, the most literal meaning of *tagasi põrkama* appeared in three sentences, and the average (non-)literalness score was 0.0. The most non-literal meaning 'to be scared' received an average (non-)literalness rating of 4.0 and was the most common meaning (appearing in seven sentences). The third meaning 'to stand back' occurred in two sentences, and was considered to be slightly more literal.

The difference for the following 10 PVs – *otsa panema* 'to add', *sisse laskma* 'to let somebody in', *sisse vaatama* 'to look inside/visit something for a moment',

[43] `http://www.eki.ee/dict/ekss/index.cgi?Q=sisse+kutsuma` (accessed 10.11.2018).
[44] `http://www.eki.ee/dict/ekss/index.cgi?Q=lahti+siduma` (accessed 10.11.2018).

tagasi minema ‘to go back’, *tagasi vaatama* ‘to look back’, *üle käima* ‘to go/walk over’, *üle tooma* ‘to carry something/adapt/change the location of something’, *üles keerama* ‘to wind up/provoke’, *vastu kajama* and *kaasa tõmbama* ‘to persuade to join/pull along’ – was 3.76. For most of these PVs, the sentences had more non-literal than literal readings; exceptions were the PVs *sisse laskma* and *sisse vaatama*. Most of the sentences containing *sisse laskma* expressed the meaning ‘to let somebody/something in’, and the degrees of (non-)literalness varied based on the subjects, objects and adverbials. One sentence expressed the meaning ‘to try a new gun by shooting’, which was evaluated as being much more non-literal. The most common meaning of *sisse vaatama* is ‘to look inside something’; another meaning expresses ‘to visit something for a moment’. However, the annotators identified additional meanings for these PVs. For example, the annotators did not evaluate one sentence (see example (40)) as being similar to the sentences expressing the aforementioned meanings. The reason for this was most likely the fact that the annotators could not decide whether one could really visit an e-shop or look inside it. The meaning of one sentence (see example (41)) was not evaluated as being fully literal because the subject was an abstract entity.

(40) Soovi-des saa-da ülevaade-t tema-∅ teos-te-st, **vaata-si-n**
wish-GER get-INF overview-PRT s/he-GEN work-PL-ELA look-PST-1SG
sisse Interneti-∅ raamatupoodi-∅.
inside Internet-GEN bookstore-ILL
‘Wishing to get an overview of his work, I looked in the e-bookstore.’

(41) Vaesus **vaata-b** ukse-st ja akna-st **sisse**,
poverty-GEN watch-3SG door-ELA and window-ELA inside
keegi pea-b selle-∅ eest vastuta-ma.
somebody must-3SG this-GEN for be responsible for-SUP
Lit. ‘Poverty looks inside from the door and the window, somebody has to take responsibility for it.’
‘Poverty is everywhere and somebody has to take responsibility for it.’

In addition to these two PVs, *tagasi vaatama* ‘to look back’, *üle käima* ‘to go/walk over’ and *üle tooma* also have one or two dominating meanings (most of the sentences had similar ratings), and one or two meanings that occurred less often and had a significantly different rating from the majority. These meanings might or might not be differentiated by the EED. For example, the EED provides one meaning for the PV *üle tooma*[45], but the annotators identified three meanings. One sentence expressed the most literal meaning ‘to carry something’ ((non-)literalness score 1.33). The fully literal usage (average rating 5.0) of this PV was ‘to adapt a storyline’, and it appeared in two sentences. The remainder of the sentences had a meaning similar to the one in the EED – ‘to change the location where some activity is carried out, e.g. ‘to change the place of employment’.

The average scores for (non-)literalness were distributed more equally amongst sentences for PVs such as *otsa panema* ‘to add’, *tagasi minema* ‘to go back’, *üles*

[45] `http://www.eki.ee/dict/ekss/index.cgi?Q=üle+tooma` (accessed 10.11.2018).

keerama ‘to wind up/provoke’, *vastu kajama* and *kaasa tõmbama* ‘to persuade to join/pull along’. For example, *üles keerama*[46] has four meanings in the EED, but more in the dataset. The most literal was ‘to roll up’, which appeared in two sentences with a (non-)literalness degree of 1.33. The meaning of ‘to turn up (higher)’ was also rated as being more literal than non-literal, with an average (non-)literalness score of 2.0. The meanings ‘to get agitated’ and ‘to turn someone against someone else’ were considered to be fully literal, with a rating of 5.0. These meanings appeared in five sentences. The meaning ‘to wind up (the clock)’ appeared in two sentences, with an average (non-)literalness rating of 3.67. In addition, there was one sentence that received a (non-)literalness rating of 4.33 (see example (42)), and one sentence was evaluated as having a score of 4.0 (see example (43)). Therefore, the annotators indicated six degrees of (non-)literalness for the PV *üles keerama*.

(42) Asi **keera-b** end **üles**.
thing turn-3SG itself up
‘The thing winds itself up.’

(43) Akadeemik Kapitsa on leid-nud, et kuigi
academician Kapitasa be.3SG found-PST.PTCP that although
inimkonna-∅ kasvukõver kasva-b, tõuse-b ta ikka
mankind-GEN growth curve increase-3SG rise-3SG it still
kaldu, kuid ta võib-b ka äkki täiesti **üles**
inclined but it might-3SG also suddenly totally up
keera-ta ja siis on kriis küll käes.
turn-INF and then be.3SG crisis indeed due
‘Academician Kapitsa has found that although the growth curve of population increases, the rise of it is inclined, it can also suddenly turn completely straight up and then the crisis is indeed here.’

The PV *vastu kajama* also had more meanings in the dataset that are given in the EED[47]. While the EED only suggests one meaning for this PV, the annotators identified more. The EED meaning of *vastu kajama* is ‘to sound like an echo’. Four sentences expressing this meaning were evaluated as having an average (non-)literalness score of 1.33. The other 11 sentences were assessed as having a score of 4.0, and expressed relatively non-literal similar meanings ‘to get a response’ (as in example (44)) or ‘to appear’ (as in example (45)).

(44) USA-∅ kaitseministri-∅ sõna-d **on** **kaja-nud**
USA-GEN Minister of Defence-GEN word-PL be.3PL echo-PST.PTCP
vastu kogu maailma-s.
back entire world-INE
‘The words of the United States Secretary of Defence have echoed back all over the world.’

[46] http://www.eki.ee/dict/ekss/index.cgi?Q=üles+keerama (accessed 10.11.2018).
[47] http://www.eki.ee/dict/ekss/index.cgi?Q=vastu+kajama (accessed 10.11.2018).

(45) Ta **kaja-b** **vastu** sajandivahetuse-∅ kultuurilise-∅
s/he echo-3SG back turn of the century-GEN cultural-GEN
kuldajastu-∅ muusika-s, arhitektuuri-s, kirjanduse-s ja
golden age-GEN music-INE architecture-INE literature-INE and
maalikunsti-s.
painting-INE

Lit. 'She/he echoes back in the music, architecture, literature and painting of the cultural golden age during the turn of the century.'

'His influence can be found in the music, architecture, literature and painting of the cultural golden age during the turn of the century.'

Six PVs had a difference of 3.33 between the maximum and minimum ratings – *ette kandma* 'to report/to serve', *kõrvale tõrjuma* 'to displace/push aside', *maha ajama* 'to drive/push/shave off/remove', *maha võtma* 'to take down', *välja kurnama* 'to wear out/filter' and *välja riputama* 'to hang out'. The PVs *kõrvale tõrjuma*, *maha võtma* and *välja kurnama*, according to Figure 9, have at least three meanings. For example, while *kõrvale tõrjuma* has been assigned two meanings in the EED[48], the annotators identified three meanings – the meaning 'to ward off' is more literal than are the others, the least literal is 'to supplant', and the one that is slightly more literal than that is the meaning 'to knock out a team'. The EED grouped the second and third meanings into one. It is important to note that the most literal meanings appeared in only one sentence, while other meanings occurred numerous times.

The scores for the PVs *ette kandma* 'to report/to serve', *maha ajama* 'to drive/push/shave off/remove', *välja riputama* 'to hang out' were more evenly distributed amongst the sentences. This means that different meanings appeared equally often or they were similar to each other. For example, the PV *maha ajama* appeared in 15 sentences. The (non-)literalness ratings varied from 1.33 to 4.67. The EED[49] presents four meanings for this PV, two of which are slang terms. The first meaning is somewhat general and expresses 'to knock/push something downwards', in which the particle *maha* expresses the direction. This meaning was considered the most literal by the annotators, and six sentences expressing this meaning were evaluated with an average degree of 1.33. One sentence (see example (46)) expressed a similar meaning, but the annotators determined it to be slightly more non-literal (with a score of 2.0). The reason might have been that the bus driver's activity was not physical but verbal, and the annotators sensed a difference. The meaning remains literal at the binary level.

(46) Bussijuht **aja-s** sõitja-d bussi-st **maha** ja
Bus driver drive-PST.3SG passenger-PL bus-ELA off and
käski-s astu-ma haka-ta.
command-PST.3SG step-SUP start-INF

'The bus driver kicked the passengers out of the bus and ordered them to start walking.'

[48] `http://www.eki.ee/dict/ekss/index.cgi?Q=kõrvale+tõrjuma` (accessed 10.11.2018).
[49] `http://www.eki.ee/dict/ekss/index.cgi?Q=maha+ajama` (accessed 10.11.2018).

The second meaning of the PV *maha ajama* in the EED is ‘to remove something from somewhere’ and it was evaluated as having a score of 4.33 (four sentences). Nevertheless, for the three sentences that expressed the removal of a piece of clothing, the average was evaluated as 3.67. Hence, the annotators thought that ‘to remove a piece of clothing’ was a more literal meaning than was ‘to remove hair or a beard’. One sentence (see example (47)) expressed a meaning that the EED labelled as ‘characteristic of oral speech or slang’ – ‘to talk (and finish[ing] it)’, and the average (non-)literalness score for this meaning was 4.67.

(47) Mu-lle meeldi-b väikelinna-∅ lihtne elu, kus kõik
I-ALL like-3SG borough-GEN simple life where all
kõik-i tunne-vad ning hommikuti poe-s pika-d
all-PL.PRT know-3PL and in the mornings shop-INE long-PL
jutu-d **maha** **ae-ta-kse**.
talk-PL off drive-IMPS-PR
Lit. ‘I like the simple life of the small town, where everybody knows everybody and long conversations are driven off in the shop in the mornings.’
‘I like the simple life of the small town, where everybody knows everybody and long conversations are held in the shop in the mornings.’

Figure 10 illustrates the average (non-)literalness ratings across the PVs with high variations (2.33–3). The first eight PVs had a variation of 3.0 – they were *eemale tõukama* ‘to push away/scare off/repel’, *ette võtma* ‘to undertake/set out/embark upon’, *järele vaatama* ‘to watch someone/check or investigate’, *juurde lõikama* ‘to cut/add (land)’, *maha minema* ‘to get off’, *üle uhtuma* ‘to flush/wash’, *üles peksma* ‘to beat/wake up somebody’ and *vastu särama* ‘to shine/reflect’. According to the distribution of the average (non-)literalness ratings, it can be inferred that the PVs *juurde lõikama* and *üles peksma* had one meaning that was more frequent than others, one meaning that was fully non-literal and one that, compared to the other meanings, was relatively literal. Nonetheless, *juurde lõikama* had two meanings that were evaluated as having an average score of 4.0 – ‘to add land to somebody or for something’ and ‘to cut pieces of cloth for sewing’. The fully non-literal meaning was ‘to gain something’ and the most literal was ‘to add something by cutting’. The most frequent meaning of *üles peksma* was ‘to beat up’, the fully non-literal was ‘to wake up somebody’ and the most literal was ‘to beat upwards’. Hence, the distribution of the average (non-)literalness scores may imply how many different meanings the PV had, but this is not a rule. The most difficult task in the automatic detection of (non-)literalness is distinguishing among the meanings of the same PVs with similar degrees of (non-)literalness.

The scores were more equally distributed for other PVs. For example, the PV *ette võtma* ‘to undertake/set out/embark upon’ was a constituent of 12 sentences. Six sentences expressed the meaning ‘to embark upon something’ with a (non-)literalness rating of 4.33, one sentence expressed the meaning ‘to take charge up front’ with a rating of 3.33, and three sentences expressed the meaning ‘to lift something in front’. Two sentences with (non-)literalness scores of 3.0 (see example (48)) and 3.67 (see example (49)) could express the meanings ‘to embark upon something’ or ‘to lift something in front’. As the annotators interpreted

Figure 10: (Non-)literalness scores across the PVs with high variations in the ratings.

the sentences based on their own opinions, it is not possible to verify the exact meaning that they evaluated.

(48) Supp ja praad on söö-dud, nüüd **võta-me**
soup and main course be.3PL eat-PST.PTCP now take-1PL
ette magustoidu-∅.
ahead dessert-GEN

Lit. 'The soup and the main course have been eaten, now we take ahead the dessert.'

'The soup and the main course have been eaten, now we eat dessert.'

(49) Nüüd tule-b **võt-ta** **ette** kuuma-d kartuli-d, lisa-da
Now must-3SG take-INF ahead hot-PL potato-PL, add-INF
ne-i-le tripsuke võid ja kast-a kartuliampsu-∅
they-PL-ALL a little butter-PRT and dip-INF mouthful of potato-PRT
enne suhupanemis-t sibulasoolvee-∅ sisse
before putting in mouth-PRT onion brine-GEN inside

Lit. 'Now you must take ahead hot potatoes, add a little butter to them and dip a mouthful of potato into the onion brine before putting in mouth.'

'Now take hot potatoes, add a little butter to them and dip a mouthful of potato into the onion brine before eating.'

Even though the average scores were distributed evenly, not all of the meanings of the PV were represented. For example, the PV *järele vaatama* has three meanings in the EED. The first one, 'to watch someone leaving/passing/going' was included in five sentences, and was evaluated as having a (non-)literalness score of 1.0. The second meaning 'to check or investigate' was assessed as having a score of 4.0, and it appeared in seven sentences. The third meaning (which is labelled as infrequent in the EED), 'to look after' was not present in the dataset.

The difference between the minimum and maximum (non-)literalness ratings was 2.67 for the following nine PVs – *külge jääma* 'to stick/get used', *läbi vaatama* 'to look through/examine', *maha käima* 'to go down/run down/go/degenerate', *mööda käima* 'to bypass', *üles soojenema* 'to warm up', *ümber tõmbama* 'to put something around somebody or something/encircle', *välja ilmuma* 'to debouch/emerge/appear unexpectedly', *vastu põrutama* 'to snap back at somebody' and *vastu rääkima* 'to talk back/dispute'. Most of the meanings in the dataset tended to be more non-literal than literal. For example, *läbi vaatama* was assigned three different average scores. The most literal meaning was 'to watch through someone' (one sentence, with an average score of 1.0). The second meaning was to 'to examine luggage' with an average (non-)literalness score of 2.0 (two sentences). The rest of the sentences (9) received a score of 3.67, and the meaning was 'to look through' – the objects could be animate (such as a patient) or inanimate (such as a document).

The most frequent meanings were not always non-literal. For example, the EED provides two meanings for *mööda käima*[50] – 'to move by' and 'to head by'.

[50] `http://www.eki.ee/dict/ekss/index.cgi?Q=mööda+käima` (accessed 10.11.2018).

The subject of the first meaning is something that moves (for example, a human or a car), while the subject of the second meaning is usually something that does not move itself (such as a road or a bus line). Both meanings were evaluated as being more literal than non-literal, while one sentence (see example (50)) was assigned a (non-)literalness degree of 3.67. The meaning of the PV in the sentence is figurative, and could be interpreted as 'to ignore'.

(50) See arengutee on üks ne-i-st rikkus-te-st, mille-st
this development path be-3SG one they-PL-ELA wealth-PL-ELA what-ELA
innuka-d progressi-∅ eest võitleja-d tuimalt **mööda käi-vad**
eager-PL progress-GEN for fighter-PL rigidly by walk-3PL
Lit. 'This development pathway is one of those wealth that the eager fighters for the progress rigidly walk by.'
'This development pathway is one of those benefits that the eager fighters for the progress rigidly ignore.'

The variation in the average (non-)literalness scores was 2.33 for five PVs – *edasi jõudma* 'to get ahead/come out on top', *ette valmistama* 'to prepare', *juurde tõmbama* 'to engage', *kaasa tooma* 'to bring something or someone/cause something' and *üle kaaluma* 'to weigh again'. *Edasi jõudma* has two meanings in the EED[51] – 'to move on (spatially or in time)' and 'to progress'. Both meanings appeared in the dataset, with the second meaning being more frequent. In addition, the PV *juurde tõmbama* 'to engage' has one meaning in the EED[52], but the annotators identified three meanings. Based on the nature of the activity, the meaning of 'pulling' was evaluated as being more literal than was the meaning of 'engaging'. The third meaning of the PV was not clear, but it seemed to be influenced by the unusual animacy of the subject and object. While the distributions of the average (non-)literalness scores for the PVs *ette valmistama* 'to prepare' and *kaasa tooma* 'to bring something or someone/cause something' reflected the multiple meanings of these PVs (these are also suggested in the EED), the high variation for *üle kaaluma* 'to weigh again' might seem surprising. In fact, the EED[53] gives only one meaning for this PV – 'to scale something again (to check for something)'. This meaning was assessed as having a degree of 2.0 for two sentences. A score of 4.33 was assigned to evaluate the sentence that expressed the meaning 'to reconsider something'. As one of the meanings of *kaaluma* is 'to consider', this interpretation is not surprising.

Figure 11 shows 32 PVs with moderate variations in the average (non-)literalness ratings for the relevant sentences. The difference between the minimum and maximum (non-)literalness scores for these PVs was 1–2.

The PVs *ette lugema* 'to read out/recite', *ette tegema* 'to do beforehand', *välja jääma* 'to stay out', *välja pääsema* and *vastu kostma* 'to reply' were rated as having a difference in the maximum and minimum ratings of 2.0. All the meanings of the PVs *ette lugema* and *ette tegema* were considered to be more

[51] `http://www.eki.ee/dict/ekss/index.cgi?Q=edasi+jõudma` (accessed 10.11.2018).
[52] `http://www.eki.ee/dict/ekss/index.cgi?Q=juurde+tõmbama` (accessed 10.11.2018).
[53] `http://www.eki.ee/dict/ekss/index.cgi?Q=üle+kaaluma` (accessed 10.11.2018).

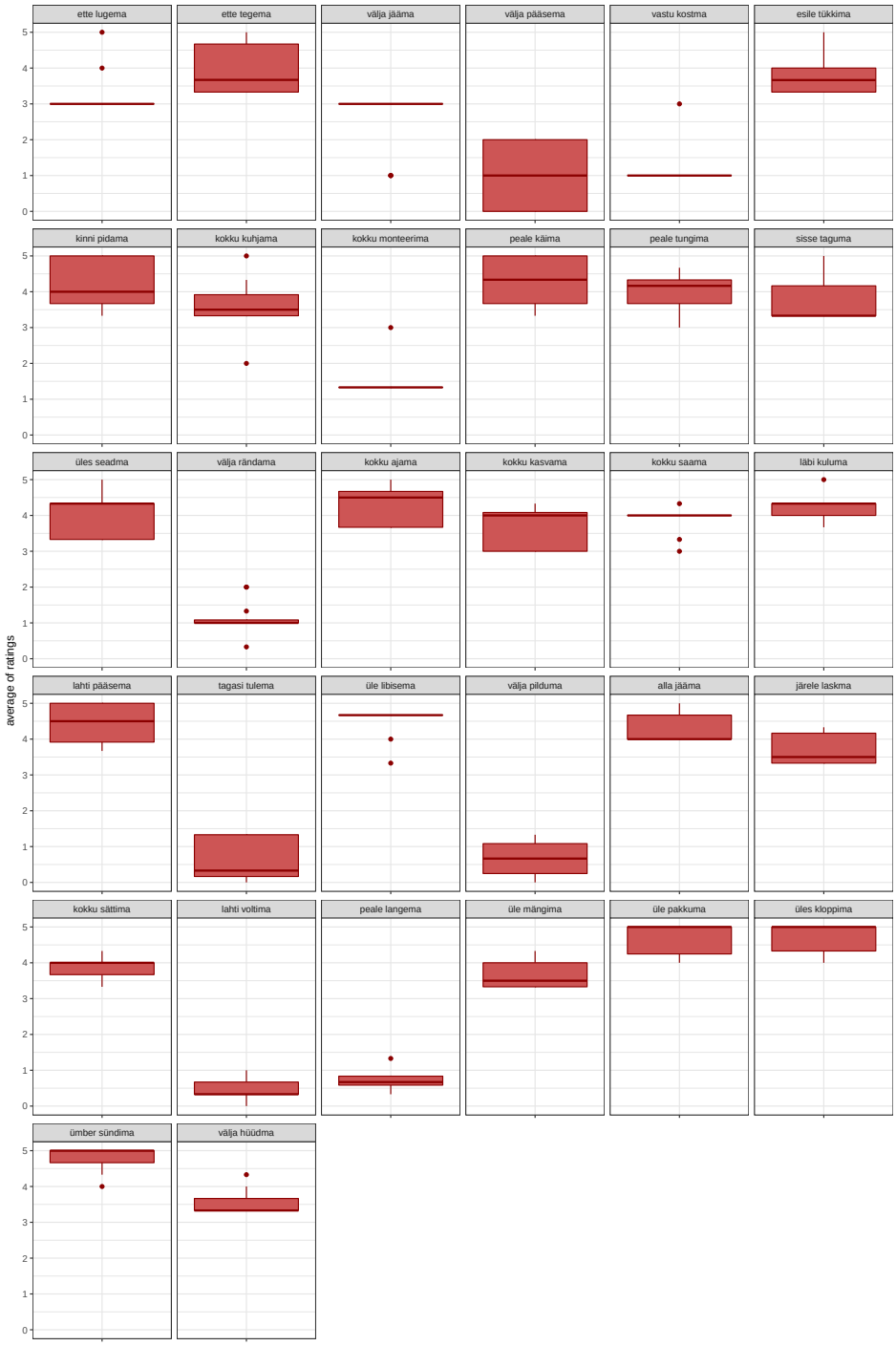

Figure 11: (Non-)literalness scores across PVs with moderate variations in their ratings.

non-literal than literal. Both PVs have multiple meanings in the EED[54,55] and in the dataset. The average scores for literalness reflect that the meanings had different degrees of compositionality. Although *välja jääma* has one meaning in the EED[56] – 'to leave out' – the annotators did not evaluate all 13 sentences in the same way. Most of these sentences (10) received an average (non-)literalness score of 3.0, and the more specific meaning of the PV in these sentences was 'to be excluded' (as in example (51)). The remaining three sentences received an average (non-)literalness rating of 1.0, and the meaning of the PV in these sentences was 'to stay outside' (as in example (52)).

(51) On kuul-da nurina-t, et mõne-d olulise-d
be.3SG hear-INF grumble-PRT that some-PL important-PL
nime-d on **välja** **jää-nud**
name-PL be.3PL out stay-PST.PTCP.
'There has been a complaint that some important names have been left out.'

(52) Kraaniosa ja toru pann-a-kse karpraua-∅
part of the tap and pipe put-IMPS-PRS channel iron-GEN
sisse, **välja** **jää-b** ainult nupp.
inside out stay-3SG only button
Lit. 'The tap and pipe will be put inside the channel iron, only the button will stay out.'
'The tap and pipe will be put inside the channel iron, only the button will be outside.'

Unlike the PVs in the same group, most of the meanings of the PVs *välja pääsema* and *vastu kostma* 'to reply' were literal. For example, the meaning of *välja pääsema* in the EED[57] is 'to get out of somewhere'. The annotators identified different degrees of (non-)literalness for this PV based on the meaning of the verb, but the meaning of the PV was the same in general.

The difference was 1.67 for the following nine PVs – *esile tükkima* 'to jut/dominate', *kinni pidama* 'to stick to/slow down/detain', *kokku kuhjama* 'to stack up', *kokku monteerima* 'to assemble/edit video', *peale käima*, *peale tungima*, *sisse taguma* 'to beat in(to)', *üles seadma* and *välja rändama*. The most frequent meanings of the PVs *kokku monteerima* and *välja rändama* were literal. Both PVs have one meaning in the EED[58,59] and all the sentences expressed these meanings. The variation was created by the use of figurative language (for example, 'cars are emigrating' versus 'eels are emigrating').

Based on the average (non-)literalness ratings, it is possible that the PVs *esile tükkima* 'to jut/dominate', *kinni pidama* 'to stick to/slow down/detain', *kokku*

[54] http://www.eki.ee/dict/ekss/index.cgi?Q=ette+lugema (accessed 10.11.2018).
[55] http://www.eki.ee/dict/ekss/index.cgi?Q=ette+tegema (accessed 10.11.2018).
[56] http://www.eki.ee/dict/ekss/index.cgi?Q=välja+jääma (accessed 10.11.2018).
[57] http://www.eki.ee/dict/ekss/index.cgi?Q=välja+pääsema (accessed 10.11.2018).
[58] http://www.eki.ee/dict/ekss/index.cgi?Q=kokku+monteerima (accessed 10.11.2018).
[59] http://www.eki.ee/dict/ekss/index.cgi?Q=välja+rändama (accessed 10.11.2018).

kuhjama 'to stack up', *peale käima*, *peale tungima*, *sisse taguma* 'to beat in(to)' and *üles seadma* had two or more non-literal meanings. All these PVs, except for *kokku kuhjama*, have more than one meaning in the EED. For example, all three meanings that the EED suggests for *üles seadma*[60] were present in the dataset. According to the annotators, the fully non-literal meaning (with an average (non-)literalness score of 5.0) was 'to pose (something)' (two sentences). The meaning 'to nominate a candidate' (four sentences) was considered to be slightly more literal, with an average (non-)literalness rating of 4.33. The meaning 'to install/set up something' (with an average score of 3.33) appeared in six sentences. One sentence (see example (53)) expressed the same meaning, but was assigned a score of 4.33. The difference here might have been due to the fact that the objects and the place for the activity were unusual, and the annotators indicated this in their ratings. The compositionality degrees of (non-)literalness varied slightly for *kokku kuhjama* 'to stack up', depending on the abstractness of the object (for example 'material' versus 'a pile of dollars').

The difference between the minimum and maximum average (non-)literalness ratings for the sentences with the following eight PVs – *kokku ajama* 'to herd together/gather', *kokku kasvama* 'to grow together', *kokku saama* 'to meet', *läbi kuluma* 'to wear out', *lahti pääsema* 'to break loose/get free', *tagasi tulema* 'to come back', *üle libisema* 'to gloss/pass over' and *välja pilduma* 'to throw something out of somewhere' – was 1.33. While the average (non-)literalness scores were distributed equally amongst the sentences for PVs such as *kokku ajama*, *kokku kasvama*, *lahti pääsema*, *tagasi tulema* and *välja pilduma*, there were some PVs that had one frequent meaning and few infrequent meanings with different degrees of (non-)literalness. For example, the most frequent meaning of *kokku saama* – 'to meet' – received a (non-)literalness rating of 4.0 and appeared in six sentences. The similar meaning also appeared in one more sentence (see example (54)), but one annotator identified it as being more non-literal than the others because it expressed the possibility that disabled people could use computers. Another meaning – 'to gather' – occurred in four sentences (with an average (non-)literalness rating of 4.0) and 'to get dirty' in one sentence (with a (non-)literalness rate of 3.33).

(53) Esmajärjekorra-s on kava-s internetti-∅ **üles** **sea-da**
first priority-INE be-3SG plan-INE Internet-ILL up set-INF
sisukorra-d ja numbri-te kokkuvõtte-d.
table of contents-PL and issue-PL.GEN summary-PL
Lit. 'The first priority is to set up the tables of contents and summaries of the issues to the web.'
'The first priority is to upload the tables of contents and summaries of the issues to the web.'

(54) Keskuse-s **saa-vad** **kokku** puude-ga inimese-d ja arvuti-d.
centre-INE get-3PL together disability-COM human-PL and computer-PL
Lit. 'Disabled people and computers get together in the centre.'
'Disabled people can learn to use computers in the centre.'

[60] `http://www.eki.ee/dict/ekss/index.cgi?Q=üles+seadma` (accessed 10.11.2018).

The PV *üle libisema* occurred 13 times and mainly expressed the same meaning given in the EED[61] – 'to pass over (not to delve into something)'. The annotators assigned the 11 sentences containing this PV a (non-)literalness degree of 4.67. The slightly more literal meaning of *üle libisema* received a (non-)literalness rating of 4.0 (see example (55)), and the meaning as expressed in example (56) obtained a (non-)literalness rating of 3.33.

(55) Hollandlase-∅ Pier Sienema-∅ loo-st ükskõikse-∅ eestlase-∅
Dutchman-GEN Pier Sienema-GEN story-ELA indifferent-GEN Estonian-GEN
silma-d hommikusöögilaua-s aga **üle ei libise**-∅.
eye-PL breakfast table-INE however over NEG slip-CONNEG
'The eyes of indifferent Estonian at the breakfast table do not slip over the article about the Dutchman Pier Sienema.'

(56) Mõne-d pea-d lõika-d ära, aga ega ka nende-lt, kelle-st vikat
some-PL head-PL cut-2SG off but nor also they-PL.ABL who-ELA scythe
üle libise-s, erilist-∅ koostöö-d loo-ta ole-∅.
over slide-PST.3SG particular cooperation-PRT hope-INF be-CONNEG
'Some heads you cut off, but you cannot expect much cooperation from the ones the scythe slid over.'

All meanings of the PVs *tagasi tulema* 'to come back' and *välja pilduma* were considered to be somewhat more literal than non-literal. While the EED provides multiple meanings for *tagasi tulema*[62], there is only one meaning for *välja pilduma*[63] – 'to throw something out of somewhere'. However, the annotators noticed a difference in the meanings of this PV. For example, the average compositionality score for the sentence in example (57) was 0, but the average score for the sentence in example (58) was 1.33. The difference of scores appeared because the sentence in example (58) did not specify the location from which the trolling spoon was thrown.

(57) Kui hakka-d autoakna-st rämpsu-∅ **välja pildu-ma**, siis
when start-2SG car window-ELA garbage-PRT out throw-SUP then
kujuta-∅ ette, mis tunne on kraavi-st sodi-∅ korja-ta
imagine-IMP what feeling be.3SG trench-ELA trash-PART pick up-INF
'When you start throwing garbage out of the car window, imagine the feeling of picking it up from the trench.'

(58) Koshura **pillu-b** käe-ga landi-∅ **välja** ja
Koshura throw-3SG hand-COM trolling spoon-GEN out and
keri-b tamiili-∅ kohvipurgi-le tagasi
scroll-3SG fishing line-GEN coffee can-ALL back
'Koshura throws the trolling spoon with his hand and scrolls the fishing line back to the coffee can.'

[61] http://www.eki.ee/dict/ekss/index.cgi?Q=üle+libisema (accessed 10.11.2018).
[62] http://www.eki.ee/dict/ekss/index.cgi?Q=tagasi+tulema (accessed 10.11.2018).
[63] http://www.eki.ee/dict/ekss/index.cgi?Q=välja+pilduma (accessed 10.11.2018).

The difference for the sentences containing the PVs *alla jääma* 'to be run over/be conquered/loose', *järele laskma* 'to loosen', *kokku sättima* 'to set/put together', *lahti voltima* 'to unfold/ unwrap', *peale langema* 'to attack', *üle mängima* 'to overplay/outplay', *üle pakkuma* 'to exaggerate', *üles kloppima* 'to fix/beat/fluff', *ümber sündima* 'to be reborn' and *välja hüüdma* 'to cry/shout out' was 1.0. Some PVs have only one meaning in the EED (such as *alla jääma*, *järele laskma*, *kokku sättima*, *lahti voltima*, *peale langema*, *ümber sündima*), but the annotators sensed slightly different degrees of (non-)literalness in their meanings. For example, the meaning of *lahti voltima*[64] is literally 'to unfold something'. The annotators interpreted it slightly differently if the sentences described 'unfolding a letter' or 'unwrapping the battery'. PVs such as *üle mängima*, *üle pakkuma*, *üles kloppima* and *välja hüüdma* have multiple meanings suggested in the EED ('to cry/shout out', 'to announce'), and at least two of them were represented in the dataset. For these PVs, the annotators suggested more meanings than were presented in the EED. For example, the meanings of *üle mängima*[65] are 'to overplay' and 'to outplay'. Although the sentences containing the PV *üle mängima* expressed these meanings in general, the annotators specified more narrow meanings within these two (broad) meanings.

Twenty-five PVs with low variation are illustrated in Figure 12. Most of the PVs had one most frequent meaning and one or two meanings that occurred less frequently. For example, the EED provides two meanings for the PV *üle pingutama*[66] – 'to overreach' and 'to overdo'. The annotators evaluated the first meaning as having a score of 3.33 (three sentences) and second as having a score of 4.0 (ten sentences).

Üle andma appeared in 13 sentences that expressed the meaning given in the EED[67] – 'to give/hand over something to someone'. The slight difference in the degrees of (non-)literalness was due to the object – the meaning of the PV is more literal if the object is more concrete, such as 'gift' or 'picture' versus 'investigation'. Based on the average (non-)literalness scores, the annotators noted at least three to seven meanings of *vastu võtma* 'to accept/welcome/admit' presented in the EED[68]. All the meanings of *vastu võtma* were evaluated as being more non-literal than literal.

In some cases, the average scores for (non-)literalness were distributed more equally among sentences containing the PV. For example, the PV *vastu hakkama* appeared in 16 sentences. Two meanings of this PV are presented in the EED – 'to resist' and 'to detest'. The annotators have clearly differentiated between these two meanings and evaluated the first one as having an average (non-)literalness of 3.67 (eleven sentences) and the second as having a score of 4.33 (five sentences). Based on the distribution of the meanings across sentences, it may be the case that the meaning 'to resist' is more common. However, the automatic discrimination between these two meanings is most likely an extremely challenging task. The PV *üle võtma* has two meanings – 'to take over the possession of something'

[64] `http://www.eki.ee/dict/ekss/index.cgi?Q=lahti+voltima` (accessed 10.11.2018).
[65] `http://www.eki.ee/dict/ekss/index.cgi?Q=üle+mängima` (accessed 10.11.2018).
[66] `http://www.eki.ee/dict/ekss/index.cgi?Q=üle+pingutama` (accessed 10.11.2018).
[67] `http://www.eki.ee/dict/ekss/index.cgi?Q=üle+andma` (accessed 10.11.2018).
[68] `http://www.eki.ee/dict/ekss/index.cgi?Q=vastu+võtma` (accessed 10.11.2018).

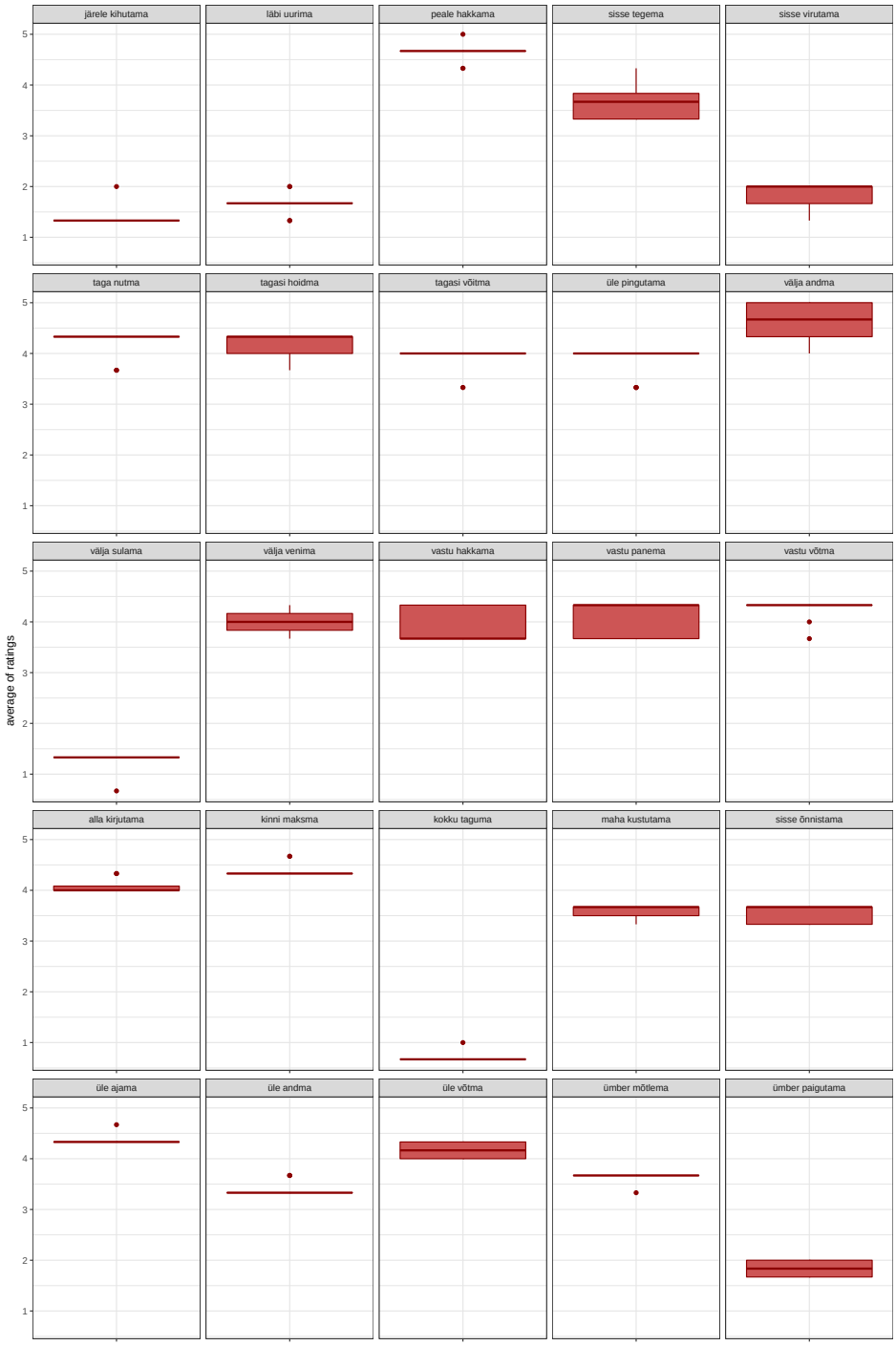

Figure 12: (Non-)literalness scores across the PVs with low variations in the ratings.

and ‘to take over something by imitating it’. The (non-)literalness scores of these meanings were 4.0 and 4.33, respectively (both were represented by six sentences).

These examples suggest that there might be a slight difference in the degrees of (non-)literalness of the different meanings of the same PV. These differences might not play a role in communication amongst humans who speak the same language. Moreover, when the meanings are translated in the same way into another language, the degrees of (non-)literalness do not make a difference. In some cases, different translations are needed and the degrees of (non-)literalness might be useful.

Figure 13 shows 56 PVs that appeared in sentences that the annotators evaluated as having the same average (non-)literalness score. It is important to note that the number of sentences for each PV was not equal and that the order of the PVs reflects the number of sentences. For example, the PVs *üles kasvama* ‘to grow up’ and *välja kuulutama* ‘to announce’ appeared in 13 sentences, while *välja kostma* ‘to be heard’ and *välja saagima* ‘to saw out’ only appeared in one sentence.

Only one PV was considered to be fully literal – all five sentences containing *läbi sadama* ‘to leak when raining’ received a (non-)literalness evaluation of 0.0. Nevertheless, as mentioned in Section 4.3.3, some of the sentences containing this PV were omitted because the annotators did not agree. The meanings of the PVs *üles lugema* ‘to list’, *ette nägema* ‘to foresee/stipulate/see ahead’, *üle trumpama* ‘to have the best of/circumvent’, *alla laadima* ‘to download’, *ette heitma* ‘to reproach/blame’, *alla käima*, *kaasa sündima*, *alt minema* ‘to fail or to be deceived’, *kokku kõlksuma* ‘to match’, *välja suitsetama*, *vastu raiuma* ‘to object/to dispute’, *maha jahtuma* ‘to cool down’, *kokku kiskuma*, *ära tegema* ‘to win somebody/finish doing something’ and *välja saagima* ‘to saw out’ were evaluated as being fully non-literal. While most of these PVs had at least one (non-literal) meaning in the EED, *välja saagima* had one literal meaning. However, example (59) shows that the sentence containing this PV was fully non-literal. Other sentences that expressed more literal meanings caused disagreement among annotators and were omitted from the dataset.

(59)

Vahetevahel	**sae-b**	Jaan	Alavere	viiuli-st	**välja**	mõne-∅
sometimes	saw-3SG	Jaan	Alavere	violin-ELA	out	some-GEN
idamaise-∅	soolo-∅.					
Eastern-GEN	solo-GEN					

Lit. ‘Jaan Alavere sometimes saws an Eastern solo out from the violin.’
‘Jaan Alavere sometimes plays an Eastern solo on the violin.’

It might be suggested that the PVs in Figure 13 have only one meaning, but this is not necessarily so. Some of the meanings of these PVs led to disagreement amongst the annotators, or were not selected for the evaluation at all (see Section 4.3.3). However, in some cases, single meaning of the PV given in the EED was what annotators actually evaluated. For example, the sole meaning of the PV *üles kasvama* that annotators assessed as having a (non-)literalness score of 2.0 for all 13 sentences corresponded to the meaning in the EED – ‘to grow up’. The PV *välja kuulutama* appeared in 13 sentences and has one meaning ‘to announce (and sometimes establish) something’ in the EED – it received an average

Figure 13: (Non-)literalness scores across PVs with no variations in their ratings.

(non-)literalness rating of 3.33 from the annotators. The meaning of *üles lugema* is 'to list', and all 12 sentences expressing this meaning were evaluated as having a (non-)literalness score of 5.0. The sole meaning of the PV *välja heitma* is 'to expel', and all sentences containing this PV received an average (non-)literalness rating of 4.0.

Some PVs are highly polysemous (and frequent), but only one of their meanings was represented in the dataset. For example, *ette nägema* was the third most frequent PV in the dataset, and it has four meanings in the EED[69]. Nonetheless, all 11 sentences containing this PV were assigned the same average score (5.0). When considering these sentences, at least two different meanings were present – 'to anticipate' and 'to designate', which are both clearly non-literal. This is a good example of how a PV can have several meanings that have the same degree of (non-)literalness. Differentiating amongst meanings of the same PV that share their degrees of compositionality definitely poses an additional challenge for the NLP models.

In general, the PV usages in the dataset could be divided as follows: a) PVs that have only one meaning (although they are not necessarily monosemous), b) PVs with several meanings that have varying degrees of compositionality, and c) PVs that have multiple meanings with the same, or very similar degrees of compositionality.

In summary, most of the PVs in this dataset had several meanings with varying degrees of compositionality. This finding illustrates that the traditional descriptions of Estonian have presented somewhat generic divisions of Estonian PVs (described in Section 2.4) based on the authors' personal opinions. Rätsep (1978), for example, used himself as informant. At the same time, Estonian is lacking a systematic analysis of polysemy and of the idiomaticity of Estonian PVs whereby a thorough overview of the semantics of PVs could be presented. Nevertheless, the current corpus-based study emphasises what was proposed previously by Muischnek (2006: 12) and Veismann and Sahkai (2016: 272), namely that semantic compositionality is a scalar property of PVs (and other MWEs).

It was shown earlier that there is no (statistically significant) correlation between the frequency of the PV and its components and the degree of compositionality assessed by human annotators (see Section 4.2.5). In order to determine whether frequency had an impact on the assessments of the literalness of the PVs, the effect of frequency on the compositionality judgements was revisited. Table 3 provides information about how PV, verb and adverb frequencies correlated with the (non-)literalness ratings across different frequency classes of the PVs. The correlations are expressed by the ρ values. The table presents also p-values to state the statistical significance of each correlation.

The most frequent PV occurred 35,929 times and the least frequent sixteen times[70]. PVs that occurred more than 1,000 times were considered to be frequent, and PVs that occurred less than 100 times were considered to be infrequent.

It is clear that, regardless of the frequency class of the PV, the (non-)literalness rating correlated somewhat weakly with the frequency of the PV and its components. Therefore, the frequency of the PV, adverb and particle did not influence the

[69] `http://www.eki.ee/dict/ekss/index.cgi?Q=ette+nägema` (accessed 10.11.2018).

[70] The frequencies were calculated based on the newspaper subcorpora of ERC.

Table 3: Effect of frequency on the (non-)literalness ratings. ρ = Spearman's rank correlation coefficient, p = corresponding p-value.

set of PVs	PV		adverb		verb	
	ρ	p	ρ	p	ρ	p
all	0.07	0.10	0.00	0.96	0.06	0.19
without frequent	0.08	0.15	0.04	0.53	0.08	0.16
without infrequent	0.07	0.16	−0.02	0.65	0.04	0.45
without frequent and infrequent	0.10	0.17	0.02	0.75	0.07	0.33
frequent	0.09	0.23	−0.09	0.20	−0.04	0.62
infrequent	0.01	0.93	0.07	0.53	0.01	0.34

literalness ratings of the PVs. However, as the correlations were not statistically significant at the $p = 0.05$ level or based on (non-)literalness ratings, it cannot be claimed that frequency does or does not influence the compositionality ratings of Estonian PVs in general.

4.3.5 Summary of the literalness ratings

The literalness ratings for Estonian PVs were initially necessary in order to develop a classifier to predict the literal versus the non-literal usage of Estonian PVs. The averaged literalness ratings were applied to determine the values of the target class of the classifier suggested in Chapter 6.

Three annotators with linguistic backgrounds evaluated 1,838 sentences on a six-point scale. The evaluation of the annotations showed that the ratings for 1,481 sentences could be used for further research because they did not cause substantial disagreement amongst the annotators. The analysis of the excluded sentences suggested that the main reasons for disagreement were the polysemy of the components, insufficient context being provided within the sentences, and the subjectivity of the task.

The overview of the PVs suggested that 144 evaluated PVs had several meanings, often with different degrees of compositionality. This finding has important implications because it suggests that, instead of the binary division of PVs, the semantic compositionality should be treated as a scalar property of the PVs in order to provide a comprehensive description of the formation of the meaning of Estonian PVs. In addition, no statistically significant correlation between human judgement and frequency was found.

4.4 Comparison of the compositionality and literalness ratings

This section compares the compositionality ratings (see Section 4.2) and the (non-)literalness ratings (see Section 4.3). The datasets were intended to collect annotations about the degree of compositionality of Estonian PVs. The fundamental differences in the datasets are discussed in the following section.

One of the main differences between the datasets was that the compositionality dataset had one score per PV regardless of the number of meanings of the PV, while the (non-)literalness ratings took the polysemy of the PVs into account. Accordingly, the (non-)literalness ratings were assigned based on the context (the sentences) in which the PVs appeared. Thus, the particular meaning of the PV that was evaluated was generally clear. However, the compositionality scores did not indicate the particular meaning of the PV, but provided a general score for the PV. However, it was assumed that the annotators assessed the predominant (frequent) meanings of the PVs. Overall, the differences resulted in a situation in which some PVs had a different number of scores. For example, the PV *kinni minema* 'to close/go to prison' had one compositionality score, but three (non-)literalness ratings.

The second difference was the scale of measurement according to which the degree of compositionality was evaluated. The compositionality ratings were assigned on a scale from 1–5 (from fully non-compositional to fully compositional), but the (non-)literalness ratings were assessed using a scale from 0–5 (from fully compositional to fully non-compositional).

The third difference was the way in which the annotations were collected. The compositionality ratings were crowdsourced, and the aim was to obtain as many evaluations as possible. The backgrounds of the annotators are not known. The (non-)literalness scores were assigned by three annotators who all had backgrounds in linguistics. Furthermore, the annotators had more time to understand the task and could ask questions during the annotation process. They also had the option of changing their judgements before submitting their scores.

The differences were mainly due to the different aims of the datasets. The compositionality rating was created to evaluate the DSM models that provided vector representations of words. It was thus crucial to collect one rating per word, and the use of an even-numbered scale was not required. The literalness scores were collected for the binary classification of the literal versus the non-literal usage of PVs. In this regard, an even-numbered scale proved useful. In addition, it was necessary to identify the different meanings of PVs, as the classification of the PVs was context-based. In summary, the datasets were distinct because the ratings were created for different purposes (see Sections 4.2.1 and 4.3.1).

Distinctions among datasets imply that there might be differences in the ratings of the degrees of compositionality of the PVs. In order to determine the range of differences, the degree of compositionality of the same PVs in the two datasets was compared by examining the PVs that received similar assessments in both datasets. An overview of these PVs is presented in Table 4. PVs with a high degree of compositionality had a (non-)literalness rating of 0–1 and a compositionality rating of 4–5; a moderate degree of compositionality is reflected in a (non-)literalness

rating of 2–3 and a compositionality rating of 3. PVs with low degrees of compositionality received a (non-)literalness rating of 4–5 and a compositionality rating of 1–2.

Table 4: PVs with similar compositionality degrees in both datasets.

PVs with high compositionality degree
edasi jõudma 'to get ahead/come out on top'; *eemale tõukama* 'to push away/scare off/repel'; *ette andma* 'to put something in front of somebody/feed/specify'; *ette jõudma* 'to get ahead/outstrip/outdistance'; *ette sattuma* 'to run across or meet somebody or something on the way'; *juurde tulema* 'to approach/accrue'; *järele kihutama* 'to chase'; *järele vahtima* 'to stare after somebody'; *kokku monteerima* 'to assemble/edit video'; *kokku taguma* 'to hit something together/put something together in a hurry'; *kokku valguma* 'to join/melt together'; *lahti siduma* 'to untie/unbind'; *lahti voltima* 'to unfold/ unwrap'; *läbi laskma* 'to let through/pretermit'; *läbi tuhnima* 'to ransack/scour'; *läbi tungima* 'to penetrate/go right through'; *läbi uurima* 'to explore/go through something'; *läbi vaatama* 'to look through/examine'; *maha minema* 'to get off'; *maha põlema* 'to burn down'; *maha suruma* 'to suppress/bottle up/allay'; *maha tõmbama* 'to cross off/out/pull down'; *mööda käima* 'to bypass'; *sisse kallama* 'to pour in/drink up'; *sisse kutsuma* 'to invite in', ; *sisse torkama* 'to stick in'; *sisse vaatama* 'to look inside/visit something for a moment'; *tagant tõukama* 'to push from behind/boost'; *tagasi minema* 'to go back'; *tagasi tulema* 'to come back'; *tagasi vaatama* 'to look back'; *välja ajama* 'to send off/out'; *välja ilmuma* 'to debouch/emerge/appear unexpectedly'; *välja jääma* 'to stay out'; *välja paiskama* 'to throw something out from somewhere/blurt out something'; *välja pilduma* 'to throw something out from somewhere'; *välja pistma* 'to stick out'; *välja pääsema* 'to get out of somewhere'; *välja rändama* 'to emigrate'; *välja sulama* 'to melt out of something'; *üle tooma* 'to carry something/adapt/change the location of something'; *ümber paigutama* 'to relocate'
PVs with moderate compositionality degree
ette kandma 'to report/to serve'; *ette küsima* 'to ask in advance'; *läbi kaaluma* 'to weigh up/consider'; *maha kõmmutama* 'to shoot dead'; *maha käima* 'to go down/run down/go/degenerate'; *peale tungima* 'to invade'; *välja kuulutama* 'to announce'; *välja registreerima* 'to check/sign out'; *üles kasvama* 'to grow up'; *üles peksma* 'to beat/wake up somebody'; *üles seadma* 'to set up'; *üles soojenema* 'to warm up'; *ümber kehastuma* 'to incarnate'; *ümber mõtlema* 'to change mind'
PVs with low compositionality degree
kinni maksma 'to pay/stump up'; *läbi viima* 'to conduct/pass through'; *taga kihutama* 'to encourage/chase'; *vahele kukkuma* 'to fall between/get caught'; *vastu põrutama* 'to snap back at somebody'; *vastu raiuma* 'to object/to dispute'; *välja nägema* 'to appear to your eyes/see outside'; *üles kloppima* 'to fix/beat/fluff'

Sixty-four of the 157 PVs were evaluated as having a similar degree of compositionality in both datasets. It is important to note that most PVs had multiple

(non-)literalness scores, but only one is reflected in the comparison. While the meanings of the PVs were not determined in the compositionality dataset, it can be assumed that the meanings reflected by the degrees of compositionality of the PVs in Table 4 were the most predominant. For example, the annotators of the (non-)literalness dataset identified two readings for the PV *eemale tõukama* 'to push away/scare off/repel'. One meaning ('to push away') was assigned an average score of 0 (fully compositional), while the second meaning ('to scare off') was assigned a score of 3.0 (more non-compositional than compositional). According to the compositionality ratings, it was the most compositional PV in the dataset. Hence, the predominant meaning of *eemale tõukama* is fully compositional interpretation of 'to push away'. The PV *maha suruma* 'to suppress/bottle up/allay' received four (non-)literalness ratings – three of them pointed to the meanings being more non-compositional than compositional. One meaning ('to press down', *down* expressing direction) is fully compositional. The compositionality rating for *maha suruma* was 3.5, indicating that the annotators did not think that the meaning was fully compositional, but that it was considered to be more compositional than non-compositional. Hence, it can be assumed that the compositional meaning of *maha suruma* was not as predominant as it was for *eemale tõukama*.

The PVs with moderate degrees of compositionality were often PVs with multiple meanings in the EED and (non-)literalness ratings. For example, *ette kandma* 'to report/to serve' has three meanings in the EED[71], the annotators of the (non-)literalness dataset identified four, and it received a compositionality rating of 3.3. *Maha käima* 'to go down/run down/go/degenerate' has three meanings in the EED[72], and received two (non-)literalness ratings and a compositionality rating of 2.5. Hence, it can be assumed that none of these meanings was predominant. However, not all PVs with moderate compositionality have several meanings. For example, the (non-)literalness score for the PV *välja registreerima* was 3.33, and the compositionality rating was 3.3. With one meaning in the EED[73] and one (non-)literalness rating, it can be claimed that some PVs fall in the middle of the compositionality scale.

The PVs with low degrees of compositionality have meanings that are not the sum of the meaning of their components. It can be speculated that these PVs have one fully non-compositional meaning. For example, *vastu raiuma* has one meaning in the EED[74], and was assigned one (non-)literalness rating, indicating that the meaning 'to object/to dispute' is fully non-compositional. The PV *välja nägema* has two meanings in the EED[75] – 'to appear to your eyes' and 'to see outside' – and it received two (non-)literalness scores (0.33 and 5) but, according to the compositionality rating of 1.9, the non-compositional meaning was predominant. Furthermore, 13 of the 14 sentences containing *välja nägema* in the (non-)literalness dataset conveyed the non-compositional meaning of this PV.

Overall, the compositionality and (non-)literalness ratings both contained in-

[71] `http://www.eki.ee/dict/ekss/index.cgi?Q=ette+kandma` (accessed 10.11.2018).

[72] `http://www.eki.ee/dict/ekss/index.cgi?Q=maha+käima` (accessed 10.11.2018).

[73] `http://www.eki.ee/dict/ekss/index.cgi?Q=välja+registreerima` (accessed 10.11.2018).

[74] `http://www.eki.ee/dict/ekss/index.cgi?Q=vastu+raiuma` (accessed 10.11.2018).

[75] `http://www.eki.ee/dict/ekss/index.cgi?Q=välja+nägema` (accessed 10.11.2018).

formation about the degree of compositionality of Estonian PVs. As the purposes of the datasets were not the same, they were created differently. Hence, the datasets are not fully comparable. Nevertheless, 64 PVs had similar degrees of compositionality in both datasets. As a result, the approximate position of these PVs on a scale from compositional to non-compositional can be claimed with a higher degree of confidence than can the other PVs.

4.5 Abstractness/concreteness ratings for Estonian

In this section, the dataset of abstractness/concreteness ratings for Estonian lemmas[76] is described. The purpose and creation of this dataset are explained in the first subsection. The overview of the content is presented in the second section.

4.5.1 Purpose and creation

Turney et al. (2011) hypothesised that the metaphorical or literal usage of a word was predictable from the degree of abstractness of its context. Therefore, it has been discussed that the abstractness of the surrounding words helps to predict whether the meaning of an expression is compositional (literal) or non-compositional (non-literal) (e.g. Tsvetkov et al. 2014). In order to study the influence of the abstractness of context in the usage of Estonian PVs, a dataset of abstractness/concreteness ratings for Estonian lemmas was required.

Estonian abstractness/concreteness ratings were created following the methods of Köper and Schulte im Walde (2016a), who presented a collection of 350,000 lemmatised German words rated according to four psycholinguistic affective attributes. Although they had several affective ratings for German previously, they used an algorithm from Turney and Littman (2003) and word representations to create large-scale abstractness ratings for German in order to increase the number of available training instances. As Köper and Schulte im Walde (2016a) demonstrated in their paper, the automatically created ratings correlated highly with human ratings. The authors translated the English abstractness ratings from Brysbaert et al. (2014) to create the dataset. The ratings for 37,058 English words and 2,896 two-word expressions were obtained from over 4,000 participants by means of a norming study using Internet crowdsourcing for the data collection. The authors made a distinction between experience-based meaning acquisition and language-based meaning acquisition, and stated that experiences must not be limited to the visual modality. They used a five-point rating scale on which 1 denoted an abstract, language-based meaning and 5 denoted a concrete, experience-based meaning. (Brysbaert et al. 2014)

The English-Estonian MT dictionary[77] was used to translate the words from English to Estonian for the Estonian dataset. After the missing translations and two-word expressions were removed, the list contained abstractness/concreteness ratings for 24,891 words. The word embeddings for words from a 170-million token corpus were calculated; based on the word similarity, 243,674 Estonian

[76] The dataset is freely accessible at `https://github.com/eleriaedmaa/compositionality`.

[77] `http://www.eki.ee/dict/ies/index.cgi` (accessed 13.03.2017).

lemmas received abstractness ratings on a scale [0, 10][78] ranging from abstract to concrete.

The final version of the dataset is at the time of writing the largest resource for Estonian that contains information about the abstractness/concreteness of Estonian lemmas.

4.5.2 Analysis of the abstractness ratings

Of the 243,674 lemmas represented in the abstractness/concreteness dataset, approximately 75% are nouns, 9% are adjectives, 5% are numerals, 4% are verbs, 1% are adverbs[79]. The rest of the lemmas are abbreviations and other POS. Table 5 shows the 10 most concrete nouns, adjectives, verbs and adverbs, as well as their abstractness/concreteness ratings.

Table 5: Overview of the most concrete lemmas.

noun	**rating**	**adjective**	**rating**
reklaamtahvel 'billboard'	9.90	*250grammine* '50-gram'	8.98
küpsisepakk 'pack of cookies'	9.85	*päikesekuivatatud* 'sun-dried'	8.92
auto 'car'	9.80	*sissetehtud* 'conserved'	8.66
videomagnetofon 'video tape recorder'	9.66	*25kilone* '25-kilogram'	8.65
apelsin 'orange'	9.57	*koduküpsetatud* 'home-baked'	8.62
jääkaru 'polar bear'	9.52	*kurvasilmne* 'sad-eyed'	8.56
tomat 'tomato'	9.51	*leelisevaba* 'alkali-free'	8.52
päevalilleseeme 'sunflower seed'	9.48	*nööbitav* 'buttoned'	8.49
baarilett 'bar counter'	9.47	*üheksakorruseline* 'nine-storey'	8.45
banaan 'banana'	9.44	*kaheksakilone* 'eight-kilo'	8.44
verb	**rating**	**adverb**	**rating**
ribastama 'to cut into strips'	8.69	*praetult* 'fried'	8.29
röstima 'to toast'	8.62	*pakitult* 'wrapped'	8.28
riisuma 'to rake'	8.51	*jaapanipäraselt* 'Japanese-styled'	8.27
idanema 'to sprout'	8.44	*kooritult* 'peeled'	8.20
kärbatama 'to wither'	8.40	*lebavalt* 'reposed'	8.01
sülitama 'to spit'	8.28	*jahtunult* 'cooled down'	7.89
paneerima 'to flour'	8.21	*kalligraafiliselt* 'calligraphically'	7.84
suitsetama 'to smoke'	8.12	*prantsuspäraselt* 'French-styled'	7.81
raseerima 'to shave'	8.10	*lõigatult* 'cutted'	7.78
viilutama 'to slice'	8.08	*hiinapäraselt* 'Chinese-styled'	7.76

The most concrete lemmas are nouns, with an average abstractness of more than nine. The most concrete nouns are *reklaamtahvel*, *küpsisepakk* and *auto*.

[78]The explanation of this assignment of rating scores for each word and how the scores were rescaled numerically within [0, 10] was provided by Köper and Schulte im Walde (2016a).

[79]The statistics are based on the analysis provided by the Vabamorf morphological analyser.

Most of the concrete adjectives express a very specific numeral value, such as *250grammine*, *25kilone* and *üheksakorruseline*. Concrete verbs also express specific actions, such as *ribastama*, *riisuma* and *raseerima*. It is interesting that many of them are connected somehow to cooking. The most concrete adverbs are *praetult* and *pakitult*, expressing extremely definite forms.

Table 6 presents the most abstract nouns, adjectives, verbs and adverbs in the dataset. The most abstract lemma is *selliselt* with a score of 0.00, but its synonyms *niimoodi*, *niiviisi* and *nõnda* also have ratings indicating highly abstract meanings. Most of the abstract verbs are relatively idiosyncratic, such as *tunduma*, *eeldama*, *lootma* and *eelistama*, expressing the subject's feelings and opinions. The most abstract nouns mainly express human qualities that are not generally measurable, such as *ausus*, *aatelisus* and *headus*. The abstract adjectives also express highly subjective concepts – examples include *jõukohane*, *otstarbekas* and *üliolulline*.

Table 6: Overview of the most abstract lemmas.

noun	**rating**	**adjective**	**rating**
lõplikkus 'finiteness'	0.47	*jõukohane* 'feasible'	0.65
ausus 'honesty	0.47	*esmatähtis* 'overriding'	0.65
ratsionaalsus 'rationality	0.49	*otstarbekas* 'expedient'	0.67
omakasupüüdmatus 'selflessness'	0.50	*väheusutav* 'incredible'	0.71
jumalikkus 'divinity'	0.50	*väljendamatu* 'inexpressible	0.71
aatelisus 'idealism'	0.60	*kantilik* 'Kantian'	0.72
headus 'goodness'	0.60	*ebareaalne* 'unrealistic'	0.73
usutavus 'believability'	0.60	*üliolulline* 'crucial'	0.74
igavikulisus 'perpetuity'	0.61	*ontoloogiline* 'ontological'	0.74
püüd 'endeavor'	0.64	*mõistusevastane* 'irrational'	0.76
verb	**rating**	**adverb**	**rating**
välistama 'to exclude'	0.69	*selliselt* 'like that'	0.00
arvestama 'to take into account'	0.80	*niimoodi* 'like that'	0.10
võima 'can, may'	0.85	*niiviisi* 'like that'	0.25
usaldama 'to trust'	0.86	*lõpmatult* 'infinitely'	0.31
vihjama 'to hint'	0.91	*nõnda* 'so, thus'	0.35
tunduma 'to feel'	0.92	*ometi* 'yet'	0.35
eeldama 'to assume'	0.93	*seepärast* 'therefore'	0.36
lootma 'to hope'	0.96	*küllalt* 'enough'	0.38
eelistama 'to prefer'	0.96	*seetõttu* 'therefore'	0.40
sundima 'to force'	0.97	*paraku* 'unfortunately'	0.41

The overview of the most concrete and abstract lemmas allows to assume that the quality of the dataset is not particularly low. However, the evaluation of the dataset is challenging at present because Estonian lacks a similar resource for affective ratings. Nevertheless, it is possible to highlight some aspects that should be taken into account prior to using it in any other research.

The dataset includes polysemous lemmas, and sense disambiguation is not

employed. This means that the dataset does not include information evaluating the sense of the polysemous words. For example, the dataset does not indicate which meaning of the highly polysemous verb *saama* to get/receive/have' is evaluated. In addition, information about the POS is not provided; thus, homonymous lemmas are not identifiable. For example, the lemma *või* 'butter/or/so' has an abstractness/concreteness rating of 6.27, but it is not possible to detect which POS or meaning the rating represents.

Although the resource has its drawbacks, it still meets the general conditions suggested by Brysbaert et al. (2014) – concrete words refer to things and actions one can experience directly through one of the five senses (smell, taste, touch, hearing or sight), while abstract words refer to meanings that can be defined by other words. All the other words fall between these two extremes.

To conclude, these data must be interpreted with caution because the evaluation of the abstractness/concreteness ratings is not provided due to the lack of resources and competence. The aim of the dataset was to obtain relatively valid abstractness/concreteness ratings in order to use them for the classification task introduced in Chapter 6. The appropriate gold standard of abstractness/concreteness ratings should be based on human judgements, and must be preceded by exhaustive study of psycholinguistic and memory research. It is hoped that such a resource of affective ratings for Estonian will be created in the near future.

5 DETECTING THE COMPOSITIONALITY OF PARTICLE VERBS

In this chapter, the experiment for detecting the compositionality of Estonian PVs by applying DSMs is described[80]. The theoretical background to the distributional semantics, semantic similarity and applied methods was provided in Section 3.2.

The approach of detecting the degree of compositionality of PVs by measuring the similarity between the PV and its components was adopted from Bott and Schulte im Walde (2014), who hypothesised that the similar context of a PV and its base verb implied a similarity in their meanings. Thus, the similarity score must indicate which PVs are more compositional than others. The approach itself is not new and has been used to detect, for example, English phrasal verbs (McCarthy et al. 2003) and noun compounds (Reddy et al. 2011b). The verbal component of the compositional PV is considered to be a substantial and meaningful component of the PV (Erelt et al. 1993), and the particle adds a connotation (for example, perfectivity or state). This definition implies that the compositionality of the PV can be detected by measuring the similarity between the verb and PV.

The experiments introduced in this section are a continuation of the pilot study to detect the degree of compositionality of the Estonian PVs using word embedding models (Aedmaa 2017)[81]. The comparison of the results from the previous research and of the current study are presented in Section 5.4.3. The multi-sense embedding models were not used for the task previously and are somewhat experimental. However, the sense representations were suggested previously in the compositionality studies on other languages (e.g. Reddy et al. 2011a; Kober et al. 2017; Köper and Schulte im Walde 2017b). This work shares the main idea of these studies, which was to determine the intended meaning of the linguistic unit and apply it in order to detect the compositionality. The results of employing multi-sense embeddings for compositionality predictions of Estonian PVs are discussed in comparison to previous studies in Section 5.5.3.

The structure of the chapter takes the form of six sections. The first Section 5.1 provides information about the applied models and describes how these models were evaluated. Section 5.2 provides an introductory comparison of the compositionality predictions of the word and multi-sense embedding models. The primary results are presented, and the effect of frequency on the models trained with default settings is discussed. The impact of parameters on the quality of predictions is explored in Section 5.3. The results of the best word and multi-sense embedding models are then analysed in Sections 5.4 and 5.5, respectively. The detection of the degrees of compositionality of the PVs via the use of DSMs is discussed and conclusions are drawn in Section 5.6.

[80] All the trained embeddings are publicly available at `https://github.com/eleriaedmaa/embeddings`.

[81] A brief overview of the results of this study is provided in Section 2.4.1.

5.1 Experimental setup and evaluation

In the following experiments, the degree of compositionality of the Estonian PVs was determined by the similarity between the PVs and their base verbs. The similarity is expressed by the value of the CS (see Section 3.2.2), which reflects the similarity between two vectors that were created using word2vec (see Section 3.2.2.1) or SenseGram (see Section 3.2.2.2). The similarities between vectors were retrieved using Gensim.

All the models were trained[82] on lemmatised etTenTen corpus introduced in Section 4.1. As the components of the PVs did not always appear in the same order adjacent to each other, the corpus was modified prior to training the embeddings. Specifically, the POS information was used to detect the co-occurrence of an adverb and a verb in the same clause. If they were not adjacent to each other, the position of the adverb was changed to precede a verb. In the sentences in which the verb was followed by an adverb, the order of the PV components was changed in such a way that an adverb would precede a verb. After all the components of the potential PVs were placed in the same order (side by side), the adverbs and verbs were concatenated using an underscore. Therefore, the format of the PVs is 'adverb_verb' in the input corpora files. Overall, the goal of reordering the PV components was to be able to calculate the embeddings for as many PVs as possible. The PVs were treated as single units, and the vector of a verb did not include the context of the PV and represents the meaning it possesses individually.

As mentioned previously (see Sections 3.2.2.1 and 3.2.2.2), word2vec was applied to train models for word vectors, and SenseGram for multi-sense embeddings. Word2vec and SenseGram can both use two architectures, namely CBOW and Skip-gram. In addition, a set of parameters was used for configuration prior to the training of these models. Despite the fact that there are general recommendations concerning how the parameters should be set, this depends largely on the data and on the task. Hence, how the parameters influenced the quality of the models predicting the compositionality of Estonian PVs is unknown. Accordingly, after the evaluation of the default models is presented in Section 5.2), the influence of the parameters on the results is examined in Section 5.3). The parameter configurations for the default models are as follows: 300 dimensions, a window size of 10, a minimum-count threshold of 10, and 5 iterations.

The evaluation of the models was conducted by comparing the results to the human-annotated datasets. The comparison is expressed by the value of Spearman's rank correlation coefficient (ρ). The machine-created rankings of the PVs were based on the CS values between the PVs and their verbs. A higher CS value indicates a greater degree of compositionality. In addition, for each correlation, probability value (p-value, p) is given.

Sense embeddings imply that one linguistic unit may have multiple vector representations. For the assessment, the PVs were ranked based on the CS value between the sense vectors of the PVs and the base verbs. In the event of having multiple senses, the most probable (first) sense vector was taken into account. This choice was motivated by the fact that, when collecting the human annotations of PV

[82]This work was carried out in part in the High Performance Computing Center of the University of Tartu.

compositionality, it was expected that they would evaluate the most predominant meaning. Thus, it was presumed that the ranking containing the most probable meanings of PVs would correlate best with the human judgements. The analysis of the results is presented in Section 5.5, in which some examples showing how this choice could influence the results are discussed.

The human-annotated ranking (PVComp) was based on the compositionality ratings described in detail in Section 4.2. The PVs were ranked according to the average compositionality scores allocated by human annotators. The second ranking (PVLit) was according to the (non-)literalness ratings described in Section 4.3. The PVs were ranked based on the medians of the average literalness scores. The average literalness scores for the different meanings of the same PV were combined in order to avoid, preference for one meaning while still giving more weight to the more frequent meaning. In fact, as the sentences for the PVLit dataset were extracted automatically, it is highly likely that the more frequent meanings were represented in more sentences. Hence, without proving that predominant meanings are always more frequent than are others, this was anticipated when creating the datasets and rankings used for the evaluation of the computational models.

The highly compositional PVs were given a high score in the PVComp dataset, but a low score in the PVLit dataset. As the PVComp dataset was created in order to evaluate the distributional models, it contained ratings for 157 PVs. These are the PVs that did not cause difficulties for the human annotators (see Section 4.2.6). The PVLit ranking was extracted from the (non-)literalness ratings dataset (see Section 4.3). The goal was to obtain median (non-)literalness scores for the same 157 PVs that formed part of the PVComp dataset. However, as some PVs caused disagreement amongst the annotators (see Section 4.3.3), only 133 PVs were given scores. Therefore, the assessment was conducted on the PVComp ranking of 157 PVs and the PVLit ranking of 133 PVs. The correlation between the two rankings was $\rho = -0.50$ which was statistically significant ($p < 0.05$). The negative correlation was caused by the difference in the scales, as explained above.

5.2 An introductory evaluation of the compositionality predictions

In this section, an overview of the compositionality predictions of the models trained to learn word and multi-sense embeddings with default settings is presented. The goal is to present a preliminary analysis of the models, and to compare the results of the CBOW and Skip-gram models. Table 7 presents correlations (ρ) between the human-annotated rankings and the predictions of the models. Both CBOW and Skip-gram were trained to learn word and multi-sense embeddings with the following default parameter configurations – 300 dimensions, a window size of 5, a minimum-count threshold of 10 and 5 iterations.

The correlation coefficients suggest that both word embedding models had a similar correlation ($\rho = 0.16$) to that of the PVComp. The correlation with the Skip-gram model's predictions was statistically significant ($p < 0.05$), but the correlation with the CBOW model predictions was not. The predictions of both

Table 7: Results for the models trained with default settings, ρ – Spearman's rank correlation coefficient, p – p-value.

word embeddings								
model	dimensions	window	minimum-count	iterations	PVComp		PVLit	
					ρ	p	ρ	p
CBOW	300	5	10	5	0.16	0.05	−0.42	<0.05
Skip-gram					0.16	<0.05	−0.39	<0.05
multi-sense embeddings								
model	dimensions	window	minimum-count	iterations	PVComp		PVLit	
					ρ	p	ρ	p
CBOW	300	5	10	5	0.13	0.10	−0.26	<0.05
Skip-gram					0.06	0.47	−0.17	0.05

word embedding models correlated moderately and statistically significantly with the PVLit. The CBOW model predictions had a marginally stronger correlation ($\rho = -0.42$).

The predictions of the multi-sense embedding models correlated weakly with the human-annotated rankings, as opposed to the predictions of the word embedding models. The correlations with the PVComp dataset were relatively weak, and were statistically not significant at the $p = 0.05$ level. In fact, the predictions of the Skip-gram model did not correlate statistically significantly with the human-annotated rankings. Hence, the best predictions were provided by the CBOW model ($\rho = -0.26$), and were statistically significant ($p < 0.05$). In addition, Wilcoxon signed-rank test[83] (Rey and Neuhäuser 2011) shows that the differences between all models are statistically significant at the $p = 0.05$ level.

The effect of frequency on human compositionality judgements was analysed in Sections 4.2 and 4.3. No significant correlation between PV frequency and human compositionality judgements was detected. In the following section, the influence of the PV frequency on both word and multi-sense embedding models' compositionality predictions is studied. Figure 14 shows the correlations between PV frequency and the compositionality predictions of the word and multi-sense models.

In Figure 14, it can be seen that, in comparison to the word embedding models, multi-sense embeddings produced higher CS values – the range of values produced by the word embedding models was from 0.13 to 0.84, while the range of values produced by the multi-sense embedding models was from −0.08 to 0.99. The range of CS scores was particularly broad for the multi-sense embedding model that used the CBOW architecture. Thus, it can be assumed that this model predicted better the set of PVs (or their most probable meanings) that were more compositional than others.

[83]Wilcoxon signed-rank test is used from now on to determine whether the results for different word and multi-sense embedding models are statistically different from each other.

In comparison to the Skip-gram models, the predictions of the CBOW models correlated weakly with the PV frequency. The effect of frequency was the weakest for the CBOW word embedding model ($\rho = -0.15$) and the strongest for the Skip-gram multi-sense embedding model ($\rho = -0.42$). The CBOW multi-sense embedding model correlated slightly more strongly with the frequency ($\rho = -0.18$) than did the CBOW word embedding model. In comparison to the Skip-gram multi-sense embedding model ($\rho = -0.40$), the association between frequency and the predictions of the Skip-gram word embedding model was slightly weaker. The predictions of all the models correlated significantly with the PV frequency ($p < 0.05$). All the correlation coefficient values were negative, indicating that the CS value decreases as the PV frequency increases. Therefore, the models, particularly the Skip-gram models, tended to evaluate the infrequent PVs as being more compositional than the frequent ones.

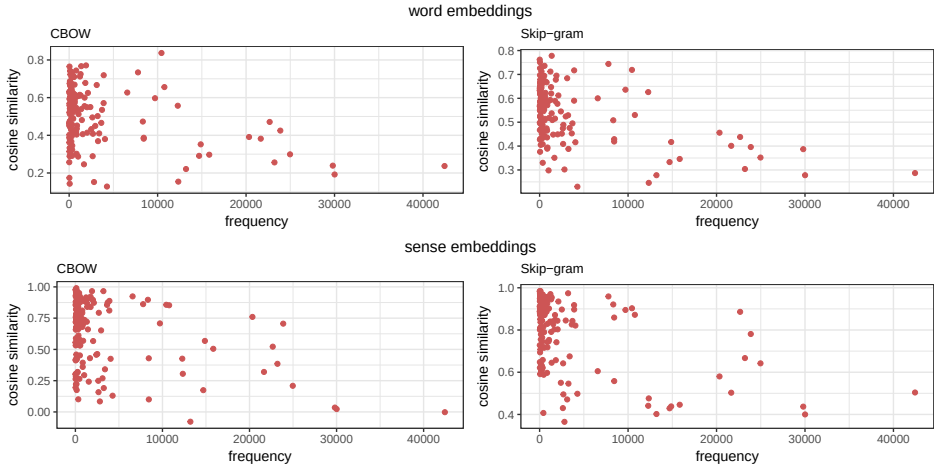

Figure 14: The correlation between the frequency and compositionality predictions of the word and multi-sense embedding models.

Overall, the preliminary results suggested that the word and multi-sense embedding models trained using the CBOW architecture provided slightly better compositionality predictions than did the Skip-gram architecture. As the differences between models are statistically significant, it is thus important what settings are used for training embeddings. The CBOW models were influenced less by the PV frequency. Nevertheless, the primary results suggested that the CS values of the word and multi-sense embeddings could be used to predict the degree of compositionality of the PVs, but the results must be improved. Therefore, additional experiments need to be conducted in order to develop a model that provides the best possible results for predicting the compositionality of Estonian PVs. These experiments are described in the following sections of this chapter. The influence of the PV frequency on the compositionality predictions is also studied in Sections 5.4.2 and 5.5.2 by investigating the correlation between PV frequency and the system compositionality predictions across different frequency ranges.

5.3 Impact of parameters on compositionality predictions

In order to study the influence of the parameters, the word and multi-sense embedding models were trained using various parameter configurations. Word2vec was used for the word embedding models. The SenseGram models used the output (word embeddings) of the word2vec models to train multi-sense embeddings. In other words, the models trained with word2vec and SenseGram used the same parameter settings. The parameters under investigation were described in Section 3.2.2.1, including the number of dimensions, the size of the context window, the minimum-count threshold and the number of training iterations.

5.3.1 Impact of the number of dimensions

In this section, the impact of the number of dimensions on the prediction quality of the word and multi-sense embeddings is analysed. Based on the general recommendations (see Section 3.2.2.1), it was hypothesised that the compositionality predictions of models trained with a higher vector dimensionality would have a stronger correlation with the human-annotated rankings than would the models trained with lower dimensionality. While the default models had a vector size of 300, CBOW and Skip-gram models with dimensions of 150 and 450 were trained for learning word and multi-sense embeddings. Table 8 shows the results of the word embedding models. The ρ values indicate a correlation between human-annotated rankings and the compositionality predictions of the models trained with various dimensionality values.

Table 8: The influence of the number of dimensions on word embedding models, ρ – Spearman's rank correlation coefficient, p – p-value.

model	dimensions	PVComp		PVLit	
		ρ	p	ρ	p
CBOW	150	0.17	<0.05	–0.44	<0.05
	300	0.16	0.05	–0.42	<0.05
	450	0.17	0.05	–0.43	<0.05
Skip-gram	150	0.13	0.10	–0.38	<0.05
	300	0.16	<0.05	–0.39	<0.05
	450	0.15	0.07	–0.39	<0.05

The correlation values demonstrate that the number of vector dimensions had a slight impact on the results of the models. However, the Wilcoxon signed-rank test indicates that the differences between all the possible pairs of models are statistically significant at the $p = 0.05$ level. Therefore, it can be concluded that the number of dimensions influenced the results of word embedding models.

The correlations between the CBOW model's predictions and the PVComp were similarly weak despite the number of dimensions. The only statistically

significant ($p < 0.05$) correlation was between the predictions of the CBOW model trained with 150 dimensions and the PVComp. The predictions of the CBOW models had moderate correlations with the PVLit ranking, and the correlations were also statistically significant ($p < 0.05$). Minimum changes in the correlation coefficient values indicate that the best results were obtained when the CBOW model was trained with 150 dimensions.

Both human-annotated rankings had slightly weaker correlations with the Skip-gram model's predictions than they did with the CBOW model's predictions. Nonetheless, the number of dimensions did not influence the results to a significant extent. The predictions of the Skip-gram model trained with a default value of dimensions (300) correlated with the PVComp statistically significantly ($p < 0.05$). In comparison to the PVLit, the Skip-gram models trained with 300 and 450 dimensions had the same results. Therefore, there was no reason to train the model with 450 dimensions instead of 300, but training with 150 dimensions might have produced less accurate predictions.

Regardless of the model used, the compositionality predictions correlated better with the PVLit ranking than they did with the PVComp. The correlations between the compositionality predictions and the PVLit were statistically significant ($p < 0.05$). The CBOW model for word embeddings should be trained with a vector size of 150 and the Skip-gram model with at least size 300, whereas other parameters maintained their default values.

Table 9 shows the results of the CBOW and Skip-gram models trained to learn multi-sense embeddings. The ρ values indicate the correlations between the human-annotated rankings and the compositionality predictions of the multi-sense embedding models trained with 150, 300 and 450 dimensions.

Table 9: The influence of the number of dimensions on the multi-sense embedding models, ρ – Spearman's rank correlation coefficient, p – p-value.

model	dimensions	PVComp		PVLit	
		ρ	p	ρ	p
CBOW	150	0.10	0.21	−0.33	<0.05
	300	0.13	0.10	−0.26	<0.05
	450	0.11	0.19	−0.30	<0.05
Skip-gram	150	0.05	0.54	−0.20	<0.05
	300	0.06	0.47	−0.17	0.05
	450	0.04	0.61	−0.18	0.05

The predictions of the CBOW model trained with the smallest number of dimensions correlated with the human-annotated rankings better than did any other model ($\rho = -0.33$). While the correlations with the PVComp were somewhat weak and statistically not significant at the $p = 0.05$ level, the predictions correlated more strongly and statistically significantly ($p < 0.05$) with the PVLit. However, the difference between the CBOW models was not statistically significant ($p < 0.05$).

The predictions of the Skip-gram models also had weak and statistically non-significant correlations with the PVComp ranking at the p = 0.05 level regardless of the number of dimensions of the models. In addition, the relationship to the PVLit dataset was weaker than was that of the CBOW models. The Skip-gram model trained with 150 dimensions obtained the strongest correlation ($\rho = -0.20$), which was not as strong as was the correlation between the 150-dimensional CBOW model and the PVLit. Wilcoxon's signed-rank test indicated that the Skip-gram multi-sense embedding models differed significantly from each other at the p = 0.05 level.

In summary, the size of the vector had a weak impact on the results. The CBOW model for learning word and multi-sense embeddings should be trained with 150 dimensions as opposed to dimensions of 300 or 450. However, the number of dimensions has insignificant impact on the multi-sense embeddings trained with CBOW. The Skip-gram model obtained the best results when it was trained with around 300 dimensions for learning word embeddings, and 150 dimensions for learning multi-sense embeddings. Overall, to obtain the best predictions, the word and multi-sense embeddings models should be trained using CBOW architecture and 150 dimensions. Therefore, the assumption that a higher dimensionality implied better results was inaccurate.

5.3.2 Impact of the window size

The influence of the window size on the quality of word and multi-sense embeddings is examined in this section. Based on the suggestions of word2vec's authors, it was expected that the default setting (window size of 5) worked best for the CBOW model, but should be increased up to 10 for the Skip-gram model (see Section 3.2.2.1). In addition to the default model with a window size of 5, both the CBOW and the Skip-gram models were trained with window sizes of 10 and 15. Table 10 shows the correlations (ρ) between the predictions of the word embedding models trained with various window sizes of 5, 10 and 15 and the human-annotated rankings.

Table 10: The influence of the window size on word embedding models, ρ – Spearman's rank correlation coefficient, p – p-value.

model	window size	PVComp		PVLit	
		ρ	p	ρ	p
CBOW	5	0.16	0.05	−0.42	<0.05
	10	0.17	<0.05	−0.44	<0.05
	15	0.17	<0.05	−0.46	<0.05
Skip-gram	5	0.16	<0.05	−0.39	<0.05
	10	0.15	0.06	−0.39	<0.05
	15	0.15	0.07	−0.38	<0.05

The correlation values indicate that the window size influenced slightly the results of the models. However, the Wilcoxon signed-rank test between all the model pairs indicates that the models differed statistically significantly from each other at the $p = 0.05$ level. The CBOW model-based rankings had a relatively weak correlation with the PVComp ranking, but the correlations for models trained with a window size of 10 and 15 were statistically significant ($p < 0.05$). The strongest correlation ($\rho = -0.46$) was found between PVLit and the predictions of the CBOW model that was trained with a window size of 15. Although the differences in the correlation of the values of models trained with various window size values were not significant, a larger window ensured better results.

The predictions of the Skip-gram models had somewhat weaker correlations with human-annotated rankings than did the predictions of the CBOW models. A model trained with a window size of 10 or 15 did not achieve better results than did a model trained with a window size of 5. Therefore, smaller window sizes resulted in stronger correlations. In summary, relatively small changes in the correlation coefficient values indicated that the window size did not have a remarkable impact on the results; however, the Skip-gram model should be trained with a window size of 5 or 10, but not with a window size of 15.

Table 11 presents the results of the multi-sense embedding models trained with window sizes of 5 (default), 10 and 15. The ρ values indicate the correlation between the human-annotated rankings and the predictions of the multi-sense embedding models trained with various window sizes.

Table 11: The influence of the window size on multi-sense embedding models, ρ – Spearman's rank correlation coefficient, p – p-value.

model	window size	PVComp		PVLit	
		ρ	p	ρ	p
CBOW	5	0.13	0.10	–0.26	<0.05
	10	0.14	0.08	–0.29	<0.05
	15	0.15	0.07	–0.35	<0.05
Skip-gram	5	0.06	0.47	–0.17	0.05
	10	0.06	0.47	–0.16	0.07
	15	0.03	0.75	–0.16	0.07

The context window influenced the results of the CBOW models – as the window size increased, so did the strength of the correlations. Therefore, the model trained with a window size of 15 produced the best predictions. However, the Wilcoxon signed-rank test indicates that while all the other models differed significantly ($p < 0.05$) from each other, no statistically significant difference between the CBOW models trained with windows sizes of 10 and 15 was detected. The correlations between the predictions and the PVComp were weak and statistically not significant at the $p = 0.05$ level, but the predictions correlated significantly ($p < 0.05$) with the PVLit ranking. In fact, the model trained with a window size

of 15 obtained the strongest correlation with the PVLit in comparison to any other multi-sense embedding model introduces thus far.

The Skip-gram models produced predictions that had weaker correlations with the human-annotated rankings than did the CBOW models. The correlations between the predictions of the Skip-gram models and the human-annotated rankings were weak and statistically not significant at the $p = 0.05$ level. The impact of the window size on the results was marginal, although models trained with a window size of 5 obtained the strongest correlations.

The impact of the context window depended on whether the CBOW or the Skip-gram was used – the results of the CBOW models were influenced more by the window size value than were the predictions of the Skip-gram model. The CBOW models should be trained with large window size, while the Skip-gram models predicted the compositionality more accurately when a narrow context was used. This finding was contrary to the earlier assumption that the CBOW model would prefer narrower context than would the Skip-gram architecture. However, the predictions were more accurate when the models were trained using the CBOW architecture instead of the Skip-gram algorithm. For the best results, the models should be trained with a window size of 15 rather than 5 or 10. However, no statistically significant difference between the CBOW multi-sense embedding models trained with window sizes of 10 and 15 was detected.

5.3.3 Impact of the minimum-count threshold

The minimum-count threshold indicates the frequency of the words that were taken into account when training the embeddings. As some of the PVs in the study were relatively infrequent, it was expected that, in order to obtain embeddings for all the PVs, the threshold should be lower than the default setting (10) – at least down to 5. Nonetheless, the quality of the embeddings did not necessarily improve when the word-count threshold was low. This section explores how the minimum threshold of word count influenced the compositionality predictions of the word and multi-sense embedding models.

In addition to a minimum-count threshold of 10, both the CBOW and the Skip-gram models were trained with thresholds of 5 and 15. Table 12 shows the results of the word embedding models. The ρ values indicate the correlations between the human-annotated rankings and the compositionality predictions of models trained with various thresholds of word count.

The influence of the minimum-count threshold on the results was marginal. The predictions of the CBOW model correlated weakly with the PVComp, and the correlations were not statistically significant at the $p = 0.05$ level. The correlations with the PVLit were the same for all the models regardless of the minimum-count threshold value. However, the correlations ($\rho = -0.42$) were statistically significant ($p < 0.05$). The minimum-count threshold value did not influence the quality of the predictions of the CBOW model. Furthermore, the Wilcoxon signed-rank test indicates that the differences between the CBOW models trained with different minimum-count thresholds were not statistically significant at the $p = 0.05$ level.

Table 12: The influence of the minimum-count threshold on the word embedding models, ρ – Spearman's rank correlation coefficient, p – p-value.

model	minimum-count	PVComp		PVLit	
		ρ	p	ρ	p
CBOW	5	0.15	0.07	−0.42	<0.05
	10	0.16	0.05	−0.42	<0.05
	15	0.15	0.07	−0.42	<0.05
Skip-gram	5	0.13	0.11	−0.37	<0.05
	10	0.16	<0.05	−0.39	<0.05
	15	0.13	0.12	−0.39	<0.05

For the Skip-gram models, the strongest correlation with the PVComp was obtained by the model trained with a word-count threshold of 10. In comparison to the PVLit, the model trained with a frequency count of at least 10 obtained the best predictions, and a higher threshold did not help to improve the results. As the correlations between the prediction and both human-annotated rankings were only statistically significant ($p < 0.05$) for the Skip-gram model trained with a frequency count of 10, it may be concluded that this is the most suitable value for training the Skip-gram model.

The minimum-count threshold value for both models should not be lower than 10. At the same time, the frequencies of the target PVs and the verbs need to be taken into account. For example, the models trained with thresholds of 5, 10 and 15 did not provide embeddings for one, three and five PVs, respectively. Therefore, in order to obtain predictions for each PV, the threshold should be 4 because this was the frequency of the most infrequent PV in the dataset. Overall, the word-count threshold had a very weak impact on the results, but the differences between the Skip-gram models trained with various thresholds are statistically significant.

The results of the multi-sense embedding models are presented in Table 13. The ρ values express the correlations between the human-annotated rankings and the predictions of the models trained with minimum-count thresholds of 5, 10 and 15.

The minimum-count threshold impacted slightly on the predictions of the multi-sense embedding models. The predictions of the CBOW model correlated weakly with the PVLit ranking, and the correlations were not significant at the $p = 0.05$ level unless the model was trained with a minimum-count threshold of 5. The correlations between the predictions of the CBOW model and the PVLit were moderate and statistically significant ($p < 0.05$). The model with a minimum-count threshold of 5 obtained the strongest correlation. However, the differences between the CBOW models are not statistically significant at the $p = 0.05$ level.

The predictions of the Skip-gram models correlated even more weakly with the human-annotated rankings, and the relationships were not statistically significant. However, the model trained with a threshold of 10 obtained slightly better results than did the models trained with thresholds of 5 and 15. Nevertheless, the impact

Table 13: The influence of the minimum-count threshold on the multi-sense embedding models, ρ – Spearman's rank correlation coefficient, p – p-value.

model	minimum-count	PVComp		PVLit	
		ρ	p	ρ	p
CBOW	5	0.17	<0.05	−0.32	<0.05
	10	0.13	0.10	−0.26	<0.05
	15	0.09	0.26	−0.30	<0.05
Skip-gram	5	0.03	0.74	−0.16	0.07
	10	0.06	0.47	−0.17	0.05
	15	0.03	0.71	−0.17	0.08

of the word-count threshold on the predictions was marginal.

In summary, the minimum-count threshold had a very weak impact on the results, particularly on the predictions of the word embedding models. The CBOW models trained to learn word embeddings achieved better results than did the other models. Taking slight differences in the results into account, the word embeddings should be trained using a CBOW model with a default threshold of 10, and the multi-sense embeddings should be trained using a CBOW model with a threshold of 5. However, Wilcoxon's test has confirmed that the minimum-count threshold does not have a significant impact on the CBOW models. In order to obtain embeddings for all the PVs, it was necessary to set the word-count threshold at less than 5.

5.3.4 Impact of the number of iterations

The impact of the number of the training iterations on the quality of the word and multi-sense embeddings is studied in this section. Based on the suggestions of word2vec's authors, it was suggested that the default setting (five iterations) should be increased to obtain better results (see Section 3.2.2.1). In addition to models with the default parameter configurations trained using five iterations, the CBOW and Skip-gram models were trained with 10 and 20 iterations.

Table 14 presents the results of the word embedding models. The ρ values indicate the correlation between the human-annotated rankings and the predictions of the models trained with various numbers of iterations. The number of the training iterations had an impact on the predictions – in comparison with the PVComp and PVLit rankings, the CBOW model trained with 20 iterations achieved the best results. In comparison with any other introduced model, the predictions of this model correlated more strongly with both of the human-annotated rankings. Hence, when the settings of the other parameters retain their default values, the CBOW models should be trained with 20 iterations. However, the Wilcoxon signed-rank test shows that the difference between the CBOW models trained with 10 and 20 iterations was not statistically significant at the $p = 0.05$ level.

Table 14: The influence of the number of iterations on the word embedding models, ρ – Spearman's rank correlation coefficient, p – p-value.

model	number of iterations	PVComp		PVLit	
		ρ	p	ρ	p
CBOW	5	0.16	0.05	−0.42	<0.05
	10	0.17	<0.05	−0.44	<0.05
	20	0.19	<0.05	−0.46	<0.05
Skip-gram	5	0.16	<0.05	−0.39	<0.05
	10	0.15	0.06	−0.42	<0.05
	20	0.18	<0.05	−0.39	<0.05

The results of the Skip-gram models were slightly poorer compared to the results of the CBOW models. The predictions of the model trained with 20 iterations achieved the strongest correlation with the PVComp but, in comparison to the PVLit, the best results were obtained by the model trained with 10 iterations. While the differences in the results were not considerable, the Skip-gram models should be also trained with a greater number of iterations than the default of 5. In addition, there are statistically significant differences between all the Skip-gram models.

The results of the same models trained to learn multi-sense embeddings are shown in Table 15. The ρ values express the correlation between the human-annotated rankings and the predictions of the models trained with 5, 10 and 20 iterations.

Table 15: The influence of the number of iterations on the multi-sense embedding models, ρ – Spearman's rank correlation coefficient, p – p-value.

model	number of iterations	PVComp		PVLit	
		ρ	p	ρ	p
CBOW	5	0.13	0.10	−0.26	<0.05
	10	0.14	0.09	−0.31	<0.05
	20	0.17	<0.05	−0.34	<0.05
Skip-gram	5	0.06	0.47	−0.17	0.05
	10	0.06	0.45	−0.22	<0.05
	20	0.05	0.54	−0.21	<0.05

The number of iterations influenced the results of the multi-sense embedding models somewhat. The difference in training the CBOW model with 5 and 20 iterations was notable. In addition, the model trained with 20 iterations worked

better than did the model trained with 10 iterations. The statistically significant impact of the number of iterations was confirmed also by the Wilcoxon signed-rank test – the difference between the CBOW models trained with 5 and 20 iterations was statistically significant at the $p = 0.05$ level.

The correlations of the models with the PVComp were not statistically significant at the $p = 0.05$ level unless the model was trained with 20 iterations – the correlation of $\rho = 0.17$ was statistically significant ($p < 0.05$). The same result was achieved by the CBOW model trained with a minimum-count threshold of 5. In comparison to all the introduced models, the correlation between the predictions of this model and the PVLit was not the strongest (the strongest correlation ($\rho = -0.35$) was achieved by the model trained with a window size of 15) but, taking the correlation coefficient values and both human-annotated rankings into account, it can be concluded that this model was slightly better than was any other system.

The predictions of the Skip-gram models had a weaker correlation with the PVComp and PVLit rankings than did those of the CBOW models. The correlations with the PVComp were not statistically significant at the $p = 0.05$ level. In comparison to the PVLit, the best predictions were made by the model trained with 10 iterations ($\rho = -0.22$). Nevertheless, compared to the results of the best CBOW model, the Skip-gram models performed significantly worse. However, differences between Skip-gram models trained with various iterations were statistically significant.

In summary, the number of iterations has a statistically significant impact on the results. For training word and multi-sense embeddings, the predictions of the models trained with CBOW architecture and 20 iterations were the best. This finding is consistent with the assumption that the models should be trained using more iterations than the default 5. The word embedding models performed better in a general sense. The CBOW model trained with 20 iterations obtained the overall best results of all the models. In order to improve the results of the word and multi-sense embeddings, some additional experiments were conducted, as explained in the next section.

5.3.5 Towards higher-quality embeddings

The previous sections illustrated that the parameters had a relatively low impact on the results. However, changing the values of parameters in comparison with the default models can result in better predictions. For example, the models trained with 20 iterations outperformed the models that were trained with 5 iterations in terms of results. Therefore, some further experiments were conducted in order to determine the best parameter settings for the task of predicting the compositionality of Estonian PVs.

The goal of the additional experiments was to improve the performance of the word and multi-sense embedding models suggested above in line with the results of previous experiments and earlier work. All of the results of the additional experiments using word embedding models are presented in Table 16, and the evaluation of the multi-sense embedding models is presented in Table 17.

Table 16: Results of the additional word embedding models, ρ – Spearman's rank correlation coefficient, p – p-value.

model	dimensions	window	min count	iterations	PVComp		PVLit	
					ρ	p	ρ	p
CBOW	150	15	10	20	0.15	0.06	−0.45	<0.05
		5			0.18	<0.05	−0.44	<0.05
	100	5	10	20	0.16	<0.05	−0.44	<0.05
	750				0.15	0.06	−0.43	<0.05
	300	15	10	20	0.16	<0.05	−0.46	<0.05
		30			0.15	0.07	−0.42	<0.05
		1			0.15	0.06	−0.35	<0.05
	300	5	2	20	0.16	0.05	−0.43	<0.05

Previous experiments (see Sections 5.3.1–5.3.4) demonstrated that the models trained with the CBOW architecture obtained better results than did the models trained using Skip-gram. In comparison to the human-annotated rankings, the best results were achieved by the word embedding model trained with default parameters (300 dimensions, a window size of 5 and a minimum-count frequency of 10), but with 20 iterations instead of the default 5. The predictions of the best model (see Table 14) correlated less strongly with the PVComp ($\rho = 0.19$) than they did with the PVLit ($\rho = -0.46$). The same correlation with the PVLit was also obtained by the CBOW model trained with 300 dimensions, a minimum-count threshold of 10 and 10 iterations, but with a window size of 15 (see Table 10). In addition, although the model trained with 150 dimensions did not obtain results that were as good as those obtained by the best models, it worked better than did the models trained with a greater number of dimensions. The results of the main experiments were adopted in order to train the additional word embedding models, as described in the following section.

The CBOW model was trained with 150 dimensions, a window size of 15 and 20 iterations. The only studied parameter that retained its default value was the minimum-count threshold (10). The correlation coefficient values reported in Table 16 demonstrate that using these settings for training word embeddings did not result in better predictions than those of the model trained with 20 iterations (see Table 14). Therefore, using the best parameter configurations of the main experiment did not help to improve the overall best results of the word embedding model.

The CBOW model was also trained with 150 dimensions and 20 iterations because, when trained with 5 iterations, the model obtained better results than when trained with a greater number of dimensions (see Table 8). The results in Table 16 indicate that the CBOW model trained with 150 dimensions and 20 iterations did not provide better predictions than did those obtained from the best word embedding model in the main experiment (see Table 14). Therefore, training the CBOW model with 150 dimensions (instead of 300) did not improve the overall best result.

The CBOW model was also trained with a window size of 15 and 20 iterations because the same model trained with 5 iterations produced better predictions than did the models trained with a more restricted context size (see Table 10). The results in Table 16 suggest that training the model with a window size of 15 and 20 iterations did not produce better results than did the same model trained with a window size of 5. Therefore, it can be assumed that, when the model is trained with 20 iterations, the other values of the studied parameters should retain their default values. However, in order to verify this claim, some further parameter configurations were examined.

Earlier experiments demonstrated that, as the number of dimensions increased, the results of the CBOW models improved. Therefore, the CBOW model was also trained with 100 dimensions and 20 iterations. The results demonstrated that training the word embeddings with a vector dimensionality of 100 did not provide better predictions than did training the model with 150 dimensions. Therefore, training the model with a lower dimensionality than 150 is not considered.

Cordeiro et al. (2016a) tested different DSMs for their predictions of the compositionality of nominal compounds, and claimed that word2vec models performed better when trained with a higher number of dimensions rather than with a low vector dimensionality. The sanity checks by Cordeiro (2017) showed that the best predictions were made by the models trained with a dimensionality of between 750 and 1,000. In order to determine whether a significantly higher dimensionality number led to better embeddings for the detection of the compositionality of Estonian PVs, the CBOW model was trained with 750 dimensions, a window size of 5, a minimum-count threshold of 10 and 20 iterations. The ρ indicates that increasing the dimensionality to 750 did not improve the results. The correlations with the PVComp dataset were not statistically significant at the $p = 0.05$ level. In comparison to the PVLit, the CBOW model performed less well than when trained with 100, 150 or 300 dimensions (see Table 14). In addition, the results re-emphasise that the number of dimensions had a marginal impact on the predictions of compositionality. Therefore, the additional experiments confirmed the earlier observations (see Section 5.3.1) that the impact of the amount of dimensionality on the results was insignificant.

Earlier experiments demonstrated that a low window size value was not better than was a high value for training word embeddings with CBOW (see Section 5.3.2). The analysis by Cordeiro et al. (2016a) proposed that a window size of 1 was better than were higher values (4 and 8). Therefore, additional experiments were conducted. The CBOW model was trained with window sizes of 1 and 30, while the other parameters were as follows: 300 dimensions, a minimum-count frequency of 10 and 20 iterations.

The ρ values in Table 16 indicate that the model trained with a window size of 30 produced poorer compositionality predictions than did the same model trained with a window size of 15. Therefore, the window size for the CBOW model should not be as high as 30. At the same time, based on the results of the model trained with a window size of 1, the context should not be so narrow that only one word from either side of the target word is taken into account. Therefore, the window size for the CBOW model should be somewhere between 5 and 15. There was not a significant difference regardless of whether the model was trained with

a window size of 5 or 15 but, based on the results, a window size of 5 worked slightly better when the model was trained with 20 iterations and 300 dimensions.

The default configuration for the minimum-count threshold is often 0, 1 or 5. It has been argued that using a minimum-count threshold as high as 50 does not improve the results (Cordeiro 2017). Although the predictions of the models trained with a minimum-count thresholds of 5, 10 and 15 were the same (see Table 12), when the model was trained with word-count threshold of 5 or higher, this did not provide predictions for all the PVs because the most infrequent PV in our dataset appeared four times. Therefore, in order to determine whether a threshold lower than 5 was better for the compositionality prediction, the CBOW model was trained with a threshold of 2 and 20 iterations.

As the correlation coefficient value presented in Table 16 indicates, the model trained with a minimum-count threshold of 2 did not obtain better results than did the models trained with a threshold of 10 (see Table 14). The correlation values between the predictions and the PVComp rating ($\rho = 0.16$) and the PVLit rating ($\rho = 0.43$) were comparable with results of the models trained with minimum-count thresholds of 5, 10 and 15, and 5 iterations. Nevertheless, the minimum-count threshold did not have a strong impact on the results. This finding was not unexpected because the results of the models trained with 5 iterations indicated that the lower threshold did not ensure better predictions (see Section 5.3.3). However, the threshold should be such that the model provides predictions for most of the infrequent PVs.

Compared to the default settings of the models, the best model in the main experiments differed in the number of iterations (20 instead of 5). Therefore, all the additional models were trained with 20 iterations. Based on the results in Table 14, it could be assumed that the results would be even better when the model was trained with a greater number of iterations. However, as the results did not improve significantly when the model was trained with 20 iterations instead of with 5, it is assumed that the improvement would not be substantial; therefore, no additional models were trained. Furthermore, increasing the iteration number would also require longer training time. The findings of Cordeiro (2017), who concluded that using 15 iterations instead of 100 for word2vec models yielded better results, supported this decision.

In the multi-sense embedding models in the main experiment, the highest statistically significant correlations ($\rho = 0.17$) in comparison with the PVComp were obtained by the CBOW models trained with a minimum-count threshold of 5 (see Table 13) and 20 iterations (see Table 15). The predictions of the model trained with a window size of 15 had the strongest correlation with the PVLit (see Table 11). In addition, the results demonstrated that 150 dimensions were more suitable for training multi-sense embeddings than were dimensions of 300 or 450 (see Table 9). These findings were adjusted to train the additional models in order to improve the overall results of the multi-sense embedding models.

Similarly to the additional experiments for training word embeddings, the CBOW model was trained with different combinations of the best parameter configurations. Therefore, the first additional model was trained with 150 dimensions, a window size of 15, a minimum-count threshold of 5 and 20 iterations. Table 17 shows that the predictions of this CBOW model correlated more strongly with

the PVLit ($\rho = -0.36$) than did the best multi-sense embedding model introduced earlier ($\rho = -0.34$, see Table 15). However, the predictions of this model correlated with the PVComp less strongly ($\rho = 0.14$) than did the predictions of the best model ($\rho = 0.17$). Therefore, combining all of the best parameter settings of the main experiments did not improve the results significantly. The CBOW models that were trained with 150 dimensions, a window size of 15, minimum-count thresholds of 5 and 10, and 20 iterations did not achieve better results than did the best multi-sense embedding models in the main experiment (see Table 15). Nonetheless, training a model with 150 dimensions, a window size of 5, a minimum-count threshold of 5 and 20 iterations increased the quality of the predictions in comparison to both human-annotated rankings – with the PVComp $\rho = 0.20$ and with the PVLit $\rho = -0.37$. In comparison to the best word embedding model, the correlation with the PVComp was even stronger (see Table 14). It is interesting that the same model trained with a minimum-count threshold of 10 produced noticeably poorer predictions.

Table 17: Results of the additional multi-sense embedding models, ρ – Spearman's rank correlation coefficient, p – p-value.

model	dimensions	window	min count	iterations	PVComp		PVLit	
					ρ	p	ρ	p
CBOW	150	15	5	20	0.14	0.08	−0.36	<0.05
			10		0.17	<0.05	−0.32	<0.05
		5	5		0.20	<0.05	−0.37	<0.05
			10		0.13	0.12	−0.30	<0.05
	100	5	10	20	0.12	0.10	−0.28	<0.05
	750				0.19	0.06	−0.34	<0.05
	300	1	10	20	0.07	0.39	−0.24	<0.05
		15			0.17	<0.05	−0.29	<0.05
		30			0.12	0.15	−0.32	<0.05
	300	5	5	20	0.21	<0.05	−0.31	<0.05
			2		0.17	0.04	−0.36	<0.05

When the models were trained with 5 iterations, using 150 dimensions resulted in better predictions than did using 300 or 450 dimensions. However, the model trained with 150 dimensions and 20 iterations did not obtain better results than did the same model trained with 300 dimensions. In order to determine whether changing the number of dimensions improved the results, one model was trained with 100 dimensions and another with 750 dimensions. It can be seen in Table 17 that the results of the model trained with 100 dimensions were worse than were the results of the same model trained with 150 dimensions. Therefore, the CBOW model for learning multi-sense embeddings should not be trained with a dimensionality lower than 300.

Training the CBOW model with 750 dimensions and 20 iterations produced similar results as did training the same model with 300 dimensions (see Table 15). However, the correlation with the PVComp dataset was stronger but statistically insignificant at the $p = 0.05$ level. The correlation with the PVLit was the same as that obtained by the best model. Nonetheless, as the correlation between the predictions and the PVComp was not significant, it cannot be concluded that using 750 dimensions is better than is using 300 dimensions. These results imply that the gold standard influences the results and that the architecture thereof must be well thought out. As a reminder, the main goal of the datasets was not the evaluation of the multi-sense embeddings. The experiment assessing the multi-sense embedding models using different (non-)literalness scores for different meanings is described and discussed in Section 5.5.3.

When the multi-sense embedding models were trained with 5 iterations, the best results were obtained by the CBOW model trained with a window size of 15 (see Table 11). Therefore, the CBOW model was also trained with a window size of 15 and 20 iterations. However, the results of this model were not better than were the results of the model trained with a window size of 5 and 20 iterations (see Table 15). In order to determine how the results of the multi-sense embedding models changed when very narrow or relatively wide contexts were applied, the CBOW models were trained with window sizes of 1 and 30. The correlation coefficient values suggested that using a narrow context was not reasonable when training multi-sense embeddings because such models produced predictions that correlated much more weakly with the human-annotated rankings than did the predictions of the model trained with a window size of 5 (see Table 15). Moreover, the model trained with a wider context did not perform better than did the model trained with a window size of 5. Therefore, when training a model with 20 iterations, the window size should not be smaller than 5, but it should not be greater than 10.

The results of training models with 5 iterations suggested that the best predictions were made when the word-count threshold of the CBOW model was 5. Therefore, the CBOW model was trained with a word-count threshold of 5 and 20 iterations. In addition, in order to obtain predictions for all the PVs and to determine how a minimum-count threshold of less than 5 influenced the results, the CBOW models were trained with thresholds of 2 and 5. The predictions of the model trained with a threshold of 5 correlated more strongly with the PVComp ($\rho = 0.21$) than did the predictions of any other word or multi-sense embedding model that was introduced. The correlation between the predictions of this model and the PVLit was not comparable with the best results. Training the model with a threshold of 2 did not help to improve the results. Therefore, regardless of the number of dimensions, the optimal minimum-count threshold for the multi-sense embedding model trained with CBOW is 5.

As with the word embedding models, the running time for multi-sense embedding models increases as the number of iterations increases. Therefore, no additional models with a greater number of iterations than 20 were trained.

In summary, the predictions of the additional word embedding models did not outperform the results of the best model in the main experiment. The best word embedding model was trained with CBOW architecture, a window size value of 5, a minimum-count threshold of 10 and 20 iterations. The correlation between

the predictions of this model and the PVComp is ($\rho = 0.19$) was significantly lower than was the correlation between the predictions and the PVLit ($\rho = -0.46$). Running additional experiments to train multi-sense embeddings helped to obtain better results than those of the best multi-sense embedding model in the main experiment. The overall strongest correlation with the PVComp ($\rho = 0.21$) was achieved by the predictions of the model trained with 300 dimensions, a window size of 5, a word-county threshold of 5 and 20 iterations. The best predictions in comparison with the PVLit ($\rho = -0.37$) were made by the model using the same setting except for the number of dimensions, which was set to 300. These results suggest that the outcome depends on the human-annotated gold standard used for the evaluation of the results. As the datasets were not developed for the assessment of the multi-sense embedding models, the results need to be interpreted with care.

5.3.6 Summary of the impact of the parameters on the predictions

The comparison of the compositionality predictions and the human-annotated datasets revealed that the studied parameters had a statistically significant impact on the results of the models. Of the parameters studied, the number of iterations has the strongest impact on the results, while the models, especially the ones trained with the CBOW architecture, were least influenced by the minimum-count threshold.

The experiments suggested that, for the compositionality predictions, the CBOW architecture is better than is Skip-gram. The difference between the architectures was more significant when looking at the multi-sense embedding models. In addition, the predictions of the multi-sense embedding models seemed to be influenced more strongly by the parameter values than were the word embedding models. The results of the additional multi-sense embedding models suggested that combining different values of the parameters can result in better predictions than can using default settings. For example, training a model with a minimum-count threshold of 5 and 20 iterations resulted in better predictions than when using default values of 10 and 5, respectively. Also, the difference between these models was statistically significant.

To obtain more substantial conclusions, more word and multi-sense embedding models with various parameter configurations would need to be trained and compared. However, as there were several aspects in addition to training parameters that influenced the work of the models described, the results of the best word and multi-sense embedding models are described, analysed and discussed in the next sections.

5.4 Compositionality predictions using word embeddings

The results of the models introduced in the previous section demonstrated that word embedding models achieved better results than did the multi-sense embedding models in comparison to the human-annotated datasets. In this section, the results of the best word embedding model are analysed, and the association

between frequency and the compositionality predictions of the word embedding model is explored.

5.4.1 Analysis of the results of the best word embedding model

In this section, an overview and analysis of the results of the best word embedding model are presented. Section 5.3 showed that the best word embedding model was the CBOW model trained with 300 dimensions, a window size of 5, a minimum-count threshold of 10 and 20 iterations. The predictions of this model correlated statistically significantly ($p < 0.05$) with both the human-annotated rankings – the correlation with the PVComp was $\rho = 0.19$ and the correlation with the PVLit was $\rho = -0.46$.

Table 18 lists the 10 most compositional PVs according to the CS value of the model and presents their frequency, their average compositionality ratings (PVComp) and median (non-)literalness rating (PVLit).

Table 18: The ten most compositional PVs according to the best word embedding model.

CS	PV	frequency	PVComp	PVLit
0.841	*läbi* 'through' *vaatama* 'to look' 'to look through/examine'	10,440	3.7	3.67
0.768	*maha* 'down' *rahunema* 'to calm' 'to calm down'	1,369	2.8	1.33
0.758	*ette* 'in advance' *valmistuma* 'to prepare' 'to prepare'	555	3.7	3.33
0.744	*vastu* 'against' *küsima* 'to ask' 'to ask in return'	794	3.9	3.33
0.740	*maha* 'down' *müüma* 'to sell' 'to sell off'	7,781	2.5	NA
0.730	*edasi* 'forward' *müüma* 'to sell' 'to sell on'	1,224	4.2	NA
0.727	*välja* 'out' *loosima* 'to draw lots' 'to raffle off'	1,907	3.5	NA
0.727	*üle* 'down' *küsima* 'to ask' 'to ask again'	1,296	2.8	NA
0.726	*kokku* 'together' *koguma* 'to gather' 'to gather'	3,908	4.1	NA
0.709	*tagasi* 'back' *minema* 'to go' 'to go back'	10,769	4.6	1.33

The CS values indicated that the most compositional PV was *läbi vaatama* 'to look through/examine'. The PVComp evaluators evaluated the PVComp as

being more compositional than non-compositional, but the PVLit annotators evaluated it as being relatively non-compositional. The PVLit score reflects the (non-)literalness of the meaning 'to examine' because sentences with the reading 'to look through something' were evaluated as being more compositional (with a score of 2). As there were more sentences that expressed the meaning 'to examine', it can be assumed that this meaning was more frequent and thus predominant.

None of the most compositional PVs was among the ten most compositional PVs according to the degree of compositionality assigned by the human annotators. However, *maha rahunema* 'to calm down' and *tagasi minema* 'to go back' had relatively low scores in the PVLit, indicating high compositionality. It can be seen that half of the PVs did not have a rating in the PVLit. This means that the sentences with these PVs caused inconsistency among the annotators, and that the compositionality of these PVs was difficult to assess.

It is interesting that both PVs containing the verb *müüma* 'to sell' were amongst the most compositional PVs. The annotators of the literalness ratings did not agree about the compositionality of these PVs. The annotators of the PVComp found the PV *maha müüma* 'to sell off' to be much more non-compositional than the PV *edasi müüma* 'to sell on'. The model's predictions for these PVs were very similar. However, the meanings of these PVs are related closely to the meaning of the verb *müüma* therefore, it is not surprising that the model found the contexts of these PVs to be similar to the context of the verb. Of the 10 most compositional PVs, the CS value of the PV *maha müüma* differed most from the average compositionality score suggested by the human annotators.

Table 19 shows the 10 least compositional PVs according to the CS values of the PV and the verb. The frequency, average compositionality rating (PVComp) and median (non-)literalness rating (PVLit) are also presented.

The least compositional PV, according to the CS values provided by the model, was *ette heitma* 'to reproach/blame'. This PV was also evaluated as being fully non-compositional by the annotators of the PVLit. The annotators of the PVComp found the compositionality of this PV to be in the middle of the scale ranging from fully non-compositional to compositional.

Other PVs (such as *ette nägema* 'to foresee/stipulate/see ahead', *vastu raiuma* 'to object/to dispute', *ette kandma* 'to report/to serve' and *üles kloppima* 'to fix/beat/fluff') were also evaluated as being fully non-compositional by the annotators of the PVLit dataset. The average compositionality ratings for these PVs were around 3. Thus, it can be assumed that the annotators were of the opinion that the PVs had multiple meanings, with some being more compositional than others. For example, the meaning 'to see ahead' of the PV *ette nägema* was considered to be more compositional than was the meaning 'to foresee'.

Both groups of human annotators agreed about the compositionality of the PV *üles kloppima* which was considered to be one of the most non-compositional PV according to the PVComp rating. The PVs *vastu raiuma* and *läbi viima* 'to conduct/pass through' were also among the ten most non-compositional PVs according to both compositionality ratings. The PV *alla tingima* 'to bargain/beat down' was evaluated as being more compositional than non-compositional by the human annotators. However, the automatic predictions for this PV indicated that it would be non-compositional. The discrepancy might have been caused by the fact

Table 19: The ten least compositional PVs according to the best word embedding model.

CS	PV	frequency	PVComp	PVLit
0.113	*ette* 'forward' *heitma* 'to throw' 'to reproach/blame'	4,274	2.8	5.0
0.122	*vastu* 'against' *pidama* 'to hold/keep' 'to withstand'	12,340	3.2	4.33
0.132	*ette* 'forward' *nägema* 'to see' 'to foresee/stipulate/see ahead'	30,012	3.5	5.0
0.192	*alla* 'down' *tingima* 'to bargain/haggle' 'to bargain/beat down'	62	4.1	2.0
0.207	*vastu* 'against' *raiuma* 'to chop' 'to object/dispute'	30	2.4	5.0
0.209	*ette* 'forward' *kandma* 'to carry' 'to report/serve'	2,812	3.3	5.0
0.218	*kinni* 'to' *pidama* 'to hold/keep' 'to stick to/slow down/detain'	13,211	4.2	4.0
0.239	*läbi* 'through' *viima* 'to carry' 'to conduct/pass through'	29,804	2.4	4.67
0.243	*esile* 'forth' *tükkima* 'to intrude' 'to jut/dominate'	53	3.5	3.67
0.257	*üles* 'up' *kloppima* 'to fluff' 'to fix/beat/fluff'	33	2.0	5.0

that the verb *tingima* 'to bargain/haggle' can appear independently in a context that does not concern merchandising, which is usually the context of this PV.

Some PVs that the model predicted would be the most compositional and the most non-compositional were evaluated as being ambiguous and some were not. Most of the compositional PVs were not considered ambiguous; three PVs have two meanings and one PV (*tagasi minema*) 'to go back' has three meanings in the EED. While the two most non-compositional PVs *ette heitma* 'to reproach/blame' and *vastu pidama* 'to withstand' are not polysemous according to the EED, the PV *ette nägema* 'to foresee/stipulate/see ahead' has four meanings and the PV *kinni pidama* 'to stick to/slow down/detain' has six. Thus, it could be assumed that non-ambiguous PVs tend to have higher CS scores than do ambiguous PVs. In order to determine whether there was a correlation between the CS value and ambiguity, the number of senses presented by the EED for the PVs, adverbs and verbs were compared. The Pearson's correlation coefficient value ($r = -0.18$) indicated that, as the number of senses increased, the CS value decreased. In addition, the PVs with non-ambiguous verbs tended to have higher CS values ($r = -0.21$). The correlations between the number of senses (of PVs and verbs) and the CS values were statistically significant at the $p = 0.05$ level. The cor-

relation between the number of senses of the adverbs and the CS values was 0.15, indicating that the PVs with ambiguous particles obtained higher CS values. This correlation was not statistically significant. Therefore, it can be claimed that ambiguous PVs with ambiguous verbs tended to have lower CS values and were therefore more likely to be evaluated as being more non-compositional than were non-ambiguous PVs.

The predictions of the word embedding models were evaluated across PVs regardless of the number of meanings they may have. The compositionality ratings were collected in the same way for polysemous and non-ambiguous PVs. In order to determine how well the word embeddings worked for non-ambiguous PVs, the correlation between the predictions of the model and the PVComp ratings of the 82 PVs that have only one meaning (according to the EED) was calculated. The correlation coefficient value of $\rho = 0.21$ indicated that the results were slightly better than they were when the ambiguous PVs were included ($\rho = 0.19$), but the correlation was not significant at the $p = 0.05$ level.

The same predictions that were compared to the PVLit were the median compositionality ratings of the PVs. In this case, the correlation between the non-ambiguous PVs and the compositionality predictions was $\rho = -0.51$, which is statistically significant ($p < 0.05$). Therefore, the model performed slightly better when predicting the compositionality of non-ambiguous PV than it did when predicting polysemous PVs[84].

Overall, the word embedding models could predict the compositionality of Estonian PVs despite their ambiguity. However, it is evident that the models worked better for non-ambiguous PVs, as they provided only one representation per word. The predictions of the word embedding models were influenced by the ambiguity of the PVs because the models tended to assign lower CS values to ambiguous PVs than to non-ambiguous ones. Another possible influence on the results – frequency – is studied in the next section.

5.4.2 Effect of frequency on the word embeddings

In this section, the association between frequency and the compositionality predictions of the word embedding model is examined. In addition, the predictions are compared to the human annotations across different PV frequency sets to explore whether the compositionality of frequent PVs was more challenging to predict or not. The latter question was motivated by previous work (e.g. Bott and Schulte im Walde 2014; Cordeiro 2017) in which it was hypothesised that it would be challenging to predict the compositionality of MWEs with high frequency. For the first goal, the correlation between PV frequency and the CS value was calculated. In

[84]Note that the overall results of the experiments are slightly different when the evaluation is of non-ambiguous PVs. Of the models introduced, the best results were obtained by the CBOW model trained with 300 dimensions, a window size of 15, a minimum-count threshold of 10 and 20 iterations. The predictions of this model correlated more strongly with the PVComp ($\rho = 0.24$) and with the PVLit ($\rho = -0.59$). Thus, it is evident that the word embedding models performed better when predicting the compositionality of non-ambiguous PVs. However, polysemous PVs are unquestionably more challenging for the automatic processing than are non-ambiguous ones; therefore, they were not excluded from this study.

order to determine whether the frequency influenced the quality of the predictions, correlations between PV frequency and CS values across different frequency sets were studied based on the results of the best word embedding model (CBOW architecture, 200 dimensions, a window size of 5, a word-count threshold of 10 and 20 iterations). Table 20 presents the correlation coefficient values for system predictions, human-annotated datasets and PV frequency across the frequency ranges of PVs.

The first insight into the effect of frequency on the compositionality predictions (see Section 5.2) proposed that the frequent PVs would tend to receive lower CS values than would the infrequent ones. The frequency information presented in Tables 18 and 19 indicates a similar claim. Four PVs with a frequency higher than 10,000 were amongst the 10 most non-compositional PVs, while two of the 10 most compositional PVs had frequencies greater than 10,000. However, there were no infrequent PVs (with a frequency of less than 100) amongst the 10 most compositional PVs, but there were four amongst the 10 most non-compositional PVs. This allows us to hypothesise that very frequent PVs tended to receive low CS scores, while frequency did not have as strong an impact on the CS scores of infrequent PVs.

Figure 15 shows the effect of frequency on the compositionality predictions of the model for the 50 most frequent and infrequent PVs. It can be seen that there was a negative correlation between the CS values and frequency for the frequent PVs and a positive correlation for infrequent PVs. Therefore, the most frequent PVs and the most infrequent PVs tended to have low CS values. The correlations were statistically significant ($p < 0.05$) for both frequency ranges. At the same time, the correlation between frequency and compositionality predictions of the PVs with moderate frequency was very weak ($\rho = 0.05$), and was not statistically significant. Therefore, frequency influenced the predictions of frequent and infrequent PVs, but not the predictions of PVs with moderate frequency.

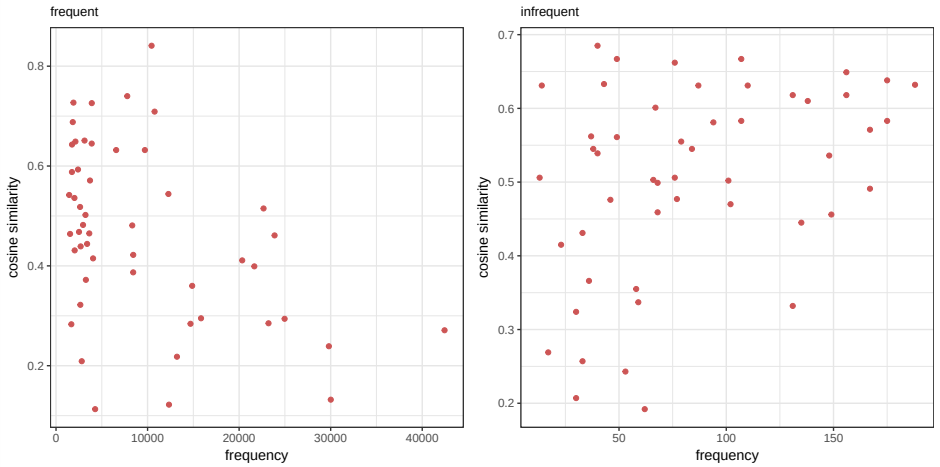

Figure 15: Correlations between frequency and compositionality predictions for the 50 most frequent and infrequent PVs.

The best model trained for word embeddings differed from the default CBOW model only in terms of the number of iterations. The default CBOW model was trained with 5 iterations, while the best model had 20 iterations. The overall correlation between frequency and the CS values of the best model was relatively weak ($\rho = -0.10$), and was not statistically significant at the $p = 0.05$ level. Compared to the default CBOW model ($\rho = -0.15$), the association between frequency and compositionality predictions was weaker. Therefore, the number of iterations decreased the effect of frequency on the system's compositionality predictions.

The effect of frequency on the compositionality predictions of the system depended on the frequency class of the PVs. In addition, the quality of the predictions depended on the frequency of the PVs. It can be seen in Table 20 that the predictions of the model correlated better with the compositionality scores of PVs that were very frequent or infrequent than with the PVs with moderate frequency. For example, the compositionality scores for the PVs with moderate frequency in the PVComp had no correlation with the compositionality predictions, while the infrequent PVs had a correlation of $\rho = 0.30$. The correlation of $\rho = 0.24$ between frequent PVs and the predictions was also stronger compared to the overall correlation, but it was not statistically significant. Therefore, based on the PVComp, it can be concluded that the prediction of the compositionality of infrequent PVs is less challenging than is the prediction of the compositionality of PVs with greater frequency.

Table 20: Predictions of the best word embedding model for different frequency sets of the PVs.

set of PVs (frequency)	PVComp		PVLit		PV frequency	
	ρ	p	ρ	p	ρ	p
all (13–42,441)	0.19	<0.05	–0.46	<0.05	–0.10	0.21
without frequent (13–1,369)	0.14	0.17	–0.37	<0.05	0.18	0.08
without infrequent (202–42,441)	0.13	0.21	–0.45	<0.05	–0.29	<0.05
without frequent and infrequent (202–1,369)	–0.03	0.82	–0.30	<0.05	0.05	0.74
frequent (1,439–42,441)	0.24	0.09	–0.54	<0.05	–0.42	<0.05
infrequent (13–188)	0.30	<0.05	–0.45	<0.05	0.37	<0.05

The evaluation of the predictions based on the PVLit scores shows that the predictions were most accurate for frequent PVs, but the predictions also correlated well with the compositionality scores for infrequent PVs and were statistically significant at the $p = 0.05$ level. Therefore, the word embedding models predicted

the compositionality of PVs with high and low frequency better than they did the PVs with moderate frequency.

In summary, the frequency of a PV had a moderate impact on the compositionality predictions of the best word embedding model. Frequency influenced the predictions of infrequent and frequent PVs. The model tended to assign a low compositionality score to frequent and infrequent PVs, whereas the compositionality of the PVs with moderate frequency was not influenced strongly by the frequency. At the same time, the compositionality predictions for infrequent and frequent PVs tended to be better than were the predictions for the PVs with moderate frequency. These main findings are compared to the results of previous research in the next section.

5.4.3 Comparison with previous research

In this section, the main results of the experiments with word embedding models are compared to previous work using DSMs to predict the compositionality of MWEs.

The first DSM model predicting the compositionality of Estonian PVs was introduced by Aedmaa (2017). The PVComp dataset was used for the evaluation, but the embeddings were trained on different (smaller) corpora using different parameter configurations. The CBOW model was trained with 200 dimensions and with a window size of 10. Therefore, the results are not fully comparable. However, the correlations between the predictions and the human-annotated scores for all the PVs ($\rho = 0.27$) were stronger than the best model in this study achieved in comparison to the PVComp, but weaker in comparison to the PVLit.

It has been argued previously that the compositionality of frequent Estonian PVs is more difficult to predict than is the compositionality of infrequent PVs (Aedmaa 2017). The same was found regarding German PVs by Bott and Schulte im Walde (2014), who explored the role of the frequency of PVs and their base verbs in compositionality predictions. Due to issues with data sparsity, these authors argued that the compositionality of high frequency verbs was easier to predict than was that of low-frequency verbs. However, the results showed that, in addition to low-frequency PVs, the compositionality of frequent PVs was also challenging to predict. Cordeiro (2017), who compared different DSMs to predict the compositionality of English, French and Portuguese MWEs, was not able to demonstrate whether the compositionality of frequent expressions was easier to predict than was that of less frequent MWEs. In contrast to earlier findings, however, the results of the current work propose that the compositionality of frequent and infrequent PVs is easier to predict than is the compositionality of PVs with moderate frequency.

In addition, Cordeiro (2017) demonstrated that the compositionality predictions correlated well with the frequency of the compounds in the corpus. The correlation between the word2vec CBOW model and compound frequency ($\rho = 0.50$) signifies that the frequent MWEs tend to obtain higher CS values. The results of this thesis suggest that the positive correlation between frequency and compositionality predictions only appears amongst infrequent PVs. Therefore, the results of the previous research and of this study are partly similar.

In summary, the results of the word embedding models predicting the compositionality of Estonian PVs are not completely comparable with any of the previous studies. However, the comparison to similar work suggests that the predictions of the word embedding models are affected by frequency, but that the task and other aspects mitigate the extent of the frequency effect.

5.5 Compositionality predictions using multi-sense embeddings

Applying multi-sense embedding models to predict the compositionality of Estonian PVs was more experimental than was using the word embeddings. Therefore, no separate dataset for the evaluation of the multi-sense embedding models was created. However, this section analyses the results of the multi-sense embedding models, investigates the effect of frequency on the results and outlines the challenges of using multi-sense embedding models to predict the compositionality of PVs.

5.5.1 Analysis of the results of the best multi-sense embedding model

In this section, an overview and an analysis of the results of the best multi-sense embedding model are presented. While the overall highest correlation with the PVLit was obtained by the best word embedding model, the predictions of the multi-sense embedding model trained with 300 dimensions, a window size of 5, a minimum-count threshold of 5 and 20 iterations correlated better with the PVComp than did any other suggested model. The correlation between the predictions of this model and the PVComp was $\rho = 0.21$ and the correlation with the PVLit was $\rho = -0.31$. Table 21 introduces the most compositional PVs according to the CS value of the model and presents their frequency, their average compositionality ratings (PVComp) and their median (non-)literalness rating (PVLit).

CS values indicate that the most compositional PVs were *vastu kajama* 'to sound like an echo', *välja loosima* 'to raffle off' and *lahti voltima* 'to unfold/ unwrap'. These PVs were not evaluated as being fully compositional by the PVComp evaluators, but rather as more compositional than non-compositional. While the PV *välja loosima* caused disagreement amongst the PVLit annotators, the score of 0.33 for *lahti voltima* indicates that the annotators evaluated it as also being compositional. On the other hand, the PV *vastu kajama* was evaluated as being relatively non-compositional.

The PVComp scores indicate that most of the compositional PVs were annotated as being rather compositional as opposed to being non-compositional by the annotators. The scores for the PVLit were more mixed. The biggest difference between the predictions and the human judgements was for the PVs *alla laadima* 'to download' and *välja saagima* 'to saw out', which the annotators annotated as being fully non-compositional. This might have been because the meaning of the PV *alla laadima* is relatively specific, but the verb *laadima* has several meanings – 'to load' and 'to charge'. The PV *välja saagima* was rated as being fully non-compositional because there was only one sentence that did not cause

Table 21: The ten most compositional PVs according to the best multi-sense embedding model.

CS	PV	frequency	PVComp	PVLit
0.973	*vastu* 'against' *kajama* 'to echo' 'to echo back'	107	3.0	4.0
0.972	*välja* 'out' *loosima* 'to draw lots' 'to raffle off'	1,907	3.5	NA
0.964	*lahti* 'open' *voltima* 'to fold' 'to unfold'	67	4.3	0.33
0.959	*läbi* 'through' *kaaluma* 'to weigh' 'to consider'	429	2.7	3.67
0.956	*alla* 'down' *laadima* 'to load' 'to download'	3,221	3.0	5.0
0.955	*välja* 'out' *saagima* 'to saw' 'to saw out'	76	4.4	5.0
0.954	*üles* 'up' *tursuma* 'to bloat' 'to swell up'	14	4.1	NA
0.940	*ümber* 'around' *reastuma* 'to align' 'to change a lane'	40	3.6	NA
0.937	*välja* 'out' *tahuma* 'to hew' 'to hew out'	40	3.7	NA
0.937	*maha* 'down' *minema* 'to go' 'to get off'	1,739	3.5	2.67

disagreement amongst the annotators, and it happened to be a sentence conveying a non-literal meaning.

The 10 least compositional PVs according to the CS values for the PV and the verb are shown in Table 22. The frequencies, average compositionality ratings (PVComp) and median (non-)literalness ratings (PVLit) are also presented. The most non-compositional PVs according to the CS value provided by the model were *läbi viima* 'to conduct/pass through' and *ette heitma* 'to reproach/blame'. The annotators of the PVLit ranking evaluated these two PVs as having low compositionality. The average compositionality scores (PVComp) show that the compositionality of *läbi viima* was somewhat more non-compositional than compositional, and the score for *ette heitma* suggested that the PV was neither fully compositional nor non-compositional, but had moderate compositionality. This was likely due to the fact that the PV's meaning of 'to conduct' is more non-compositional than is the meaning 'to pass through'.

While the PVLit scores for other PVs indicated that the PVs were relatively non-compositional, the scores for *alla tingima* 'to bargain/beat down' and *välja ilmuma* 'to debouch/emerge/appear unexpectedly' suggested that the PVs were compositional. The PVs were evaluated as being more non-compositional than

compositional by the annotators of the PVComp. As both PVs have only one meaning but the verb has more, this might explain why the scores differed. For example, the model predicted that the PV *alla tingima* would be non-compositional, probably because the context of the PV was different from the context in which the verb *tingima* in meaning 'to bargain/haggle' appeared. The contexts of *alla tingima* and *tingima* were presumably similar. As the evaluation was conducted on the most probable meanings, the result was apparently influenced by this. Therefore, the evaluation of the multi-sense embedding model was conducted using meanings other than the most probable ones.

Table 22: The ten least compositional PVs according to the best multi-sense embedding model.

CS	PV	frequency	PVComp	PVLit
−0.023	*läbi* 'through' *viima* 'to carry' 'to conduct/pass through'	29,804	2.4	4.67
−0.007	*ette* 'forward' *heitma* 'to throw' 'to reproach/blame'	4,274	2.8	5.0
0.020	*kinni* 'to' *pidama* 'to hold/keep' 'to stick to/slow down/detain'	13,211	4.2	4.0
0.024	*ette* 'forward' *kandma* 'to carry' 'to report/to serve'	2,812	3.3	5.0
0.028	*ette* 'forward' *valmistama* 'to prepare' 'to prepare'	14,692	3.9	3.67
0.052	*alla* 'down' *tingima* 'to bargain/haggle' 'to bargain/beat down'	62	4.1	2.0
0.092	*välja* 'out' *ilmuma* 'to appear' 'to appear unexpectedly/emerge'	3,117	3.8	2.0
0.135	*kaasa* 'along' *tooma* 'to bring' 'to bring along/imply'	24,971	4.1	4.33
0.147	*ühte* 'together' *hoidma* 'to keep' 'to stick together'	131	2.9	4.33
0.153	*ette* 'forward' *nägema* 'to see' 'to foresee/stipulate/see ahead'	30,012	3.5	5.0

More precisely, the best multi-sense embedding model was evaluated once more in comparison to the PVLit dataset. Unlike the main evaluation, instead of calculating the median literalness scores for each PV, the average (non-)literalness scores assigned by the annotators for each meaning were used (the ones that represented each PV in the (non-)literalness dataset (Aedmaa 2018). In addition, the intended meanings suggested by SenseGram's WSD mechanism were applied in place of the most probable meanings of the PV and verb. Therefore, the meanings of the PV and verb were detected for each sentence via the WSD mechanism, and the CS values were calculated for the vectors of the intended

meanings. Sentences containing the same meanings of the same PV, the average literalness score and the CS value were merged. As a result, the average (non-)literalness and CS scores for 438 meanings were calculated and the rankings based on these scores were compared. The comparison of literalness and the CS scores suggested a weaker correlation between the rankings ($\rho = -0.29$) than the same model obtained in the original evaluation ($\rho = -0.34$). While the differences in the results were not significant, the evaluation method described was unreasonably complicated and costly. However, a specially designed dataset must be created for a better evaluation of multi-sense embedding models in the future.

Amongst the 10 least compositional PVs, there were seven that were also the 10 least compositional according to the results of the best word embedding model (see Table 19). At the same time, there were only two PVs that were the most compositional according to the best word and multi-sense embedding models. The similarities could have been caused by the fact that the word embeddings were applied to train the multi-sense embeddings. However, compared to the results of other introduced models, the same PVs tended to obtain the lowest CS values across the models trained with different architectures and parameter configurations. For example, the PVs *ette kandma* 'to report/to serve', *ette nägema* 'to foresee/stipulate/see ahead', *ette valmistama* 'to prepare', *kinni pidama* 'to stick to/slow down/detain' and *läbi viima* 'to conduct/pass through' were also amongst the 10 least compositional PVs according to the predictions of the Skip-gram model trained with the default settings (300 dimensions, a window size of 5, a minimum-count threshold of 5 and 5 iterations).

In summary, multi-sense embedding models can be used for detecting the compositionality of Estonian PVs. However, the multi-sense embedding models do not provide as good predictions as word embedding models. The second, but more sophisticated evaluation of the model, showed that the further research of using multi-sense embeddings for predicting the compositionality of Estonian PVs is needed. The reasons for the modest performance of the multi-sense embeddings for the compositionality detection task are further discussed in Section 5.5.3. The association between frequency and the predictions of multi-sense embedding model is studied in the next section.

5.5.2 Effect of frequency on the multi-sense embeddings

The association between frequency and the predictions of the multi-sense embedding models is explored in this section. For this task, the correlation between the frequency of the PV and the CS value was calculated. As with the word embedding models, the compositionality predictions were compared to the human annotations across different PV frequency sets to determine how well the model predicted the compositionality of PVs with various frequencies. The best multi-sense embedding model studied here was the one trained with CBOW architecture, 300 dimensions, a window size of 5, a minimum-count threshold of 5 and 20 iterations. Table 23 presents the correlation coefficient values for the system's predictions, the human-annotated datasets and the PV frequency across the frequency ranges of the PVs.

The first insight into the effect of frequency on the compositionality predictions of multi-sense embedding models demonstrated (see Section 5.2) that, in comparison to Skip-gram, the frequency correlation was weaker in the CBOW's predictions. As the best word embedding model also used the CBOW architecture, it can be assumed that the effect of frequency was similar for the best multi-sense embedding model. The frequency information presented in Tables 21 and 22 suggests that frequent PVs tended to have low CS values. At the same time, low compositionality was also assigned to the infrequent PVs. However, based on the findings of the word embedding models introduced earlier, it can be hypothesised that the multi-sense embedding models' compositionality predictions for frequent PVs were also influenced more by the frequency than by the compositionality predictions for infrequent PVs.

Figure 16 illustrates the effect of frequency on the compositionality predictions using the multi-sense embedding model for the 50 most frequent and infrequent PVs. It can be seen that the correlation between frequency and the CS values was negative for the frequent PVs, while there is no correlation for the infrequent PVs. The results in Table 23 show that the correlation between the frequency and CS values for the frequent PVs ($\rho = -0.38$) was much stronger for the frequent PVs than it was for the infrequent PVs ($\rho = 0.07$). The correlation for the frequent PVs was statistically significant; thus, it can be concluded that the CS values for the frequent PVs tended to be low. Therefore, the model predicted that the frequent PVs would be non-compositional. This finding is the same as that of the best word embedding model (see Section 5.4.2). There were some infrequent PVs with low CS values, but there were more infrequent PVs with high CS values.

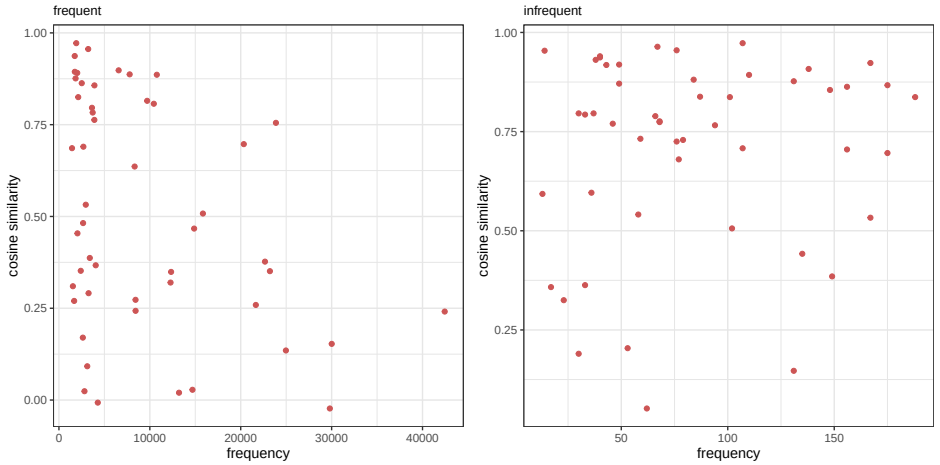

Figure 16: Correlation between frequency and compositionality predictions for the 50 most frequent and infrequent PVs.

The overall correlation between frequency and the CS values produced by the best model ($\rho = -0.26$) was moderate. As it was statistically significant, it can be claimed that, as the frequency increased, the CS values of the multi-sense

embedding model tended to decrease. As mentioned previously, this tendency was particularly strong for frequent PVs.

The quality of the compositionality predictions varied across the frequency sets of the PVs. In Table 23, it can be seen that the predictions for the frequent (particularly for the infrequent PVs) were better than were the predictions for the PVs with moderate frequency. For example, the correlation between the predictions of infrequent PVs and the PVComp was much stronger ($\rho = 0.33$) than was the correlation between the PVs with moderate frequency and the PVComp ($\rho = 0.06$). Similarly, in comparison to the PVLit dataset, the predictions for frequent and infrequent PVs were better than they were for PVs with moderate frequency. It can be thus concluded that the systems predicted the compositionality of infrequent and frequent PVs well. The multi-sense embedding model predicted the scores for infrequent PVs particularly well, while the frequency was associated weakly with the predictions of infrequent PVs.

Table 23: Predictions of the best multi-sense embedding model for different frequency sets of PVs.

set of PVs (frequency)	PVComp		PVLit		PV frequency	
	ρ	p	ρ	p	ρ	p
all (13–42,441)	0.21	<0.05	−0.31	<0.05	−0.26	<0.05
without frequent (13–1,369)	0.19	0.05	−0.26	<0.05	−0.02	0.82
without infrequent (202–42,441)	0.16	0.11	−0.30	<0.05	−0.35	<0.05
without frequent and infrequent (202–1,369)	0.06	0.69	−0.18	0.24	0.02	0.91
frequent (1,439–42,441)	0.18	0.20	−0.32	<0.05	−0.38	<0.05
infrequent (13–188)	0.33	<0.05	−0.35	<0.05	0.07	0.61

To conclude, in comparison to the best word embedding model, the effect of frequency on the results of multi-sense embedding was similar. As concluded in Section 5.4.2, the frequency was clearly associated with the predictions of the frequent PVs. The compositionality of frequent and infrequent PVs was predicted more accurately than was the compositionality of PVs with moderate frequency.

5.5.3 Discussion of using multi-sense embeddings for the compositionality predictions

The motivation for using multi-sense embeddings was to take the word polysemy into account; thus obtaining compositionality predictions that were more accurate. As the results of the word and multi-sense embeddings demonstrated (see Sections

5.2 and 5.3), multi-sense embeddings performed more poorly than did word embeddings in which there was one vector for each word regardless of the real number of meanings that this word might have. This section analyses two main reasons for the poor performance of the multi-sense embeddings.

The first reason for the modest performance of the multi-sense embeddings when predicting compositionality was the quality of distinguishing the different senses of words. The SenseGram worked relatively well when differentiating the meanings of nouns. For example, using pre-trained models for English[85] demonstrates how the SenseGram is able to distinguish between the two main meanings of the word 'mouse' (although the SenseGram suggests more meanings for this word). The word sense disambiguation (WSD) mechanism revealed that the word 'mouse' in the sentence 'It was believed that a mother eating a mouse would help heal the baby who was ill'[86] is 'a small rodent', while the 'mouse' in the sentence 'The mouse turns backwards and forwards and left and right movements of the hand into equivalent electronic signals that, in turn, are used to move the pointer'[87] refers to 'a device to move the cursor on a computer screen'. Similarly, based on the multi-sense embeddings trained for Estonian in the current research, SenseGram's suggested two meanings for the Estonian word *hiir* 'mouse'. The WSD mechanism detected correctly that the first meaning was used in the sentence *Hiir on väike loom, kes armastab juustu* 'A mouse is a small animal that loves cheese' and the second meaning in the sentence *Seda juhtmevaba hiirt on mugav kasutada* 'This wireless mouse is comfortable to use'.

Verbs were suggested as being the most difficult POS for WSD problems (e.g. Cabezudo et al. 2015). Therefore, the multi-sense embeddings for verbs might not support high-quality compositionality predictions. For example, the model produced only one meaning for ambiguous verbs such as *ajama* 'to drive/run', *käima* 'to go/walk' and *minema* 'to go'. However, the distinct meanings of some verbs were relatively clear based on their nearest neighbours. For example, SenseGram provided five meanings for the verb *mängima* – the most likely meaning is 'to play an instrument', the second most likely meaning expresses 'to play a game or sport', the third meaning indicates 'to play in a cinema or theatre', the fourth meaning is 'to have fun' and the fifth expresses 'to play a (ball) sport'. On the other hand, the nearest neighbours of some verbs do not provide information to indicate clearly which meanings are intended. For example, the differences between the two meanings of the verb *võtma* 'to take' could not be clarified by considering the nearest neighbours of the senses. Similarly, the meanings of the PVs were not presented well. In addition, there were cases in which the sentences expressed exactly the same meaning of the PV, but the system evaluated them as expressing different meanings. For example, examples (60) and (61) containing the PV *üle pakkuma* both express the meaning 'to exaggerate', but the SenseGram suggested that the PV had different meanings. Furthermore, the exact meanings identified by the SenseGram could not be clarified based on the nearest neighbours of the PV.

[85]The pre-trained models for English, German, and Russian are downloadable at `http://ltdata1.informatik.uni-hamburg.de/sensegram/` (accessed 21.11.2018).

[86]`https://en.wikipedia.org/wiki/Mouse` (accessed 21.11.2018).

[87]`https://en.wikipedia.org/wiki/Computer_mouse` (accessed 21.11.2018).

(60) Nüganen **ei** **paku-∅** **üle** liigutus-te-ga
Nüganen NEG offer-CONNEG over movement-PL-COM
Lit. 'Nüganen does not offer over with movements.'
'Nüganen does not exaggerate with his movements.'

(61) Nüüdse-ks on selgu-nud, et see hirm **ol-i**
current-TRL be.3SG appear-PST.PTCP that this fear be-PST.3SG
üle **paku-tud**.
over offer-PST.PTCP
Lit. 'It is clear now, that the fear was over offered.'
'It is clear now, that the fear was exaggerated.'

The number of PV senses the SenseGram suggested did not correlate with the number of senses presented in the EED. Approximately 40% of the PVs had the same number of senses in the EED as suggested by the SenseGram. However, most of these PVs were those that had one meaning in the dictionary. Therefore, the SenseGram did not provide more meanings for the PVs that had a high number of meanings in the EED. For example, *läbi laskma* 'to let through/pretermit', *vastu võtma* 'to accept/welcome/admit' and *kinni pidama* 'to stick to/slow down/detain' have eight, seven, and six meanings, respectively, in the EED, but the best multi-sense embedding model produced one, two and three meanings for these PVs, respectively. Similarly, *välja kuulutama* 'to announce', which has one meaning in the EED, was given six meanings by the model. However, the PV frequency did not have an impact on the number of senses the SenseGram model produced.

The second reason for the poor performance of the multi-sense embeddings might have been the fact that the human-annotated datasets that were used were not designed to evaluate the compositionality predictions of the multi-sense embeddings. More specifically, the PVComp did not differentiate among the senses – each PV had one score that was compared to the CS score for the most likely senses of the PV and the verb vector. Therefore, the comparison of the rankings was based on the assumption that both scores reflected the degree of compositionality of the most predominant meaning; hence, they were comparable. Therefore, the comparison between the predictions and the PVComp was tentative.

When comparing the predictions for the 48 PVs that the best multi-sense model provided one meaning and the PVs that have also one meaning in the EED with the compositionality scores, the correlation is $\rho = 0.18$. It is not higher than was the correlation between the PVComp and the predictions for all the PVs ($\rho = 0.21$), and it was not statistically significant at the $p = 0.05$ level either. Of the 75 ambiguous PVs (the PVs that have more than one meaning in the EED), the multi-sense embedding model provided more than one meaning for 31 PVs. The predictions for those (ambiguous) PVs correlated more weakly with the PVComp ($\rho = 0.08$) than did the predictions for non-ambiguous PVs ($\rho = 0.18$). However, the correlation is not statistically significant. In conclusion, the multi-sense embedding model predicted the compositionality of non-ambiguous PVs substantially better than it did the compositionality of ambiguous ones.

The PVLit distinguished among the meanings of the PVs, and was thus more suitable for the evaluation of the predictions of the multi-sense embedding models.

However, one value was used for the efficient evaluation – the CS value for the most likely senses of the PV and verb was compared to the median of the (non-)literalness scores for the PV. This approach made the evaluation of each model straightforward, but also limited. The additional evaluation of the multi-sense embedding model (see Section 5.5.1) proposed that the results of the multi-sense embedding model were not better when relying on the WSD mechanism of the SenseGram; however, the evaluation was probably more precise. In future research, a gold standard designed to evaluate the predictions of the multi-sense embedding models should be created and employed.

Some work has been done in terms of applying multi-sense embeddings in compositionality studies. For example, Cheng and Kartsaklis (2015) proposed a successful deep compositional distributional model in which syntax-aware, multi-sense word vectors were applied to detect paraphrasing. Kober et al. (2017) investigated whether the disambiguation of word senses was strictly necessary, or whether the meanings of a word (in context) could be disambiguated through composition alone. The performance of single-vector and multi-sense vector models were evaluated using a phrase similarity task. The authors found that single-sense vector models performed as well or better than did multi-sense vector models. These results provide a general background to the studies of multi-sense embeddings to explore compositionality, but are not comparable to the results of the current study.

However, some studies in which multi-sense embeddings have been used to predict compositionality of MWEs have been conducted. For example, Salehi et al. (2015) used among other vector-space models a multi-sense skip-gram model for the first attempt to predict the compositionality of English and German MWEs (that is, English and German noun compounds and English verb particles) with embeddings. They found that single word embeddings are empirically slightly superior to multi-sense embeddings. However, with the dataset containing English verb particles, the multi-sense skip-gram model obtained correlation of $r = 0.51$. The result is not directly comparable with our observations, but offers an overview of how well multi-sense embeddings work for predicting the compositionality of English MWEs. In addition, among other tasks, Köper and Schulte im Walde (2017a) explored multi-sense representations for the prediction of the compositionality of PVs and the literal versus the non-literal usage of PVs. The overall best result of the compositionality predictions improved when the multi-sense embeddings were applied. The correlation between the predictions of the best model and the human-annotated dataset was $\rho = 0.32$. Therefore, the overall results of the compositionality predictions for Estonian and German PVs were similar. However, it is important to note that the experimental setup in the present study differed substantially that of Köper and Schulte im Walde (2017a). For example, they only applied multi-sense learning to their target words. Although multi-sense embeddings were not used to detect the literal versus the non-literal usage of Estonian PVs, the literalness dataset was used to assess the predictions of multi-sense embedding models. A similar dataset for German PVs (Köper and Schulte im Walde 2016b) was used by Köper and Schulte im Walde (2017a) to assess multi-sense embeddings used to predict the literal versus the non-literal usage of German PVs. In order to use the literalness ratings for the evaluation of

the models for Estonian, the intended meaning was identified using SenseGram's WSD mechanism. For the German PVs, the intended meanings were selected by computing a CS value between a verb vector and the vector of all the context words in the sentence. The most similar multi-sense embedding was then selected for the prediction. This approach could be used as an alternative evaluation of the compositionality predictions of Estonian PVs in the future.

In general, the need to distinguish between word senses in DSMs has been noted. The results did not show that the multi-sense embedding models worked better than dud the word embedding models for the detection of compositionality, but further research is definitely necessary. To point out one important direction, the evaluation of such models for predicting the compositionality of Estonian PVs must be revised based on the issues discussed earlier in this section.

5.6 Summary and discussion of detecting the compositionality of particle verbs

This section discusses the main results when using DSMs to predict the compositionality of Estonian PVs. In addition, the impact of the human compositionality ratings on the results are discussed, and the work of the models used for predicting the compositionality of ambiguous and non-ambiguous PVs is compared.

Like other MWEs, PVs form a continuum from fully compositional units to fully non-compositional ones based on their degrees of compositionality. In order to model the compositionality of PVs, the semantic similarity between the PV and its base verb was measured using vector representations trained with DSMs. Two kinds of DSMs were utilised to predict the compositionality of PVs – DSMs that learnt one representation per word and DSMs that were able to learn multiple representations per word. Word embeddings were trained using word2vec, and these word representations were applied to train multi-sense embeddings via SenseGram.

The impact of four parameters – the number of dimensions, the window size, the minimum-count threshold and the number of iterations – on the compositionality predictions of word and multi-sense embeddings was investigated. The influence of the window size and iterations were stronger than was the impact of the number of dimensions and the minimum threshold of word count. In comparison with word embedding models, the multi-sense embedding models are less influenced by the change of parameter values. The comparison of different models revealed that the difference between the models trained with default parameter settings and the best word and multi-sense embedding models was statistically significant. Thus, some guidelines concerning how to improve the quality of embeddings were provided.

For the word and multi-sense embeddings, two word2vec architectures – CBOW and Skip-gram – were trained and compared. The CBOW models performed slightly better for training the word embeddings than did the Skip-gram models. For the multi-sense embeddings, the difference in the predictions between CBOW and Skip-gram architectures were significant. Therefore, depending on the parameter configurations, the Skip-gram architecture might be as good as the CBOW model for training word embeddings, but the CBOW architecture should

be preferred when training multi-sense embeddings. In addition, the frequency of the PVs influenced the predictions of the Skip-gram models more than it did the predictions of the CBOW models.

The best results – the highest correlation with the human compositionality judgements – were achieved by the word embedding model using the CBOW architecture and trained with 300 dimensions, with a window size of 5, a minimum-count threshold of 10 and 20 iterations. The multi-sense embedding model trained with the same parameters, except for the word-count threshold of 5, provided the best predictions of all the multi-sense embedding models that were introduced. Therefore, somewhat surprisingly, the multi-sense embeddings did not outperform the predictions of the word embedding models. However, it is important to bear in mind that the predictions of this model correlated more strongly with the compositionality ratings (PVComp) than did those of any other model. Nevertheless, there are number of clear reasons for the modest performance of the multi-sense embedding models. Most importantly, the human-annotated datasets were not designed to evaluate multi-sense embedding models; thus, the evaluation was somewhat tentative. Therefore, studies of the use of multi-sense embeddings in compositionality studies must continue in the future.

Both the word embedding and the multi-sense embedding models predicted the compositionality of frequent and infrequent PVs much more accurately than they did the compositionality of PVs with moderate frequency. As it was expected that the vectors of frequent PVs would be more representative because more data are associated with them, the result for frequent PVs was not surprising. The frequent PVs tend to be more polysemous; therefore, it could be expected that the compositionality of these PVs would be more difficult to predict than would the compositionality of other PVs. The frequent PVs studied and their verbal components were indeed more polysemous than were the other PVs and their base verbs[88], but the predictions were good for both sets of PVs. However, of the 50 frequent and infrequent PVs, 33 and 18 were ambiguous, respectively, and the predictions for frequent and infrequent PVs were similarly good; thus, it can be assumed that the compositionality of frequent PVs was predicted well due to the representative vectors, and the compositionality of infrequent PVs was predicted well because they were often non-ambiguous. This conclusion is illustrated well by the fact that the multi-sense embedding model's predictions of infrequent, non-ambiguous PVs correlated with the PVLit much more strongly ($\rho = -0.52$) than did the predictions of infrequent, ambiguous PVs ($\rho = -0.17$). At the same time, the word embedding model's predictions of frequent PVs were more similar regardless of the number of senses the PVs had – the correlation between the predictions and the PVComp for ambiguous PVs was $\rho = 0.22$, and $\rho = 0.30$ for non-ambiguous PVs.

The predictions for frequent and infrequent PVs correlated moderately with the frequency of the PVs. The most frequent and most infrequent PVs tended to be assigned lower CS scores than did the slightly more infrequent/frequent ones. The multi-sense embedding model's predictions of infrequent PVs had a

[88]Average number of senses among frequent PVs is 2.4 and the average number of senses of infrequent PVs is 1.5; the average number of verbs of frequent PVs is 8.1, average number of verbs of infrequent PVs is 4.24.

weak correlation with frequency, but the predictions were still relatively good. Hence, it can be said that the quality of the predictions for infrequent PVs was not influenced by the frequency. However, the PVs with moderate frequency were not influenced by the frequency, and were not predicted as well as were the infrequent and frequent PVs. There were 23 ambiguous PVs and 27 non-ambiguous PVs with moderate frequency – the models predicted the compositionality of the ambiguous PVs better than they did the compositionality of non-ambiguous PVs. However, as there was an almost equal number of ambiguous and non-ambiguous PVs, the overall quality of the predictions was not high. Moreover, it can be assumed that the vectors were not as representative as were the vectors of the frequent PVs; therefore, the quality of the predictions suffered. Furthermore, when dividing the PVs with moderate frequency into two categories – frequent and infrequent ones – the predictions of frequent PVs correlated better with the human judgements than did the predictions of infrequent PVs. To conclude, the vector representations of PVs with moderate frequency were not good enough and could not compensate for the difference between the predictions of ambiguous and non-ambiguous PVs.

The compositionality predictions of the DSMs trained to learn word and multi-sense embeddings were evaluated against two human-annotated rankings. The first was based on the averaged compositionality scores of PVs (see Section 4.2), while the other ranking was based on the averaged literalness scores for the PV meanings (described in Section 4.3). The compositionality ratings were collected in order to evaluate the results of the word embedding models. However, the predictions of the word embedding models did not correlate strongly with the human compositionality judgements – the highest correlation that the best word embedding models attained was $\rho = 0.19$. The most probable reason for the relatively low correlation was that both ambiguous and non-ambiguous PVs were included in the study, but the predictions and human-annotated datasets only included one score for a PV regardless of the number of senses it might have. However, the correlation between the predictions and the compositionality scores did not improve significantly when only the compositionality of the non-ambiguous PVs was predicted. The correlation between the predictions and the compositionality scores for the non-ambiguous PVs was $\rho = 0.21$. At the same time, the correlation amongst the non-ambiguous PVs was $\rho = 0.16$. None of these correlations was statistically significant. The best multi-sense embedding models achieved similar, but slightly weaker correlations with the compositionality ratings – $\rho = 0.21$. The most likely senses that the model would predict were used for the evaluation. However, the difference between the predictions for ambiguous and non-ambiguous PVs was not significant for the multi-sense embedding models – the correlation between the predictions for non-ambiguous PVs and compositionality scores was weak ($\rho = 0.18$), and almost non-existent ($\rho = 0.08$) for ambiguous PVs.

Another human-annotated dataset containing literalness (compositionality) scores for PVs was originally designed for the task of classifying the literal versus the non-literal usage of Estonian PVs. However, the scores distinguished amongst different meanings of the PVs, and could therefore have been more suitable for the evaluation of the multi-sense embedding models. However, the predictions of both models correlated better with these scores. The best word embedding model obtained a correlation of $\rho = -0.46$, and the best multi-sense embedding model

obtained a correlation of $\rho = -0.35$. However, the evaluation was conducted based on one score for each PV, which was an average literalness score assigned to one PV by the human annotators. The reason for such an evaluation was that the detection of the intended meanings of each sentence was complicated and the result would be dependent on the quality of the WSD mechanism of SenseGram, while the quality of its use for Estonian data was unknown. Compared to the literalness dataset, the word embedding model predicted the non-ambiguous words significantly better ($\rho = -0.51$) than it did ambiguous ones ($\rho = -0.39$); the difference was not as substantial for the multi-sense embedding model ($\rho = -0.34$ versus $\rho = -0.31$).

The study of word and multi-sense embeddings for predicting the compositionality of Estonian PVs confirmed the complexity of the task of compositionality detection. The predictions of the word embedding models correlated relatively well with the compositionality scores, but the results depended on the frequency of the PVs and their ambiguity. In addition, the evaluation of the multi-sense embedding models was poor, probably because they were evaluated against human judgements that were not collected for the assessment of such models. Therefore, the evaluation of the multi-sense embedding models could be improved by using different methods and resources.

6 DETECTING THE LITERAL AND NON-LITERAL USAGE OF PARTICLE VERBS

This section describes the experiments on the identification of PVs via automatic detection of the literal versus the non-literal usage of Estonian PVs. A theoretical overview of supervised learning and the methods applied is presented in Section 3.3.

Previous research on the automatic detection of non-literal language usage has mainly been undertaken in English and German. While general indicators for identifying non-literal languages have been studied extensively, there is little work on language-specific directions. For example, Tsvetkov and Wintner (2014) studied language-independent features such as abstractness and imageability, semantic categories from WordNet and vector space word representations to detect metaphors in English, Spanish, Farsi and Russian. Köper and Schulte im Walde (2016b) distinguished between the literal and non-literal usage of German PVs using standard features such as affective ratings and unigrams combined with PV-specific information. The present study combines standard, PV- and language-specific features in order to detect the non-literal meanings of Estonian PVs.

Aedmaa et al. (2018) described the first experiment for the automatic detection of the non-literal usage of Estonian PVs. The basic model was adapted from Köper and Schulte im Walde (2016b) and was constructed including language-specific features. Compared to the earlier study, the present research is more exhaustive in terms of the description of the automatic detection of the literal versus the non-literal usage of Estonian PVs. This research analyses the impact of the features on the results, and introduces frequency as predictive of non-literal language.

This chapter is organised in the following way. In Section 6.1, the experimental setup and evaluation are described. This is followed by an overview of the features suggested being useful for the detection of the literal versus the non-literal usage of Estonian PVs in Section 6.2. The results of the experiments and the usefulness of each feature are analysed in Section 6.3. Section 6.4 studies frequency as a predictor of the (non-)literal usage of Estonian PVs. The results are discussed and conclusions drawn in Section 6.5.

6.1 Experimental setup and evaluation

The aim of the experiments was to develop a classifier that could predict the literal versus the non-literal usage of Estonian PVs. The model was trained on a dataset that included (non-)literalness ratings for 184 Estonian PVs in 1,481 sentences (see Section 4.3). The selection of PVs and sentences is described in the same section. Based on the averaged (non-)literalness ratings, the sentences were divided into two categories, namely literal and non-literal. Literal sentences are sentences with an average (non-)literalness score of 0–2.33, and non-literal sentences have a score of 2.67–5. The classifier predicted the class of the sentences using a set of features discussed in Section 6.2.

Although there are many machine learning models, the random forest classifier was utilised because this method has been proven to work well for linguistic data.

For example, the random forest classifier has been applied successfully to the detection of the compositionality of German PVs (e.g. Köper and Schulte im Walde 2016b). The random forest learning method is described in Section 3.3.1.

Waikato Environment for Knowledge Analysis (Weka[89]) software (Frank et al. 2016) was used for the machine learning algorithms. As it contains tools for data preparation, classification, regression, clustering, association rules mining, and visualisation, some data pre-processing was also conducted using Weka. Specifically, the missing values were addressed using the NumericCleaner filter[90].

The evaluation of the models was conducted using 10-fold cross-validation (see Section 3.3.3). The values for the overall classification accuracy and the F_1 scores for each model are presented. The F_1 score is a widely used measure in machine learning, and expresses the harmonic mean of recall and precision (see Section 3.4.3). The F_1 scores for non-literal (F_1 n-lit) and literal (F_1 lit) classes are provided. The results are compared to the majority baseline. The majority baseline was calculated by applying the Zero Rule method, which predicts the non-literal class as the result of all predictions (simply because it has more observations than does the literal class). Weka class ZeroR[91] was applied to determine the majority baseline value.

6.2 Features

This section introduces a set of features that were suggested to detect the literal versus the non-literal usage of Estonian PVs. The values of these features were annotated for all 1,481 sentences that were classified based on the average (non-)literalness ratings described in Section 4.3. Annotated sentences formed the dataset of (non-)literal ratings for Estonian PVs (Aedmaa 2018). Some of the features were adopted from previous studies of the automatic prediction of metaphorical and literal language usage, mainly from Turney et al. (2011) and Köper and Schulte im Walde (2016b). Other features were motivated by the specific features of the Estonian language. The introduced features were applied later (see Section 6.3) in different combinations to develop the best classifier for the automatic detection of the literal versus the non-literal usage of Estonian PVs.

6.2.1 Abstractness ratings

In this section, four abstractness features – the average abstractness ratings of all words, the average abstractness ratings of all nouns (excluding proper nouns), the abstractness ratings of the subjects and the abstractness ratings of the objects – are introduced. The abstractness and concreteness ratings have been used in previous work on the detection of non-literal language usage. The present research is based on the hypothesis that the degree of abstractness of the context influences the

[89] `https://www.cs.waikato.ac.nz/~ml/weka/index.html` (accessed 02.03.2017).

[90] `http://weka.sourceforge.net/doc.dev/weka/filters/unsupervised/attribute/NumericCleaner.html` (accessed 02.05.2018).

[91] `http://weka.sourceforge.net/doc.dev/weka/classifiers/rules/ZeroR.html` (accessed 24.10.2018).

literalness of the word. The selection of abstractness features in this work was largely motivated by two studies. Turney et al. (2011) presented an algorithm to classify a word sense as metaphorical or literal in a given context, and Köper and Schulte im Walde (2016b) distinguished between the literal and the non-literal usage of German PVs.

Unlike previous studies, the average abstractness ratings for all words are included as one of the features. This decision was motivated by the fact that there are no reports on this feature in the related research, but it represents adequately the aforementioned hypothesis that the usage of a word depends on the degree of abstractness of other words in the context (see Section 4.5). In addition, Köper and Schulte im Walde (2016b) reported that features such as the average abstractness rating of proper names/verbs/adjectives/adverbs added few or no additional information to their dataset. Thus, it was decided to create one instead of four features, including words with different POS. This feature indicates the average abstractness rating of all lemmas for each sentence that can be found in the dataset of abstractness/concreteness ratings.

The average abstractness rating of all the nouns was adopted from Turney et al. (2011), whose targets were adjective-noun combinations and verbs. In their experiment, the abstractness of the nouns had the largest weight in predicting whether the target verb was used metaphorically or not. Moreover, according to the information obtained, the abstractness of the nouns was one of the most salient features for distinguishing between the literal and the non-literal usage of German PVs. Overall, features that depended on nouns, such as the common nouns in the context, and nouns marking a subject and an object were more useful than were the features that contained information about other POS. (Köper and Schulte im Walde 2016b) Accordingly, the abstractness ratings for subjects and objects as potentially helpful features for predicting the literal versus the non-literal usage of Estonian PVs were also studied.

Apart from the evidence provided in previous work, the impact of the degree of abstractness of the surrounding words is also presumable when examining Estonian data. For example, examples (62)–(65) illustrate that literal sentences (see examples (62) and (64)) include more concrete words than do non-literal sentences (see examples (63) and (65)). According to the abstractness/concreteness dataset (see Section 4.5), the words *sõber* 'friend' (with an abstractness score of 5.9) and *koer* 'dog' (with an abstractness score of 8.5) are much more concrete than are the words *surm* 'death' (with an abstractness score of 4.2) and *viha* 'anger' (with an abstractness score of 1.7).

(62) **Sõber** jooks-is mu-lle järele.
friend run-PST.3SG I-ALL after
'A friend ran after me.'

(63) **Surm** jooks-is ta-lle järele.
death run-PST.3SG he-ALL after
Lit. 'The death ran after him.'
'Death was following him.'

(64) Mees suru-s **koera-∅** maha.
man push-PST.3SG dog-GEN down
'The man pushed the dog down.'

(65) Mees suru-s **viha-∅** maha.
man push-PST.3SG anger-GEN down
Lit. 'The man pushed his anger down.'
'The man suppressed his anger.'

Related research and evidence from the data led to the assumption that the abstractness of the context of PVs could help to predict the literal versus the non-literal usage of Estonian PVs. Four abstractness features proposed in this section are included in the feature space, and their impact is studied in depth in Section 6.3.

6.2.2 Cases of subject and object

Estonian distinguishes between 'total' subjects in the nominative case and 'partial' subjects in the partitive case. Partial subjects are not in subject-predicate agreement (Erelt et al. 1993). For example, the subjects *külaline* 'guest' and *naine* 'woman' are in the nominative case in examples (66) and (68). The subjects are in the partitive case in examples (67) and (69). It was observed that subject case assignment might correlate with (non-)literal readings. For example, the meaning of the PV *juurde tulema* is non-literal in example (67), but literal in example (66).

(66) **Külaline** tule-b juurde.
guest come-3SG by
Lit. 'The guest is coming by.'
'The guest is approaching.'

(67) **Külalis-i** tule-b juurde.
guest-PL.PRT come-3SG by
Lit. 'The number of guests is coming by.'
'The number of guests is increasing.'

(68) **Naine** lähe-b peatuse-s maha.
woman go-3SG stop-INE off
Lit. 'The woman goes off at the stop.'
'The woman gets off at the stop.'

(69) **Vett-∅** ei läi-nud maha.
water-PRT NEG go-PST.PTCP down
Lit. 'Water did not go down.'
'Water did not fall down.'

Similarly, a 'total' object in Estonian receives the nominative or genitive case, and a 'partial' object receives the partitive case. For example, the object in example (70) is in the genitive case (*supi*), but the object in example (71) is in the partitive case (*mida*). There is a difference in the meaning of the PV *ette võtma* in these sentences – the meaning of the PV is more literal in example (70) than it is in example (71). These kinds of examples found in the text corpora are evidence that the object case might influence the literal versus the non-literal usage of PVs.

(70) Tüdruk võt-tis ette **supi-∅**.
girl take-PST.3SG front soup-GEN
'The girl took the soup in front of her.'

(71) **Mi-da** koos ette võt-ta?
what-PRT together ahead take-INF
Lit. 'What should we take ahead together?'
'What should we do together?'

The impact of the subject/object case on the (non-)literal meaning has not been examined in theoretical linguistics. Therefore, it is not possible to provide further evidence regarding the associations between the subject/object case and the formation of a PV's meaning. Nevertheless, the first study of the automatic detection of the literal versus the non-literal usage of Estonian PVs demonstrated that the subject case contained information that was useful for the task (Aedmaa et al. 2018). Both features are thus included in the current, more exhaustive study, which tests whether the case distribution predicts (non-)literal language usage. The impact of the case features is analysed in Sections 6.3.4.5 and 6.3.4.6.

6.2.3 Subject and object animacy

When nouns or verbs are used metaphorically, either semantic or syntactic principles are often violated. For example, the verb 'to eat' requires an animate subject and an edible direct object; thus, the sentence 'Tom ate the apple' is semantically acceptable, but the sentence 'The desktop printer ate the paper' is not. The usage of the verb 'to eat' is literal in the first sentence and non-literal in the second sentence. (Glucksberg et al. 2001)

Similarly, the meaning of the PV can determine the animacy of its subject(s) and object(s). For example, if the PV meaning requires an animate subject but the subject is inanimate, the meaning of the sentence might become non-literal. For example, the PV in examples (72) and (73) is the same (*sisse kutsuma* 'to invite in'), but the subject *sõber* 'friend' is animate and the sentence is literal in the first sentence, while the subject *maja* 'house' in example (73) is inanimate and the sentence is non-literal. Similarly, the subject *naine* 'woman' in example (74) is animate and the sentence is literal, while the subject *välimus* 'appearance' in example (75) is inanimate and the sentence is non-literal.

(72) **Sõber** kutsu-s mu-∅ sisse.
friend invite-PST.3SG I-GEN inside
'The friend invited me in.'

(73) **Maja** ei kutsu-∅ sisse.
house NEG invite-CONNEG inside
Lit. 'The house does not invite in.'
'The house doesn't look inviting.'

(74) **Naine** tõuka-s mehe-∅ eemale.
woman push-PST.3SG man-GEN away
'A woman pushed a man away.'

(75) Poe-∅ **välimus** tõuka-s mehe-∅ eemale.
shop-GEN appearance push-PST.3SG man-GEN away
Lit. 'The appearance of the shop pushed the man away.'
'The appearance of the shop made the man go away.'

The impact of object animacy on the meaning of the PVs is less intuitive. However, some evidence for the significance of object animacy can be found in the Estonian language. Firstly, the literal meaning of the PV *läbi põletama* is 'to fuse something' and the object is expected to be inanimate. The non-literal meaning of the PV is similar to the verb *läbi põlema* 'to burn out', and requires an animate object. For example, the object in example (76) is inanimate and the meaning of the PV is literal, while the object in example (77) is animate and the meaning of the PV is non-literal.

(76) Mees põleta-s **kaitsme-∅** läbi.
man burn-PST.3SG fuse-GEN out
Lit. 'The man burned the fuse out.'
'The man burned the fuse.'

(77) Mees põleta-s **enese-∅** läbi.
man burn-PST.3SG himself-GEN out
Lit. 'The man burned himself out.'
'The man had a burnout.'

Based on examples (76) and (77), one can assume that the abstractness scores already indicate the (non-)literal usage of the PV – the inanimate words in the literal sentence are concrete, and the animate words in the non-literal sentences are abstract. Nevertheless, as in examples (72) and (73), the subject of the literal sentence can represent a more abstract (and animate) word than can the

concrete inanimate subject of the non-literal sentence. Thus, it is argued that the abstractness ratings are also insufficient to express the animacy of the words.

The association between subject/object animacy and (non-)literalness has not been examined in depth previously. Nevertheless, there is sufficient evidence in the data to indicate that the subject and object animacy provide useful information for predicting the literal versus the non-literal usage of Estonian PVs. Furthermore, the first experiment to predict the (non-)literalness of Estonian PVs (see Aedmaa et al. 2018) suggested that the animacy of the subject and object deserved thorough research as classification features for predicting a PV's (non-)literal language usage. The impact of the animacy features is broken down in Sections 6.3.4.7 and 6.3.4.8.

6.2.4 Case government

Case government is a phenomenon in which the lexical meaning of the base verb affects the grammatical form of the argument, for example, the predicate determines the case of the argument (Erelt et al. 1993). Thus, the case of the argument depends on the meaning of the PV. For example, in example (78), the PV *läbi minema* 'to go through' is literal and requires an argument that answers the question 'from where?'. Hence, the argument has to be in the elative case. In example (79), the PV has the non-literal meaning 'to succeed' and does not require any additional arguments. Similarly, the meaning of the PV *järele vaatama* differs in examples (80) and (81). The meaning of the PV 'to follow somebody with one's eyes' requires an argument in the allative case in example (80). The PV argument is in the elative case, suggesting that the PV is used in the sense of 'to look something up' in example (81). Moreover, the degree of (non-)literalness of the different meanings of the same PV vary.

(78) Ta läk-s **metsa-st** läbi.
s/he go-PST.3SG forest-ELA through
'S/he went through the forest.'

(79) Mu-∅ ettepanek läk-s läbi.
I-GEN proposal go-PST.3SG through
Lit. 'My proposal went through.'
'My proposal was successful.'

(80) Ta vaata-s **ema-le** järele.
s/he look-PST.3SG mother-ALL after
Lit. 'S/he looked after mother.'
'Her/his eyes followed mother.'

(81) Ta vaata-s **arvuti-st** järele.
s/he look-PST.3SG computer-ELA after
Lit. 'S/he looked after from the computer.'
'S/he looked the information up on the computer.'

Aedmaa et al. (2018) proposed that the case argument provided information to predict the literal versus the non-literal usage of Estonian PVs; hence, the feature has been analysed in depth in the current study. The impact of case government on the classification task is described in Section 6.3.4.9. There are fourteen cases[92] in Estonian; thus, the argument is one of those cases or does not appear at all. Note that, as the cases of the subject and object are distinct features, only the cases of other types of arguments, such as adverbials and modifiers, are annotated as values of this feature.

6.3 Results for the classification of literal and non-literal usage

This section provides an overview of the results of the automatic classification of the literal versus the non-literal usage of PVs as follows: First, the feature selection experiments are carried out to ascertain which features are automatically selected as the most relevant. In the second section, we describe how each studied feature operates independently of the effect of any other feature is described. The results of the feature combinations with different sizes are then provided. The impact of each feature is analysed in ablation study with a focus on the sentences classified incorrectly by the models. The overall classification accuracy and f-scores (F_1) for both classes (literal and non-literal) are presented for each model.

6.3.1 Results for the feature selection

The results of attribute selection are presented in this section. The goal of attribute selection was explained previously in Section 3.3.2. As relatively few features were studied in this thesis, automatic attribute selection could have been avoided. However, as one of the purposes of the feature selection process is a better understanding of the data and features (Chandrashekar and Sahin 2014), the attribute selection can provide an appropriate introduction to the following analysis. The attribute selection was implemented using three different techniques for the feature selection available via Weka – a correlation-based approach, Information Gain, and a learner-based feature selection.

6.3.1.1 Correlation-based feature selection

Correlations refer to the Pearson correlation coefficient between each attribute and the output (literal and non-literal usage). A correlation is high when the value is close to 1 or −1, and low when it is close to 0. A choice of attributes can be made based on the correlation value. The correlation was calculated using Weka's CfsSubsetEval[93] technique, which requires the use of the Greedy Stepwise search method. The CfsSubsetEval evaluates the work of a subset of features by

[92] Nominative, genitive, partitive, illative, inessive, elative, allative, adessive, ablative, translative, terminative, essive, abessive, comitative.

[93] `http://weka.sourceforge.net/doc.dev/weka/attributeSelection/CfsSubsetEval.html` (accessed 10.05.2018).

considering the individual predictive ability of each feature, together with the degree of redundancy amongst them. The correlation feature selection (CFS) is based on the hypothesis that a good feature set contains features that are correlated significantly with the output class, yet uncorrelated with each other (Hall 1998). 10-fold cross-validation was implemented, and the results of the experiment are shown in Table 24.

Table 24: Results for the correlation-based feature selection.

number of folds	feature
0	particle
10	verb
10	unigrams
0	the average abstractness of words
5	the average abstractness of nouns
0	subject abstractness
3	object abstractness
0	subject case
0	object case
0	subject animacy
0	object animacy
0	case government

CFS revealed four features that appeared at least in one fold in the best subset of features, namely the verb, the unigram, the average abstractness of the nouns and the object's abstractness. Verbs and unigrams were among the best features in all 10 folds, the average abstractness of nouns in five folds, and the object abstractness in three folds. It can therefore be assumed that verbs and unigrams did not correlate well with each other, but they correlated well with the predictable class – the literal versus the non-literal usage of PVs.

6.3.1.2 Information Gain

Another method of selecting features is to calculate the information gain (entropy) for each attribute in relation to the class. A value of 0 indicates no information and 1 indicates maximum information. The attributes that contributed more information had a higher information gain value, while the attributes with a lower score did not add as much information. Weka's InfoGainAttributeEval[94] class was employed using the Ranker search method. This approach evaluates the worth of the features by measuring the information gain with regard to the class (the literalness of the sentences). The results of the 10-fold cross-validation are presented in Table 25.

[94] `http://weka.sourceforge.net/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html` (accessed 11.11.2017).

Table 25: Results for the information gain.

average merit	feature
0.093 +− 0.003	particle
0.367 +− 0.007	verb
0.164 +− 0.006	unigrams
0.034 +− 0.003	the average abstractness of words
0.069 +− 0.006	the average abstractness of nouns
0.007 +− 0.003	subject abstractness
0.015 +− 0.002	object abstractness
0.001 +− 0.000	subject case
0.003 +− 0.001	object case
0.003 +− 0.001	subject animacy
0.001 +− 0.000	object animacy
0.032 +− 0.002	case government

The results of the information gain are in accordance with the results of the correlation-based feature selection (described in Section 6.3.1.1) – the verb and unigram features contributed more information than did the other features. In addition, compared to the other features, the particle and the average abstractness of the nouns contributed more information than the other features.

6.3.1.3 Learner-based feature selection

The third feature selection technique used was a learner-based one. As the random forest classifier was implemented for the classification task, the same method was used to evaluate the dataset with different subsets of attributes. The WrapperSubsetEval[95] technique in Weka was adopted, and the BestFirst[96] search method[97] was applied. The results of the 10-fold cross-validation that was used to estimate the accuracy of the learning scheme for the set of features if presented in Table 26.

The learner-based feature selection technique indicated 10 features that appeared in at least one fold in the best subset of features. Particles, verbs, unigrams, subject animacy and object animacy constituted to the best subset of features in all 10 folds. The average abstractness of the nouns and case government were included in the best subset in nine folds, the subject abstractness and subject case in three folds, and the object case in one fold. According to the learner-based feature selection, the average abstractness of words and object abstractness are not relevant for detecting the literal versus the non-literal usage of Estonian PVs.

[95] `http://weka.sourceforge.net/doc.dev/weka/attributeSelection/WrapperSubsetEval.html` (accessed 11.11.2017).

[96] `http://weka.sourceforge.net/doc.dev/weka/attributeSelection/BestFirst.html` (accessed 11.11.2017).

[97] The GreedyStepwise search method was also tested – the results were the same.

Table 26: Results for the learner-based feature selection.

number of folds	feature
10	particle
10	verb
10	unigrams
0	the average abstractness of words
9	the average abstractness of nouns
3	subject abstractness
0	object abstractness
3	subject case
1	object case
10	subject animacy
10	object animacy
9	case government

6.3.1.4 Summary of the feature selection results

The aim of the feature selection was to test different feature selection techniques and ascertain the features that automatic feature selection methods select as the most relevant ones.

Both filter methods – the correlation-based feature selection and the information gain – chose the verb and unigram as the most relevant features. The correlation-based method also selected the average abstractness of the nouns and object abstractness as somewhat relevant features. The information gain suggested that the particle and the average abstractness of nouns contributed more information than did the other features (except for verbs and unigrams).

The wrapper method – a learner-based feature selection – chose more features than did the filter methods. According to this method, the most relevant features are particle, verb, unigrams, subject animacy, object animacy, average abstractness of nouns and case government. The average abstractness of words and object abstractness was not selected once and are therefore not relevant for detecting literal vs. non-literal usage of Estonian PVs.

In summary, all the feature selection methods highlighted the most relevant features for the dataset, and provided information that was useful for a better understanding of the data. Feature selection should be used whenever the feature set is large and the aim is to identify the most useful features within a short period. The results of the feature selection are not adopted directly for the current research because the aim was to inspect the impact of the studied features on the classification results. However, the results of the feature selection are visited in comparison with the classification results of the features and their combinations in the following sections.

6.3.2 Results for the independent features

The aim of the classification experiments was to determine the best combination of features in order to distinguish between the literal versus the non-literal usage of Estonian PVs. The first phase aimed to determine how well the studied features worked independently of information obtained from other features. The results of these experiments are presented in this section.

Table 27 shows how well the studied features performed individually. Three different settings are presented for the unigrams. The first represents a setting in which all the unigrams (1,518 unique lemmas) were independent features. The second setting reveals the results for the unigrams' features when only unigrams with a frequency greater than 5 were taken into account. After running experiments with various settings, the unigram feature that considered only those unigrams with a minimum-count threshold of 6 provided the best results. The threshold is optimal because only a few unigrams were omitted. For the third result, the unigram feature was implemented as one feature instead of considering every unigram to be with a frequency of more than 5 as an independent feature[98]. The unigram feature was implemented as the output of a random forest classifier when the result was either a literal or a non-literal sentence. This also helped to counter the data sparseness. In NLP, data sparsity is a term referring to the phenomenon of not having sufficient data to model language accurately. This means that accurate observations about the distribution and pattern of language cannot be made because of insufficient data. (Allison et al. 2006) The same setting is referred to hereafter whenever discussing the unigram feature.

The unigram feature received the highest overall accuracy of 82.8% over a baseline of 74.0%. It classified the greatest number of non-literal sentences correctly, with an f-score of 89.3. The base verb predicted the highest number of literal sentences correctly, achieving an F_1 score of 59.4. The case government and abstractness of the object also outperformed the majority baseline. However, the f-score for non-literal usage was higher than the baseline (85.1) for the unigram and base verb features. All the features, except for the subject case, object case, subject animacy and object animacy, exceeded the majority baseline for literal sentences. Of the abstractness features, the abstractness rating of the object gave the best results. For the literal usage, the highest f-score was for the average abstractness of nouns.

In comparison with the feature selection described in 6.3.1, the results are not surprising because feature selection results indicated that verb and unigram are more useful features than others. However, based on the feature selection, it could have been assumed that particle, average abstractness of nouns, subject animacy and object animacy perform better than the rest of the features. However, the correlation-based feature selection (see Section 6.3.1.1) indicated that the average abstractness of nouns and object abstractness are more relevant features than other abstractness features. As an independent feature, the object abstractness achieved greater f-score for non-literal usage than did any other abstractness feature. The average abstractness of nouns obtains higher f-score for literal usage than any other feature (except verb and unigram).

[98]Hence the singular form 'unigram'.

Table 27: Classification results for independent features.

	feature type	size	accuracy %	F_1 n-lit	F_1 lit
	majority baseline	0	74.0	85.1	0.0
1	particle	1	74.0	84.6	17.6
2	base verb	1	82.0	88.4	**59.4**
3	unigram	1,518	81.2	88.6	46.3
3	unigram, f>5	406	**82.8**	**89.3**	56.0
3	unigram, f>5	1	**82.8**	**89.3**	56.0
4	the average abstractness of words	1	69.0	80.4	26.3
5	the average abstractness of nouns	1	68.7	79.9	30.0
6	rating of the PV subject	1	72.4	83.4	17.7
7	rating of the PV object	1	74.3	84.9	15.2
8	subject case	1	74.0	85.1	0.0
9	object case	1	74.0	85.1	0.0
10	subject animacy	1	74.0	85.1	0.0
11	object animacy	1	74.0	85.1	0.0
12	case government	1	74.1	84.9	9.9

Overall, these results suggest that the base verb and unigram features are the best single performing features for predicting the literal versus non-literal usage of PVs. Considering that multiple features can have a similarly important effect on the results of the models, the need for further testing with combinations of features is self-explanatory. The next section introduces the results of the models in which information from multiple features was combined.

6.3.3 Results of the combinations of features

The most substantial results of the models combining information from several features are presented in this section. The aim was to examine the role of each studied feature in predicting the correct class of literal and non-literal sentences. In order to determine the importance of each feature, the results of the models are reported according to their feature space size; that is, the number of features used in the model. The aim of the first section is to determine whether the single best-performing features – the unigram and base verb – worked in combination better than did any other 2-feature combinations. Therefore, all the models containing a maximum of two features were compared. The development of the best model continued by combining the two most salient features with other suggested features. Accordingly, the best random forest classifier predicting the literal versus the non-literal usage of Estonian PVs is suggested in this section.

6.3.3.1 Fundamental features

As demonstrated previously (see Section 6.3.2), the best individual features for classifying the literal versus the non-literal usage of PVs are the base verbs and the unigram feature. The goal of this section is to ascertain whether these two features provide better predictions together than any other 2-feature combinations do. Table 28 presents the results of the best models using a maximum of two features.

Table 28: Results for the best 2-feature classifiers.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–2	2	**85.8**	**90.7**	**69.7**
1, 3	2	83.3	89.6	56.9
2, 3	2	84.7	90.3	64.1
2, 10	2	83.2	89.0	64.2
2, 12	2	84.3	89.8	66.0
3, 7	2	83.1	89.4	58.2

The model merging the best individual features – the base verb (2) and unigram (3) – obtained an accuracy of 84.7%. Therefore, the combination of these two features worked better than the features did separately. However, the best predictions were made by the combination of the particle (1) and verb – the overall accuracy was 85.8%, the f-score for non-literal sentences was 90.7, and 69.7 for literal sentences. Compared to the verb alone, the particle added information that increased the F_1 for literal usage by more than 10 points. These results demonstrate that information about the PV components is sufficient to classify a high number of literal and non-literal sentences correctly. Therefore, this setup forms an essential combination for the study, and these features are included in all the studied combinations introduced from this point onwards[99].

From the results of other combinations, it can be seen that the second-best F_1 lit score (66.0) was obtained by the combination of verb and case government (12). Thus, it can be assumed that information about the case government is particularly useful for predicting literal sentences. At the same time, the unigram feature seemed to contribute less to predicting the literal usage.

Taken together, these results suggest that using only the information about particles and verbs – the components of the PVs – results in a high number of correctly classified sentences. The influence of the particle and verb on the classification results is discussed further in Section 6.3.4.2.

[99]All models without one or both of these two features were also trained, and the results reinforced the importance of particle and verb information.

6.3.3.2 Results for the combinations of all features

This section explores the models containing all the possible feature sets of different sizes. In order to determine how the models with various feature set sizes perform, the results of the best-performing models for each feature set size from 3 to 12 are presented. All the suggested models used information about the particle and the verb.

Table 29 presents the results of the best 3-feature classifiers. The accuracy of the essential combination (1–2, particle and verb) was 85.8% (see Table 28). The accuracy of this model increased when adding the average abstractness of nouns (5), subject animacy (10), object animacy (11) or case government (12) to the feature set. The F_1 n-lit of the combination of fundamental features was 90.7 which increased only when the particle and verb were combined with the object animacy (91.2) or case government (91.1). The highest F_1 score for literal usage (72.6) was achieved by the combination of the particle, verb and average abstractness of nouns (1–2, 5). This score outperformed the result of the particle-verb combination by 2.9 points. In combination with the particle and verb, the unigram feature (3), average abstractness of words (4), subject case (8) and object case (9) did not improve the overall best result.

Table 29: Results for the best 3-feature classifiers.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–3	3	84.5	89.8	67.8
1–2, 4	3	83.7	88.9	68.7
1–2, 5	3	86.0	90.6	**72.6**
1–2, 6	3	85.4	90.2	70.8
1–2, 7	3	85.6	90.4	70.4
1–2, 8	3	85.1	90.3	68.6
1–2, 9	3	84.3	89.7	67.7
1–2, 10	3	85.9	90.7	70.9
1–2, 11	3	**86.6**	**91.2**	72.0
1–2, 12	3	86.4	91.1	71.5

Of the models with feature size 3, the highest overall accuracy and F_1 for non-literal usage were achieved when the fundamental features were combined with object animacy (11). The most successful 3-feature model predicting the literal usage was the one in which the particle, verb and average abstractness of nouns were combined. These findings, while preliminary, suggest that, in addition to the particle and verb, the average abstractness of nouns and object animacy are the most relevant features for predicting the literal versus the non-literal usage of Estonian PVs.

Table 30 shows the results of the best 4-feature classifiers. The highest overall accuracy (87.0%) was obtained by the combination of the particle, verb, average abstractness of nouns and subject animacy (1–2, 5, 10). Compared to the best 3-feature combination, the accuracy was 0.4% higher. The same setup also obtained the highest f-scores – the F_1 for predicting non-literal usage was 91.3, and 74.0 for literal usage. This outcome was a little surprising because the best 3-feature classifier (see Table 29) included the object animacy (11) instead of the subject animacy (10). Thus, it can be hypothesised that the average abstractness of nouns provided information that combined better with information about subject animacy, not object animacy. Furthermore, when the average abstractness of nouns was added to the best 3-feature combination (1–2, 11), the overall accuracy did not change, but the f-score for non-literal usage decreased and increased for literal usage.

Table 30: Results for the best 4-feature classifiers.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–4	4	83.9	89.1	68.8
1–3, 5	4	86.0	90.6	72.6
1–2, 4–5	4	86.8	91.2	73.7
1–3, 6	4	85.3	90.1	71.0
1–2, 5–6	4	86.3	90.8	73.0
1–3, 7	4	84.6	89.7	69.2
1–2, 5, 7	4	85.6	90.3	71.5
1–3, 8	4	85.2	90.3	69.0
1–2, 5, 8	4	86.5	90.9	73.5
1–3, 9	4	83.4	89.0	66.4
1–2, 5, 9	4	86.4	90.9	73.1
1–2, 8–9	4	84.3	89.6	68.1
1–3, 10	4	86.0	90.7	71.5
1–2, 5, 10	4	**87.0**	**91.3**	**74.0**
1–3, 11	4	85.6	90.5	70.0
1–2, 5, 11	4	86.6	91.0	73.6
1–2, 10–11	4	86.4	91.0	72.5
1–3, 12	4	85.8	90.6	70.8
1–2, 5, 12	4	86.7	91.1	73.6

The animacy features and the average abstractness of nouns seemed to be the features that contributed more to the classification of the literal versus the non-literal usage of the PVs. The comparison with the best 3-feature classifiers indicated that using four features instead of three improved the results. This outcome also added certainty to the notion that, in combination with the fundamental

features, the average abstractness of nouns was a more influential feature than were the other suggested features.

The correlation-based feature selection (see Section 6.3.1.1) suggested that four features (the verb, the unigram, the average abstractness of the nouns and the object's abstractness were the only relevant features for the task. However, the accuracy of the model combining these four features was 82.8%, which is not comparable to the accuracy of the best 4-feature combination. Therefore, the correlation-based feature selection could be not sufficient to ascertain the best classifier for detecting the literal versus the non-literal usage of PVs.

The best results of the 5-feature classifiers are presented in Table 31. The best 5-feature model is not surprising when considering the results of the best 4-feature classifier. The combination of the fundamental features, average abstractness of nouns, subject animacy and case government (1–2, 5, 10, 12) produced 0.9% greater accuracy (87.9%) than did the best 4-feature combination. The same setup also obtained the best f-scores – 91.0 for non-literal and 75.4 for literal usage.

Table 31: Results for the best 5-feature classifiers.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–5	5	86.2	90.8	72.3
1–3, 5–6	5	86.2	90.7	72.6
1–2, 4–6	5	86.7	91.2	72.9
1–2, 5–7	5	86.1	90.7	72.0
1–3, 5, 8	5	87.1	91.4	74.4
1–2, 4–5, 9	5	85.8	90.5	71.1
1–2, 5–6, 9	5	86.2	90.8	72.1
1–3, 5, 9	5	85.4	90.2	71.2
1–2, 5, 8–9	5	86.6	91.0	73.4
1–3, 5, 10	5	87.8	91.9	75.2
1–2, 4–5, 10	5	86.6	91.1	72.8
1–2, 4–5, 11	5	86.8	91.2	73.3
1–2, 5, 7, 11	5	86.4	90.9	72.7
1–2, 5, 8, 11	5	87.2	91.5	74.8
1–2, 5, 10–11	5	87.8	91.9	**75.4**
1–2, 5–6, 12	5	86.6	91.0	73.1
1–2, 5, 7, 12	5	86.2	90.8	72.5
1–2, 5, 8, 12	5	87.5	91.6	75.2
1–2, 5, 9, 12	5	87.1	91.4	74.2
1–2, 5, 10, 12	5	**87.9**	**92.0**	**75.4**

The improvement for the non-literal usage was 0.7 and 1.4 for literal usage compared to the best 4-feature model. The increase in scores was produced by

information from the case government feature. Hence, it is possible to suggest that information about case government is important for predicting the literal versus the non-literal usage of PVs with a high degree of accuracy. Another 5-feature model obtained the same f-score for literal usage as the best one. Compared to the latter, this model included the object animacy feature instead of case government. This result demonstrates that both animacy features are good predictors of the literal usage of PVs.

The results of the best 6-feature models are shown in Table 32. The best-performing model 1–2, 5, 10–12 combined features that were previously demonstrated to provide more useful information than did other features – the average abstractness of nouns, subject animacy, object animacy and case government. This combination of features worked better than did the best 5-feature classifier – the accuracy increased by 0.5%, the f-score for non-literal usage by 0.3 points, and the f-score for literal usage by 0.9 points.

Table 32: Results for the best 6-feature classifiers.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–6	6	86.4	91.0	72.2
1–2, 4–7	6	86.0	90.7	71.0
1–3, 5–6, 8	6	87.4	91.6	74.9
1–2, 4–6, 9	6	86.4	91.0	72.1
1–3, 5–6, 9	6	85.8	90.5	71.4
1–3, 5, 8–9	6	86.4	90.9	72.8
1–3, 5–6, 10	6	87.6	91.8	74.8
1–3, 5, 8, 11	6	87.0	91.3	74.3
1–3, 5, 10–11	6	87.8	91.9	75.3
1–2, 4–5, 10, 12	6	87.9	92.0	75.2
1–2, 5, 7, 10, 12	6	87.4	91.7	74.0
1–2, 5–6, 8, 12	6	87.5	91.7	74.5
1–2, 5, 8, 10, 12	6	87.7	91.9	74.9
1–2, 5, 9–10, 12	6	88.3	**92.3**	75.9
1–2, 5, 10–12	6	**88.4**	**92.3**	**76.3**
1–3, 5, 10, 12	6	87.7	91.8	75.1

The classifier using features 1–2, 5, 9–10, 12 obtained the same f-score for non-literal usage (92.3) as did the best model. The former included the object animacy instead of the object case. An implication of this is the possibility that, in comparison to the object animacy, the object case does not provide information that is as good for predicting literal usage. The results also indicate that the object case might be another useful feature to improve the prediction quality.

The classification results of the best models with feature set size 7 are presented

in Table 33. The best 7-feature model 1–2, 5, 8–10, 12 performed better than did the best 6-feature combination. The best combination obtained accuracy of 88.7%, an f-sore for non-literal usage of 92.5 and an f-score for literal usage of 76.6. Compared to the features that belonged to the best 6-feature classifier, the best 7-feature set included the subject and object cases, but excluded the object animacy. Therefore, it is possible that, when the case information is provided, the object animacy does not add any new information to the classification.

Table 33: Results for the best 7-feature classifiers.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–7	7	86.2	90.9	71.5
1–2, 4–8	7	86.1	90.9	70.9
1–6, 8	7	86.6	91.2	72.3
1–6, 9	7	85.9	90.6	71.3
1–3, 5–6, 8–9	7	86.4	90.9	72.5
1–5, 7, 10	7	86.8	91.3	72.6
1–3, 5–6, 8, 10	7	87.6	91.8	74.8
1–3, 5–6, 8, 11	7	87.0	91.4	73.8
1–3, 5, 8, 10–11	7	88.1	92.1	75.7
1–2, 4–5, 9–10, 12	7	87.8	92.0	74.6
1–2, 4–5, 10–12	7	88.1	92.2	75.2
1–3, 5, 9–10, 12	7	88.0	92.1	75.3
1–3, 5, 10–12	7	88.5	92.3	76.5
1–3, 5–6, 10, 12	7	88.0	92.1	75.3
1–2, 5–6, 10–12	7	88.1	92.1	75.6
1–2, 5, 7, 10–12	7	87.6	91.8	74.2
1–2, 5, 8, 10–12	7	88.5	92.4	**76.6**
1–2, 5, 8–10, 12	7	**88.7**	**92.5**	**76.6**

The combination 1–2, 5, 8, 10–12 obtained the same f-score for literal usage as did the best 7-feature classifier. The setup contained the average abstractness of nouns, subject case, subject and object animacy, and case government. In other words, the object animacy feature was included instead of the object case. This agrees with the earlier observation, which showed that the object animacy provided information that was more useful than was the object case for the prediction of literal sentences. Furthermore, it is possible that the object case and object animacy contained contradictory information that, when combined, decreased the number of correct predictions of literal sentences.

The claim that object case and object animacy contain contradictory information is supported by the results of the best 8-feature classifiers suggested in Table 34. The model 1–2, 5, 8–12 combined both of these features. While the overall

accuracy (88.1%) and the f-score for non-literal usage (92.2) were the highest amongst all the 8-feature classifiers, the F_1 lit (76.0) was not. The models without the object case feature – 1–3, 5, 8, 10–12 – obtained the highest f-score (76.1) of all the 8-feature classifiers.

Table 34: Results for the best 8-feature classifiers.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–8	8	86.2	90.9	72.0
1–2, 4–9	8	84.9	90.0	69.2
1–5, 7–9	8	85.6	90.4	70.5
1–6, 8–9	8	86.2	90.9	71.8
1–7, 9	8	86.2	90.8	72.0
1–6, 8, 10	8	86.9	91.3	73.2
1–6, 8, 11	8	86.8	91.3	73.0
1–3, 5, 8–11	8	87.5	91.7	75.0
1–2, 4–6, 10–12	8	87.6	91.8	74.6
1–3, 5–6, 10–12	8	87.8	91.9	75.2
1–2, 4–5, 7, 10–12	8	87.0	91.5	73.0
1–2, 4–5, 8, 10–12	8	88.0	92.0	75.5
1–2, 5, 7, 9–12	8	87.4	91.7	74.2
1–2, 5–6, 8, 10–12	8	87.6	91.8	74.7
1–3, 5, 8–10, 12	8	87.4	91.7	74.6
1–3, 5, 8, 10–12	8	88.1	**92.2**	**76.1**
1–2, 5, 8–12	8	**88.2**	**92.2**	76.0

Compared to the best 7-feature classifier, the overall highest accuracy for the best 8-feature model was 0.5% lower. While the comparison of the 7- and 8-feature combinations may show that it is not advisable to combine more than seven features in order to obtain the best predictions, the results of the 9-feature models shown in Table 35 demonstrate that this is not a valid conclusion.

More specifically, the best 9-feature classifier predicted the literal versus the non-literal usage better than did the best 8-feature combination. In fact, the best 9-feature combination 1–3, 5, 8–12 obtained the same accuracy and f-score for non-literal usage as did the best 7-feature combination. The f-score for literal usage was slightly higher (76.8) for the model that combined the unigram feature, the average abstractness of nouns, the subject case, the object case, subject animacy, object animacy and case government in comparison to the same model without the unigram and object animacy features. This finding indicates that object animacy provided useful information for predicting literal usage.

Table 35: Results for the best 9-feature classifiers.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–9	9	85.4	90.4	69.8
1–2, 4–10	9	86.2	90.9	71.7
1–5, 7–10	9	86.5	91.1	72.1
1–6, 8–10	9	86.8	91.3	72.9
1–6, 8–9, 11	9	86.3	90.9	71.8
1–8, 11	9	87.0	91.5	73.2
1–5, 8–11	9	87.0	91.4	73.3
1–7, 10–11	9	87.1	91.5	73.2
1–2, 4–5, 7–8, 10–12	9	87.2	91.6	73.6
1–3, 5, 8–12	9	**88.7**	**92.5**	**76.8**
1–3, 5–6, 8, 10–12	9	88.0	92.0	75.5
1–5, 8, 10–12	9	87.8	91.9	74.7

Table 36 shows the results of the best 10-feature classifiers. The results are not as good as those produced by the best 9-feature classifiers – the overall accuracy was 88.3%, F_1 n-lit was 92.3 and F_1 lit was 75.7. The best 10-feature classifier included all the features except for the average abstractness of words and the object abstractness. The comparison with the 9-feature models demonstrates that subject abstractness mainly undermined the predictions of the literal usage of PVs – the difference between the F_1 lit scores of the best 9- and 10-feature combinations was 1.1 points. Hence, adding information about the abstractness of words, subject and object did not improve the results.

Table 36: Results for the best 10-feature classifiers.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–10	10	86.4	91.0	72.0
1–6, 8–11	10	87.1	91.5	73.4
1–7, 9–11	10	86.4	91.0	71.9
1–8, 10–11	10	86.8	91.3	72.5
1–9, 11	10	85.8	90.6	70.6
1–3, 5–6, 8–12	10	**88.3**	**92.3**	**75.7**
1–5, 7–8, 10–12	10	87.6	91.8	74.2
1–5, 8–12	10	87.6	91.9	74.4
1–7, 10–12	10	87.1	91.5	73.1

This claim is supported by the results of the best 11-feature classifiers, and the model that included all 12 features. The scores for these setups, presented in Table 37, demonstrate that combining 11 or 12 features did not result in predictions that were as good as those provided by the models in which the average abstractness of words, abstractness of subject and object were not presented. Of the 11-feature classifiers, the combination that included any feature other than the object case obtained the greatest accuracy (87.6%). This result was better than the one obtained by the model including all the suggested features – the accuracy of the latter was 87.4%, the F_1 n-lit was 91.7 and the F_1 lit was 73.6. Compared to the best models, the difference in total accuracy was 1.3%.

Table 37: Results for the best 11- and 12-feature classifiers.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–11	11	86.2	90.9	71.2
1–2, 4–12	11	87.2	91.6	73.3
1–3, 5–12	11	87.2	91.6	73.5
1–4, 6–12	11	86.5	91.1	72.0
1–5, 7–12	11	87.1	91.5	72.9
1–6, 8–12	11	86.8	91.3	72.5
1–7, 9–12	11	87.2	91.6	73.3
1–8, 10–12	11	**87.6**	**91.8**	**74.0**
1–9, 11–12	11	86.3	91.0	71.4
1–10, 12	11	87.0	91.4	72.6
1–12	12	87.4	91.7	73.6

The results of all the possible 11-feature combinations suggest that the exclusion of features such as case government (12), the average abstractness of nouns (5) and subject animacy (10) resulted in less accuracy than did the removal of other features. Thus, it can be hypothesised that these features are the most influential for the classification task. At the same time, as mentioned previously, the abstractness features – the average abstractness of words (4), the subject abstractness (6) and the object abstractness (7) – could be suggested as being irrelevant in terms of the classification task. The impact of all the features on the classification results is discussed in Section 6.3.4.

The results in this section suggest that the best model combined information from nine features. The overall accuracy of this setup was 88.7%; the f-score for non-literal sentences was 92.5 and the f-score for literal sentences was 76.8. The classifier outperformed a relatively high majority baseline accuracy of 74.0%. A very similar result was also suggested by the learner-based feature selection (see Section 6.3.1.3). The only difference is the subject abstractness that was suggested by the feature selection but does not belong to the best model. Therefore, the results of the classification task show that the learner-based feature selection was successful indicating the relevance of the features.

In order to see how significant the differences amongst the best models across feature set sizes are, the statistical significance of the results of the best models is explored. Table 38 introduces the statistical significance of the differences in the performances of the best-performing combinations across feature set sizes. The statistical significance was calculated using the χ^2 test. The '+' indicates a highly statistically significant (p <0.001) difference, and the '*' indicates a statistically significant (p<0.05) difference. The differences that were not statistically significant are marked with '-'.

Table 38: Statistical significance of the differences in the performances of the best models.

		acc	0	1	2	3	4	5	6	7	8	9	10	11	12
0	baseline	74.0%													
1	3	82.8%	+												
2	1–2	85.8%	+	*											
3	1–2, 11	86.6%	+	+	-										
4	1–2, 5, 10	87.0%	+	+	-	-									
5	1–2, 5, 10, 12	87.9%	+	+	-	-	-								
6	1–2, 5, 10–12	88.4%	+	+	*	-	-	-							
7	1–2, 5, 8–10, 12	88.7%	+	+	*	-	-	-	-						
8	1–3, 5, 9–12	88.2%	+	+	-	-	-	-	-	-					
9	1–3, 5, 8–12	88.7%	+	+	*	-	-	-	-	-	-				
10	1–3, 5–6, 8–12	88.3%	+	+	*	-	-	-	-	-	-	-			
11	1–8, 10–12	87.6%	+	+	-	-	-	-	-	-	-	-	-		
12	1–12	87.4%	+	+	-	-	-	-	-	-	-	-	-	-	

All the best combinations, regardless of the feature set size, outperformed the majority baseline, and the differences in performance were statistically highly significant. The differences between the results of the best independent feature (the unigram) and the results of other combinations (that had more than one feature) were also statistically significant. Therefore, the classification of the literal versus the non-literal usage of PVs should be conducted by including more than one feature.

As the combination of fundamental features (particles and verbs) produced a relatively high accuracy rate (85.8%), the differences between the combination's results and the results of combinations with a greater number of features are not considered significant. However, four combinations outperformed this result with a statistically significant difference at the $p = 0.05$ level. Therefore, it can be assumed that the particles and verbs provided predictions that could be improved by adding context features. In this particular case, combining these features with up to eight other features resulted in predictions that were significantly improved. Hence, for the successful automatic detection of Estonian PVs, it is crucial to take information about context abstractness, the cases of the subject and object, subject animacy, object animacy and case government into account. However, the results of the combinations with three and more features did not alter to the extent

that the differences in the performances of the best models would be statistically significant.

To conclude, the best model detected the literal versus the non-literal usage of Estonian PVs at 88.7% accuracy. This proposed model combined information concerning nine features – the particle, verb, unigram, the average abstractness of nouns, the subject case, the object case, subject animacy, object animacy and case government. Prior to a further discussion of the features presented in Section 6.3.4, the results are compared to the outcome of the attribute selection – the automatic selection of the features that are most relevant to the problem.

6.3.4 Analysis of the impact of the features

This section concerns the effect of the studied features on the results. The focus is on the features that constituted the best classifier, but insight into the features that did not belong to the best model is also presented.

Based on the results presented earlier (see Section 6.3.3.2), the 9-feature model was proposed as the best classifier for detecting the literal versus the non-literal usage of Estonian PVs. The model classified literal sentences correctly with an f-score of 76.8, and non-literal sentences with an f-score of 92.5; the overall accuracy was 88.7 over a baseline of 74.0. It is clear that all nine features – the particle, verb, unigram, the average abstractness of nouns, subject case, object case, subject animacy, object animacy and case government – affected the results and contributed information that was relevant for the classification task. However, as the model combining only verb and particle reaches accuracy 85.8%, then it might be suspected that it is overfitted model[100], i.e. model works well on training data because it memorizes the most likely class for given PV and context information is rarely used. The best way to examine whether the model is overfitted or not is to test it on new data. Considering that it is costly, further (and more sophisticated) evaluation could also verify whether the model is overtrained or not. However, 57% of the PVs appear only in literal or only in non-literal sentences. Therefore, it is inevitable that a high amount of sentences are correctly classified only based on the information of particle and verb. When looking at PVs that appear at least once in literal and at least once in non-literal sentence, the model combining particle and verb does not outperform a majority baseline (both 65.3%), but the accuracy for the 9-feature model is 5% higher. Although it does not prove that the model is not overfitted, the result indicates that the majority of information comes from the PV itself, but context features provide information that leads to more precise predictions. For exhaustive conclusions, the models should be tested on unseen data.

A set of 8-feature models was trained in order to determine the impact of each feature. One of the nine features was omitted from each model. The results are compared to the outcome of the best model. Table 39 shows the results of these models.

[100] In machine learning models are often over- or underfitted. Building a model that is too complex for the amount of available information is called overfitting, and choosing too simple model is called underfitting. Overfitted model works well on the training set but it is not able to generalize to new data. Underfitted model works also poorly on the training set. (Müller and Guido 2016)

Table 39: Impact of the features on the predictions.

features	size	accuracy %	F_1	
			n-lit	lit
majority baseline	0	74.0	85.1	0.00
1–3, 5, 8–12	9	88.7	92.5	76.8
without particle	8	84.9	90.0	69.2
without verb	8	83.0	88.7	65.6
without unigram	8	88.2	92.2	76.0
without average abstractness of nouns	8	86.4	91.0	72.8
without subject case	8	87.4	91.7	74.3
without object case	8	88.2	92.2	76.1
without subject animacy	8	87.3	91.6	74.5
without object animacy	8	87.4	91.7	74.6
without case government	8	87.5	91.7	75.0

The models in which fundamental features – particles or verbs – were not present, unquestionably, obtained the worst results. Without the particle, the model attained an accuracy of 84.9%; without the verb, the accuracy was 83.0%. The latter is comparable to the result that the unigram feature achieved independently (82.8%, see Table 27). This outcome reinforces the choice of the particle and verb as the most influential features.

Of the other features, the impact of the average abstractness of nouns was the strongest. This feature increased the accuracy by 2.3%, the f-score for non-literal usage from 91.0 to 92.5 and the f-score for literal usage from 72.8 to 76.8. Hence, the aforementioned assumption concerning the average abstractness of nouns as being one of the most important features for the task (see Section 6.3.3.2) is confirmed.

The accuracy of the models without the subject case (8), subject animacy (10), object animacy (11) or case government (12) was 1.2–1.4% lower than that of the best model. Therefore, their impact could be labelled as moderate. The impact of the unigram (3) and object case (9) were the weakest on the results of the best model – they added 0.5% to the accuracy. Overall, these features are considered to be relevant to the classification of literal versus non-literal sentences. The further analysis of the effect of these features on the results is provided in Sections 6.3.4.2–6.3.4.9.

The accuracy of the best model indicated that 11.3% of the sentences were classified incorrectly. This means that, of the 1,481 sentences, 1,313 were classified accurately and 168 were classified inaccurately by the best system. As there were many more non-literal sentences in the dataset, it is reasonable to suggest that it would be more difficult to identify literal sentences than it would be to identify non-literal usage. The F_1 score also confirmed this – the F_1 for non-literal usage was 92.5 and 76.8 for literal usage. Of the 168 incorrectly classified sentences, 107 were literal and 61 were non-literal. Table 40 shows the 73 PVs that appeared in the sentences that were classified inaccurately by the best classifier.

The values in the column ‘sentences’ show the number of sentences that received an incorrect classification.

Table 40: Overview of the PVs in sentences that were classified incorrectly.

sentences	PVs
7	*välja riputama* ‘to hang out’
6	*lahti siduma* ‘to untie/unbind’, *läbi tulema* ‘to come through’, *sisse kallama* ‘to pour in/drink up’
5	*ette võtma* ‘to undertake/set out/embark upon’, *läbi laskma* ‘to let through/pretermit’, *tagasi minema* ‘to go back’, *välja pistma* ‘to stick out’
4	*eemale tõukama* ‘to push away/scare off/repel’, *ette andma* ‘to put something in front of somebody/feed/specify’, *läbi vaatama* ‘to look through/examine’, *välja ajama* ‘to send off/out’, *välja minema* ‘to go out’, *üles võtma* ‘to take something up/start something (song, conversation)/record’
3	*läbi lendama* ‘to fly through/fail’, *läbi põletama* ‘to fuse something/to burn something out’, *maha tõmbama* ‘to cross off/out/pull down’, *sisse vaatama* ‘to look inside/visit something for a moment’, *vastu rääkima* ‘to talk back/dispute’, *välja paiskama* ‘to throw something from somewhere/blurt out’, *välja võtma* ‘to take out’, *üle uhtuma* ‘to flush/wash’, *üles keerama* ‘to wind up/provoke’, *üles lööma* ‘to dress up/toss (upward)’, *üles soojenema* ‘to warm up’
2	*edasi jõudma* ‘to get ahead/come out on top’, *ette sattuma* ‘to run across or meet somebody or something on the way’, *ette vaatama* ‘to foresee/look ahead’, *juurde tõmbama* ‘to engage’, *kokku monteerima* ‘to assemble/edit video’, *kokku valguma* ‘to join/melt together’, *kõrvale tõrjuma* ‘to displace/push aside’, *läbi valgustama* ‘to x-ray/dissert’, *maha käima* ‘to go down/run down/go/degenerate’, *maha minema* ‘to get off’, *maha võtma* ‘to take down’, *otsa panema* ‘to add’, *sisse kutsuma* ‘to invite in’, *vahele pistma* ‘to interlard a conversation with/stick between something’, *vastu kajama* ‘to sound like an echo’, *välja jääma* ‘to stay out’, *välja tulema* ‘to get out (of)/turn up/come up with’, *üle kaaluma* ‘to weigh again’
1	*ette kandma* ‘to report/to serve’, *ette valmistama* ‘to prepare’, *juurde lõikama* ‘to cut/add (land)’, *juurde tulema* ‘to approach/accrue’, *kaasa tooma* ‘to bring something or someone/cause something’, *kinni minema* ‘to close/go to prison’, *kokku kuhjama* ‘to stack up’, *külge jääma* ‘to stick/get used’, *läbi tungima* ‘to penetrate/go right through’, *läbi viima* ‘to conduct/pass through’, *maha ajama* ‘to drive/push/shave off/remove’, *maha suruma* ‘to suppress/bottle up/allay’, *mööda käima* ‘to bypass’, *sisse laskma* ‘to let somebody in’, *sisse taguma* ‘to beat in(to)’, *tagant tõukama* ‘to push from behind/boost’, *tagasi põrkama* ‘to bounce back’, *tagasi vaatama* ‘to look back’, *vahele kukkuma* ‘to fall between/get caught’, *vastu kostma* ‘to reply’, *vastu põrutama* ‘to snap back at somebody’, *vastu särama* ‘to shine/reflect’, *välja ilmuma* ‘to debouch/emerge/appear unexpectedly’, *välja kurnama* ‘to wear out/filter’, *välja nägema* ‘to appear to your eyes/see outside’, *välja sülgama* ‘to spit out’, *üle käima* ‘to go/walk over’, *üle tooma* ‘to carry something/adapt/change the location of something’, *üles peksma* ‘to beat/wake up somebody’, *ümber tõmbama* ‘to put something around somebody or something/encircle’

The most 'difficult' PV for the classifier was *välja riputama* 'to hang out', which was represented in 12 sentences – eight literal and four non-literal. The classifier misclassified four literal and three non-literal sentences. Six sentences containing the PVs *lahti siduma* 'to untie/unbind', *läbi tulema* 'to come through' and *sisse kallama* 'to pour in/drink up' received incorrect predictions. Three literal and three non-literal sentences containing the PVs *lahti siduma* and *sisse kallama* received false predictions, while the class of only one non-literal sentence was predicted correctly. Of the five non-literal sentences containing the PV *läbi tulema*, two were predicted correctly, while all three literal sentences received incorrect predictions. While most of the PVs had more correctly classified sentences than they did incorrectly classified ones, there were some PVs that were more complicated. For example, of the seven sentences containing the PV *läbi laskma* 'to let through/pretermit', five were classified inaccurately, the class of all three sentences containing the PV *üle uhtuma* 'to flush/wash' was inaccurate, and four of the eight sentences containing the PV *välja minema* 'to go out' received incorrect predictions. The reasons for false predictions of some of these PVs are discussed in the following sections where the results of the classification task are analysed.

Firstly, three studied features that did not contribute to improve the prediction accuracy in combination with other features – the average abstractness of nouns, subject abstractness and object abstractness – are analysed as irrelevant features in Section 6.3.4.1. Each salient feature – the particle, verb, unigram, average abstractness of nouns, subject case, object case, subject animacy, object animacy and case government are then examined. Each feature is explored and their (dis)advantages highlighted. In addition, the misclassified sentences are discussed in the qualitative analysis. As the PVs that only appeared with their literal or non-literal meanings were (with one exception) classified correctly, the focus is only on the PVs that had at least one non-literal and one literal meaning in the dataset. Although the reasons for some sentences being misclassified were not always obvious, the main obstacles to the automatic prediction of the literal versus the non-literal usage of Estonian PVs are highlighted.

6.3.4.1 Irrelevant features

The results of the classification task and feature selection suggested that some of the studied features are irrelevant for detecting the literal versus the non-literal usage of PVs. While the feature selection (see Section 6.3.1) stressed two of them, namely the average abstractness of words and object abstractness, the results of the classification task (see Section 6.3) demonstrated that the subject abstractness was also not a relevant feature, as it was not part of the best classifier.

Figures 17, 18 and 19 illustrate the average abstractness scores of words, subjects and objects for the PVs that appeared at least once with a literal and once with a non-literal meaning. Most of these PVs (except for *järele vaatama* 'to watch someone/check or investigate', *kaasa tooma* 'to bring something or someone/cause something' and *taga kihutama* 'to encourage/chase') had at least one incorrectly predicted sentence. In addition, the PV *sisse taguma* 'to beat in(to)', which only had a non-literal meaning, was added because of having one inaccurately predicted sentence. Missing plots indicate that it was not possible to

find values for the subject and/or object. The reason was mainly that the sentences were lacking a subject and/or an object. In a few cases, no abstractness score for the subject/object was available in the abstractness/concreteness dataset.

The **average abstractness score for words** tended to be less than 5.0 for most of the sentences. It is thus likely that, in order to be a good feature for distinguishing between literal and non-literal usage, the literal sentences would have a higher score (the context is more concrete) than would the non-literal sentences (the context is more abstract). In some cases, the average abstractness of words differentiated well between the literal and non-literal meanings. For example, *ette sattuma* 'to run across or meet somebody or something on the way', *juurde lõikama* 'to cut/add (land)', *kaasa tõmbama* 'to persuade to join/pull along', *kokku kuhjama* 'to stack up', *kõrvale tõrjuma* 'to displace/push aside', *läbi tungima* 'to penetrate/go right through', *läbi lendama* 'to fly through/fail', *välja ilmuma* 'to debouch/emerge/appear unexpectedly', *välja paiskama* 'to throw something from somewhere/blurt out', *välja nägema* 'to appear to your eyes/see outside' and *üles soojenema* 'to warm up' had higher average scores for literal usage than they did for non-literal usage. However, in most of the cases, the feature could not distinguish between literal and non-literal PV usage. In fact, the medians of literal usage were even lower for some PVs than they were for those with for non-literal usage. For example, the words in the literal sentences with the PVs *ette valmistama* 'to prepare', *ette kandma* 'to report/to serve', *ette vaatama* 'to foresee/look ahead', *juurde tõmbama* 'to engage', *kokku valguma* 'to join/melt together', *läbi tulema* 'to come through', *läbi laskma* 'to let through/pretermit' and so forth are more abstract than are the words in the non-literal sentences. Surprisingly, for the PV *taga kihutama* 'to encourage/chase', which appeared only in correctly predicted sentences, the average abstractness of the words did not distinguish between literal and non-literal meanings. As mentioned previously (see Section 6.2.1), it is particularly difficult to assess the abstractness of words that are not nouns. This leads to a situation in which many words received medium scores (not indicating strong abstractness/concreteness) and, despite the literalness of the sentences, the average scores became very similar to each other. Therefore, the average abstractness of all the words in the sentence did not provide helpful information to distinguish between PV meanings.

It seems possible that the poor performance of the average abstractness of words was due to the automatic creation of the dataset of abstractness/concreteness ratings. As stated previously (see Section 4.5), different POS and meanings were not identified in the dataset; thus, the abstractness scores might not be accurate. However, this result reflected the findings of Köper and Schulte im Walde (2016b), who found that the features relying on adverbs, adjectives and verbs did not provide additional information for distinguishing between the literal and the non-literal usage of German PVs. Therefore, the poor performance of the average abstractness of words cannot have been caused entirely by the method used to create the dataset

The **abstractness of the subject** was usually greater for literal sentences and lower for non-literal sentences – in Figures 17, 18 and 19 it can be seen that subjects tended to be less abstract for literal sentences than they were for non-literal sentences; examples include *kaasa tõmbama* 'to persuade to join/pull

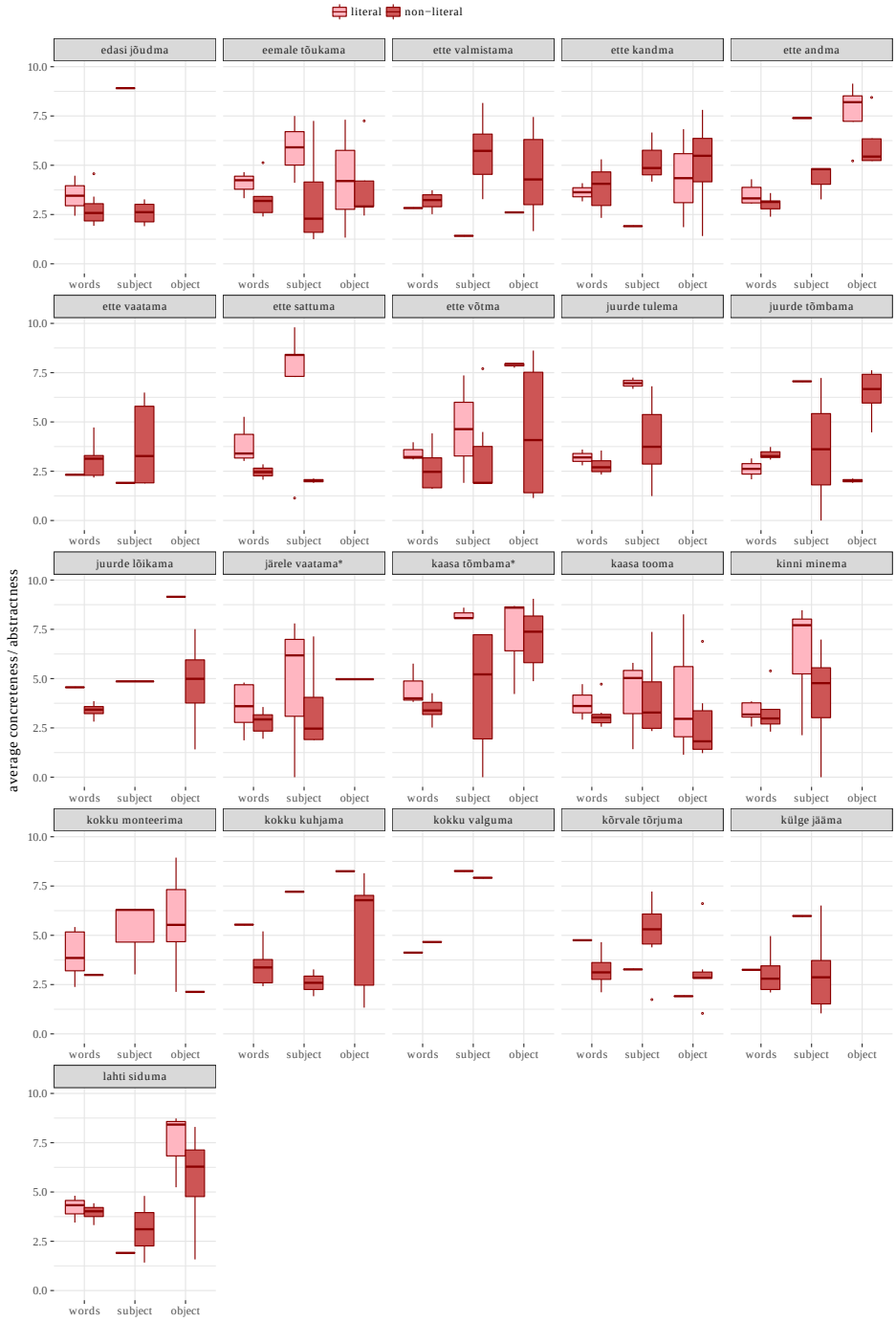

Figure 17: The average abstractness of words, subjects and objects of PVs with the particles *edasi* 'forward', *eemale* 'away', *ette* 'ahead', *juurde* 'by', *järele* 'after', *kaasa* 'along', *kinni* 'to', *kokku* 'together', *kõrvale* 'aside', *külge* 'to' and *lahti* 'open'.

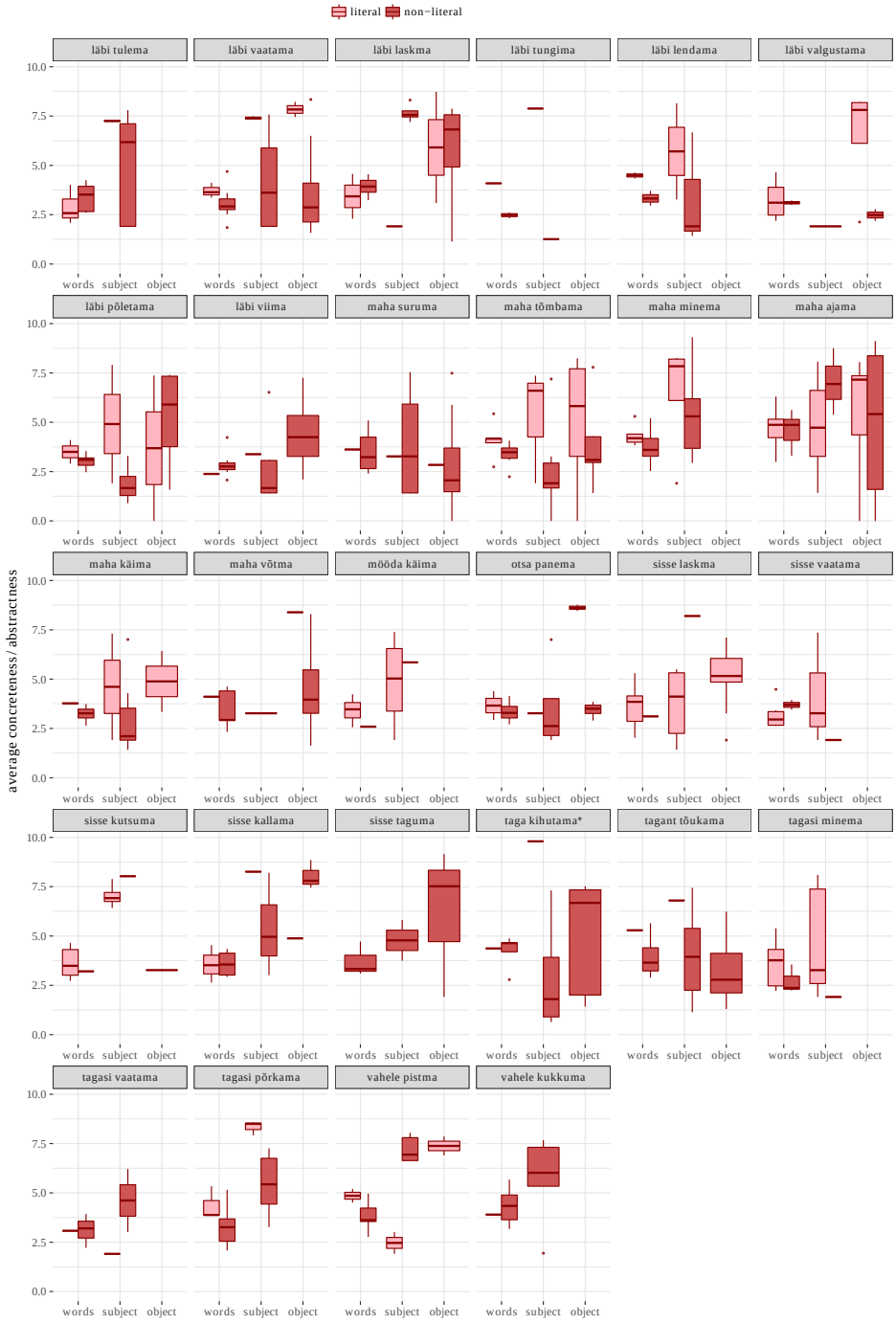

Figure 18: The average abstractness of words, subjects and objects across PVs with the particles *läbi* 'through', *maha* 'down/off', *mööda* 'along', *otsa* 'out', *sisse* 'in', *taga* 'behind', *tagant* 'from behind', *tagasi* 'back' and *vahele* 'between'.

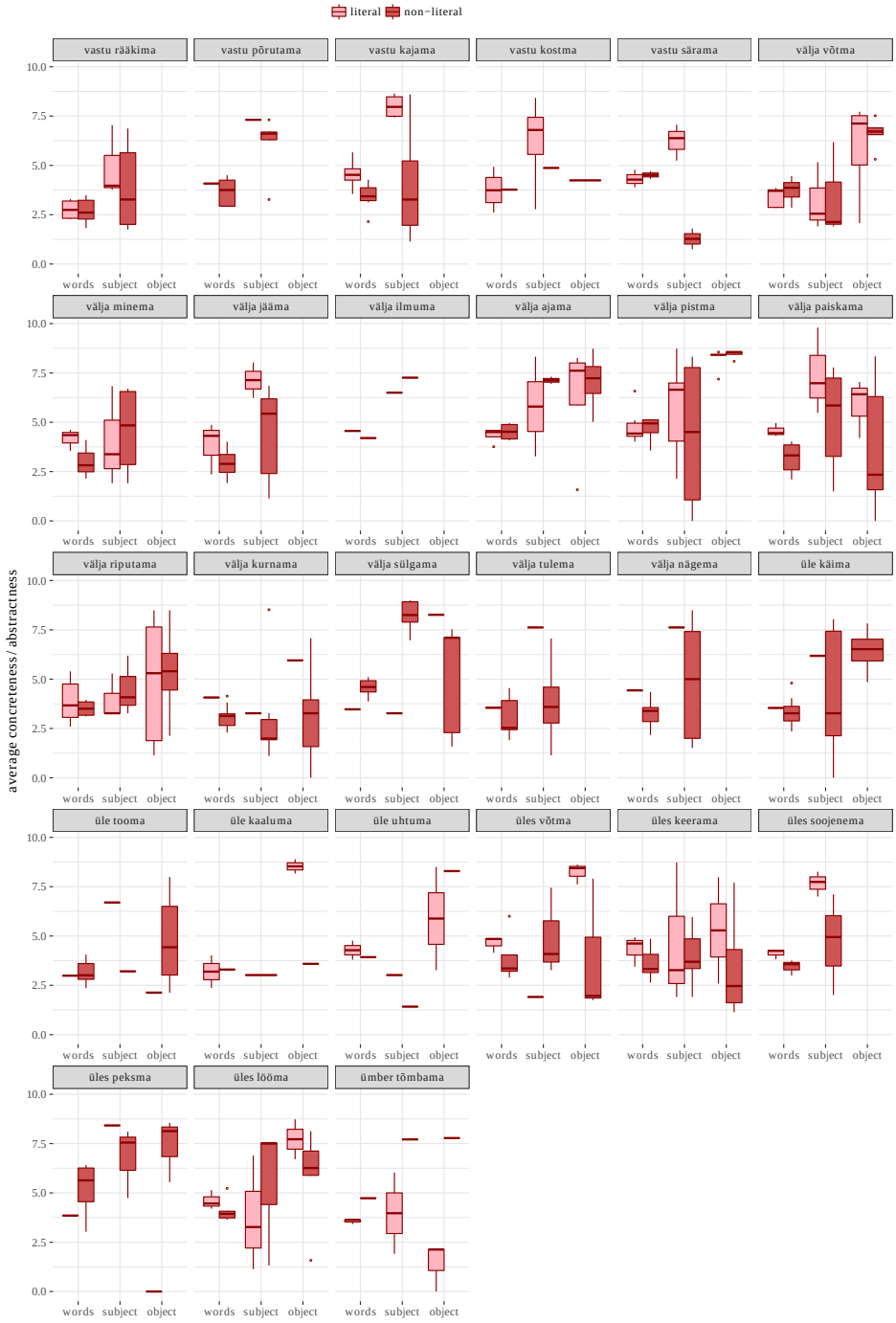

Figure 19: The average abstractness of words, subjects and objects across PVs with the particles *vastu* ‘against’, *välja* ‘out’, *üle* ‘over’, *üles* ‘up’ and *ümber* ‘around’.

along', *kokku kuhjama* 'to stack up', *läbi viima* 'to conduct/pass through', *taga kihutama* 'to encourage/chase', *tagasi põrkama* 'to bounce back' and so on. More specifically, as can be seen from example (82), a relatively concrete subject (*peitel* 'chisel', with a score of 8.03) led to correct automatic prediction. In fact, all three literal sentences containing the PV *tagasi põrkama* had concrete subjects, but one was still predicted incorrectly by the best model. This occurred because the subject's abstractness was not taken into account, and the context (the average abstractness of nouns) was very low. Accordingly, the subject abstractness feature can work very well in some cases, but not for many PVs.

(82) Peitel **põrka-b** tuhmi-∅ kolksu-ga **tagasi**.
chisel bounce-3SG dull-GEN clatter-COM back
'The chisel bounces back with a dull clatter.'

One reason that subject abstractness is not a good feature is illustrated by the sentences in examples (83) and (84). Pronouns mark subjects quite frequently in the sentences, but the abstractness scores for the pronouns were low. For example, the abstractness score for the word *mina* 'I' was 1.91, and 3.27 for *tema* 's/he' 3.27. Thus, when the subject in literal sentences was a pronoun, it might have led to an incorrect prediction.

(83) Siis haara-b ta raami-∅ sisse aseta-tud võrgu-∅, **kurna-b**
then grab-3SG s/he frame-GEN inside placed net-GEN filter-3SG
olluse-∅ **välja**, aseta-b selle-∅ pappaluse-le ning
substance-GEN out put-3SG this-GEN cardboard stand-ALL and
tupsuta-b svammi-ga kuiva-ks.
swab-3SG sponge-COM dry-TRL
'Then she grabs the net placed inside the frame, filters out the substance, puts it on the cardboard and swabs it with the sponge.'

(84) Päeva-l, kui ärka-si-n, teg-i-n end-∅ korda-∅,
day-ADE when wake up-PST-1SG do-PST-1SG myself-PRT order-PRT
õhtu-l **läk-si-n** **välja** või kelle-legi külla-∅.
evening-ADE go-PST-1SG out or someone-ALL village-ILL
'During the day, after I woke up, I cleaned myself up, in the evening I went out or visited somebody.'

The impact of the third irrelevant feature – **the abstractness of the object** – was not obvious (see Section 6.2.1), and the results of the classification proved that object abstractness did not help to differentiate between literal and non-literal PV usage. As can be seen in Figures 17, 18 and 19, there were quite a few sentences in which the PV did not have an object abstractness score, such as *ette sattuma* 'to run across or meet somebody or something on the way', *kinni minema* 'to close/go to prison', *läbi tungima* 'to penetrate/go right through', *tagasi vaatama* 'to look back', *vastu särama* 'to shine/reflect' and *välja tulema* 'to get out (of)/turn

up/come up with'. Therefore, the object abstractness did not provide information for many sentences, and thus could not have altered the results significantly.

A few PVs exhibited a considerable differences in the abstractness of the object in literal and non-literal sentences – the best examples are *juurde lõikama* 'to cut/add (land)', *kokku monteerima* 'to assemble/edit video', *kokku kuhjama* 'to stack up', *maha võtma* 'to take down', *otsa panema* 'to add', *välja sülgama* 'to spit out' and *üle kaaluma* 'to weigh again'. It is interesting that the sentences including these PVs that had high object abstractness scores were frequently classified incorrectly. For example, *välja sülgama* appeared in one literal sentence and in five non-literal sentences. The object abstractness of the literal sentence was 8.26, which is higher than was the abstractness of the object in other (non-literal) sentences. Nonetheless, the prediction for this sentence was incorrect.

The finding that subject and object abstractness were not relevant for the task is contrary to the findings in Köper and Schulte im Walde (2016b), which suggested that the noun-based abstractness features worked well for distinguishing between the literal and the non-literal usage of German PVs. However, these features were not among the three best features that these authors explored.

Taken together, the average abstractness of words, the subject abstractness and the object abstractness are three suggested features that do not belong to the best classifier. The reasons for the poor results for these features are as follows – the automatic quality of the abstractness/concreteness dataset, the difficulty of evaluating words that are not nouns (such as adverbs, adjectives, verbs and pronouns) and insufficient information about the features (for example, many sentences lacked an object).

6.3.4.2 Particles and verbs

The results of the classification task (see Section 6.3) demonstrated that particles and verbs formed the best 2-feature combination, which was called a fundamental combination. As the verb carries the main meaning of the PV, it has a stronger impact on the results than do any other features, including the particle. This section provides an overview of the impact of the particle and verb features on the results of classifying the literal versus the non-literal usage of PVs. The impact of particle and verb features on the results is discussed.

The best suggested model classifies 1,313 sentences correctly. The same model without the particle information predicted the class of 1,258 sentences correctly; without the verb information, the model classified 1,229 instances correctly. When both features were excluded, the model classified 1,116 sentences correctly. The accuracy of the latter model was 1.4% higher than the majority baseline (74.0%). Therefore, both features contributed to improving the classification results.

Figure 20 illustrates the distribution of literal and non-literal usage across particles. Most of the particles appeared in literal as well as in non-literal sentences, and were not helpful for the task. The particles that only appeared in non-literal sentences – *esile* 'forth', *alt* 'from under', *ühte* 'together', *ära* 'away/out/off'

and *ligi* 'near' – were predicted correctly by the best model[101]. As an example, the same model without particle information failed to classify one non-literal sentence containing the PV *ära võtma* 'to take away'. The reason was that the average abstractness score of the nouns (8.31) indicated that the sentence was literal.

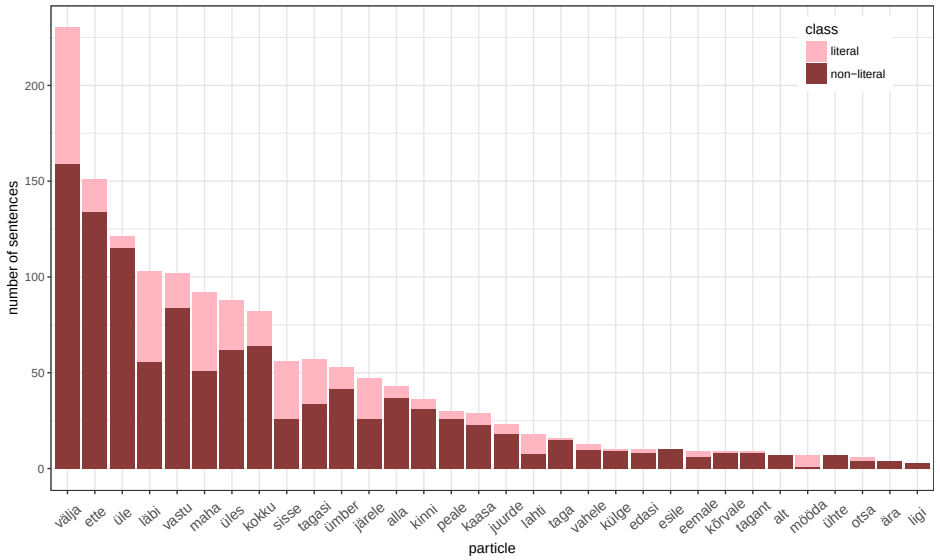

Figure 20: Distribution of literal and non-literal usage across particles.

In some cases, the particle provided information that resulted in a correct prediction despite the fact that the values of other features indicated the opposite. For example, the sentence in example (85) had a high average abstractness score (7.89), which indicates that the sentence is literal. The unigram feature also classified the sentences as being literal. However, when the particle information was included, the model made the correct prediction when classifying the sentences as non-literal.

(85) Pealtnäha tundu-b keeruline laul-da, mööda lava-∅
seemingly seem-3SG complicated sing-INF along stage-PRT
ringi tantsi-da ning samas **võt-ta** **välja**
around dance-INF and at the same time take-INF out
keerulis-i noot-e viiuli-l.
complicated-PL.PRT note-PL.PRT violin-ADE

Lit. 'It seems complicated to sing, dance and take out difficult notes on the violin at the same time.'

'It seems complicated to sing, dance and play difficult pieces on the violin at the same time.'

[101]It is important to bear in mind that this finding cannot be generalised by saying that these particles can only occur in PVs with non-literal meanings because all these particles appeared in one PV (except for *ära*, which was a component of two PVs).

At the same time, information about the particle was not always sufficient for the model to classify the sentence correctly. For example, the example (86) was evaluated as being literal by the annotators, but the model classified it as non-literal regardless of whether information about the particle was included or not.

(86) Kuidas ne-i-d **juurde** **tõmma-ta**, see kogemus
how they-PL-PRT by pull-INF this experience
on ta-l filigraanne.
be.3SG s/he-ADE filigrane
'How to pull them by, s/he has filigrane experience.'
'She knows well how to engage them.'

There were also sentences in which the particle information led to incorrect predictions. For example, the class of example (87) was predicted correctly as being non-literal by the model that did not use information about the particle. The best model (which included particle features) predicted that the sentences would be literal. The incorrect prediction was most likely because this sentence has similar feature values to literal sentences containing the same PV (see example (88)).

(87) Bayern ja ManU **tõrju-si-d** FC Barcelona-∅
Bayern and ManU repulse-PST-3PL FC Barcelona-GEN
alagrupiturniiri-l **kõrvale**.
subgroup tournament-ADE aside
Lit. 'Bayern and Man Utd repulsed FC Barcelona aside at the group stage.'
'Bayern and Man Utd eliminated FC Barcelona at the group stage.'

(88) Tema hüppa-s ette, vehki-s kä-te-ga ja
he jump-PST.3SG ahead wave-PST.3SG hand-PL-COM and
tõrju-s mu-∅ raja-lt **kõrva-le** pehme-le lume-le.
repulse-PST.3SG I-GEN track-ABL aside soft-ALL snow-ALL
Lit. 'He jumped forward, waved with his hands and repulsed me aside from the track onto the soft snow.'
'He jumped forward, waved his hands and pushed me off the track onto the soft snow.'

The distribution of literal and non-literal usage across verbs is illustrated in Figure 21. The verbs on the left appeared more frequently than did the verbs on the right. Fifty-six verbs appeared in literal and non-literal sentences, 49 only appeared in non-literal sentences and 15 only appeared in literal sentences. Therefore, more than half of the verbs only appear in one kind of sentence, which was why the models including information about the verb predicted significantly more sentences correctly than did the models that excluded information about the verb.

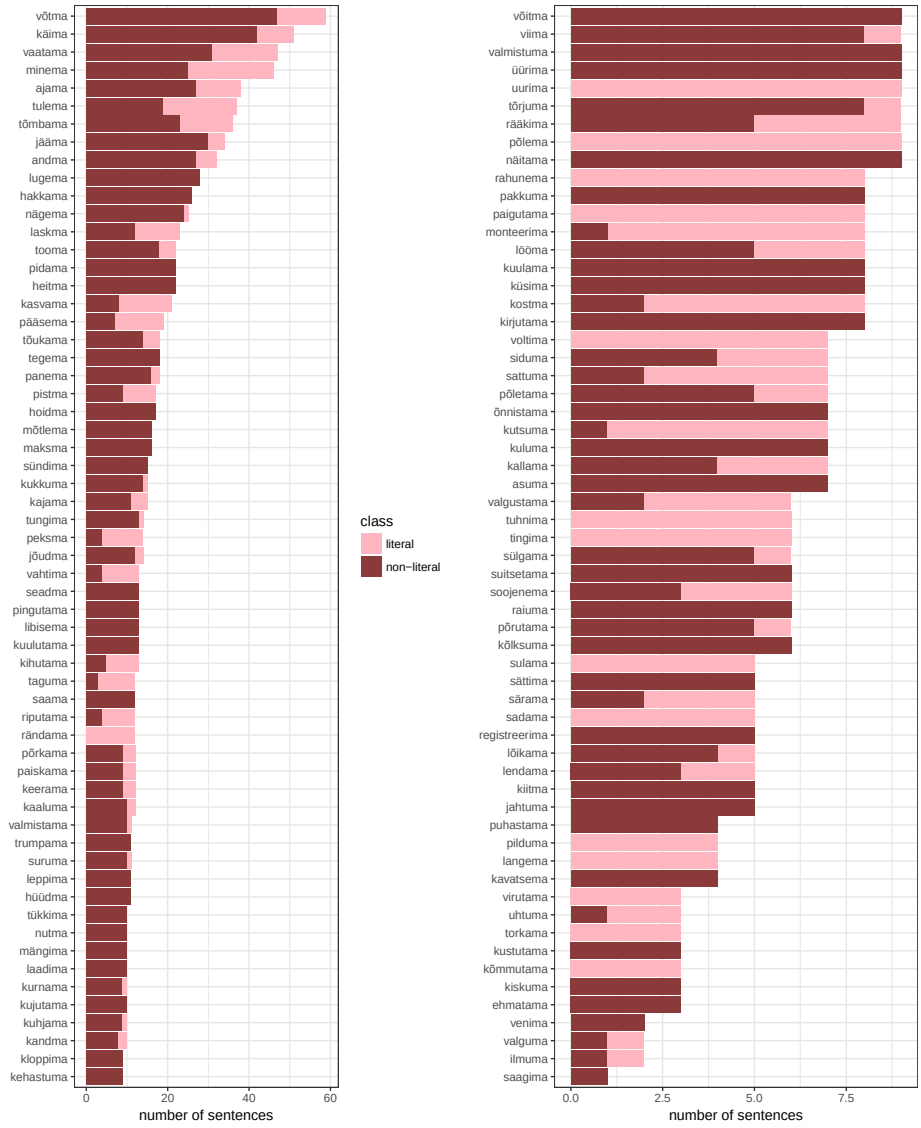

Figure 21: Distribution of literal and non-literal usage across verbs.

Some frequent verbs, such as *lugema* ‘to read’ and *hakkama* ‘to start’ only appeared in non-literal sentences. However, they were components of only two or three PVs; thus, the interpretation that these verbs only appeared in PVs with non-literal meanings is not conclusive. Nonetheless, the verb is a strong indicator of the literal versus the non-literal usage of PVs.

Information about the verb was often not sufficient when the verb appeared once in a literal sentence and more than twice in a non-literal sentence. For example, one literal and nine non-literal sentences contained the verb *kuhjama* ‘to heap up’. Even though the average abstractness of nouns and other features

indicated that a sentence was more literal than were other sentences containing this verb, the model classified the literal sentence as being non-literal. The misclassification was for the same reason, which is that literal sentences containing the verbs *kukkuma* 'to fall', *valmistama* 'to prepare' and *kurnama* 'to filter' and non-literal sentences containing the verbs *kostma* 'to sound' and *sattuma* 'to come upon' occurred.

Nevertheless, there were verbs that appeared in both literal and non-literal sentences, and which were classified correctly. In these cases, it can be assumed that the verb provided useful information, but the correct prediction was made in combination with other features. For example, there were nine literal and four non-literal sentences containing the verb *vahtima* 'to stare' in the dataset. The model that used information about the verb predicted the class of all these sentences correctly, but the same model without the verb feature did not. Therefore, information about the verb not only contributed to differentiating the verbs that appeared exclusively in literal or in non-literal sentences, but also identified the verbs that could occur in both literal and non-literal sentences.

In summary, particles alone do not work well for predicting the (non-)literalness of the PVs because most of the particles were components of the PVs that had both literal and non-literal meanings. Verbs provide information that is central to the correct prediction of the literal versus the non-literal usage of the PVs. As demonstrated previously (see Section 6.3.3.1), when using only information about the verb, 82% of sentences were predicted correctly. Particle and verb features combined predicted around 86.0% of the sentences correctly. The model containing particle and verb information often misclassified sentences containing verbs that were mainly used in non-literal meanings; these were used with a literal meaning, or vice versa, in one or two sentences. As a result, a further study in which the training data contain equal amounts of literal and non-literal sentences is needed.

In addition, in the event of having literal and non-literal sentences with similar feature values, the model tended to classify literal sentences as being non-literal and non-literal sentences as being literal. In order to avoid this, more data should be analysed.

6.3.4.3 Unigram feature

The best independent feature for classifying sentences according to the literalness of the PV was the unigram, with a frequency greater than 5 (see Section 6.3.2). The earlier analysis proposed that the impact of the unigram feature was relatively weak compared to the other features (see Table 39), but it still contributed to achieve the best results for the classification task (see Section 6.3.3.2). This section explains how this feature helped to distinguish literal from non-literal PV usage. Some instances in which the feature failed to lead to a correct prediction are also provided.

In terms of correctly classified sentences, the best classifier (that included the unigram feature) correctly classified seven sentences more than did the same model without information about the unigram. For example, the sentences presented in examples (89)–(90) were predicted incorrectly by the model with a feature set that

did not include the unigram feature, but were predicted accurately by the model that included information about the unigram. With regard to a sentence (89) that was assessed as being non-literal by the human annotators and based on the unigram feature, the best model predicted its class correctly. Example (90) was evaluated as being literal by the human annotators, and the unigram feature also predicted that it would be literal. Therefore, the best model produced the correct prediction with the help of the unigram feature.

(89) Kui Anvar me-i-d Eesti-sse kutsu-s, siis **vaata-si-n**
when Anvar we-PL-PRT Estonia-ILL invite-PST.3SG then look-PST-1SG
kaardi-lt **järele**, et aa, selline koht siis.
map-ABL after that oh such place then

Lit. 'When Anvar invited us to Estonia, then I looked after from the map and thought, 'Oh that place'.'

'When Anvar invited us to Estonia, then I looked it up on the map and thought, 'Oh that place'.'

(90) Kas siit **lähe-b** tee **välja** muuseumi-∅ eeskotta-∅ või
whether from here go-3SG road out museum-GEN lobby-ILL or
sisse hauakambri-sse?
inside tomb-ILL

Lit. 'Does this path from here go out to the museum lobby or into the tomb?'

'Does this path go to the museum lobby or the tomb?'

However, information about the unigram was not always sufficient for the model to achieve the correct prediction. For example, the sentence in example (91) was evaluated as being non-literal by the annotators and based on the unigram feature. Nevertheless, the model misinterpreted the data and classified this sentence as literal. The sentence in example (92) was classified as literal by the annotators and based on the unigram feature. The model still predicted that the sentences would be non-literal. Therefore, the values of the other features had a stronger influence, thus causing the model to make an incorrect prediction. For example, the example (91) has a relatively concrete context because it contains concrete nouns such as *litt* 'litas', *dollar* and *euro*, which indicate that it would be a literal sentence. As the average abstractness of nouns was more influential than was the unigram feature, the incorrect prediction was inevitable.

(91) Leedu kavatse-b **sidu-da** liti-∅ **lahti** dollari-st ning
Lithuania plan-3SG bind-INF litas-GEN loose dollar-ELA and
sidu-da selle-∅ ümber euro-ga.
bind-INF this-GEN over euro-COM

Lit. 'Lithuania plans to bind the litas loose from the dollar and over bind it with the euro'

'Lithuania plans to unpeg the litas from the dollar and peg it to the euro.'

(92) Kui me palgi-∅ sealt **läbi** **lase-me**, siis lauamaterjali-∅ müü-a
If we log-GEN from there through let-1PL, then lumber-PRT sell-INF
on juba palju keerulise-m kui palki-∅.
be.3SG already more complicated-COMP than log-PRT
'If we let logs through it, then the lumber would be more difficult to sell than the logs.'

The unigram feature helped to predict the correct class of one (see example (93)) for three literal sentences containing the PV *vastu särama* 'to shine/reflect'. The best model classified one literal sentence containing this PV incorrectly, but predicted the class of two of them correctly. The correct prediction of the sentence in example (93) mainly occurred because the unigram feature predicted that it would be literal – the values of the other features were more characteristic of non-literal usage or did not distinguish between literal and non-literal usage at all; for example, the verb *särama* 'to shine' and the adverb *vastu* 'against' can appear in non-literal and in literal sentences.

(93) Kilomeetri-te kaupa **sära-vad** kõrbeliiva-lt **vastu**
kilometer-PL.GEN by shine-3PL sand of desert-ABL back
kõikvõimaliku-d ilutule-d.
various-PL lampion-PL
Lit. 'Kilometers of various lampions are shining back from the desert sand.'
'Kilometers of various lampions are shining from the desert sand.'

In fact, some sentences were predicted incorrectly because the unigram feature provided information that caused the model to produce an inaccurate outcome. An incorrect prediction may have been caused solely by the unigram feature or the unigram in combination with other features. For example, the unigram feature predicted that the sentence in example (94) would be non-literal, while the annotators evaluated it as being literal.

(94) Alumise-le korruse-le **on** **lõiga-tud** rida
lower-ALL floor-ALL be.3SG cut-PST.PTCP row
akna-i-d **juurde** ja see muuda-b endise-∅
window-PL-PRT by and this change-3SG former-GEN
riietehoiu-∅ hubase-ks baariruumi-ks.
cloakroom-GEN cosy-TRL barroom-TRL
Lit. 'An additional row of windows has been cut for the ground floor and it changes the former cloakroom into cosy barroom.'
'A row of windows has been added to the ground floor and it changes the former cloakroom into cosy barroom.'

Overall, the unigram feature was part of the best model, but it did not have a remarkable impact on the results. There were some sentences that were predicted correctly solely because of information about the unigram, but the unigram feature did not often combine well with other features to produce more correct predictions

6.3.4.4 Average abstractness of nouns

Previous observations (see Table 39) allowed the claim that, after particles and verbs, the average abstractness of nouns was the most influential feature when classifying the literal versus the non-literal usage of Estonian PVs. Details regarding how this feature helps to distinguish between the literal versus the non-literal usage of PVs are presented in this section. In addition, a few examples illustrating how and why the feature failed to make correct predictions are included.

In fact, the model without the average abstractness of nouns correctly predicted 33 sentences fewer than did the best classifier. Therefore, the average abstractness of nouns contributed significantly to the differentiation between literal and non-literal usages. For example, while the combination of features without the average abstractness of nouns predicts three non-literal and four literal sentences containing the PV *ette andma* 'to put something in front of somebody/feed/specify' incorrectly, the best model predicted one non-literal and three literal sentences incorrectly. The problematic non-literal sentence had a high average abstractness rating for nouns (7.33), and one literal sentence had a relatively low score (3.97). The other two incorrect literal sentences had high scores but, despite this, the outcome was incorrect.

Figures 22, 23 and 24 show 75 PVs that appeared with literal and with non-literal meanings in the dataset. Most of these PVs occurred at least once in incorrectly classified sentences (PV that are not labelled with an asterisk (*)). In addition, although the PV *sisse taguma* 'to beat in(to)' only appeared in non-literal sentences, it is included in Figure 23 because one of the sentences containing this PV was classified inaccurately.

It can thus be suggested that the nouns in the literal sentences were more concrete (with a higher abstractness score) than were those in the non-literal sentences. Examples are the three PVs that had literal and non-literal meanings, but no incorrectly classified sentences – *järele vaatama* 'to watch someone/check or investigate', *kaasa tõmbama* 'to persuade to join/pull along' and *taga kihutama* 'to encourage/chase'. The medians of the average abstractness of nouns were higher for the literal usage than they were for non-literal usage, indicating that the nouns in the literal sentences were more concrete than were the nouns in the non-literal sentences. This was also the case for many other PVs. For example, the correctly classified literal sentences containing the PV *tagasi minema* 'to go back' had average abstractness scores of 6.48, 7.31, 6.56, 5.97, 6.28 and 7.41, which means that the nouns in these sentences were more concrete than were those in the inaccurately classified literal sentences (4.24, 2.81).

It is clear that none of the features produced perfect outcomes. For many PVs, the average abstractness of nouns in the literal and non-literal sentences was too similar, or there were some outliers that caused incorrect predictions. For example, Figure 22shows that there was one non-literal sentence with the PV *eemale tõukama* 'to push away/scare off/repel' that contained many more concrete nouns than any other non-literal sentence. This sentence was the only incorrectly predicted non-literal sentence containing this PV; thus, it can be suggested that the reason was the significantly high abstractness score of the nouns within it (7.31). The same was also the case for one non-literal sentence containing the PV *maha*

Figure 22: The average abstractness of nouns across PVs containing the particles *edasi* ‘forward’, *eemale* ‘away’, *ette* ‘ahead’, *juurde* ‘by’, *järele* ‘after’, *kaasa* ‘along’, *kinni* ‘to’, *kokku* ‘together’, *kõrvale* ‘aside’, *külge* ‘to’ and *lahti* ‘open’.

Figure 23: The average abstractness of nouns across PVs containing the particles *läbi* 'through', *maha* 'down/off', *mööda* 'along', *otsa* 'out', *sisse* 'in', *taga* 'behind', *tagant* 'from behind', *tagasi* 'back' and *vahele* 'through'.

Figure 24: The average abstractness of nouns across PVs containing the particles *vastu* ‘against’, *välja* ‘out’, *üle* ‘over’, *üles* ‘up’ and *ümber* ‘around’.

minema ‘to get off’, which had an extremely concrete context (9.31, see example (95)). Despite the fact that other features indicated non-literal usage, the model’s prediction was not accurate.

(95) Kuuse-d **lähe-vad** kõik **maha**.
spruce-PL go-3PL all down
Lit. ‘All spruces will go down.’
‘All spruces will be cut down.’

Moreover, the following two sentences (see examples (96) and (97)) had a high average abstractness score for nouns (6.60 and 7.40, respectively), but this did not guarantee a correct prediction – the model predicted that these two literal sentences would be non-literal. It is therefore likely that the values of other features led to an incorrect result for these kinds of sentences.

(96) Hea küll, et minu-∅ kabineti-s ja tööaja-l, kuid
good enough that I-GEN office-INE and working time-ADE but
nad oleks või-nud vähemalt ukse-∅ sulge-da ja
they be-COND might-PST-PTCP at least door-GEN close-INF and
sildi-∅ **välja** **riputa-da**, et toimu-b nõupidamine
sign-GEN out hang-INF that take place-3SG meeting
‘It’s alright to have meetings in my office during working hours, but they should have at least closed the door and hung out the sign that a meeting was taking place.’

(97) Kord võtt-is ta Kaju-l nati-st kinni ja
once take-PST.3SG s/he Kaju-ADE scruff-ELA to and
riputa-s tolle-∅ üle viienda-∅ korruse-∅
hang-PST.3SG that-GEN over fifth-GEN floor-GEN
rõduääre-∅ **välja**.
edge of the balcony-GEN out
‘Once he took Kaju and hung him over the fifth-floor balcony.’

It is correct to say that, the average abstractness of the nouns alone rarely resulted in a correct prediction when the other features did not support the distinction between literal and non-literal language usage. For example, the feature values of the sentences containing the PV *vastu põrutama* ‘to snap back at somebody’ were very similar. Although one literal sentence (see example (98)) containing this PV had a higher average abstractness of nouns than did the others (this can also be seen in Figure 23), the other features did not support the distinction at all, and the model made an incorrect prediction.

(98) Kui lõunaslaavlane eestlas-t küünarnuki-ga lõ-i,
when South Slav Estonian-PRT elbow-COM hit-PST.3SG
põruta-s meie-∅ mees kartmatult **vastu**.
knock-PST.3SG we-GEN man fearlessly back
‘When South Slav hit Estonian with an elbow, our man hit back fearlessly.’

Having concrete nouns in the sentence did not necessarily imply that the class of the sentence was predicted correctly. For example, one literal sentence containing the PV *läbi tulema* 'to come through' had a high average abstractness of nouns (7.98), but the sentence was still classified incorrectly as being non-literal. In fact, of the eight sentences containing *läbi tulema*, two correctly classified non-literal sentences had high average abstractness scores (7.85 and 7.22). Figure 23 shows that both the literal and the non-literal sentences containing this PV had relatively high abstractness scores, and they were similar to each other.

The similarity in abstractness scores for nouns in literal and in non-literal sentences was very common amongst the PVs in incorrectly classified sentences. This is noticeable when looking at the abstractness scores for sentences containing the PV *sisse vaatama* 'to look inside/visit something for a moment', for example (see Figure 23). Of the eight sentences (six literal and two non-literal), three sentences (one literal and two non-literal) were classified incorrectly by the classifier. The average abstractness of the nouns in literal sentences varied from 2.24 to 8.48, and the scores for non-literal sentences ranged between 4.72 and 6.51. Hence, the scores for the literal and the non-literal sentences did not help to distinguish between literal and non-literal usage.

There were even some PVs in literal sentences with more abstract nouns than non-literal ones, such as *juurde tõmbama* 'to engage', *läbi laskma* 'to let through/pretermit' and *välja võtma* 'to take out'. However, not all the sentences containing these kinds of PVs were predicted falsely, and the weight of the other features was also important. For example, for the PV *juurde tõmbama*, both literal sentences received false predictions – while one had a very low abstractness score (1.67) and the misprediction was obvious, the other sentence had a higher average abstractness score than did any other literal sentence containing this PV (5.85). These observations suggest that, as the score was still similar to the scores for the non-literal sentences (that were all predicted correctly) and the other features did not add sufficiently strong evidence for the computer to classify the sentence as literal, an incorrect prediction occurred.

In addition, as the sentence in example (99) suggests, literal sentences do not always include concrete nouns – the abstractness score for *Holland* was 6.57, 5.72 for *turniir*, 1.48 for *jõud* and 2.84 for *konkurent*. As all the other sentences with these PVs were non-literal, it is understandable that the automatic prediction of this sentence would be erroneous

Hence, when the abstractness scores for literal and non-literal sentences were similar, the predictions were similar because there were insufficient data to train the computer to make a different prediction. As the concrete nouns did not always appear in the literal sentences or abstract nouns in non-literal sentences, incorrect predictions were inevitable.

(99) Hollandi-∅ turniiri-l **suru-s** ta lausa jõu-ga kõik
Dutch-GEN tournament-ADE press-PST.3SG s/he straight strength-COM all
konkurendi-d **maha**.
competitor-PL down

'In the tournament in the Netherlands, he pushed all the competitors down using pure force.'

Taken together, even when the average abstractness of the nouns provided appropriate information to support a correct classification, the machine clearly did not make a correct prediction because a) there were insufficient similar data (sufficient sentences with the same classification and score), and b) the values of other features suggested a different classification. Nevertheless, there were many sentences in which the average abstractness of the nouns worked well and helped to distinguish between the literal and the non-literal usage of PVs.

6.3.4.5 Subject case

The results of the classification task and the observations about the impact of the subject case on the results of the classification of the literal versus the non-literal usage of PVs (see Table 39) demonstrated that the subject case had a moderate impact on the results. In this section, the influence of the subject case is discussed and illustrated using some examples from the data.

The distribution of the subject case across all the literal and non-literal sentences is illustrated in Figure 25, and in the sentences that the model classified correctly in Figure 26. Of the 385 literal sentences, 27.0% lacked a subject. Most of the sentences (72.2%) had the subject in nominative case, and 0.8% had the subject in partitive case. The distribution was very similar for the non-literal sentences – of 1,096 sentences, 29.7% had no subject, and the subject was in the nominative case in 69.7% of the sentences and in the partitive case in 0.6% of the sentences. Therefore, the usefulness of subject case was not obvious in all the sentences.

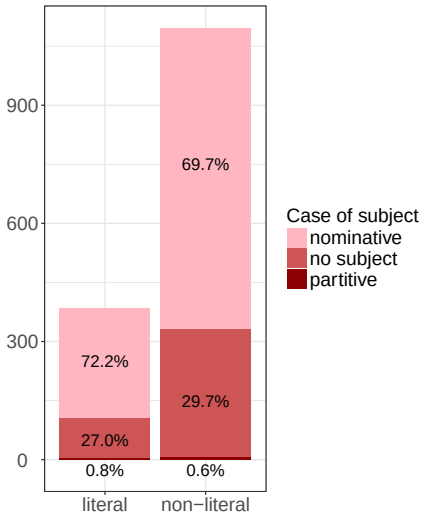

Figure 25: Distribution of the subject case across all the sentences.

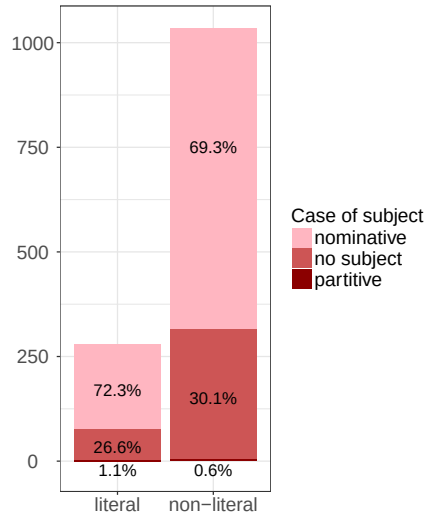

Figure 26: Distribution of the subject case across the correctly classified sentences.

Approximately 70% of both the literal and non-literal correctly classified sentences had the subject in the nominative case, and around 1% had the subject in the partitive case. The biggest difference was in the number of sentences without a subject – 26.6% of the literal and 30.1% of the non-literal sentences did not have a subject. Although the difference in the number of sentences without a subject was 3.5%, the subject case did not seem to identify many literal and non-literal sentences.

However, the best model classifies 18 more sentences correctly than did the same classifier that did not include information about the subject case. For example, the best combination did not predict the class for one literal sentence containing the PV *ette kandma* 'to report/to serve' with the subject in nominative case, while the model without the subject case feature predicted another literal sentence (with subject in nominative case) incorrectly as well as one non-literal sentence without a subject. In fact, it is not clear why one literal sentence received a false prediction while other did not because the values of the other features in these sentences were very similar. The only difference was in the value of the average abstractness of nouns – 6.39 for inaccurately predicted sentences and 6.83 for correctly predicted one.

For some PVs, the subject case clearly helped to distinguish between literal and non-literal usage. For example, with the help of the information provided by the subject case, the following sentences received a correct prediction (see examples (100–101)). The subject in example (100) is in the nominative case, and the sentence itself was evaluated as being literal. The subject in example (101) is in the partitive case, and the meaning of the PV is non-literal. It can thus be assumed that this difference ensured the correct prediction.

(100) Üks minu-st veidi vane-m poiss **tul-i** **juurde** ja
One I-ELA a little old-COMP boy come-PST.3SG by and
ütle-s venna-le, mis vahi-d siin.
say-PST.3SG brother-ALL what gaze-2SG here
Lit. 'One boy, who was a little bit older than me, came by and told my brother what he was staring here.'
'One boy, who was a little bit older than me, approached and asked my brother what he was staring at.'

(101) Ega ma teleka-st saate-i-d eriti ei vaata-gi,
Nor I TV-ELA telecast-PL-PRT in particular NEG watch-CONNEG.CL
jälgi-n ainult, kas uus-i kanale-i-d on
observe-1SG only whether new-PL.PRT channel-PL-PRT be.3PL
juurde **tul-nud**
by come-PST.PTCP
Lit. 'I do not watch much telecasts from TV, I just observe whether any new channels have come by.'
'I do not watch TV much, I just check whether any new channels have been added.'

However, as some non-literal sentences containing the PV *juurde tulema* 'to approach/accrue' had the subject in nominative case while some literal sentences

had the subject in the partitive case, the subject case did not predict them all correctly. For example, despite the nominative subject, the literal sentence (see example (102)) received an incorrect prediction even when the information about the subject case was provided.

(102) Aeg-ajalt **tul-i** Tõnis **juurde**, ütle-s: "Vaata-∅
now and then come-PST.3SG Tõnis by say-PST.3SG look-IMP
se-da pilti-∅, kas sa märka-d seal sellis-t
this-PRT picture-PRT do you notice-2SG there this kind-PRT
veidra-t ja põneva-t detaili-∅?".
weird-PRT and exciting-PRT detail-PRT

Lit. 'Every now and then Tõnis came by and said, "Look at this picture, do you notice there this kind of weird and exciting detail?"'

'Every now and then Tõnis came to me and said, "Look at this picture, do you notice there this kind of weird and exciting detail?"'

In some sentences, the subject case provided information that helped the model to make a correct prediction, but did not distinguish between literal and non-literal usage. For example, the model that included all the features from the best-feature set, except for the subject case, predicts the sentences in examples (103–104) incorrectly as being non-literal. When the subject case was included, these sentences received correct predictions. As both sentences were evaluated as being literal, the subject case did not help to distinguish between literal and non-literal usage but, in combination with another feature, it provided evidence that these sentences were more literal than they were non-literal. For these kinds of sentences, it is very difficult to say why two sentences were predicted as being non-literal when all the other sentences with *järele vahtima* 'to stare after somebody' were predicted to be literal.

(103) Jah, ega nee-d inimese-d **vahti-si-d** ju päris pikka-∅
Yes nor this-PL human-PL stare-PST-3PL of course quite long-PRT
aega-∅ **järele**, kui autorooli-s naine ol-i.
time-PRT after when driving wheel-INE woman be-PST.3SG

Lit. 'Yes, these people stared quite a long time after, when a woman was driving a car.'

'Yes, these people were staring quite a long time, when a woman was driving a car.'

(104) Siis on, mi-da mees-te-l daami-∅ möödu-des
Then be.3SG what-PRT man-PL-ADE lady-GEN pass by-GER
pea-d pööra-tes **järele** **vahti-da**.
head-PRT turn-GER after stare-INF

Lit. 'Then there is something for men to turn heads to and stare after when a lady passes by.'

'Then there is something for men to look at when a lady passes by.'

In summary, while the distribution of the subject case across the literal and non-literal sentences did not provide evidence that the subject case distinguished between literal and non-literal PV usage successfully in all cases, it still had a moderate influence on the results.

6.3.4.6 Object case

The results of the classification task and the observations about the impact of the object case on the results of the classification of the literal versus the non-literal usage of PVs (see Table 39) demonstrated that the object case had a relatively weak effect on the results. The influence of the object case is comparable to the impact of the unigram feature. This feature is still part of the model, and this section describes how the object case contributed to the differentiation between the literal and the non-literal usage of PVs. In addition, some examples showing why the feature did not work well are provided.

The best model predicted seven more sentences correctly than did the combination of the same features without the information about the object case. For example, in the sentences containing the PV *ette võtma* 'to undertake/set out/embark upon', the object case contributed to predicting the correct class of one non-literal sentence that had a partial object. While other features helped to predict the class of other non-literal sentences with objects in the partitive case correctly, it was necessary to have information about the object case of this sentence in order to make a correct prediction.

Figure 27 illustrates the distribution of the case of object across all sentences. The object is in genitive case in 14.0% of literal sentences and in 12.1% of non-literal sentences. 16.9% of literal sentences and 15.2% of non-literal sentences have object in nominative case. The percentage of sentences without the object is also similar – 51.7% of literal and 49.2% of non-literal sentences do not have an object. The biggest difference is in sentences with partial object – 17.4% of literal and 23.5% of non-literal have object in partitive case. Even though there are differences in the case of objects between literal and non-literal usage, it is still possible to suggest that the object case does not affect many sentences.

Figure 28 shows the distribution of the object case across all the sentences. The object was in the genitive case in 14.0% of the literal sentences and in 12.1% of the non-literal sentences – 16.9% of the literal sentences and 15.2% of the non-literal sentences had objects in the nominative case. The percentage of sentences without an object was also quite equal – 51.7% of the literal and 49.2% of the non-literal sentences did not have an object. The biggest difference was in sentences with a partial object – 17.4% of the literal and 23.5% of the non-literal sentences had objects in the partitive case. Even though there were differences in the object case between literal and non-literal usage, it is still possible to suggest that the object case did not affect many sentences.

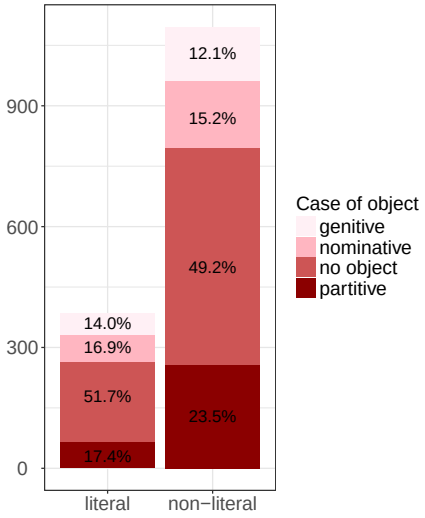

Figure 27: Distribution of the object case across all the sentences.

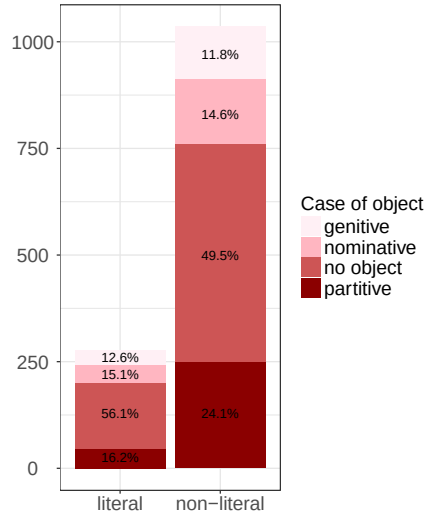

Figure 28: Distribution of the object case of object across the correctly classified sentences.

Although there were no obvious differences between the object case in the literal and in the non-literal sentences, there were still some sentences in which the object case contributed to differentiating between literal and non-literal usage. For example, the following literal sentence (see example (105)) did not have an object, while all the other sentences containing this PV were non-literal and had partial objects. Hence, the lack of an object was useful for predicting the correct class of this sentence.

(105) Veel ebameeldiva-m ol-i see, et tiheda-∅ liikluse-ga
more unpleasant-COMP be-PST.3SG this that dense-GEN traffic-COM
tänava-l sõit-is vilkuri-te-ga auto **kihuta-des** mu-l
street-ADE drive-PST.3SG flashing light-PL-COM car speed-GER I-ADE
taga nagu suure-l kurjategija-l.
behind like great-ADE criminal-ADE
'More unpleasant was that, at the busy street, the car with flashing lights sped up behind me like I was some great criminal.'

In some cases, the literal and non-literal sentences had objects in different cases, but this information was not sufficient to lead to a correct prediction. For example, two sentences containing *üle uhtuma* 'to flush/wash' were literal (see example (106)), and one sentence was non-literal (see example (107)). Literal sentences had objects in the nominative case and non-literal sentences had them in the genitive case. The model mispredicted all three sentences, and the reason might have been that the values of other features were too similar to each other and the impact of the object case was insufficient to generate the correct prediction.

(106) **Uhu-∅** puhastatud kala pealt ja seest vee-ga **üle**.
wash-IMP cleaned fish off and from in water-COM over
'Wash the cleaned fish inside and out.'

(107) Natalja mäleta-b aga teistsugus-t Jezhovi-∅ –
Natalja remember.-3SG yet different-PRT Jezhov-PRT –
hella-∅ isa-∅, kes **uhtu-s** tütre-∅
tender-PRT father-PRT who wash-PST.3SG daughter-GEN
kinki-de-ga **üle** ning mängi-s tema-ga õhtuti.
gift-PL-COM over and play-PST.3SG s/he-COM in the evenings
Lit. 'Natalja remembers a different Jezhov – gentle father, who washed his daughter with many gifts and played with her in the evenings.'
'Natalja remembers a different Jezhov – gentle father, who gave many gifts to his daughter and played with her in the evenings.'

In addition, there were cases in which the object case did not distinguish between literal and non-literal usage. For example, the PV *läbi valgustama* 'to x-ray/dissert' appeared in four correctly predicted literal and two incorrectly predicted non-literal sentences. Three sentences had objects in the nominative case and three sentences in the partitive case – one non-literal sentence had a partial object (see example (108)), while another had a total object (see example (109)). These sentences convey the same meaning and, as the object case was different, it is possible to suggest that the object case did not distinguish between the literal and the non-literal usage of *läbi valgustama*.

(108) Festivali-∅ idee ol-i **läbi** **valgusta-da** erineva-te
festival-GEN concept be-PST.3SG through light-INF different-PL.GEN
maa-de rahvusvähemus-te probleeme-∅.
country-PL.GEN national minority-PL.GEN problem-PL.PRT
Lit. 'The concept of the festival was to X-ray the problems of national minorities of different countries.'
'The concept of the festival was to deal with the problems of national minorities of different countries.'

(109) Etenduse-∅ käigus **valgusta-ta-kse** **läbi** ka kareda-ks
play-GEN in the process of light-IMPS-PR through also harsh-TRL
kulunud idee-d.
ablated idea-PL
Lit. 'The play also X-rays worn-out ideas.'
'The play also deals with worn-out ideas.'

The analysis of this section shows that the object case predicted the correct class of relatively few sentences. However, the distribution of the case object across correctly predicted instances demonstrates that the objects in the partitive case helped to identify different meanings of the PVs. Overall, there were sentences that received the correct prediction solely because of the object case and sentences in which the object case operated in combination with other features.

6.3.4.7 Subject animacy

The results of the classification task and the examination of the impact of subject animacy on predicting the literal versus the non-literal usage of PVs (see Table 39) suggested that subject animacy had a moderate influence on the results. This section describes how the subject animacy helped to distinguish between the literal and non-literal usage of PVs in more detail. In some of the sentences, the subject animacy failed to predict the correct classification. These instances are also discussed.

The information about the subject animacy helped in the correct prediction of 20 more sentences than did the feature set that did not include the subject animacy feature. For example, the best model predicts the class of one literal sentence containing the PV *üle tooma* 'to carry something/adapt/change the location of something' incorrectly. The model that did not use the subject animacy feature predicted the same literal sentence (and a non-literal sentence) incorrectly. This means that the information about the non-literal sentence having no subject (and therefore no subject animacy) helped to predict the correct class of this sentence.

Figure 29 shows the distribution of subject animacy across all the sentences. The subject could be animate, inanimate or absent. There were slight differences between the literal and non-literal sentences. The biggest difference was in the number of sentences with an animate subject – 51.7% of the literal sentences and 44.9% of the non-literal sentences had an animate subject, while 21.3% of the literal and 25.4% of the non-literal sentences had inanimate subjects, and the remainder of the sentences did not have a subject.

Figure 30 shows the distribution of subject animacy across the correctly predicted sentences. The differences between the literal and non-literal sentences were more substantial – 52.5% of the literal sentences and 44.8% of the non-literal sentences had animate subjects, 20.9% of the literal and 25.1% of the non-literal sentences had inanimate subjects and the remainder of the sentences did not have a subject.

The comparison of the distributions of subject animacy across all the sentences and across the correctly classified sentences indicated subject animacy was somewhat helpful for distinguishing between literal and non-literal PV usage. For example, of the three literal and three non-literal sentences containing the PV *üles soojenema* 'to warm up', one non-literal and two literal sentences received a false prediction. The correctly predicted non-literal sentences had animate subjects and the literal sentence had an inanimate subject. Hence, the reason that the non-literal sentences with inanimate subjects (see example (110)) and the literal sentence with an animate subject (as in example (111)) received erroneous predictions was the value of the subject animacy. The second literal sentence had an inanimate subject; hence, it was probably classified incorrectly due to having a similar average abstractness of nouns as the non-literal sentences.

(110) Nüüd on sõprussuhte-d **üles** **soojene-nud**.
now be.3SG friendship-PL up warm up-PST.PTCP
Lit. 'The friendship has warmed up now.'
'The friendship has been restored now.'

(111) Lume-le visatud ahven külmu-b kiiresti, kuid visa-∅ hinge-ga
snow-ALL thrown perch freeze-3SG quickly but tough-GEN spirit-COM
elukas hakka-b **üles** **soojene-des** kohe liiguta-ma.
creature start-3SG up warm up-GER immediately move-SUP
'The perch thrown on the snow freezes quickly, but when warming up, the tough-spirited creature starts moving immediately.'

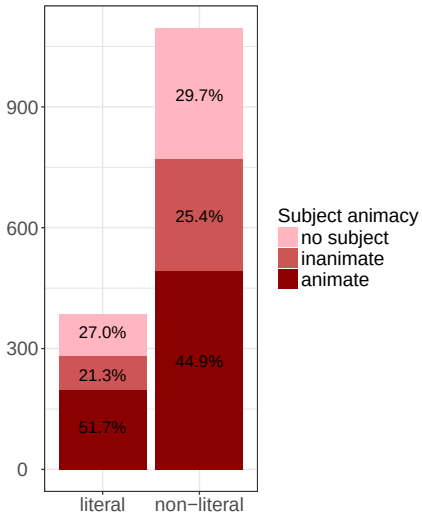

Figure 29: Distribution of subject animacy across all the sentences.

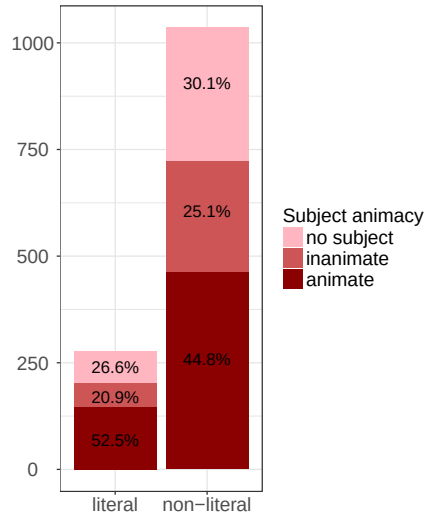

Figure 30: Distribution of subject animacy across the correctly classified sentences.

For some PVs, two meanings (literal and non-literal) were differentiated appropriately by the subject animacy. For example, the PV *kaasa tooma* has a literal meaning 'to bring something or someone' and a non-literal meaning 'to cause something'. The literal sentences have animate subjects, and the non-literal sentences have inanimate objects. The classifier assigned the correct class to all the sentences except for one literal one (see example (112)). In this sentence, differentiation was based on the subject's animacy, which did not provide sufficient information; thus, the automatic prediction was false. The reason might have been the relatively low average abstractness of the nouns.

(112) Keegi ei ol-nud tema-∅ tead-a varem Bentley-st
someone NEG be-PST.PTCP s/he-GEN know-INF before Bentley-ELA
Eesti-sse kiitus-t **kaasa** **too-nud**.
Estonia-ILL commendation-PRT along bring-PST.PTCP
Lit. 'As far as he knew, nobody from Estonia had brought a commendation from Bentley with them before.'
'As far as he knew, nobody from Estonia had been awarded a commendation from Bentley before.'

Although the subject case may be able to distinguish between literal and non-literal meanings, the classifier did not necessarily make a correct prediction. For example, all five of the correctly predicted non-literal sentences containing the PV *välja sülgama* 'to spit out' had an inanimate subject (as in example (113)), but the inaccurately classified literal sentences had an animate one (see example (114)). Of the other features, only the object case supported the distinction between literal and non-literal usage. Nonetheless, these two features did not convey sufficient information and the classifier produced an incorrect prediction.

(113) Kaia Kanepi on pikk ja sihvakas tüdruk, kelle-∅
Kaia Kanepi be.3SG tall and slim girl who-GEN
kohta interneti-∅ otsingusüsteemi-d **sülga-vad** **välja** iga-s
about Internet-GEN search engine-PL spit-3PL out every-INE
keele-s fakt-e
language-INE fact-PL.PRT
Lit. 'Kaia Kanepi is a tall and slim girl, about whom the Internet search engines spit out facts in every language.'
'Kaia Kanepi is a tall and slim girl, about whom the Internet search engines give out facts in every language.'

(114) Sellise-∅ hulga-∅ veini-de hindamine osutu-s
this kind-GEN amount-GEN wine-PL-GEN evaluation turn out-PST.3SG
tõsise-ks katsumuse-ks ka kogenud degustaatori-te-le,
serious-TRL challenge-TRL also experienced wine taster-PL-ALL
ehkki nad maitstud veini-d **välja** **sülga-si-d**
although they tasted wine-PL out spit-PST-3PL
'Although they spat out the wines they tasted, it was a serious challenge even for experienced wine tasters to evaluate so many wines.'

In summary, the distribution of subject animacy across the correctly classified sentences indicated slight differences between literal and non-literal sentences. The analysis of the results demonstrated that there were sentences that were only predicted correctly because of the information about subject animacy. In some cases, the value of the subject animacy did not lead to correct predictions because information about the animacy was not supported by the values of other features.

6.3.4.8 Object animacy

The results of the classification task and the considerations of the impact of the object animacy on the predictions of the literal versus the non-literal usage of PVs (see Table 39) demonstrated that object animacy had a moderate influence on the results. This section describes how the object animacy helped to distinguish between literal and non-literal usages of PVs in more detail. In addition, some examples showing how and why the feature failed are provided.

The best model predicted 18 more sentences correctly than did the model with the same feature set excluding information about the object animacy. For

example, the non-literal sentence containing the PV *tagant tõukama* 'to push from behind/boost' (that the model that did not use information about the object animacy was unable to classify correctly) was predicted correctly by the best model. Hence, object animacy contributed useful information for the correct classification of both literal and non-literal sentences.

Figure 31 shows the distribution of object animacy across all the sentences. The object could be animate, or inanimate or non-existent. There were very few differences between the literal and non-literal sentences – 51.7% of the literal sentences and 45.0% of the non-literal sentences did not have an object, 37.9% of the literal and 39.1% of the non-literal sentences had inanimate objects and the remainder of the sentences had animate objects.

Figure 32 shows the distribution of object animacy across the correctly predicted sentences. The differences between literal and non-literal sentences were greater – 56.1% of the literal sentences and 49.5% of the non-literal sentences did not have an object, 33.1% of the literal and 38.7% of the non-literal sentences had inanimate objects and the remainder of the sentences had animate objects.

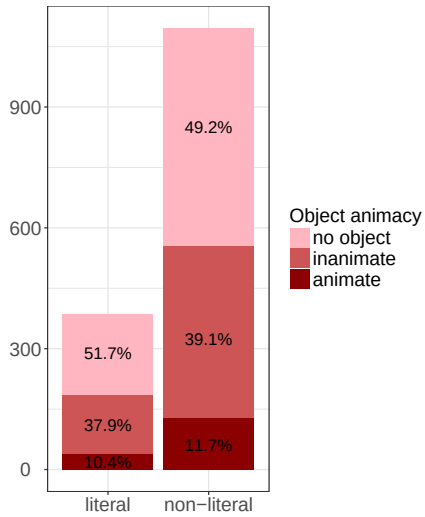

Figure 31: Distribution of object animacy across all the sentences.

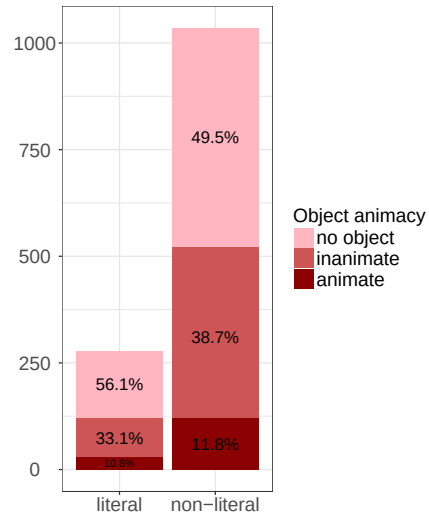

Figure 32: Distribution of object animacy across the correctly classified sentences.

The distributions of object animacy across all the sentences and across the correctly classified sentences indicated that, although the feature did not work perfectly, it was somewhat helpful for distinguishing between literal and non-literal usage. For example, object animacy distinguished literal from non-literal sentences containing the PV *kaasa tõmbama* 'to persuade to join/pull along'. In addition to the high average abstractness of the nouns in the literal sentences, the differences in the object animacy values in the literal and non-literal sentences also supported the correct prediction of all the sentences containing this PV. Of the 11 sentences, all three literal sentences contained inanimate objects (as in example

(115)), meaning that this information helped to distinguish literal from non-literal usage. However, as two non-literal sentences also had inanimate objects (as in example (116)), other features also provided information that supported the correct predictions.

(115) Põhjaminev suur laev **tõmba-b** tema-ga külgneva-∅ väikese-∅
foundering big ship pull-3SG s/he-COM adjacent-GEN little-GEN
laeva-∅ nagunii **kaasa**.
ship-GEN anyway along
'Big foundering ship pulls the adjoining little ship along anyway.'

(116) Kunsti-∅ kõrval **tõmma-ta-kse** **kaasa** uus-i ala-sid,
art-GEN beside pull-IMPS-PR along new-PL.PRT field-PL.PRT
eelkõige disaini-∅ ja arhitektuuri-∅
foremost design-PRT and architecture-PRT
'Besides art, new areas, in particular design and architecture, are dragged along.'
Lit. 'Besides art, new areas, especially design and architecture, are involved.'

In fact, there were sentences in which object animacy provided information that was useful for making correct predictions, but this was insufficient when other features did not support the same prediction. For example, of the 10 sentences containing the PV *maha käima* 'to go down/run down/go/degenerate', two are literal (as in example (117)) and eight non-literal (as in example (118)). The model classified the literal ones incorrectly despite the fact that they had different object animacy values from those of the non-literal sentences. Hence, the object animacy differentiated between literal and non-literal usage, but this information was not sufficient for the model to predict all the sentences correctly. These sentences also had different object case values from those of the non-literal sentences, but all the other features indicated more non-literal than literal usage – this led to incorrect predictions.

(117) Mina ise **käi-si-n** eile näiteks kümme
I myself go-PST-1SG yesterday for instance ten
kilomeetri-t kindlasti **maha**.
kilometre-PRT certainly down
Lit. 'I myself, for example, definitely walked down ten kilometres yesterday.'
'I myself, for example, definitely walked ten kilometres yesterday.'

(118) Aga kui saa-n ainult seitse tundi-∅ maga-da, siis
but if can-1SG only seven hour-PRT sleep-INF then
hakka-n **maha** **käi-ma**
start-1SG down go-SUP
'But if I can only sleep for seven hours, my energy level goes down.'

Moreover, information about the object animacy was insufficient for predicting the class of two literal sentences containing the PV *vahele pistma* 'to interlard a conversation with/stick between something'. Five non-literal sentences containing this PV did not have objects (as in example (119)), unlike the literal ones that had inanimate objects (as in example (120)). The fact that the objects in the literal sentences had different object animacy values than did those in the non-literal ones, this did not lead to the correct prediction of the literal sentences.

(119) Baaridaam Hilka **pista-b** rõõmsalt **vahele**, et Tea ja
barmaid Hilka tuck-3SG cheerfully between that Tea and
Raido käekäigu-∅ vastu tunne-vad suur-t huvi-∅
Raido hand movement-GEN for feel-3PL great-PRT interest-PRT
ka tema teismelise-d lapse-d.
also s/he-GEN teenage-PL child-PL
'Barmaid Hilda says cheerfully that her teenage children are greatly interested how Tea and Raido are doing.'

(120) **Pista**-∅ juustukang ja singitükk **vahele**.
tuck-IMP cheese straw and piece of ham between
'Tuck the cheese straw and piece of ham in between.'

In addition, the information about the object animacy did not always distinguish between literal and non-literal usage. For example, the model did not differentiate between literal and non-literal usage for sentences containing the PV *ette valmistama* 'to prepare'. There was one literal sentence (see example (121)) and eight non-literal sentences (as in example (122)) with inanimate objects, and the model failed to predict the correct class of the literal sentence because the value of the object animacy was the same for both types of sentences.

(121) Ol-i-d tegusa-d inimese-d, kes **valmista-si-d** **ette**
be-PST-3PL active-PL people-PL who prepare-PST-3PL in advance
polstri-d, kus ülemnõukogu-∅ esimehe-l ol-i
upholstery-PL where supreme council-GEN chairman-ADE be-PST.3SG
hää hõlju-da.
good float-INF
'Active people prepared the upholsteries, where it was good for the chairman of the supreme council to float.'

(122) Eesti-∅ Hoiupank **valmista-b** **ette** oma-∅ vara-∅
Estonia-GEN savings bank prepare-3SG in advance own-GEN assets-GEN
kindlustamise-∅ tellimus-t.
insurance-GEN order-PRT
'Estonian Savings Bank prepares the order of insurance for its assets.'

The analysis presented in this section suggests that, although the incorporation of object animacy in the study was less intuitive than was the inclusion of subject

animacy (see Section 6.2.3), the extent of the impact of these features was similar. Object animacy had a moderate impact, which means that there were sentences that were predicted correctly solely because of information provided by the subject animacy, as well as in combination with other features.

6.3.4.9 Case government

The results of the classification task and the observations about the impact of case government on predicting the literal versus the non-literal usage of PVs (see Table 39) demonstrated that case government had a moderate influence on the results. This section describes how this feature assisted in the distinction between the literal and non-literal usage of PVs. Furthermore, some examples are provided in order to explain why and how the feature failed to make a correct prediction. With the help of case government, the model was able to predict 17 more sentences correctly than did the model without the case government feature. For example, the model was able to assign the correct class to one of two sentences containing the PV *välja ilmuma* 'to debouch/emerge/appear unexpectedly', while the combination without the information about the argument case led to incorrect predictions.

The distribution of the argument case across the literal and non-literal usages in all the sentences is illustrated in Figure 33. Most of the sentences did not have any case government. The most common argument cases were the elative and allative. Comitative, inessive, translative, adessive and essive cases only appeared non-literal sentences. The latter appeared only once.

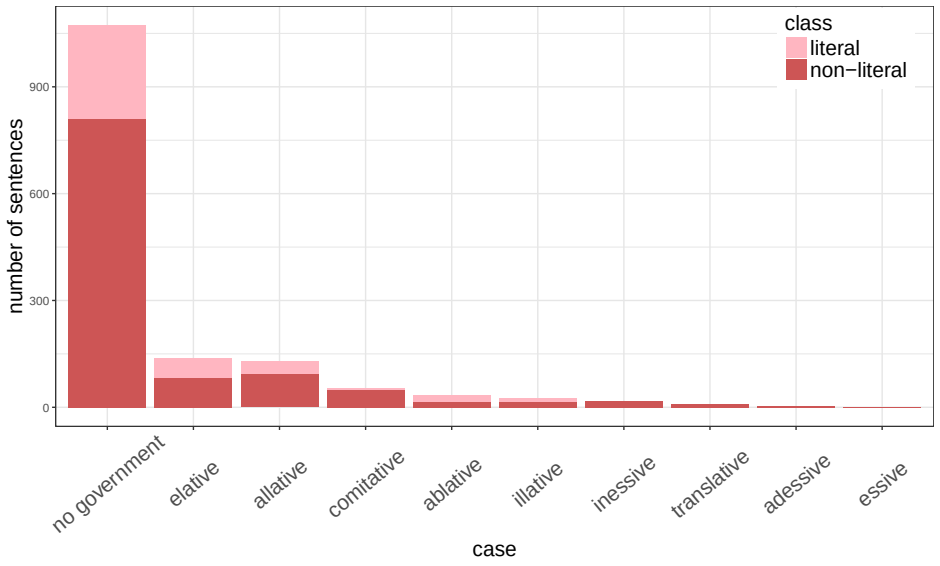

Figure 33: Distribution of the argument cases across all the sentences.

The distribution of the argument case across the literal and non-literal usages in correctly predicted sentences is illustrated in Figure 34. Most of the correctly classified sentences did not have case government. While the elative case was the most common in all the sentences, the allative was the most common in the correctly predicted sentences. This means that it was easier to predict the class of sentences with the allative case than it was of those with the elative case.

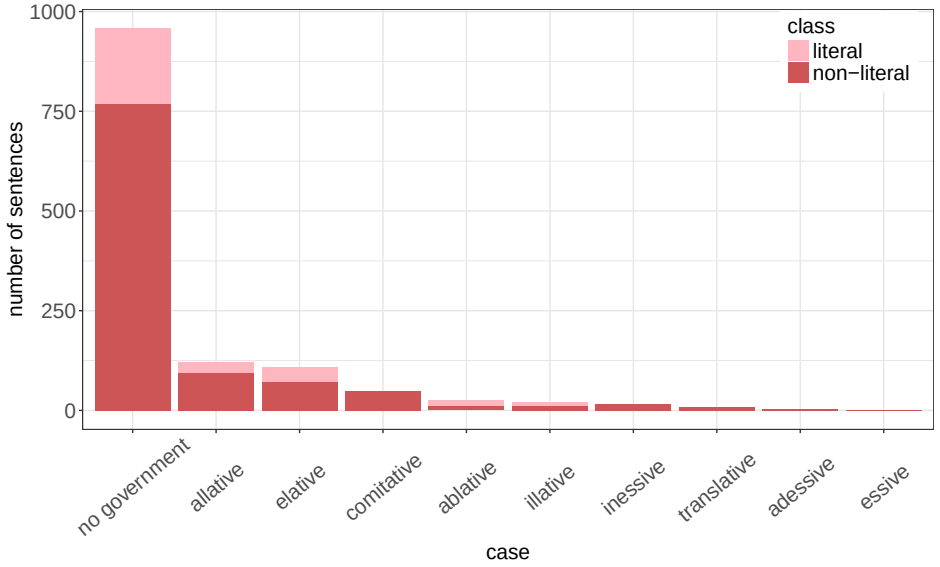

Figure 34: Distribution of the argument cases across the correctly classified sentences.

For some PVs, the difference between the literal and the non-literal meaning was very clear in terms of case government. For example, the PV *järele vaatama* 'to watch someone/check or investigate' appeared in 12 sentences – five of them were literal and seven were non-literal. The arguments in the literal sentences were all in the allative case (as in example (123)), while the arguments in the non-literal cases varied; for example, the argument in the sentence in example (124) is in the elative case.

(123) Ma sõit-si-n jalgratta-ga politseijaoskonna-st mööda ja üks
I drive-PST-1SG bicycle-COM police station-ELA past and one
politseinik **vaata-s** mu-lle pika-∅ pilgu-ga **järele**.
police officer look-PST.3SG I-ALL long-GEN look-COM after
Lit. 'I cycled past the police station and one police officer looked after me with a long look.'
'I cycled past the police station and one police officer followed me with his eyes.'

(124) Arvuti-st saa-b kohe **järele** **vaada-ta**, kes tolle-st
computer-ELA can-3SG now after look-INF who that-ELA
ringkonna-st riigikokku-∅ pääse-si-d.
region-ELA riigikogu-ILL get through-PST-3PL
Lit. 'It is possible to look after those from that region who got through to Riigikogu from the computer.'
'It is possible to look up those from that region who were elected to Riigikogu on the computer.'

Case government definitely helped to distinguish between the literal and non-literal usage in the PVs but, as with any other feature, it required supplementary information from other features. For example, the combination that did not include the case government feature (1–3, 5, 8–11) was able to classify one literal sentence in four sentences containing the PV *ümber tõmbama* 'to put something around somebody or something/encircle' correctly. The same combination with the case government feature classified three literal sentences correctly. Although the value of the case government was different for the non-literal sentence, other features did not provide sufficient information for the machine to make a correct prediction.

Case government can also (in combination with other features) lead to erroneous predictions. For example, of the 10 sentences containing *maha minema* 'to get off', two were classified incorrectly. While the reason for the faulty classification of the non-literal sentence might have been the concrete context (see Section 6.3.4.4), the misclassification of the literal sentence (see example (125)) might have been caused by the value of the case government. In fact, the elative argument appeared in both literal and non-literal sentences, but not in any other literal sentences containing this PV.

(125) **Läk-si-me** auto-st **maha** ja kohe tul-i punase-s
go-PST-1PL car-ELA down and immediately come-PST.3SG red-INE
ülikonna-s mees.
suit-INE man
Lit. 'We just went down from the car when the man in the red suit came.'
'We just got out of the car when the man in the red suit came.'

Overall, as the effect of the case government on the results was moderate, there were sentences that were only predicted correctly because of the information it provided. However, some sentences that the case government classified correctly received incorrect predictions from the best model because the values of the other features did not combine well with information about the case government.

6.3.4.10 Summary of the impact of the studied features on the classification

This analysis of the effect of the studied features suggested that nine of the twelve studied features contributed to achieving the highest accuracy when predicting the literal versus the non-literal usage of Estonian PVs. Three features that could

not provide useful information to improve the results – the average abstractness of words, subject and object abstractness – and were thus deemed irrelevant for the task. The poor effect of these features was due to a combination of reasons – the quality of the dataset of abstractness/concreteness ratings, the difficulty of evaluating the abstractness of words that were not nouns, and the lack of sentences in which an object was present

Information about a particle and a verb in a sentence is key to the successful prediction of the literal versus the non-literal usage of Estonian PVs. There were still more features that helped to improve the quality of the classifications. Amongst the remainder of the relevant features studied, the average abstractness of the nouns had the strongest impact on the results. The influence of the subject case, animacy features and case government was stronger than was the impact of the unigram and object case features. Nevertheless, all these features provided information that helped in the correct prediction of sentences that other features did not.

In fact, depending on the size of the impact, there were sentences that were only classified correctly because of the information provided by one feature. However, some of the sentences were labelled correctly because the information of the feature combined well with the predictions resulting from other features. In addition, there were sentences that are classified incorrectly due to the contradictory information provided by other features, and the more influential features led to a false conclusion. However, the reasons for the false predictions remain vague.

6.3.5 Summary of the classification of literal and non-literal usage

The experiments demonstrated that the main predictive features for distinguishing between the literal and non-literal usage of Estonian PVs were the components of the PVs – the particle and the verb. This essential combination of features detected 85.8% of the sentences successfully, while the verb alone predicted 82.0% of the sentences correctly. Although the unigram feature was the best single feature and predicted more non-literal sentences (F_1 n-lit 89.3) than did the verb (F_1 n-lit 88.4), the verb obtained a higher f-score for literal sentences (56.0 versus 59.4). Therefore, as there were fewer literal sentences in the dataset, the detection thereof was more challenging than was predicting the class of the non-literal sentences.

Of the 12 features studied – the particle, verb, unigram, the average abstractness of words, the average abstractness of nouns, subject abstractness, object abstractness, subject case, object case, subject animacy, object animacy and case government – nine contributed to achieving an accuracy rate of 88.7%. Whereas the improvement in the classification accuracy was not even 3.0% compared to the accuracy that the particle and the verb achieved, the features were important for improving the predictions of the correct class of both the non-literal and (more importantly) the literal sentences. In fact, the average abstractness of nouns, subject case, object case, subject animacy, object animacy and case government combined provided such a wealth of information that the f-score for literal usage increased up to 76.8. Briefly, this study suggests that the automatic classification of Estonian PVs benefits not only from standard features such as information about

particles, verbs and context abstractness, but also from language-specific features such as case information.

The analysis indicates that the results could be improved by creating a human-judgement based abstractness/concreteness dataset in which different meanings and the POS of words are identified. In addition, the possible and useful predictive features are not limited to the ones introduced in this study. For example, affective ratings other than abstractness scores, WordNet categories, clusters, vector-space word representations and so on have been used to train a classifier on samples of other languages (e.g. Shutova et al. 2013; Tsvetkov et al. 2014; Köper and Schulte im Walde 2016b). None of these features is argued further in this thesis but, as the interaction between frequency and compositionality has been discussed previously in this study (see Sections 4.2.4, 4.3.4 and 5.4.2) as well as by other authors (e.g. McCarthy et al. 2003; Bott and Schulte im Walde 2014), frequency as a predictor variable of the literal versus the non-literal usage of PVs is examined in the next section.

6.4 Frequency as a feature for detecting (non-)literalness

A considerable amount of literature has studied the correlation between frequency and the compositionality of MWEs. This was motivated by the results of the automatic extraction of MWEs in which the co-occurrence frequency worked well for some MWEs, such as German PVs (Krenn and Evert 2001). Therefore, some frequency-based AMs, such as PMI have been applied to predict the compositionality of PVs (e.g. Fazly and Stevenson 2006; Gurrutxaga and Alegria 2013). In recent studies, it has been demonstrated that AMs should not be used as estimators of compositionality. For example, Köper and Schulte im Walde (2016b) concluded that local mutual information (LMI) was not a successful predictor of the non-literal usage of German PVs, and Cordeiro (2017) demonstrated that PMI did not correlate with compositionality scores and should not be used as an estimator of compositionality. The results presented by Aedmaa (2016, 2017) suggested that the AMs studied (t-score, MI, X^2, log-likelihood and minimum sensitivity) should not be used to predict the degree of compositionality of Estonian PVs.

The effect of frequency on the compositionality of MWEs was also studied separately from the AMs. McCarthy et al. (2003) detected the compositionality of English phrasal verbs, and concluded that the frequency of the verb and particle did not have a significant relationship with compositionality judgements. The study of the compositionality assessments of German PVs (Bott and Schulte im Walde 2014) demonstrated that frequency determined the predictions. In fact, their model worked well for predicting the compositionality of PVs with medium frequency, but did not succeed in predicting the compositionality of infrequent and frequent PVs. Cordeiro (2017) explored the hypothesis that idiomatic MWEs should occur more frequently than should compositional ones in general human communication. As a result, it was proposed that frequent compounds were assigned higher compositionality scores. In summary, frequency has an impact on compositionality.

As noted previously (see Sections 4.2.5 and 4.3.4), frequency is associated weakly with the compositionality judgements assigned by humans. However,

due to the statistically non-significant correlations between frequency and compositionality scores, it cannot be stated that frequency does not influence the compositionality of PVs at all. Accordingly, the effect of frequency on the compositionality of Estonian PVs requires further investigation. This section discusses frequency as a predictor of the compositionality of PVs. Three features – the frequency of adverbs, the frequency of verbs and the frequency of PVs – were added to the best-performing feature sets of the main experiment. How the frequency feature combined with other features and whether they were able to improve the overall best results is explored.

6.4.1 Results for the frequency features

Each frequency feature (adverb, verb and PV frequencies) were studied as separate features. Information about the frequency was added to feature sets combining various numbers of features from the main experiment. The aim was not only to improve the results of the overall best-feature set, but also to analyse how information about the frequency combined with the information provided by other features. Therefore, the frequency features were not only added to the best-performing model; all the possible combinations of features were also trained and assessed.

Table 41 shows the best models using information about frequency. The best-performing model for each feature set size is presented for each frequency feature. In addition, some well-performing and interesting combinations are suggested. The information in the brackets shows the values for the accuracy and for f-scores of the combinations without the frequency features.

The results in Table 41 reveal that the best frequency feature was the frequency of the PV. However, in comparison to the features in the main experiment, the PV frequency did not attain greater overall accuracy than did the unigram feature (82.8%, see Table 27). At the same time, the f-score for literal usage was higher than was the score for the verb – 59.4 versus 62.8. It is therefore likely that PV frequency contributed to improving the overall best result.

The best-performing 2-feature model of the main experiment combined the particle and the verb (accuracy of 85.8%; see Table 28). When one of these foundational features was combined with information about the frequency, the results were similar. However, when the particle was combined with the PV frequency, an accuracy rate of 85.9% was obtained. This result is slightly better than that of the combination of the particle and verb. It allows us to hypothesise that PV frequency adds a similar amount of useful information to the task as does the verb.

While the results of the models with one and two features indicate that the frequency of the PV could possibly be used instead of the verb, further experiments demonstrated that, in combination with the other features, the PV frequency did not work as well as the verb. For example, the accuracy of the best 3-feature combination in the main experiment – the particle, verb and object animacy (1–2, 11) – was 86.6% (see Table 29), but the accuracy of the PV frequency combined with the particle and object animacy was 86.0%. The main drawback of using information about the frequency instead of the particle or the verb is the quality

Table 41: Results across the combinations containing frequency features. $\Delta\%$ and ΔF_1 indicate the absolute difference in percentage points (accuracy) and F_1 points between the models with and without frequency features.

features	size	accuracy		n-lit		lit	
		%	Δ%	F_1	ΔF_1	F_1	ΔF_1
majority baseline	0	74.0		85.1		0.00	
particle freq	1	74.0		84.6		17.6	
+ 2	2	85.8	+3.8	90.7	+2.3	69.7	+10.3
+ 2, 11	3	86.4	+3.7	91.0	+2.3	71.6	+9.4
+ 1–2, 5	4	85.8	−0.2	90.4	−0.2	72.3	−0.3
+ 1–2, 11	4	86.5	−0.1	91.1	−0.1	71.8	−0.2
+ 1–2, 5, 10	5	87.3	+0.3	91.6	+0.3	74.4	+0.4
+ 1–2, 5, 10, 12	6	88.3	+0.4	92.2	+0.2	76.0	+0.6
+ 1–2, 5, 10–12	7	88.4	0.0	92.3	0.0	76.5	+0.2
+ 1–2, 5, 8, 10–12	8	88.6	+0.1	92.4	+0.0	76.8	+0.2
+ 1–3, 5–6, 10–12	9	88.4	+0.1	92.3	+0.1	76.4	+0.2
+ 1–3, 5, 8–12	10	88.5	−0.2	92.4	−0.1	76.3	−0.5
+ 1–3, 5–6, 8–12	11	88.3	0.0	92.2	−0.1	75.8	+0.1
verb freq	1	81.9		88.4		59.3	
+ 1	2	85.7	+11.7	90.7	+6.1	69.5	+51.9
+ 1, 11	3	86.5	+8.6	91.1	+5.1	71.9	+25.0
+ 1–2, 5	4	85.8	−0.2	90.9	+0.3	72.4	−0.2
+ 1–2, 11	4	86.6	0.0	91.2	0.0	72.0	0.0
+ 1–2, 5, 10	5	87.6	+0.6	91.8	+0.5	74.7	+0.7
+ 1–2, 5, 10, 12	6	88.3	+0.4	92.2	+0.2	76.4	+1.0
+ 1–2, 5, 10–12	7	88.3	−0.1	92.3	0.0	76.2	−0.1
+ 1–2, 5, 8, 10–12	8	88.3	−0.2	92.3	−0.1	76.3	−0.3
+ 1–3, 5–6, 10–12	9	88.4	+0.1	92.3	+0.1	76.4	+0.2
+ 1–3, 5, 8–12	10	88.1	−0.6	92.1	−0.4	75.9	−0.8
PV freq	1	82.7		88.7		62.8	
+ 1	2	85.9	+11.9	90.8	+6.2	69.9	+52.3
+ 2, 11	3	86.4	+3.7	91.0	+2.3	71.6	+9.4
+ 1–2, 5	4	85.8	−0.2	90.5	−0.1	72.3	−0.3
+ 1–2, 11	4	86.6	0.0	91.2	0.0	72.0	0.0
+ 1–2, 5, 10	5	87.3	+0.3	91.6	+0.3	74.4	+0.4
+ 1–2, 5, 10, 12	6	88.3	+0.4	92.2	+0.2	76.2	+0.8
+ 1–2, 5, 10–12	7	88.4	0.0	92.3	0.0	76.3	0.0
+ 1–2, 5, 8, 10–12	8	**88.9**	+0.4	**92.6**	+0.2	**77.4**	+0.8
+ 1–3, 5, 8, 10–12	9	88.3	0.0	92.2	0.0	76.1	+0.1
particle freq, PV freq	2	85.8		90.7		69.6	
+ 1–2, 5–6, 8, 10–12	10	88.3	+0.7	92.2	+0.4	76.0	+1.3
+ 1–2, 5–6, 8–12	11	88.3	0.0	92.3	0.0	75.7	+0.1
verb freq, PV freq	2	85.7		90.7		69.5	
+ 1–2, 5, 8, 10–12	9	88.7	+0.2	92.5	+0.1	76.9	+0.3
+ 1–3, 5, 8–12	11	88.3	−0.4	92.3	−0.2	75.9	−0.9
particle freq, verb freq, PV freq	3	86.0		90.8		69.9	
+ 1–3, 5, 8–12	12	88.6	−0.1	92.4	−0.1	76.8	0.0
+ 1–3, 5–6, 8–12	13	88.3	0.0	92.3	0.0	76.0	+0.3
+ 1–6, 8–12	14	87.8	+1.0	91.9	+0.6	74.8	+2.3
+ 1–12	15	86.8	−0.6	91.3	−0.4	72.5	−1.1

of the prediction of literal sentences – none of the 3-feature combinations with frequency features achieved the results of the best 3-feature combination in the main experiment (an accuracy rate of 86.6%).

The results did not improve when adding one of the frequency features to the best-performing 3-feature combinations (1–2, 11 and 1–2, 5). Furthermore, the frequency of the particle decreased accuracy and f-scores. In addition, none of the 4-feature combinations achieved accuracy ratings and f-cores that were as high as those of the best 4-feature models with the original features (see Table 30). With or without the frequency features, the 4-feature combinations with the average abstractness of nouns (5) still predicted the literal usage of PVs better than did the combinations including the object animacy (11). On the other hand, when information about the frequency was available, the best f-scores for non-literal usage were achieved by the combinations that included object animacy. The greatest accuracy (86.6%) was achieved by the combinations including the verb frequency or the PV frequency combined with the particle, verb and object animacy.

The results of the best 5-feature combination with at least one frequency feature were worse than were the results of the best-performing 5-feature model in the main experiment (1–2, 5, 10, 12, see Table 31). However, as the best combinations included features such as the average abstractness of nouns, subject animacy and case government, it can be confirmed that these features were more influential than were the unigram, subject case, object case and object animacy features. Of the frequency features, in combination with other features, the verb frequency provided more useful information than did the particle frequency or the PV frequency.

The best 6-feature combination in the main experiment (1–2, 5, 10–12, accuracy of 88.4%, see Table 32) classified non-literal sentences correctly with an f-score of 92.3, and literal sentences with an f-score of 76.3. These results were very similar to those obtained by the best 6-feature combinations that included information about frequency. It is interesting that all the frequency features worked similarly in combination with the particle, verb, average abstractness of nouns, subject animacy and case government. In comparison to the other frequency features, the verb frequency provides knowledge that is better for predicting literal usage.

The best 7-feature model in the main experiment (1–2, 5, 8–10, 12) achieved an accuracy level of 88.7% (see Table 33). This result was not improved upon by any of the 7-feature combinations that included information about frequency. The best combination combined the particle frequency, particle, verb, average abstractness of nouns, subject and object animacy and case government. The results of the 8-feature combinations suggest that, if the particle frequency is replaced by the PV frequency, and the object animacy replaced by the object case in this combination, the accuracy of 88.9% is achieved. Compared to the best model in the main experiment (1–3, 5, 8–12), the accuracy was 0.2% higher. This setup classified non-literal sentences correctly with an f-score of 92.6, and literal sentences with an f-score of 77.4. Therefore, using the PV frequency instead of the unigram and object case (that is, a model containing features such as the PV frequency, particle (1), verb (2), the average abstractness of nouns (5), subject

case (8), subject animacy (10), object animacy (11) and case government (12)) resulted in the highest number of correctly predicted sentences. The results of this model compared to the results of the best model in the main experiments are discussed in Section 6.3.4.

The best-performing model in the main study (1–3, 5, 8–12) included nine features and achieved an accuracy level of 88.7% (see Table 35). The same result was obtained by the best nine-feature model containing information about the frequency. The results of the combination of the verb frequency and PV frequency + 1–2, 5, 8, 10–12 obtained the same accuracy and f-score for non-literal usage (92.5), which shows that the unigram feature and the object case could be replaced by the verb and PV frequencies to obtain the same results. This finding is not surprising because, as concluded in Section 6.3.3.2, the unigram feature and the object case were the less influential components of the best model in the main experiment.

Compared to the best 10-feature combination in the main study (see Table 36), the best 10-feature combination included the particle frequency, excluded the subject abstractness, and achieved slightly better results. Hence, the particle frequency provided more useful information than did the subject abstractness. Although subject abstractness was concluded to be irrelevant for the task of predicting the literal versus the non-literal usage of Estonian PVs (see Section 6.3.4), such a finding is not remarkable. The best 11-feature combinations obtained slightly lower scores than did the best 10-feature combination, but higher scores than did the best 11-feature combination in the main experiment (1–8, 10–12, see Table 37).

The results of the 12-feature models revealed that replacing the average abstractness of words, subject abstractness and object abstractness with the frequency features in the 12-feature model combining the original 12 features provided an improvement of 1.2% in accuracy, 0.7 points in the f-score for non-literal usage, and 3.2 points in the f-score for literal usage. Therefore, the frequency features worked better than did the irrelevant abstractness features (see Section 6.3.4). However, the combination (frequency features + 1–3, 5, 8–12) did not work quite as well as did the same feature set without the frequency features. This result indicated that, in combination with these features in the main experiment, the frequency features did not contribute to improving the quality of predicting the literal versus the non-literal usage of the PVs. The best 13- and 14-feature combinations included all the frequency features and achieved accuracy levels of 88.3% and 87.8%, respectively. Combining all the features resulted in an even lower degree of accuracy (86.8%) than did incorporating only the 12 features in the main experiment (87.4%).

Overall, the PV frequency helped to achieve the overall best results in the experiments. The effect of other frequency features was not as clear, but they could replace the irrelevant features (see Section 6.3.4.1) in some combinations. The analysis of the performance of the frequency features is presented in the following sections.

6.4.2 Analysis of the impact of the frequency features on predictions

The results described in the previous section showed that the overall classification results could be improved with the help of the PV frequency. The impact of PV frequency was studied via a comparison with the best model in the main experiment. As the role of the particle and verb frequency on the results is not clear, this section also examines the contribution of other frequency features. Therefore, the selection of all features in order to determine the role of the frequency features in comparison to the features in the main experiment is conducted first. Each feature is then analysed briefly. The third part of this section explores the influence of PV frequency on the best results.

The experiments for the feature selection for the main experiment were successful (see Section 6.3.1); thus, the 10-fold cross-validation of attribute selection using a learning scheme for all of the features was implemented. The aim was to determine how the frequency features performed in comparison to other features, particularly features with a low impact (unigram and object case). The experiment was motivated by the findings that frequency features provided information that was more influential than was the information provided by irrelevant or less influential features. The results of the attribute selection are presented in Table 42.

Table 42: Results for the learner-based feature selection for 15 features.

number of folds	attribute
10	particle
10	verb
7	unigrams
0	the average abstractness of words
9	the average abstractness of nouns
3	subject abstractness
1	object abstractness
4	subject case
2	object case
10	subject animacy
10	object animacy
10	case government
5	particle frequency
4	verb frequency
6	PV frequency

The best-performing combination in the main experiment had an accuracy rate of 88.7%, and included features such as the particle, verb, unigram, average abstractness of nouns, subject case, object case, subject animacy, object animacy and case government. The best-performing combination of experiments with frequency features received an accuracy rate of 88.9%. The latter combined

information provided by eight features – compared to the best model in main experiment, it excluded the unigram feature and object case, and included the PV frequency. The results of the attribute selection suggest that six features that the combinations shared were selected in a high number of folds – particle, verb, subject animacy, object animacy and case government in 10 folds, and the average abstractness of nouns in nine folds. The subject case, which was also a constituent of both combinations, was selected in four folds. The unigram feature and object case, which formed part of the best combination in the main experiment, were selected in seven and two folds, respectively. The PV frequency that replaced these two features in the overall best-performing combination was selected in six folds. All the other features were selected less often – particle frequency in five folds, verb frequency in four folds, subject abstractness in three folds, object abstractness in one fold and the average abstractness of words was not selected at all.

The results of the attribute selection confirmed that the PV frequency performed better than did the other frequency features. As the unigram feature did not form part of the model that produced the best overall result, it is surprising that the feature selection selected it in seven folds. On the other hand, the subject case that also provided useful information for the task was selected in four folds, which was less than the particle frequency. In general, features that did not form part of the best model were selected in a lower number of folds. Therefore, the automatic attribute selection supports the results described earlier in Section 6.4.1. It can be concluded that frequency predicted the compositionality of PVs, but that the particle, verb and PV frequency did not have the same impact.

Compared to the other frequency features, the **frequency of the particle** obtained the lowest accuracy and F_1 scores independently. This indicates that particle frequency had less of an impact on the prediction of the literal and non-literal usage of PVs than did other frequency features. Furthermore, the frequency of the particle was not part of the combination that achieved the best results. The best combination that included the particle frequency obtained an accuracy rating of 88.6%, which was lower than that obtained by the best combination in the main study. Although the frequency of the particle could not improve on the best result, it could be used instead of some other features in certain combinations. For example, the accuracy of the combination 1–3, 5, 8, 10–12 was 88.1%, while the f-score for non-literal usage was 92.2 and 76.1 for literal usage. The same combination using the particle frequency instead of the unigram feature achieved an accuracy rating of 88.6%, with the f-scores for non-literal usage 92.4 and 76.8 for literal usage. The difference between the f-scores for literal usage provides evidence that the frequency of the particle is a better feature than is the unigram for predicting the literal usage of the PVs.

The **verb frequency** received an independent accuracy rating of 81.9%, which is higher than the frequency of the particle and lower than the frequency of the PV. The results of the 2-feature combinations demonstrated that the combination of verb frequency and particle worked almost as well as did the particle and verb combination. Nonetheless, the best 2-feature combinations in conjunction with other frequency features performed better than did the one including the verb frequency. In the 3-, 4- and 5-feature combinations, verb frequency was one of the

components of the combinations with the best results, particularly when predicting the class of literal sentences. However, the overall best-feature set did not contain the verb frequency. In fact, the best combination of features that included the verb frequency obtained an accuracy rating of 88.7%. This result is the same as the best combination of features in the main experiment obtained. In fact, the results of the 9-feature combinations suggest that, together with the PV frequency, verb frequency could replace features such as the unigram and object case in the best-performing combination in the main task. Nevertheless, the result of this combination was not as good as without the verb frequency.

The **PV frequency** obtained the best results amongst the three suggested frequency features, with an accuracy level of 82.7%. It correctly classified non-literal sentences with an f-score of 88.7, and literal sentences with an f-score of 62.8. Compared to the unigram feature (82.8% and 89.3), the accuracy and the f-score for non-literal usage were not as high, but PV frequency outperformed the f-score for literal usage obtained by the verb (59.4). Therefore, of the frequency features studied, the PV frequency was the best for predicting literal usage. In addition, in combination with the particle, verb, average abstractness of nouns, subject case, subject animacy, object animacy and case government, PV frequency obtained an accuracy rating of 88.9%, which was the greatest degree of accuracy amongst all models. Accordingly, the PV frequency alone can replace the unigram feature and the object case, and attain better results.

Compared to the best model in the main experiment (the 9-feature classifier), the overall best model (the 8-feature model) did not obtain statistically significantly ($p< 0.05$) better results. However, the 8-feature classifier outperformed the baseline and all the other combinations of the studied features. It also required information from fewer features. In fact, the 8-feature classifier predicted three more sentences accurately than did the 9-feature classifier. Not all the incorrectly and correctly predicted sentences were the same. There were sentences that one classifier predicted correctly, but which the other predicted incorrectly. Specifically, the 8-feature classifier predicted 25 sentences accurately that the 9-feature classifier did not, and the 9-feature model classified 22 sentences correctly that 8-feature model did not. The discrepancies were caused by the fact that the 8-feature classifier excluded the unigram feature and object case, and included the PV frequency. Hence, there were sentences that were only predicted correctly due to the information provided by the PV frequency[102].

The distribution of the PV frequency across all the literal and non-literal sentences can be seen in Figure 35. Although the outliers are not visible in the figure, the PVs with the highest frequency appeared more than 30,000 times in the corpus in both literal and non-literal sentences. The PVs with the lowest frequency – 16 – also occurred in literal and non-literal sentences. The frequency of the PVs in literal sentences tended to be lower than it was in non-literal sentences. Accordingly, non-compositional PVs tended to occur more frequently than did compositional ones.

[102]The models classified most of the sentences correctly because they shared the majority of features. Therefore, the influence of these common features on the 8-feature classifier was similar to the influence on the 9-feature classifier. The analysis of these features in the 9-feature classifier can be found in Section 6.3.4.

Figure 36 shows the distribution of the PV frequency across the literal and non-literal sentences that were classified correctly by the 8-feature model. As with all the sentences, the PVs in literal sentences were less frequent than were the PVs in non-literal sentences. However, it is apparent that the literal sentences containing highly frequent PVs (with a frequency of more than 1,200) were not predicted correctly. Thus, it was challenging for the classifier to predict the correct class of these literal sentences. This observation indicates that one of the reasons for the incorrect prediction of literal sentences was the high frequency of the PVs.

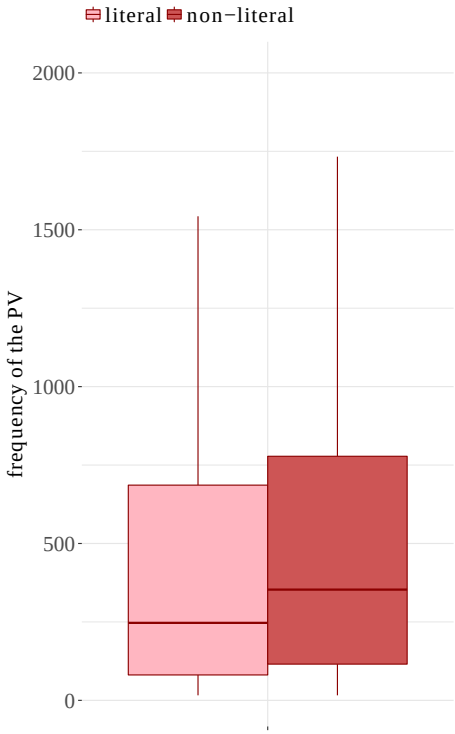

Figure 35: Distribution of PV frequency across all the sentences.

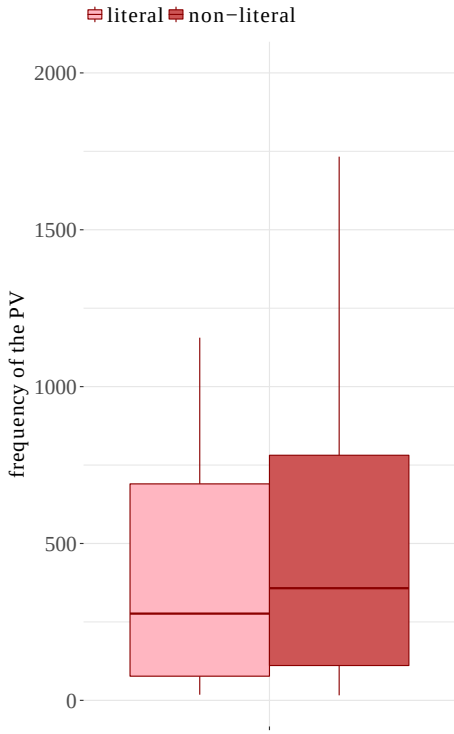

Figure 36: Distribution of PV frequency across the correctly classified sentences.

In some cases, the PV frequency differentiated between the literal and non-literal usages of PVs relatively well. For example, there were frequent PVs that only appeared in correctly classified non-literal sentences, such as *vastu võtma* 'to accept/welcome/admit' (with a frequency of 35,929), *ette nägema* 'to foresee/stipulate/see ahead' (with a frequency of 28,477) and *välja andma* 'to give out' (with a frequency of 19,743). In addition, there were infrequent PVs that only appeared only in correctly classified literal sentences, such as *järele vahtima* 'to stare after somebody' (with a frequency of 23), *lahti voltima* 'to unfold/ unwrap' (with a frequency of 28) and *alla tingima* 'to bargain/beat down' (with a frequency of 55).

The literal and non-literal usage was not identifiable via PV frequency in sentences that contained different meanings of the same PV. In order to determine which sentences benefitted most from the information about PV frequency, the sentences that the 8-feature model classified correctly were compared to the sentences that were classified incorrectly by the 9-feature model and the 8-feature model without information about the PV frequency (that is, 1–2, 5, 8, 10–12). This collation revealed sentences that were mainly classified accurately due to information about the PV frequency.

For example, the models without PV frequency wrongly predicted the class of the four sentences containing the PV *eemale tõukama* 'to push away/scare off/repel'. One of them received the correct prediction from the 8-feature model because of its relatively low frequency (221), which indicated that the sentences were literal. Compared to the other two literal sentences that were classified inaccurately despite the information about the PV frequency, the nouns in the correctly classified sentence were more abstract on average. The correct prediction was made because the PV frequency indicated that the PV would be literal, and it combined well with the average abstractness of the nouns, suggesting that the sentence would also be literal. Similarly, correct predictions were made for other sentences that were classified incorrectly by the models without information about the PV frequency. For example, one literal sentence containing the PV *kokku monteerima* 'to assemble/edit video' had similar feature values to a non-literal sentence. The PV frequency contributed to the making of a correct prediction, but only because the average abstractness of the literal sentence was slightly higher (4.4) than it was for the non-literal sentence (3.6). Other feature values were the same for these sentences.

Furthermore, there were two literal sentences containing the PV *ette andma* 'to put something in front of somebody/feed/specify' that had similar feature values. One of them was classified correctly without information about the PV frequency, while the other was not. The argument case was the only difference in their feature values. After adding information about the PV frequency, the correct prediction was made. Similarly, the sentence containing the PV *läbi vaatama* 'to look through/examine' was predicted correctly because the information about the PV frequency worked well in combination with information about object animacy and case government.

Nevertheless, there were cases in which the PV frequency was not sufficiently useful to lead to a correct prediction. For example, an incorrectly classified literal sentence containing the PV *välja pistma* 'to stick out' had similar feature values to a different, yet correctly predicted literal sentence. The difference was in the average abstractness of the nouns (the correctly predicted sentence had a lower score (6.92) than did the incorrectly predicted one (7.5)) and in subject animacy (the correctly predicted sentences had an animate subject, while the incorrectly predicted sentence had an inanimate subject). While the relatively low PV frequency and the high average abstractness of nouns indicated a high degree of literalness for the inaccurately predicted sentence, the sentence still received an incorrect classification, probably because of the inanimate subject.

Overall, similarly to other features analysed in Section 6.3.4, the frequency features did not distinguish perfectly between literal and non-literal usage. Non-

etheless, the evaluation of the models and the attribute selection demonstrated that all the frequency features performed better than did the irrelevant features in the main experiment. PV frequency was a better predictor of literal versus non-literal usage than were particle and verb frequencies. In fact, the PV frequency improved the overall classification accuracy of literal versus non-literal usages, meaning that there were sentences that the model only predicted accurately due to the PV frequency input.

6.4.3 Summary of the use of frequency to predict (non-)literalness

Previous research on the effect of frequency on the compositionality of MWEs encouraged the study of whether information about frequency would be useful for predicting the literal versus the non-literal usage of Estonian PVs. Therefore, three frequency features – the particle, verb and PV frequency – were included in the feature set, and their influences on the results were studied.

The particle and verb frequencies contributed to the classification of the Estonian PVs according to their usage. However, the overall best results in the main experiment did not improve following the addition of these features. On the other hand, particle and verb frequency features could be used as replacements for features such as unigrams or the object case in order to achieve the same results as the best-performing feature combinations in the main experiment.

The results of the experiments and attribute selection suggested that all the frequency features performed better than did the irrelevant features in the main experiment. The best frequency feature was PV frequency, which provided information that led to an improvement in the overall best result. In fact, after omitting the features that had a low impact on the best result in the main experiment – the unigram feature and the object case – the model subsequently produced a slightly higher number of correct predictions than did the best model in the main experiment. Thus, the overall best model combined eight features – the particle, verb, average abstractness of nouns, subject case, subject and object animacy, case government and PV frequency – and classified 88.9% of the sentences correctly. Non-literal sentences were classified correctly with an f-score of 92.6, and literal sentences with an f-score of 77.4.

In addition to achieving better results than those of the best-performing model in the main experiment, the model using the PV frequency feature combined fewer features. As information about the frequency is relatively easy to extract from the data compared to the linguistic information, this factor can be crucial when preparing data for detecting the literal versus the non-literal usage of Estonian PVs. A further discussion of this topic is presented in Section 6.5 as part of the discussion of the results of detecting the non-literal usage of PVs.

6.5 Summary and discussion of detecting the literal and non-literal usage of particle verbs

In this section, the classification of the literal versus the non-literal usage of Estonian PVs is finalised. In addition, a brief discussion about overfitting models

is provided. Moreover, the results are analysed from the data acquisition point of view, as the labelled data are a prerequisite for supervised learning. Therefore, in addition to determining the predictive features for the task, it is important to investigate the challenges to obtain properly labelled data.

The goal of the classification task was to develop a classifier that predicted the correct class of literal and non-literal sentences. The classes of the sentences were created based on the (non-)literalness ratings submitted by three human annotators who evaluated the meanings of the PVs in the selected sentences. A set of features that could predict the target class (literal versus non-literal sentences) was introduced, and the evaluation was performed on all possible feature combinations, examining the impact of each suggested feature.

The set of studied features contained attributes that were used previously for the detection of the non-literal usage of MWEs, as well as features that had not been applied to similar tasks previously. More specifically, the effect of the following 12 features was studied in the main experiment – the particle, verb, unigram feature, average abstractness of words and nouns, abstractness, case and animacy of PV subject and object and case government (the case of the PV argument, except for the subject and object). In addition, the influence of three frequency features was explored. The 1,481 sentences that were assigned (non-)literalness ratings and all the features studied now constitute a freely available dataset of (non-)literalness ratings for Estonian PVs (see Aedmaa 2018).

The study demonstrated that, of the 12 suggested features, nine were relevant for the task. The average abstractness of words, subject and object abstractness scores did not provide such information that, in combination with other features, improved the overall classification prediction. While context abstractness is proposed to be helpful for predicting the (non-)literal usage of MWEs, the results might have been affected by the quality of the abstractness/concreteness dataset.

Some relevant features influenced the results more than did others. The most influential features were the particle and verb – these components of the PV assisted in the correct classification of almost 86.0% of all the sentences. Other features contribute less, but still helped to improve the classification results. In combination with the particle and verb, the average abstractness of nouns had a stronger impact on the results than did other features. Those that were moderately effective included the subject case, subject animacy, object animacy and case government. The impact of the unigram feature and the object case was weak. The information provided by the latter feature could easily be replaced by information about the PV frequency in order to produce a more accurate classifier. In fact, as the information about the PV frequency was a useful feature for classifying Estonian PVs according to literal versus non-literal usage, it was concluded that frequency had an impact on the compositionality of PVs. Other frequency features – particle and verb frequency – were not irrelevant for the task, but they did not provide information that improved the overall accuracy of the classifications.

The current research on the automatic detection of the literal versus the non-literal usage of Estonian PVs is a continuation and extension of work described by Aedmaa et al. (2018). The slight revision of the dataset caused slight differences in results, but the general conclusions remain the same – while the particle-verb information classified the majority of sentences correctly, the context abstractness,

case and animacy information contributed to improving the results. The overall best accuracy proposed in the previous study was 87.9%, which is 1% lower than was the accuracy of the best model introduced in this study. However, the previous authors did not analyse the effect of frequency on the results.

As other previous research on the automatic detection of (non-)literal language usage was not carried out on Estonian, a full comparison of the results is not possible. Nevertheless, there has been some research on the automatic detection of the non-literal usage of MWEs in other languages. Therefore, the general results of those studies are compared to the outcome of the current research.

The approach of the present task is most closely related to the study by Köper and Schulte im Walde (2016b), which distinguished between the literal and the non-literal usage of German PVs. These authors demonstrated that affective ratings (including contextual abstractness) could improve the overall performance of the classifier, and noted that the features concerning nouns were more useful than were those involving other POS. This matches the findings of this study to some degree. While the average abstractness of nouns provided useful information for classifying the literal versus the non-literal usage of Estonian PVs, the subject and object abstractness (which are also noun-based features) were concluded to be irrelevant for the task. The usefulness of the abstractness scores for non-literal language usage detection was demonstrated previously by Turney et al. (2011), who detected the literal and metaphorical usage of English adjective-noun expressions, and by Tsvetkov et al. (2014), who carried out metaphor detection in the English, Spanish, Farsi and Russian languages.

In addition to the abstractness ratings, the current study shares another feature with Köper and Schulte im Walde (2016b) – unigrams. While the unigram feature attained higher accuracy independently than did any other studied feature, it had no remarkable impact on the results in combination with other features, particularly when information about the PV frequency was available. However, when detecting the literal versus the non-literal usage of German PVs, the unigram feature worked well in combination with other features, including abstractness.

The present research revealed that literal (compositional) PVs occurred less frequently than did non-literal (idiomatic) ones. This result contradicts that suggested by Cordeiro (2017), who demonstrated that idiomatic MWEs did not occur more often than did compositional MWEs. However, those results are not fully comparable with ours because their dataset was balanced and they studied compound nouns. Nonetheless, the results indicate that frequency has an impact on the compositionality of MWEs.

Whereas supervised learning requires labelled data, the concern was not only to determine the predictive features for the task, but also to acquire annotated data. Therefore, the results of the classification task described earlier are discussed from the point of view of data preparation. As explained previously, annotated data are a prerequisite for supervised machine learning. High-quality labelled data are costly – they may require human resources, money, time and tools, for example. In the following section, the results of the classification task are discussed from the point of view of data annotation. Following a general overview of how the data were prepared for the classification task, the (available) tools and datasets that could be used for less expensive labelled data are discussed.

The preparation of the dataset for classifying the literal versus the non-literal usage of Estonian PVs (Aedmaa 2018) began with the creation of the dataset of (non-)literalness ratings of Estonian PVs. The formulation of this resource was described in detail in Section 4.3. As the aim of the task was to build a classifier that predicted the correct class (literal or non-literal) of the sentences, the values of the target classes were derived via human annotation. The values of other features – the particle, verb, unigram feature, abstractness, case, animacy and frequency features – were acquired after the evaluation of the annotation was conducted. The annotation of the features was designed to be as automatic as possible.

After the sentences were assessed according to the literalness of the PVs, it was clear that information about the lemmas of the sentences in the dataset was crucial for all the features. Therefore, the first phase in the data preparation was to lemmatise all the sentences. This provided the necessary information for the particle, verb and unigram features. The information for the abstractness ratings was derived from automatically created abstractness/concreteness datasets, as introduced in Section 4.5. The abstractness of nouns also required POS tagging of the features concerning subjects and objects (that is, subject and object abstractness, and case animacy) for the syntactic analysis. The animacy of the subject and object were annotated manually after the subject and object were detected automatically. The annotation of the case government features benefitted from dependency parsing to detect the arguments. In addition, the sentences containing multiple arguments were assessed manually to determine the one that was taken into account in the annotation.

A fair number of useful NLP tools for Estonian was available for the annotation. Most of the tools are available through the Python EstNLTK module[103] (Orasmaa et al. 2016). For example, it includes the Estonian morphological analyser Vabamorf[104] that can be used to extract information about the POS and cases. As Vabamorf is highly accurate (99%) (Kaalep 1997; Kaalep and Vaino 2001), it is particularly useful for automatic data annotation. The morphological analyser provided input for the morphological disambiguation in the Estonian Constraint Grammar (CG) parser (Müürisep 2000) that helped to detect subjects, objects and other arguments. The parser also employs rules that determine clause boundaries, surface syntactic analysis and dependency relations (Muischnek et al. 2017). Also, the CG parser has a special module for identifying PVs that works with a high degree of precision and recall (both at 97.4%) (Muischnek et al. 2013). Therefore, the morphological analyser and the parser were useful tools for the automatic extraction of the information necessary for the annotation. These tools provided information about lemmas, clause boundaries, POS, cases and dependencies.

It is important to note that, as the sentences in this dataset were selected manually in order to collect (non-)literalness ratings from human annotators, automatic PV discovery was not applied here. However, it is very important to be able to discover PVs in textual data in which the sentences containing PVs alternate with those without PVs. In addition, as the results of the classification task demonstrated, particles and verbs classified more than 85% of sentences

[103] `https://github.com/estnltk` (accessed 29.05.2018).

[104] `https://github.com/Filosoft/vabamorf` (accessed 29.05.2018).

correctly. Therefore, the successful automatic discovery of PVs is important for the annotation of the data in order to detect the literalness of the PVs. Based on the previous work on the automatic detection of PVs described in Section 2.4.3, the automatic discovery of PVs could be carried out successfully using AMs. Furthermore, while the methods for statistical approaches to the automatic discovery of PVs are still at the research stage, the model for discovering PVs in the rule-based parser could be used prior the identification of PVs.

It is clear that, with the help of the existing tools, the lemmatisation, detection of clause boundaries, POS tagging and parsing are not problematic aspects of data labelling. The biggest challenges concern the abstractness and animacy features, which require a special dataset containing information about Estonian lemmas. The automatically created abstractness/concreteness dataset provides useful information, but the quality thereof is unknown. At the same time, the average abstractness of nouns is one of the most influential features for good classification results. It is thus important to create a dataset containing abstractness/concreteness ratings for Estonian lemmas in which the POS and the meanings of the lemmas are identifiable.

Other challenging features that are also important for the classification task are those concerning animacy. As there are no existing datasets containing information about the animacy of Estonian lemmas, developing such resources and other tools is necessary. For example, automatic animacy classification has reached accuracy levels of more than 90% (e.g. Orasan and Evans 2001) and several binary (e.g. Moore et al. 2013), as well multi-class animacy datasets (e.g. Bos et al. 2017) have been created. These findings are encouraging with regard to creating a tool for the automatic classification of animacy or animacy-annotated datasets for Estonian. Until these resources become available, the data need to be annotated manually.

While information about the PV arguments and their cases is provided by the parser, there is a problem with the case government feature. More specifically, when the PV has multiple adverbials in different cases, which case in which argument should be used for the annotation is not automatically detectable. This does not involve the majority of sentences, and could thus be performed manually.

Taking the results of the classification task into account, in order to train the overall best classifier, it is important to annotate the values of the following features – the particle, verb, average abstractness of nouns, subject case, subject animacy, object animacy, case government and PV frequency. With the help of the tools developed previously, the necessary information for most of the features can be acquired easily and automatically at present. The most problematic features from the point of view of automatic data labelling are the animacy features, as there are no resources available for Estonian that would guarantee the automatic annotation of these features. Some other features can be obtained automatically.

In general, the classification of the literal versus the non-literal usage of Estonian PVs suggested that certain contextual (abstractness), grammatical (cases, animacy) and statistic (frequency) features contribute to the detection of whether a PV was used with a literal or a non-literal meaning. Whereas the use of frequency and context abstractness was not surprising for the task of detecting non-literal language usage, the impact of information about the case is definitely a novel

finding. However, as data are crucial for computational research in linguistics, the use of these features can be somewhat limited. More specifically, the current study showed that, while many useful tools are already available for data acquisition, there is still a need for some resources that provide information about animacy and abstractness.

7 CONCLUSION

The present thesis used computational methods for the detection of Estonian PVs. DSMs were trained to learn word and multi-sense embeddings that were used to predict the compositionality of PVs. A supervised classifier was developed to predict the literal versus the non-literal usage of PVs. The compositionality of Estonian PVs was explored based on a representative selection of Estonian PVs containing various particles and verbs with diverse frequencies. Special attention was paid to the association between frequency and compositionality by studying distinct aspects, such as how the frequency of PVs and of their components affected human compositionality judgements and the compositionality predictions of models, as well as how well information about frequency predicted the literal versus non-literal usage of PVs.

The contributions of the thesis are both theoretical and practical. On the theoretical side, the compositionality of the PVs has been discussed and analysed in depth. The difficulty of the phenomena was illustrated via numerous examples from corpora. The human compositionality judgements were analysed and compared to claims from earlier computational and theoretical research. For example, the analysis of compositionality judgements suggested that the explanation for the formation of a PV's meaning provided in the literature thus far is limited to qualitative approaches. More specifically, the human judgements illustrated well that a binary classification of Estonian PVs is insufficient, as compositionality occurs along a continuum.

The main practical contributions are the computational models that were developed: the DSMs that predicted the compositionality of the PVs using both word and multi-sense embeddings, and the supervised classifier predicting the literal versus the non-literal usage of PVs. With this being the first exhaustive computational study exploring the compositionality of Estonian MWEs, the application of the models was described in such a way that provides general guidelines for future research, not only on the compositionality of MWEs, but also for other areas of linguistics that could employ the same methods. In addition, the proposed model for predicting the non-literal usage of PVs achieved high levels of accuracy, and could thus be employed for downstream applications, for example, machine translation. Furthermore, three novel datasets for Estonian were created, analysed and made publicly available for research.

The three main goals of the thesis were: 1) To detect the compositionality of Estonian PVs automatically, 2) to introduce and apply methods that are widely used for different tasks in NLP, and 3) to provide outcomes (including resources) that are encouraging and helpful for future computational studies of the compositionality of a wide range of MWEs and other phenomena. In order to accomplish the goals, the following research questions were proposed:

1. To what extent do human annotators agree with each other when evaluating the compositionality of PVs? What are the main reasons for disagreement?

2. How well do DSMs predict the compositionality of Estonian PVs? Which training parameters and other aspects influence the quality of word and multi-sense embeddings for detecting the degree of compositionality?

3. Which (linguistic) features predict the usage of compositional versus non-compositional PVs? How well are the values of these features acquirable automatically?

4. How does frequency affect the compositionality of Estonian PVs? How are human judgements of PV compositionality and automatic compositionality predictions associated with frequency?

5. Are widely adopted and successful computational methods suitable for detecting the compositionality of Estonian PVs? What are the drawbacks and benefits of these methods?

The answers to these questions are summarised as the main conclusions provided in the next section. The main results of this thesis are compared to the studies of the compositionality of MWEs in Section 7.2. The chapter concludes with Section 7.3, in which the main drawbacks of the current research are discussed and suggestions for future work are provided.

7.1 Main conclusions

In this thesis, the compositionality of Estonian PVs was studied by applying computational methods – DSMs trained to learn word and multi-sense embeddings to predict the compositionality of PVs, and a supervised classifier was developed to predict the literal versus the non-literal usage of PVs. These approaches were selected based on the results of previous studies of MWE processing, general tendencies in the field of NLP, and the applicability to Estonian data. This section presents the main conclusions of the study.

The evaluation of models generally involves comparing the predictions to the datasets (gold standards) that contain information provided by humans. Accordingly, two datasets contained human judgements of Estonian PV compositionality were created. Although both datasets contain information about the compositionality of Estonian PVs, the methods used to collect the compositionality judgements were dissimilar. The first dataset – compositionality ratings – was collected via crowdsourcing with the aim of obtaining one compositionality score per PV and ignoring the potential ambiguity of the PVs. Having one score per PV allowed us to compare the ratings with the compositionality predictions of word embedding models because they also produce one score per PV. The second dataset – literalness ratings – contained the compositionality scores for PV meanings because the PVs were evaluated based on a given context (a sentence). The dataset was mainly created for modelling the non-literal usage of PVs, which inspired the name of the dataset.

The assessments of the annotations of the datasets demonstrated that the compositionality of some PVs and their meanings were more difficult to assess than were other PVs and their meanings. The main reason for the disagreement was the ambiguity of the PVs and their components, particularly particles. However, the standard deviation among the compositionality ratings suggested that less than 10% of the PVs caused substantial disagreement amongst the annotators. The inter-annotator agreement on the literalness ratings implied a fair degree of

agreement for all six categories ($\kappa = 0.36$, $\alpha = 0.68$) and substantial agreement at the binary (literal versus non-literal) level ($\kappa = 0.71$, $\alpha = 0.71$). Thus, taking the difficulty and subjectivity of the task into account, it can be concluded that the annotators agreed sufficiently with each other in terms of their evaluations of the compositionality of Estonian PVs.

The automatic detection of PV compositionality was carried out using DSMs trained to learn word and multi-sense representations. Word embedding models are widely used for performing different NLP tasks, including MWE processing. Multi-sense representations are used less often, particularly for compositionality prediction, and further and more exhaustive studies are therefore needed. Both types of embeddings were applied to the same task – to predict the compositionality of Estonian PVs – and proved to be suitable for the task. The comparison of the word and multi-sense embedding models suggested that word representations were preferable for the task, but the evaluation of the multi-sense embedding models was carried out on the data that was not created for this purpose. Nevertheless, the quality of the predictions can be concluded as being fair, as the best predictions correlated rather weakly (with compositionality ratings $\rho = 0.21$, with literalness ratings $\rho = 0.46$) with the human-annotated ratings. The quality of the predictions did not vary significantly across the training parameter configurations that were studied. namely the number of dimensions, window size, minimum-count threshold and the number of iterations. The latter had the strongest impact on the results.

The algorithms employed, the types of the embeddings, the evaluation datasets, and frequency affected the results more than other parameters that were studied. The comparison of the two word2vec architectures studied – CBOW and Skip-gram – revealed that the best predictions were provided by the word embedding model using the CBOW architecture and compared to the dataset containing literalness scores. While it is difficult to explain why the CBOW model performed better than did the Skip-gram model, there are clear reasons for the differences in other situations. For example, the human-annotated datasets were not created for the same task and did not employ identical methods; therefore, there may have been a difference in the quality of these datasets. The models using multi-sense embeddings provided significantly less accurate predictions than did the word embedding models because the gold standards were not created for the assessment of the multi-sense embedding models. Therefore, the research on multi-sense embeddings needs to be continued via the application of revised evaluation methods. However, both types of models used in the current study predicted the compositionality of frequent and infrequent PVs more accurately than they did the PVs with moderate frequency. There were two main reasons for this – the compositionality of the frequent PVs was predicted well because the vectors of the frequent PVs were more representative, and the compositionality of the infrequent PVs was easier to predict because they were less ambiguous. The vectors of PVs with moderate frequency were not sufficiently representative to compensate for the poor quality of the predictions of ambiguous PVs.

The thesis investigated a set of features that, based on the previous theoretical and computational research, could predict the non-literal usage of the PVs, but which had not been applied to such a task previously. The goal of detecting the

(non-)literalness of Estonian PVs was to develop a supervised classifier to predict the class of sentences in which PVs are used either with their literal or non-literal meanings. The sentences were divided into two sets based on the average literalness scores assigned by human annotators on a 6-point scale. Twelve features that could be divided according to standard, language-independent and language-specific features were introduced. The result of the token-based classification suggested that, of the standard features – the PV components, unigrams and abstractness of the context – information about the particles and verbs was crucial for predicting the compositionality of the PVs. Of the abstractness features, only the average abstractness of the surrounding nouns had an impact that helped to improve the overall classification accuracy in combination with information about the particles and verbs. The unigram feature was the most successful for the task when information about the other features was not available. Of the language-specific case features, the subject case and case government were more influential than was the object case. In addition, information about the language-independent features, subject and object animacy, was sufficiently useful for the system to predict the class of the sentences with a degree of accuracy of almost 89%. The analysis of the frequency features demonstrated that PV frequency provided such a wealth of information that unigrams and the object case were not necessary for predicting the class of literal versus non-literal PVs in the absence of information about when information about the PV frequency was available.

The successful automatic classification of literal and non-literal sentences is possible when relevant features are annotated in the training data. Some of the features, such as unigrams and frequency, are relatively easy to acquire automatically, but there are features that are more costly to annotate. As morphological analysers and parsers are available for Estonian, information about the lemmas, cases and syntactic roles can be obtained easily. At the same time, the automatic acquisition of information about animacy is currently problematic because no appropriate resources are available. In addition, the dataset containing information about abstractness was created automatically for the current study and has not been evaluated. In brief, there are some open issues concerning data acquisition for the future. Overall, the annotation for the detection of the compositionality (literalness) of Estonian PVs is mainly obtainable automatically from the textual data; therefore, the use of supervised learning is much less expensive than is manual annotation.

In general, both the computational methods discussed are suitable for detecting the compositionality of Estonian PVs, but both approaches also have some drawbacks and benefits that need to be considered. For example, the DSMs do not require anything other than a large, lemmatised text corpus and a high-performance computer; therefore, this method is definitely a less costly method than is supervised learning. However, while the processing of the DSMs is practical, it is not clear whether DSMs can really address deeper semantic questions (Lenci 2008). Hence, the analysis of the results of the DSMs is somewhat superficial and conjectural. As expected, supervised classifiers were more accurate than were DSMs. While the acquisition of the labelled training data necessary for supervised learning can be expensive, this study showed that most of the annotations could be done automatically fro Estonian. Hence, the cost of using supervised

learning could be decreased considerably. In addition, the detailed analysis of the results of the classifier provided a deeper understanding of the compositionality of Estonian PVs than did the analysis of the results of the DSMs. For example, a set of language-specific features was suggested to predict the non-literal usage of Estonian PVs. However, it is important to note that the generalisation power for unseen PVs of the supervised model remained unexplored in the current study. Therefore, this should be definitely addressed in future work. Overall, both methods had their drawbacks and benefits – while supervised learning was more expensive, the DSMs were less accurate. An experiment combining these methods would be another potential direction for future research.

The effect of frequency on the compositionality of Estonian PVs was studied from different perspectives. The study included PVs with different frequencies, which allowed for the investigation of the degree of which the compositionality predictions depended on the frequency of the PVs. No statistically significant association between the human compositionality judgements and frequency of the PVs and their components was found. However, the compositionality predictions of DSMs were associated with frequency – the predictions of the frequent and infrequent PVs correlated with frequency in such a way that the models tended to predict that the most frequent and infrequent PVs would be less compositional than would the PVs with moderate frequency. In addition, information about the PV frequency is helpful to increase the classification accuracy of the literal versus non-literal usage of Estonian PVs. Therefore, the compositionality predictions of the computational models were affected by frequency, but the influence on the human judgements needs to be investigated further. Ways for increasing the accuracy of compositionality predictions for PVs with moderate frequencies also need to be addressed in the future.

In general, the results of the thesis are definitely novel from the perspective of the automatic processing of Estonian MWEs, specifically PVs. In addition, they provided corpus-based findings that extend the description of Estonian PV compositionality provided previously in the literature. The comparison to the related work presented in the next section demonstrates the importance of the results for the wider community.

7.2 Comparison of the results and other studies of the compositionality of MWEs

Although it focused solely on PVs, this dissertation is the first large-scale computational study of the compositionality of Estonian MWEs. Therefore, any comparison with previous work is somewhat constrained. However, the results of the first experiments for detecting the compositionality of Estonian PVs and for predicting the literal versus the non-literal usage of Estonian PVs have been published in peer-reviewed publications (see the overview in Section 2.4.3). The main conclusions of the current study are not only comparable to the work done on the compositionality of Estonian PVs, but also to the research conducted on the MWEs in other languages. The aim of the comparison is to determine if and how the automatic processing of Estonian PVs differs from the treatment of similar phenomena in other languages when applying the same methods.

The first study exploring the compositionality of Estonian PVs (Aedmaa 2017) argued that the compositionality predictions correlated weakly, but statistically significantly, with the human compositionality judgements. This analysis suggested that the compositionality of frequent PVs was more difficult to predict than was that of other PVs. This finding is in line with the conclusions of Bott and Schulte im Walde (2014), who found that the compositionality of infrequent and frequent German PV was more challenging to predict using word embedding models than was the compositionality of other PVs due to data sparseness and the high degree of ambiguity of frequent PVs. As the overall correlation between the compositionality predictions of German PVs and human judgements was moderate but statistically significant, the word embedding models generally function similarly for Estonian and German PVs. Nevertheless, both word and multi-sense embedding models predicted the compositionality of frequent and infrequent Estonian PVs with greater accuracy than they did the compositionality of PVs with moderate frequency. This result may have differed from that proposed by Aedmaa (2017) because a bigger corpus and different parameter configurations were used in the current study – Cordeiro (2017: 86) proposed that the corpus size affected the quality of DSM representations strongly. While the success of the DSM predictions depended on the frequency of Estonian and German PVs, Cordeiro (2017) demonstrated a generally weak correlation between frequency and the difficulty of predicting the compositionality of nominal compounds in English, French and Portuguese, but there are also models that have predicted the compositionality of frequent compounds with significant accuracy. In brief, the results of the effect of frequency on the DSMs' compositionality predictions are inconclusive, and require further investigation.

No statistically significant correlation between human compositionality judgements and (PV, particle and verb) frequency was detected in this study. This finding is consistent with that of McCarthy et al. (2003), who studied English phrasal verbs and claimed there was no significant relationship between compositionality judgements and the frequency of the components of phrasal verbs. Bott et al. (2016) observed a slight variation in German PV compositionality ratings and frequency bands, indicating that frequency did not affect human judgements to a significant degree. Cordeiro (2017) hypothesised that there should be a negative correlation between the compositionality score and the frequency of the compound. However, he found that frequent English and French compounds were evaluated as being more compositional than were the less frequent ones. There was a significant correlation between the compositionality ratings of Portuguese compounds and frequency. Therefore, the association between frequency and human compositionality judgements is not generally strong, but can depend on the type of the MWE and the language in question. In addition, the results may vary according to the methods used for the collection of judgements.

The inter-rater agreement on the compositionality and literalness ratings of Estonian PVs was assessed as being relatively good. The compositionality ratings were crowdsourced, and the agreement was assessed using standard deviation. After removing PVs that were difficult to evaluate (almost one third of the PVs evaluated) and outlier annotators, about 10% of the PVs were reported as having high standard deviation values, indicating high levels of disagreement amongst the

annotators. The main reason for disagreement was the ambiguity of the particles. Reddy et al. (2011b) suggested a similar number of English compound nouns that caused disagreement amongst the annotators; they hypothesised that the main reason for the differences was the ambiguity of the compounds. Cordeiro et al. (2016b), who discussed the quality of human compositionality judgements with English, French and Portuguese MWEs, stated that the proportion of high standard deviation compounds was between 14% and 19% after filtering (removing outlier annotators and annotations). The comparison revealed that regardless of the type, or language of the MWEs, the annotators did not agree about the compositionality of some MWEs. Bott et al. (2016), who collected compositionality ratings for German PVs, reported that the high average standard deviation per rating (1.82 on a 6-point scale) reflected the difficulty of the annotation task. Therefore, the evaluations of the crowdsourced dataset suggest similar conclusions – the task might be challenging due to the ambiguity of the MWEs.

Perfect agreement was also not achieved when the annotations were provided by experts and evaluated via kappa coefficients. When assessing the agreement for Estonian PVs literalness ratings, the agreement (after filtering) for six categories was $\kappa = 0.43$ and $\kappa = 0.71$ for two categories. Similar inter-rater agreement was also found amongst the annotators of German PV literalness – $\kappa = 0.35$ for six categories and $\kappa = 0.70$ for two categories. Comparable results were also found in the assessments of other MWEs. For example, three experts who evaluated Basque noun-verb expressions as being idioms, collocations or free combinations, achieved moderate agreement ($\kappa = 0.58$) (Gurrutxaga and Alegria 2013). Farahmand et al. (2015), who collected annotations for 1,048 English noun-noun compounds to be assessed based on their (non-)compositionality and (non-)conventionalisation. They also found moderate agreement for non-compositionality ($\kappa = 0.62$). Hence, regardless of the number of annotators and their proficiency, the compositionality (literalness) of MWEs is difficult to evaluate. The reasons for the disagreement are universal – the subjectivity of the task and the ambiguity of the MWEs and their components.

There is relatively little research on the use of multi-sense embeddings for detecting the compositionality of MWEs. The results of the current study are thus truly comparable only to the study detecting the compositionality of German PVs. Köper and Schulte im Walde (2017b) found that multi-sense embeddings resulted in better predictions than did word embedding models, but the correlation between human compositionality judgements and the model's predictions was still relatively weak. We found that the multi-sense embeddings did not provide better compositionality predictions for Estonian PVs than did the word embedding models. Therefore, in contrast to Köper and Schulte im Walde (2017b), the superiority of multi-sense embeddings in comparison to word embedding could not be stated categorically. Li and Jurafsky (2015), who found that multi-sense embeddings often performed as well as did word embeddings with high dimensionality, discussed the usefulness of multi-sense embeddings for NLP in general. They suggested testing embeddings in real NLP applications rather than via simple human-matching tasks. The application of multi-sense embeddings is an on-going research topic in NLP, and the current study emphasised why and how multi-sense embeddings should receive attention in future work on MWE compositionality.

Four parameters – the number of dimensions, window size, the minimum-count threshold and the number of iterations – did not have a strong impact on the predictions of Estonian PV's compositionality according to models trained to learn word and multi-sense embeddings. The most influential parameter investigated was the number of iterations – models trained with more iterations provided slightly better predictions than did the models trained with fewer iterations. No models were trained with a greater number of iterations than 20 because, based on related research (e.g. Cordeiro 2017), it was assumed that more iterations would not yield significantly better results. Furthermore, the training time increases with the number of iterations (this is particularly important when training multi-sense embeddings). Cordeiro (2017) also investigated the impact of other features. He confirmed an earlier suggestion by Lapesa and Evert (2017) that the appropriate choice of window size was task-specific, and added that the choice of the DSM may also have an effect on the window size. He suggested that the models trained with a greater rather than with a lower number of dimensions might make the best predictions, and that a low word-count threshold might lead to better results for compositionality prediction tasks as opposed to a high threshold. Bott and Schulte im Walde (2014) demonstrated that a window size of 5 is better than 1, 2, 10 or 20 for predicting the compositionality of German PVs. This finding is in line with the findings of the current study because it was found that, despite small differences, a window size of 5 was more appropriate than was a size of 1 or 30. Overall, related research can provide some general guidelines for parameter configurations, but specific values depend on the task, DSMs and other aspects. Further research is needed in order to predict the compositionality of MWEs in Estonian and other languages.

This thesis confirms the main conclusions of the first study of the automatic detection of the literal versus the non-literal usage of Estonian PVs introduced by (Aedmaa et al. 2018). Both studies showed that the most predictive features for the task were the particle, the verb, the average abstractness of the nouns, the subject case, the subject animacy, the object animacy and case government. As a novel feature, PV frequency was introduced in this study. Therefore, in addition to the standard features that are used to predict the (non-)literalness of MWEs in other languages, some novel features were introduced and their usefulness was proven. Specifically, the case features (the case of subject, the case of object and the case government) were introduced and the subject case and case government were found to be useful for the detection of the (non-)literal usage of PVs. In addition, the animacy features were suggested and included in the best model. The usefulness of the abstractness of nouns for classifying literal and non-literal PVs was not surprising because abstractness has been used for (non-)literalness detection in English previously (e.g. Turney et al. 2011; Klebanov et al. 2015). The abstractness of nouns was also one of the most salient features for predicting the (non-)literalness of German PVs (Köper and Schulte im Walde 2016b). While unigrams worked well for German, the unigram feature was only successful independently for Estonian, and not in combination with other features. However, the suggested models for detecting the literal versus the non-literal usage of Estonian and German PVs shared some features (such as abstractness ratings) and obtained similar results – the model for German attained 86.8% (baseline

64.9%) accuracy, while the model for Estonian achieved 88.9% (baseline 74.0%) accuracy. There are no studies of other languages that apply the subject case, object case, case government, subject animacy or object animacy as features for predicting the non-literal usage of MWEs.

Overall, the differences from the previous research on the compositionality of PVs (and other MWEs) emphasise the importance and necessity of conducting studies of specific MWE types in different languages. The similarities to related research demonstrate that the results were not random or specific to Estonian PVs. In addition, the comparison to other studies pointed out some aspects of Estonian PV compositionality that were not discussed here or which need further investigation. The drawbacks of the current study and possible directions for future work are presented in the next section.

7.3 Future work

This section discusses the shortcomings of the research presented in this thesis, and introduces some ideas for future work.

Firstly, the role the meaning of the particle in the meaning of the PV was not studied in this thesis. However, the semantics of particles has been explored thoroughly from a cognitive linguistics perspective (e.g. Veismann 2009). Therefore, future computational research on PV compositionality could benefit from earlier (theoretical) studies of verbal particles. For example, information about the semantic classes or prototypical usage of Estonian adverbs could be employed to improve the accuracy of the automatic compositionality predictions. In order to improve the quality of compositionality predictions, the semantics of particles should also be studied computationally. With some exceptions (e.g. Cook and Stevenson 2006; Köper and im Walde 2016), not much work has been done on the semantics of particles in other languages. Nevertheless, Bhatia et al. (2018) have proposed an approach to identify the senses of particles using WordNet (an existing lexical resource) for the classification of compositional versus non-compositional VPCs. Following a similar method requires us to investigate the suitability of the Estonian WordNet[105] for the task. Hence, future research should focus on modelling the semantics of particles.

One of the main concerns of the work presented in this thesis is the assessment of the compositionality predictions of distributional models. Firstly, it has been stated that there are no standardised extrinsic evaluation methods for evaluating word vectors; therefore, they are often assessed using computationally inexpensive and rapid word similarity. This approach has multiple flaws, such as the subjectivity of the task, a low correlation with extrinsic evaluations, the absence of statistical significance, the inability to account for polysemy and so on. It has been suggested that the word vector models should be compared based on how well they can perform on a downstream NLP task until a better solution is available (Faruqui et al. 2016). Therefore, a model using compositionality predictions for Estonian PVs to detect the transparency of their meanings in the text should be created and applied to improve the performance of other tasks, such as machine

[105] `https://www.cl.ut.ee/ressursid/teksaurus/` (accessed 10.12.2018).

translation. The assessment of the machine translation could then demonstrate whether the quality of the compositionality predictions is sufficiently high so to improve the accuracy of machine translation.

The inability to account for polysemy – one of the problems with word embeddings – has been addressed in this thesis by using multi-sense embeddings for compositionality predictions. However, the predictions were not evaluated against well-designed human judgements; therefore, further research on multi-sense embeddings requires an appropriate dataset containing human judgements. For example, a dataset similar to the Stanford contextual word similarity dataset (Huang et al. 2012) could be created for Estonian PVs. The dataset should contain human judgements about the compositionality of PVs in specific contexts. For example, how successfully the verb alone could replace the PV in the same context could be assessed – when the PV could be replaced with the verb fully while the meaning of the context remains the same, the PV's meaning is fully compositional. The creation of such a dataset would no doubt be expensive, and requires exhaustive prior research.

Another problematic aspect of the evaluation of multi-sense embeddings was that the most probable meanings were used and compared to the median score of the human ratings. Therefore, the full advantage of learning multiple senses for verbs and PVs was not taken into account. Applying SenseGram's WSD mechanism to determine intended meanings did not help to improve the results. As the assessment of the tool was not possible, it was not clear whether the results did not improve because of the poor quality of the disambiguation tool or the of multi-sense embeddings. Any future work on PV compositionality would thus benefit from a WSD system that can not only succeed if disambiguating the meanings of verbs, but also of PVs. The development and evaluation of a WSD system require a large, semantically annotated corpus in which the meanings of PVs are also labelled. At present, the semantically disambiguated corpus of Estonian (Kahusk 2011) contains 500,000 words. Furthermore, the Estonian WordNet has been under development for years. Nonetheless, no recent work on developing a WSD tool for Estonian has been published; therefore, a model performing WSD for PVs is an open issue for the future. Still, the work on a WSD system for PVs might benefit from the features that were useful for classifying the literal versus the non-literal usage of PVs.

This work focused on the identification of Estonian PVs. The studied PVs were selected from amongst the PVs that were previously discovered (detected) automatically from the text corpora using statistical AMs (Aedmaa 2014). The list combined the PVs that were discovered by at least one studied AM. Therefore, additional experiments are needed in order to determine whether combining the results of multiple AMs is a more useful approach than is using other tools and methods, such as parser (Muischnek et al. 2013). Further research on automatic PV processing should focus on binding automatic discovery and identification to model both tasks.

There are several ways to continue the work using the same methods applied here. For example, in addition to word2vec, there are numerous other types of algorithms for learning word embeddings, such as PPMI (Levy and Goldberg 2014), GloVe (Pennington et al. 2014), lexvec (Salle et al. 2016) and so on,

which could be trained to make compositionality predictions. A recent study by Wendlandt et al. (2018) demonstrated that the stability of different algorithms varied, and that multiple factors contributed to the stability of word embeddings, such as the domain, POS, vocabulary size and so on. The impact of these factors on compositionality prediction should be considered and investigated using Estonian data. For example, the Estonian Reference Corpus 2017 (Kallas and Koppel 2018) containing 1.1 billion words has been published, word and multi-sense embedding models could be retrained using a larger corpus, and the impact of the corpus size on the results could be studied.

In addition, several features have been proposed in the literature with regard to non-literal language usage and metaphor detection that could be tested for detecting the literal versus the non-literal usage of Estonian PVs. For example, among other features, Tsvetkov et al. (2014) used vector-space word representation to develop an English metaphor detection system, and Do Dinh and Gurevych (2016) introduced supervised metaphor detection combining neural network architecture with word embeddings. Employing previously trained word and multi-sense embeddings or combining them with neural networks for predicting the compositionality (literalness) of Estonian PVs could be one possible direction for future research.

The methods used for predicting the compositionality of Estonian PVs have been used widely and successfully not only in compositionality studies, but also for other NLP tasks. However, new approaches and techniques have been proposed in the interim, and could also be applied to Estonian. For example, Hashimoto and Tsuruoka (2016) demonstrated an adaptive, joint learning method for compositional and non-compositional phrase embeddings. Their method addressed the problems of learning solely compositional embeddings, such as data sparsity, and achieved state-of-the-art results in the compositionality detection task of verb-object pairs. Therefore, this suggests further work on other phrases in other languages, such as Estonian PVs. Gong et al. (2017) suggested a simple test for the compositionality of a word and phrase using the local linguistic context. As they did not use any external resources and employed only word vectors, their approach could be replicated to detect not only the compositionality of PVs, but also of other MWEs.

In summary, the automatic processing of MWEs is acknowledged as being a challenging task, mainly because of their semantic (non-)compositionality. The methods introduced in this thesis are merely a few of the approaches that have been proposed in the literature during years of intense research. There are several ways to continue the research on Estonian PVs (and other MWEs), such as modelling the semantics of particles, developing resources for a more solid assessment of the compositionality predictions of multi-sense embedding models, employing new methods such as neural networks, and many more.

8 SUMMARY IN ESTONIAN

Eesti keele ühendverbide automaattuvastus lingvistiliste ja statistiliste meetoditega

Siinne töö keskendub eesti keele üht tüüpi püsiühendite – ühendverbide – automaatsele tuvastamisele. Püsiühendite automaattöötlus on keeruline just seetõttu, et püsiühendid koosnevad rohkem kui ühest sõnast ning sõnaühendi tähendus pole tihtipeale komponentide tähenduste summa, vaid sõnade koosesinemisel tekib uus tähendus (Sag jt 2002). Püsiühendile on pakutud palju erinevaid definitsioone lähtuvalt uurimisobjektist ja -meetodist. Valiku eri definitsioonidest on esitanud näiteks Constant jt (2017). Selles töös lähtun põhimõttest, et püsiühenditel on kaks või enam komponenti ning nende semantiline kompositsionaalsus on skalaarne. See tähendab, et püsiühendeid ei ole otstarbekas tähenduse moodustumise järgi jagada kaheks: kompositsionaalseteks (tähendus on komponentide summa) ja mitte-kompositsionaalseteks (sõnaühend omandab uue tähenduse). Püsiühendid saab hoopis asetada skaalale, mille ühes otsas on ühendid, mille tähendus on selgelt selle komponentide tähenduse summa, ja teises otsas ühendid, mille tähendus pole üldse tuletatav selle komponentide tähendustest (Bannard jt 2003). Samast arusaamast on lähtunud paljud arvutilingvistilised erinevate keelte püsiühendeid käsitlevad uurimused, mis tagab võimaluse selle töö tulemusi nendega võrrelda.

Püsiühenditeks liigitatakse paljusid eri omadustega konstruktsioone. Constant jt (2017) tõid välja, et püsiühenditena on uuritud näiteks idioome, verbi ja partikli ning mitmesuguseid verbist ja noomeni(te)st koosnevaid ühendeid, mitmesõnalisi nimesid ja termineid jne. Loetelu näitab, et püsiühendeid on mitmesuguseid ning kõigi ühendite korraga uurimine väga keeruline. Seetõttu on leitud, et tõhusam on eri püsiühendeid eraldi käsitleda. Sellest kõigest tingituna vaadeldakse siinses töös vaid eesti keele ühendverbide automaatset tuvastamist. Ühendverbid koosnevad afiksaaladverbist ja verbist, on eesti keeles sagedased ja produktiivsed ning nende tähenduse kompositsionaalsus varieerub. Lisaks sellele, et püsiühendite automaatne tuvastamine on üleüldiselt komplitseeritud, lisavad ülesandele keerukust ka näiteks asjaolud, et afiksaaladverbid on tihti homonüümsed kaassõnadega ning ühendverbide komponentide järjestus ja kaugus ei ole kindlaks määratud.

Eesti keele ühendverbid on olnud nii deskriptiivsete kui ka arvutilingvistiliste uurimuste keskmes. Ühendverbi käsitlemist eesti keele alastes varajastes uurimustes tutvustas Huno Rätsep (1978), kes muuhulgas esitas ka ise põhjaliku ühendverbide kirjelduse. Rätsep jagas ühendverbid nende tähenduse moodustumise alusel kaheks: korrapärasteks (ehk kompositsionaalseteks) ja ainukordseteks (ehk mittekompositsionaalseteks). Sellist jaotust on hiljem korratud näiteks eesti keele grammatikas (Erelt jt 2013) ja eesti keele süntaksi tervikkäsitluses (Erelt jt 2017). Ühendverbide uurimist läbi ajaloo on vaadelnud ka Kadri Muischnek, kes uuris teiste verbi püsiühendite seas ühendverbe nii lingvistilisest kui ka arvutilingvistilisest vaatenurgast. Muischnek tõi välja, et püsiühendid moodustavad sellise jada, mille ühes otsas on ühendid, millele saab tähenduse omistada ainult tervikuna, ja teises otsas kollokatiivsed ühendid. (Muischnek 2006: 12) Läbi prosoodia prisma vaatlesid ühendverbe Ann Veismann ja Heete Sahkai (2016), kes kirjeldasid semantilist ainukordsust skalaarse tunnusena. Eri arvutuslikke mee-

todeid on ühendverbide automaatsel tuvastamisel rakendanud Heiki-Jaan Kaalep ja Kadri Muischnek (2002; 2006; 2008). Kristel Uiboaed (2010) tuvastas ühendverbe eesti murretest, kasutades statistilisi meetodeid. Sarnaselt on eesti keele ühendverbe tuvastanud töö autor (Aedmaa 2014), kes on avaldanud ka ainsad eesti keele ühendverbide kompositsionaalsuse automaatse tuvastamise kohta ilmunud uurimused (Aedmaa 2016; 2017; Aedmaa jt 2018).

Püsiühendeid on uuritud mõistagi ka teistes keeltes, kuid enamik töid keskendub inglise keelele. Samas on eesti keele seisukohalt oluline, et ka saksa keele ühendverbide automaatsele tuvastamisele on rohkelt tähelepanu pööratud, sest paljud eesti keele ühendverbid on kasutusele võetud just saksa keele eeskujul (Hasselblatt 1990, viidatud Erelt jt 2017 järgi). Näiteks on saksa keele ühendverbide kompositsionaalsuse tuvastamisel rakendatud klasterdamist (Kühner ja Schulte im Walde 2010), distributiivse semantika mudeleid (Bott ja Schulte im Walde 2014; Köper ja Schulte im Walde 2017a) ja klassifitseerimist (Köper ja Schulte im Walde 2016b) jms. Siinse töö meetodite valik põhinebki suuresti teiste keelte püsiühendite tuvastamist käsitlevatel uurimustel ja keeletehnoloogia põhisuundadel.

Kui varasemad meetodid kasutasid püsiühendite tuvastamiseks eelkõige sõnade koosesinemist (näiteks sõnadevahelise seose tugevuse mõõdikud), siis hiljem on fookus jäänud püsiühendite kompositsionaalsusele ja selle tuvastamisele. Edukas kompositsionaalsuse automaattuvastus tagab, et lisaks püsiühendi enda leidmisele tekstist, suudab arvuti eristada ka püsiühendi eri tähendusi. Selles töös rakendatakse kompositsionaalsuse tuvastamiseks masinõpet ehk algoritme, mis teevad andmete põhjal otsuseid. Kõige üldisemalt jagatakse need algoritmid kaheks – juhendamata ja juhendatud õppimine –, kuid eristatakse ka näiteks stiimulõpet või pooljuhendatud algoritme. Juhendamata ja juhendatud masinõppe suurim erinevus on märgendatud või märgendamata andmete kasutamine. Nimelt suudavad juhendamata algoritmid märgendamata või klassifitseerimata andmestikust tuvastada uusi struktuure. Juhendatud algoritmid kasutavad aga märgendatud andmeid ja ette on antud ka probleem, mille kohta andmestiku põhjal uusi järeldusi tehakse. See tähendab, et juhendatud õppimine vajab eelnevat tööd andmestikuga, näiteks märgendussüsteemi väljatöötamist ja mingi hulga teksti märgendamist. Seega on juhendatud algoritmide kasutamine tihtipeale kallim ehk nõuab rohkem aega ja teisi ressursse. Siinses töös määratakse juhendamata distributiivse semantika mudelitega ühendverbide kompositsionaalsust ning juhendatud klassifitseerija treenitakse, selleks et eristada ühendverbide kompositsionaalset ja mittekompositsionaalset tähendust lausetes.

Distributiivse semantika mudelite (*distributional semantic models*) (ka vektorruumi, semantilise ruumi, sõnaruumimudelite) alus on distributiivse semantika hüpotees, mille järgi kirjeldab sõna tema kontekst: statistiline sõnade jaotus (distributsioon) kontekstis (tekstis) on see, mis kujundab sõnade tähenduse (Firth 1957). Seega on üldine arusaam, et sarnase tähendusega sõnad esinevad sarnases kontekstis (Lenci 2008). Distributiivse semantika hüpotees on semantikas aluseks tervele hulgale statistilistele meetoditele, mida üldnimetusena kutsutakse distributiivseteks meetoditeks. Distributiivset semantikat üldiselt rakendatakse loomulike keele automaattöötluse ülesannete lahendamiseks, näiteks sõnatähenduste ühestamine (Schütze 1998) ja teksti segmenteerimine (Choi 2001), aga ka kognitiivteadu-

ses inimkäitumise modelleerimiseks (nt Landauer ja Dumais 1997; McDonald ja Brew (2004))[106].

Distributiivse semantika mudelite populaarsus on põhjendatav sellega, et need töötavad ainult distributiivse statistika põhjal. Seetõttu eeldab seesuguste mudelite kasutamine üksnes suurt tekstihulka ja sarnasust saab mõõta lisaks sõnadele ka näiteks fraaside või dokumentide vahel. Üldine sõna semantilise sarnasuse mudelite tööpõhimõte on järgmine: sõnade distributiivne info ehk esinemissagedus teiste elementide suhtes kogutakse kokku mitmedimensionaalse ruumi vektoritena. Uuemat tüüpi distributiivse semantika mudelite tööpõhimõte on suuresti mõjutatud neurovõrkude keelemudelitest (*neural network language models*). Seesugused mudelid ennustavad tõenäosust modelleerides konteksti põhjal sõnajada järgmist sõna. Seesuguste sõnade vektoresituste (*word embeddings*) abil saab mõõta sõnade vahelist tähenduslikku sarnasust, mis väljendub geomeetrilise kaugusena vektorruumis sõnu esitavate vektorite vahel. Vektoritevahelist kaugust saab mõõta eri mõõdikutega, siinses töös väljendatakse sarnasust koosinuskauguse abil. Kui koosinuskauguse väärtus on 1, siis vektoritega esitatud sõnad on väga sarnased, kui väärtus on –1, siis väga erinevad. Selleks et järjestada eesti keele ühendverbid nende kompositsionaalsuse järgi, leitakse koosinuskaugus ühendverbi vektori ja verbi vektori vahel, põhinedes eeldusel, mida rakendati saksa keele ühendverbide kompositsionaalsuse tuvastamisel: ühendverbi kompositsionaalsus oleneb sellest, kui sarnased on verbi ja ühendverbi tervikuna esinemise kontekstid (Bott ja Schulte im Walde 2014). Kui üldjuhul on vektorruumi mudelid võimelised koostama ühe vektori sõna kohta, vaatamata sellele, kas sõna on mitmetähenduslik või mitte, siis on arendatud ka selliseid mudeleid, mis eristavad sõnade tähendusi ja toodavad tähendusvektoreid (*multi-sense embeddings*). Siinses töös rakendatakse kompositsionaalsuse määramiseks mõlemat tüüpi vektoreid: sõnavektorid on saadud word2vec (Mikolov jt 2013a) tööriistaga ning tähendusvektorid SenseGrami (Pelevina jt 2016) kasutades. Seda, kuidas distributiivse semantika mudelid vektoreid õpivad, on võimalik mitme parameetri abil muuta. Siinses töös võrreldakse, kuidas tulemused erinevad, kui muuta treeningmeetodit, vektorruumi dimensioonide arvu, kontekstiakna suurust, kaasatavate sõnade miinimumsagedust ja treeningiteratsioonide arvu.

Sõna- ja tähendusvektorid leitakse eesti keele veebikorpusest 2013[107]. Mudelite tööd hinnatakse kahe inimmärgenduse põhjal loodud andmestiku abil. Esimene neist koguti ühisloome abil ning see sisaldab iga uurimusse kaasatud ühendverbi kohta ühte kompositsionaalsuse hinnet. See arv on inimeste antud kompositsionaalsuse hinnangute aritmeetiline keskmine. Igale ühendverbile koguti viie palli skaalal vähemalt kümne inimese hinnang. Ei uuritud inimese lingvistilist tausta ega seda, missuguse ühendverbi tähendust ta määratles. Sellist ressurssi oli vaja vektoresituste hindamiseks, mis niisamuti ei erista eri tähendusi. Teine andmestik, mida mudelite töö hindamiseks kasutati, sisaldab tähenduste kompositsionaalsuse hinnanguid. Kuna see info eraldati andmehulgast, mida märgendati klassifitseerija treenimiseks, siis selle täpsem kirjeldus esitatakse järgmises lõigus. Lõplikest andmestikest (ehk nendest, mida kasutati mudelite hindamiseks) jäetakse välja

[106]Väga põhjalikult on distributiivse semantika eri mudelite rakendusvõimalustest kirjutanud Turney ja Pantel (2010).

[107]`http://www.keeleveeb.ee/dict/corpus/ettenten/about.html` (Vaadatud 09.01.2019)

need ühendverbid, mille kompositsionaalsust oli ühel või teisel põhjusel keerukas kindlaks määrata. Ühendverbid reastatakse nii inimeste antud kompositsionaalsuse hinnangute kui ka mudelite leitud vektoritevahelise koosinuskauguse väärtuse järgi. Järjestuste sarnasust väljendatakse Spearmani korrelatsioonikordajaga.

Selleks et tuvastada, kas lauses on ühendverbi kasutatud kompositsionaalses või mittekompositsionaalses tähenduses, viiakse läbi binaarne klassifitseerimine. Klassifitseerija ise põhineb juhumetsa algoritmil, mis omakorda koosneb paljudest otsustuspuudest. Otsustuspuu on vahend, mis meenutab puu struktuuri ja mida, nagu nimigi ütleb, kasutatakse otsuste tegemiseks. Muuhulgas on otsustuspuud kasulikud andmete klassifitseerimiseks, kus mingid üksused määratakse (andmestikus märgendatud) tunnuste järgi kategooriatesse. Siinses töös klassifitseeritakse lauseid, mis ühendverbide tähenduse moodustumise alusel on andmestikus märgendatud kui kompositsionaalsed või mittekompositsionaalsed. Märgendus põhineb kolme eksperdi hinnangutel, kes hindasid kuue palli skaalal ühendverbide kompositsionaalsust 1838 lauses. Binaarseks jaotuseks arvutati iga lause hinnete aritmeetiline keskmine ning selle põhjal klassifitseeriti laused. Andmestikku jäid ainult laused, mille kõik hinded kuulusid samasse binaarsesse klassi. Neid hindeid kasutatakse ka distributiivse semantika mudelite hindamiseks. Klassifitseerija arendamiseks märgendati andmestikus ka (lingvistilisi) tunnuseid, mille kasulikkust ühendverbide tähenduse tuvastamisel hinnatakse ja mida omavahel kombineeritakse. Selgitatakse välja, missugust informatsiooni on vaja, et arendada võimalikult hästi töötav klassifitseerija.

Tööl on neli põhieesmärki: a) eesti keele ühendverbide kompositsionaalsuse automaattuvastus; b) teiste keelte püsiühendite automaattöötluses väga laialt ja edukalt kasutatud meetodite tutvustamine ja rakendamine eesti keele peal; c) tulemuste kirjeldamine sel viisil, et need julgustaksid tutvustatud meetodeid rohkem keeleteaduslikes uurimustes rohkem kasutama; d) uute ressursside loomine, et need oleksid rakendatavad ka muudel eesmärkidel peale püsiühendite automaattöötluse. Nende eesmärkideni jõutakse, vastates järgmistele uurimisküsimustele:

1. Mil määral sarnanevad inimeste hinnangud ühendverbide kompositsionaalsuse hindamisel? Mis on peamised lahkarvamuste põhjused?

2. Kui hästi töötavad distributiivse semantika mudelid eesti keele ühendverbide kompositsionaalsuse tuvastamisel? Mis parameetrid ja teised asjaolud mõjutavad sõna- ja tähendusvektorite kvaliteeti?

3. Mis (lingvistilised) tunnused viitavad sellele, kas ühendverbi tähendus lauses on kompositsionaalne või mittekompositsionaalne? Kui hästi on need tunnused automaatselt märgendatavad?

4. Mis on sageduse mõju eesti keele ühendverbide kompositsionaalsusele? Kuidas mõjutab sagedus inimeste ja mudelite ühendverbide kompositsionaalsuse hinnanguid?

5. Kas varem edukad olnud arvutuslikud meetodid sobivad eesti keele ühendverbide kompositsionaalsuse tuvastamiseks? Mis on kasutatud meetodite head ja halvad küljed?

Arvutuslike meetodite kvaliteedi mõistmiseks kasutatakse tihtipeale inimeste antud hinnanguid või märgendusi sisaldavaid andmestikke ehk kuldstandardeid. See on otstarbekas, sest üldine eesmärk on, et arvutid oleksid võimelised töötlema keelt inimese võimetele lähedase täpsusega. Tähenduse mõistmine on arvutite jaoks keerukas. Sõnade mitmetähenduslikkuse tõttu ei ole see ka alati ühesugune inimeste seas. Selleks et uurida, mil määral inimesed nõustuvad üksteisega eesti keele ühendverbide kompositsionaalsuse hindamisel, viiakse läbi ka andmestike evalveerimine. See näitab, et osa ühendverbide ja nende eri tähenduste kompositsionaalsust on keerulisem hinnata kui teiste ühendverbide ja tähenduste kompositsionaalsust. Peamine põhjus on ühendverbide ja nende komponentide mitmetähenduslikkus, eriti afiksaaladverbide puhul. Üldiselt aga on inimeste kompositsionaalsuse hinnangud sarnased, sest vähem kui 10% ühendverbidest põhjustas olulisi lahkarvamusi nende märgendajate seas, kes hindasid ühendverbide kompositsionaalsust võimalikke tähendusi eristamata. Fleissi kappa koefitsiendi[108] väärtus näitab nende kolme eksperdi seas, kes hindasid ühendverbide eri tähendusi, et kuue palli skaalal on ühtivus mõõdukas ja binaarselt (kompositsionaalne–mittekompositsionaalne) tugev. Seega vaatamata ülesande raskusele olid inimeste hinnangud ühendverbide kompositsionaalsusele sarnased.

Ühendverbide kompositsionaalsuse automaattuvastus distributiivse semantika mudelitega näitab, et nii sõna- kui ka tähendusvektorid sobivad kompositsionaalsuse tuvastamiseks. Sõna- ja tähendusvektorite omavaheline võrdlus viitab sellele, et mudelid, mis õpivad sõnavektoreid, tuvastavad kompositsionaalsust paremini kui tähendusvektorite mudelid. Samas tuleb nentida, et tähendusvektorite töö hindamine on siin uurimuses tinglik, sest kumbki kuldstandarditest ei olnud selleks loodud, vaid kohandatud üldpildi saamiseks. Nimelt on tähendusvektoreid kompositsionaalsuse tuvastamiseks rakendanud mõnel harval korral. Seetõttu vajab sobiva kuldstandardi väljatöötamine põhjalikku uurimistööd, mis sellesse töösse ei mahtunud. Kindlasti aitaks sobiva kuldstandardi ja tähendusvektorite hindamismeetodi arengule kaasa hästi töötav automaatne sõnatähenduste ühestaja, mida ei ole loodud eesti keele jaoks. Siiski saab öelda, et automaatselt tuvastatud ühendverbide kompositsionaalsuse ja inimhinnangute vahel on keskmise tugevusega korrelatsioon. Treeningmeetodil, vektorruumi dimensioonide arvul, kontekstiakna suurusel, kaasatavate sõnade miinimumsagedusel ja treeningiteratsioonide arvul ehk mudelite treeningparameetritel ei ole tulemustele tugevat mõju. Siiski kahest treeningmeetodist – pidevast järjestikuste sõnade esinemissagedusest (*continuous bag-of-words*, CBOW) ja pidevast skip-grammi mudelist (*continuous skip-gram*) – on ühendverbide kompositsionaalsuse mudeldamisel edukam esimene. Kahe treeningmeetodi põhiline erinevus on, et järjestikuste sõnade esinemissageduse mudelis ennustatakse sõna konteksti põhjal, skip-grammi mudelis aga konteksti sõna põhjal. Teistest uuritud parameetritest on iteratsioonil tugevaim mõju: suurem iteratsioonide arv tähendab paremaid tulemusi. Samas, kuna treeningaeg pikeneb iteratsioonide arvu suurenedes, pole tulemuste erinevus nii suur, et väga palju iteratsioone kasutada. Selle töö tulemused näitavad, et 20 iteratsiooni on parem kui viis või kümme.

Lausete klassifitseerimine nendes esineva ühendverbi tähenduse järgi kompo-

[108] Statistiline mõõdik, mida kasutatakse märgendajatevahelise ühtivuse hindamiseks (vt Fleiss 1971).

sitsionaalseteks või mittekompositsionaalseteks näitas, et lisaks ühendverbi komponentidele on lausetes ka teisi tunnuseid, mis aitavad tõsta klassifitseerija kvaliteeti. Töös tutvustatakse 12 tunnust: ühendverbi komponendid ehk afiksaaladverb ja verb, unigrammid, neli abstraktsusega seotud tunnust, subjekti ja objekti käänded, subjekti ja objekti elusus ning ühendverbi ja käändsõna (mis ei ole subjekt ega objekt) vaheline rektsiooniseos. Seitset tunnust võib nimetada standardseks ehk selliseks, mida on kasutatud teiste keelte püsiühendite idiomaatilise tähenduse tuvastamiseks. Ühendverbi komponendid ise sisaldavad nii palju informatsiooni, et klassifitseerija töötab 85%-lise täpsusega. Unigrammid ehk kõik lausetes sisalduvad sõnad aga ei lisa nii palju kasulikku infot, et täpsus paraneks. Neljast uuritud abstraktsuse tunnusest – kõikide lauses esinevate sõnade keskmine abstraktsus, lauses esinevate nimisõnade keskmine abstraktsus, subjekti abstraktsus ja objekti abstraktsus –, osutus kasulikus ainult nimisõnade abstraktsus. Ka subjekti ja objekti elusus, mida pole varem kompositsionaalsust puudutavates uuringutes rakendatud, aitab õigesti klassifitseerida rohkem lauseid kui süsteem ilma selle infota. Teised uuritud tunnused on seesugused, mis lähtuvad pigem eesti keele spetsiifikast. Kusjuures nendest kolmest on subjekti kääne ning ühendverbi ja käändsõna vaheline rektsiooniseos ühendverbide klassifitseerimisel mõjusamad tunnused kui objekti kääne. Lisaks tutvustatakse kolme sagedusega seotud tunnust – ühendverbi, afiksaaladverbi ja verbi sagedusi – ja selgub, et kombineerituna teiste tunnustega on kasulik ainult ühendverbi sagedus. Kõige parema tulemuse annabki klassifitseerija, kus on kombineeritud info ühendverbi komponentide, nimisõnade abstraktsuse, subjekti käände, subjekti ja objekti elususe, rektsiooni ja ühendverbi sageduse kohta. Selline süsteem klassifitseerib lauseid täpsusega 89%, mis tähendab, et parimate tulemuste jaoks on oluline ka lingvistiliste tunnuste rakendamine.

Kvaliteetse klassifitseerija saab luua vaid siis, kui on valitud head tunnused. Tihtipeale on seesuguste mudelite arendamisel vaatluse all sadu tunnuseid ja seetõttu on oluline, et andmete märgendamine oleks automaatne. Kuigi siinses töös pole vaatluse all palju tunnuseid, on olemas tööriistu, mis märgendamist automatiseerivad. Näiteks unigramme ja sagedust on suhteliselt lihtne märgendada, sest tekst on vaja vaid lemmatiseerida. Lemmatiseerimine on võimalik tänu morfoloogilisele analüsaatorile, mis lisaks aitab tuvastada ka sõnaliigid ja käänded. Samamoodi on eesti keele jaoks olemas süntaktiline analüsaator, mis teksti parsib ehk annab informatsiooni süntaktiliste rollide kohta. Keerulisemad tunnused on näiteks subjekti ja objekti elusus, mis pole praegu automaatselt märgendatavad, sest puudub ressurss, millest seda infot omandada. Info konteksti abstraktsuse kohta on kättesaadav andmestikust, mis loodi automaatselt siinse töö jaoks. See sisaldab eesti keele lemmade abstraktsuse hindeid, kuid pole ise evalveeritud. Seega on abstraktsuse info küll automaatselt kättesaadav, kuid edaspidi on vajalik selle ressursi hindamine või uue andmestiku loomine. Kokkuvõttes on mõni tunnus, mis pole automaatselt märgendatav, kuid siiski enamik vajalikust märgendusest on olemasolevate tööriistadega automaatselt omandatav.

Lisaks sagedusega seotud tunnustele lausete ühendverbide tähenduse järgi klassifitseerimisel uuritakse töös sageduse mõju kompositsionaalsusele veel mitmes aspektis. Näiteks ei leita statistiliselt olulist seost ühendverbide, afiksaaladverbide ja verbide sageduste ning inimeste antud kompositsionaalsuse hinnangute

vahel. Samas mõjutab sagedus distributiivse semantika mudelite tulemusi – väga sagedaste ja väga harvade ühendverbide kompositsionaalsust suudavad mudelid paremini ennustada kui keskmise sagedusega ühendverbide kompositsionaalsust. Samas pole kompositsionaalsus lineaarses seoses ühendverbide sagedusega, vaid mudelid kipuvad väga sagedased ja harvad ühendverbid määrama vähem kompositsionaalsemaks kui teised ühendverbid. Seega saab öelda, et arvutuslikud meetodid on kompositsionaalsuse määramisel sagedusest mõjutatud, kuid sageduse mõju inimeste hinnangutele vajab täiendavaid uurimusi. Niisamuti vajab keskmise sagedusega ühendverbide kompositsionaalsuse automaattuvastuse kvaliteet parandamist.

Uurimusest selgub, et ühendverbide kompositsionaalsuse määramine on nii juhendamata kui juhendatud masinõppe meetoditega võimalik. Samas on mõlemal tutvustatud meetodil nii häid kui ka halbu külgi, mida tuleb nende rakendamise eel arvesse võtta. Näiteks ei vaja distributiivse semantika mudelid muud kui suurt tekstihulka ja suure jõudlusega arvutit. Seetõttu on meetod odavam kui juhendatud klassifitseerimine. Siinses töös selgus sama, mida on varem ka kirjanduses mainitud (nt Lenci 2008): tihtipeale pole distributiivsete meetodite tulemused selgesti analüüsitavad ning keerukamad semantikat puudutavad küsimused jäävad vastuseta. Seega isegi kui tulemused on head, on tulemuste analüüs pinnapealne ja oletuslik. Lisaks kõigele ei suuda distributiivse semantika mudelid saavutada ühendverbide kompositsionaalsuse määramisel sama häid tulemusi kui klassifitseerimine. Juhendatud meetodid on küll kallimad, sest nõuavad märgendatud andmeid, kuid nagu siinne töö näitab, siis tihtipeale on juba olemasolevaid ressursse kasutades võimalik nende hinda vähendada. Tulemuste detailne analüüs aitab sügavuti mõista, mis tunnused mõjutavad ühendverbide kompositsionaalsust ning niiviisi on tulemused kasulikud kompositsionaalsuse lingvistiliseks kirjeldamiseks. Kokkuvõttes on mõlemal rakendatud meetodil nii häid kui halbu külgi: klassifitseerimine on täpsem ja distributiivse semantika mudelite rakendamine odavam.

Töö tulemused näitavad, et tulevikus võiks tutvustatud meetodeid nii ühendverbide kui ka teiste eesti keele püsiühendite kompositsionaalsuse tuvastamiseks kombineerida. Ka on palju teisi uurimisviise ja -vahendeid, mida sama ülesande lahendamiseks rakendada saaks. Näiteks on word2veci kõrval ka teisi sõnavektorite treenimise vahendeid (GloVe, PPMI jne). Samamoodi on võimalik kasutada veel suuremat tekstikorpust kui siinses töös kasutatud eesti keele veebikorpus, sest 2018. aastal loodi 1,1 miljardi sõna suurune eesti keele ühendkorpus 2017 (Kallas ja Koppel 2018). Lisaks sellele jäi siinses töös uurimata afiksaaladverbide tähtsus ühendverbide kompositsionaalsuse moodustumisel. Selleks võiks rakendada näiteks informatsiooni adverbide semantiliste rühmade või prototüüpse kasutuse kohta. Ka on oluline luua tulevikus andmestik, mille põhieesmärk on hinnata tähendusvektorite tööd püsiühendite kompositsionaalsuse määramisel. Töö olulisust laiemas keeletehnoloogilises plaanis aitaks välja selgitada siinse töö mudelite integreerimine mõnesse tootesse või teenusesse eesmärgiga selle kvaliteeti parandada.

Kokkuvõttes rakendati töös kahte masinõppe meetodit, mida ei ole eesti keele püsiühendite tuvastamiseks varem kasutatud. Uurimuse tulemused on tähtsad kompositsionaalsuse arvutuslikuks modelleerimiseks ja täiendavad ka varasemat

ühendverbide kompositsionaalsuse kohta avaldatud kirjandust. Olulised on ka töö käigus loodud ressursid: ühendverbide ja nende tähenduse kompositsionaalsuse hinnangud, eesti keele lemmade abstraktsuse/konkreetsuse hinnangud ning sõna- ja tähendusvektorid. Neid saab edaspidi kasutada ka teistes uurimustes, mis käsitlevad muud peale püsiühendite tuvastamise.

REFERENCES

Aedmaa, Eleri. 2014. Statistical methods for Estonian particle verb extraction from text corpus. In *Proceedings of the ESSLLI 2014 Workshop: Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations*. Tübingen, Germany, 17–22.

Aedmaa, Eleri. 2015. Statistilised meetodid ühendverbide tuvastamisel tekstikorpusest [Statistical Methods for Estonian Particle Verb Extraction from Text Corpora]. *Eesti Rakenduslingvistika Ühingu Aastaraamat* 11: 37–54.

Aedmaa, Eleri. 2016. Eesti keele ühendverbide kompositsionaalsuse määramine [Detecting the Compositionality of Estonian Particle Verbs]. *Eesti Rakenduslingvistika Ühingu Aastaraamat* 12: 5–23.

Aedmaa, Eleri. 2017. Exploring Compositionality of Estonian Particle Verbs. In *Proceedings of the ESSLLI*. Toulouse, France, 197–208.

Aedmaa, Eleri. 2018. (Non-)literalness Ratings for Estonian Particle Verbs. `http://dx.doi.org/10.15155/re-56`. `https://doi.org/10.15155/re-56`.

Aedmaa, Eleri, Maximilian Köper, and Sabine Schulte im Walde. 2018. Combining Abstractness and Language-specific Theoretical Indicators for Detecting Non-literal Usage of Estonian Particle Verbs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 9–16.

Allison, Ben, David Guthrie, and Louise Guthrie. 2006. Another Look at the Data Sparsity Problem. In *International Conference on Text, Speech and Dialogue*. Springer, 327–334.

Alpaydin, Ethem. 2009. *Introduction to machine learning*. MIT press.

Artstein, Ron and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics* 34(4): 555–596.

Attia, Mohammed A. 2006. Accommodating Multiword Expressions in an Arabic LFG Grammar. In *Advances in Natural Language Processing*, Springer, 87–98.

Baldwin, Timothy, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Association for Computational Linguistics, 89–96.

Baldwin, Timothy and Su Nam Kim. 2010. Multiword Expressions. *Handbook of Natural Language Processing* 2: 267–292.

Baldwin, Timothy and Aline Villavicencio. 2002. Extracting the Unextractable: A Case Study on Verb-particles. In *Proceedings of the 6th Conference on Natural Language Learning*. Association for Computational Linguistics, 1–7.

Bannard, Colin. 2005. Learning about the meaning of merb–particle constructions from corpora. *Computer Speech & Language* 19(4): 467–478.

Bannard, Colin, Timothy Baldwin, and Alex Lascarides. 2003. A Statistical Approach to the Semantics of Verb-particles. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Association for Computational Linguistics, 65–72.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 238–247.

Bartsch, Sabine. 2004. *Structural and Functional Properties of Collocations in English: A Corpus Study of Lexical and Pragmatic Constraints on Lexical Co-occurrence*. Gunter Narr Verlag.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3(Feb): 1137–1155.

Bhatia, Archna, Choh Man Teng, and James Allen. 2017. Compositionality in Verb-Particle Constructions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. 139–148.

Bhatia, Archna, Choh Man Teng, and James F Allen. 2018. Identifying Senses of Particles in Verb-particle Constructions. In *Multiword Expressions at Length and in Depth: Extended Papers from the MWE 2017 Workshop*. Language Science Press, volume 2, page 61.

Biemann, Chris. 2006. Chinese Whispers: An Efficient Graph Clustering Algorithm and Its Application to Natural Language Processing Problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*. Association for Computational Linguistics, 73–80.

Birke, Julia and Anoop Sarkar. 2006. A Clustering Approach for Nearly Unsupervised Recognition of Nonliteral Language. In *1th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Blaheta, Don and Mark Johnson. 2001. Unsupervised Learning of Multi-word Verbs. In *Proceedings of the ACL/EACL 2001 Workshop on the Computational Extraction, Analysis and Exploitation of Collocations*. Association for Computational Linguistics, 54–60.

Bos, Johan, Valerio Basile, Kilian Evang, Noortje J. Venhuizen, and Johannes Bjerva. 2017. The Groningen Meaning Bank. In *Handbook of Linguistic Annotation*, Springer, 463–496.

Bott, Stefan, Nana Khvtisavrishvili, Max Kisselew, and Sabine Schulte im Walde. 2016. GhoSt-PV: A Representative Gold Standard of German Particle Verbs. In *Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon*. The COLING 2016 Organizing Committee, 125–133.

Bott, Stefan and Sabine Schulte im Walde. 2014. Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb Compositionality. In *Proceedings of Ninth International Conference on Language Resources and Evaluation (LREC 2016)*. Association for Computational Linguistics, 509–516.

Bott, Stefan and Sabine Schulte im Walde. 2015. Exploiting Fine-grained Syntactic Transfer Features to Predict the Compositionality of German Particle Verbs. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*. Association for Computational Linguistics, 34–39.

Bott, Stefan and Sabine Schulte im Walde. 2017. Factoring Ambiguity out of the Prediction of Compositionality for German Multi-Word Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, 66–72.

Boukobza, Ram and Ari Rappoport. 2009. Multi-Word Expression Identification Using Sentence Surface Features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 468–477.

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

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46(3): 904–911.

Bybee, Joan et al. 2007. *Frequency of Use and the Organization of Language*. Oxford University Press on Demand.

Cabezudo, Marco Antonio Sobrevilla, Thiago Alexandre Salgueiro Pardo, et al. 2015. Exploratory Study of Word Sense Disambiguation Methods for Verbs in Brazilian Portuguese. *International Journal of Computational Linguistics and Applications* 6(1): 131–148.

Caselles-Dupré, Hugo, Florian Lesaint, and Jimena Royo-Letelier. 2018. Word2vec Applied to Recommendation: Hyperparameters Matter. *arXiv Preprint ArXiv:1804.04212* `https://arxiv.org/pdf/1804.04212.pdf`.

Chafe, Wallace L. 1968. Idiomaticity as an Anomaly in the Chomskyan Paradigm. *Foundations of Language* 109–127.

Chakraborty, Tanmoy, Dipankar Das, and Sivaji Bandyopadhyay. 2011. Semantic Clustering: An Attempt to Identify Multiword Expressions in Bengali. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. Association for Computational Linguistics, 8–13.

Chandrashekar, Girish and Ferat Sahin. 2014. A Survey on Feature Selection Methods. *Computers & Electrical Engineering* 40(1): 16–28.

Cheng, Jianpeng and Dimitri Kartsaklis. 2015. Syntax-Aware Multi-Sense Word Embeddings for Deep Compositional Models of Meaning. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1531–1542.

Choi, Freddy YY, Peter Wiemer-Hastings, and Johanna Moore. 2001. Latent Semantic Analysis for Text Segmentation. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. MIT Press.

Choueka, Yaacov. 1988. Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. In *RIAO 88:(Recherche D'Information Assistée Par Ordinateur). Conference*. 609–623.

Church, Kenneth Ward and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics* 16(1): 22–29.

Clark, Alexander, Chris Fox, and Shalom Lappin. 2013. *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons.

Cohen, Jacob. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20(1): 37–46.

Colwell, DJ and JR Gillett. 1982. Spearman versus kendall. *The Mathematical Gazette* 66(438): 307–309.

Constant, Mathieu, Gülşen Eryiğit, Johanna Monti, Lonneke Van Der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword Expression Processing: A Survey. *Computational Linguistics* .

Cook, Paul, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling Their Weight: Exploiting Syntactic Forms for the Automatic Identification of Idiomatic Expressions in Context. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*. Association for Computational Linguistics, 41–48.

Cook, Paul and Suzanne Stevenson. 2006. Classifying Particle Semantics in English Verb-particle Constructions. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Association for Computational Linguistics, 45–53.

Cordeiro, Silvio, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016a. Predicting the Compositionality of Nominal Compounds: Giving Word Embeddings a Hard Time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 1986–1997.

Cordeiro, Silvio, Carlos Ramisch, and Aline Villavicencio. 2016b. Filtering and Measuring the Intrinsic Quality of Human Compositionality Judgments. In *Proceedings of the 12th Workshop on Multiword Expressions*. 32–37.

Cordeiro, Silvio Ricardo. 2017. *Distributional Models of Multiword Expression Compositionality Prediction*. Ph.D. thesis, Federal University of Rio Grande Do Sul; Aix Marseille University.

Culicover, Peter W. 1999. *Syntactic Nuts: Hard Cases, Syntactic Theory, and Language Acquisition*. Oxford University Press on Demand.

Dahlmann, Irina and Svenja Adolphs. 2007. Pauses as an Indicator of Psycholinguistically Valid Multi-Word Expressions (MWEs)? In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*. Association for Computational Linguistics, 49–56.

Do Dinh, Erik-Lân and Iryna Gurevych. 2016. Token-level Metaphor Detection using Neural Networks. In *Proceedings of the Fourth Workshop on Metaphor in NLP*. 28–33.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1): 61–74.

Erelt, Mati. 2013. *Eesti keele lauseõpetus: Sissejuhatus. Öeldis [Estonian Syntax. An Introduction, a Predicate]*. Tartu: Tartu Ülikooli Eesti Keele osakond.

Erelt, Mati, Tiit Hennoste, Liina Lindström, Helle Metslang, Renate Pajusalu, Helen Plado, and Ann Veismann. 2017. *Eesti keele süntaks [Estonian Syntax]*. Eesti Keele Varamu. Tartu Ülikooli Kirjastus.

Erelt, Tiiu, Ülle Viks, Mati Erelt, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, and Silvi Vare. 1993. *Eesti keele grammatika II. Süntaks [The Grammar of Estonian II. Syntax]*. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Eskildsen, Søren W. and Teresa Cadierno. 2007. Are Recurring Multi-word Expressions Really Syntactic Freezes? Second Language Acquisition from the Perspective of usage-based Linguistics. In *Nordic Conference on Syntactic Freezes*. 86–99.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis, Universität Stuttgart.

Evert, Stefan and Hannah Kermes. 2003. Experiments on Candidate Data for Collocation Extraction. In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 83–86.

Evert, Stefan and Brigitte Krenn. 2001. Methods for the Qualitative Evaluation of Lexical Association Measures. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 188–195.

Farahmand, Meghdad, Aaron Smith, and Joakim Nivre. 2015. A Multiword Expression Data Set: Annotating Non-compositionality and Conventionalization for English Noun Compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*. 29–33.

Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with Evaluation of Word Embeddings using Word Similarity Tasks. *arXiv Preprint ArXiv:1605.02276* .

Fazly, Afsaneh and Suzanne Stevenson. 2006. Automatically Constructing a Lexicon of Verb Phrase Idiomatic Combinations. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*.

Fillmore, Charles J., Paul Kay, and Mary Catherine O'Connor. 1988. Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language* 64(3): 501–538.

Finlayson, Mark Alan and Nidhi Kulkarni. 2011. Detecting Multi-word Expressions Improves Word Sense Disambiguation. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. Association for Computational Linguistics, 20–24.

Firth, John Rupert. 1957. *Papers in Linguistics, 1934-1951*. Oxford University Press.

Fleiss, Joseph L. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin* 76(5): 378.

Fontenelle, Thierry. 2001. Collocation Modelling: from Lexical Functions to Frame Semantics. In *Proceedings of the ACL Workshop on Collocation*. 1–7.

Frank, Eibe, MA Hall, and IH Witten. 2016. The WEKA Workbench. *Data Mining: Practical Machine Learning Tools and Techniques* 4.

Fraser, Bruce. 1970. Idioms within a Transformational Grammar. *Foundations of Language* 6(1): 22–42.

Freckleton, Peter. 1985. Sentence Idiom in English. *Working Papers in Linguistics* 11: 153–168.

Giry-Schneider, Jacqueline. 1978. Syntax and Lexicon: Blessure (wound), Noeud (knot), Caresse (caress)... *Journal of Linguistic Calculus* 3–4: 55–72.

Glucksberg, Sam, Matthew S McGlone, Yosef Grodzinsky, and Katrin Amunts. 2001. *Understanding Figurative Language: From Metaphor to Idioms*. 36. Oxford University Press on Demand.

Goldsmith, J. 2005. Review Article of bruce nevin, Ed. *The Legacy of Zellig Harris: Language and Information Into the 21st Century* 1: 719–736.

Gong, Hongyu, Suma Bhat, and Pramod Viswanath. 2017. Geometry of Compositionality. In *AAAI*. 3202–3208.

Google. 2013. word2vec. `https://code.google.com/archive/p/word2vec/`.

Grefenstette, Gregory. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston, USA.

Gries, Stefan Th. 2008. *Phraseology and Linguistic Theory: A Brief Survey*, John Benjamins Publishing Company, 3–25.

Gries, Stefan Th. and Dagmar Divjak. 2012. *Frequency Effects in Language Learning and Processing*, volume 244. Walter De Gruyter.

Gross, Maurice. 1986. Lexicon-Grammar: the Representation of Compound Words. In *Proceedings of the 11th Coference on Computational Linguistics*. Association for Computational Linguistics, 1–6.

Gurrutxaga, Antton and Inaki Alegria. 2013. Combining Different Features of Idiomaticity for the Automatic Classification of Noun+Verb Expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions*. 116–125.

Guyon, Isabelle and André Elisseeff. 2003. An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research* 3(Mar): 1157–1182.

Hall, M. A. 1998. *Correlation-based Feature Subset Selection for Machine Learning*. Ph.D. thesis, University of Waikato, Hamilton, New Zealand.

Hamilton, William L., Jure Leskovec, and Dan Jurafsky. 2016. Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1489–1501.

Harris, Zellig S. 1970. Distributional Structure. In *Papers in Structural and Transformational Linguistics*, Springer, 775–794.

Harris, Zellig Sabbettai. 1951. *Methods in Structural Linguistics*. Chicago University Press.

Hartmann, Silvana. 2008. Einfluss Syntaktischer und Semantischer subkategorisierung Auf Die kompositionalität Von partikelverben. *Studienarbeit. Institut Für Maschinelle Sprachverarbeitung, Universität Stuttgart. Supervision: Sabine Schulte im Walde and Hans Kamp* .

Hashimoto, Chikara and Daisuke Kawahara. 2008. Construction of an Idiom Corpus and its Application to Idiom Identification based on WSD incorporating Idiom-Specific Features. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 992–1001.

Hashimoto, Kazuma and Yoshimasa Tsuruoka. 2016. Adaptive Joint Learning of Compositional and Non-compositional Phrase Embeddings. *arXiv Preprint ArXiv:1603.06067* `https://arxiv.org/pdf/1603.06067.pdf`.

Hasselblatt, Cornelius. 1990. *Das Estnische Partikelverb Als Lehnübersetzung Aus Dem Deutschen*, volume 31. In Kommission Bei O. Harrassowitz.

Hoang, Hung Huu, Su Nam Kim, and Min-Yen Kan. 2009. A Re-examination of Lexical Association Measures. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*. Association for Computational Linguistics, 31–39.

Hoey, Michael. 1997. From Concordance to Text Structure: New Uses for Computer Corpora. In *PALC*. volume 97, 2–22.

Huang, Eric H, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving Word Representations via Global Context and Multiple Word Prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 873–882.

Iacobacci, Ignacio, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. Sensembed: Learning Sense Embeddings for Word and Relational Similarity. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 95–105.

Iomdin, Boris L., Anastasiya A. Lopukhina, Konstantin A. Lopukhin, and Grigoriy V. Nosyrev. 2016. Word Sense Frequency of Similar Polysemous Words in Different Languages. *Computational Linguistics and Intellectual Technologies. Dialogue* 2016: 214–225.

Jackendoff, Ray. 1975. Morphological and Semantic Regularities in the Lexicon. *Language* 51: 639–671.

Jackendoff, Ray. 1997. Twistin'the Night Away. *Language* 73(3): 534–559.

Justeson, John S. and Slava M. Katz. 1995. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering* 1(1): 9–27.

Kaalep, Heiki-Jaan. 1997. An Estonian Morphological Analyser and the Impact of a Corpus on Its Development. *Computers and the Humanities* 31(2): 115–133.

Kaalep, Heiki-Jaan and Kadri Muischnek. 2002. Using the Text Corpus to Create a Comprehensive List of Phrasal Verbs. In *Proceedings of the Third International Conference on Language Resources and Evaluation*.

Kaalep, Heiki-Jaan and Kadri Muischnek. 2006. Multi-word verbs in a flective language: the case of Estonian. In *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts*. 57–64.

Kaalep, Heiki-Jaan and Kadri Muischnek. 2008. Multi-word verbs of Estonian: a database and a corpus. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*. 23–26.

Kaalep, Heiki-Jaan and Kadri Muischnek. 2009. Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas. *Eesti Rakenduslingvistika Ühingu Aastaraamat* 5.

Kaalep, Heiki-Jaan and Tarmo Vaino. 2001. Complete Morphological Analysis in the Linguist's Toolbox. *Congressus Nonus Internationalis Fenno-Ugristarum Pars V* 9–16.

Kahusk, Neeme. 2011. Ühestatud sõnatähendustega korpus. `https://doi.org/10.15155/1-00-0000-0000-0000-00081l`. `https://doi.org/10.15155/1-00-0000-0000-0000-00081L`.

Kallas, Jelena and Kristina Koppel. 2018. Eesti keele ühendkorpus 2017 [Estonian Reference Corpus 2017]. `https://doi.org/10.15155/3-00-0000-0000-0000-071E7L`.

Katz, Graham and Eugenie Giesbrecht. 2006. Automatic Identification of Non-compositional Multi-word Expressions using Latent Semantic Analysis. In *Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*. Association for Computational Linguistics, 12–19.

Kendall, Maurice G. 1938. A new measure of rank correlation. *Biometrika* 30(1/2): 81–93.

Klebanov, Beata Beigman, Chee Wee Leong, and Michael Flor. 2015. Supervised Word-level Metaphor Detection: Experiments with Concreteness and Reweighting of Examples. In *Proceedings of the Third Workshop on Metaphor in NLP*. 11–20.

Kober, Thomas, Julie Weeds, John Wilkie, Jeremy Reffin, and David Weir. 2017. One Representation per Word - does it make Sense for Composition? In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and Their Applications*. Association for Computational Linguistics, 79–90.

Kohavi, Ron and George H John. 1997. Wrappers for Feature Subset Selection. *Artificial Intelligence* 97(1-2): 273–324.

Köper, Maximilian and Sabine Schulte im Walde. 2016. Automatic Semantic Classification of German Preposition Types: Comparing Hard and Soft Clustering Approaches Across Features. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 256–263.

Köper, Maximilian and Sabine Schulte im Walde. 2016a. Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Köper, Maximilian and Sabine Schulte im Walde. 2016b. Distinguishing Literal and Non-Literal Usage of German Particle Verbs. In *HLT-NAACL*. 353–362.

Köper, Maximilian and Sabine Schulte im Walde. 2017a. Applying Multi-Sense Embeddings for German Verbs to Determine Semantic Relatedness and to Detect Non-Literal Language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 535–542.

Köper, Maximilian and Sabine Schulte im Walde. 2017b. Complex Verbs are Different: Exploring the Visual Modality in Multi-Modal Models to Predict Compositionality. *MWE 2017* page 200.

Krenn, Brigitte. 2000. *The Usual Suspects: Data-Oriented Models for Identification and Representation of Lexical Collocations*. Ph.D. thesis, Saarbrücken: DFKI & Universität Des Saarlandes.

Krenn, Brigitte and Gregor Erbach. 1994. Idioms and Support Verb Constructions. *German in Head-Driven Phrase Structure Grammar* 46.

Krenn, Brigitte and Stefan Evert. 2001. Can We Do Better Than Frequency? A Case Study on Extracting PP-verb Collocations. In *Proceedings of the ACL Workshop on Collocations*. 39–46.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Kühner, Natalie and Sabine Schulte im Walde. 2010. Determining the Degree of Compositionality of German Particle Verbs by Clustering Approaches. In *Proceedings of the 10th Conference on Natural Language Processing*. 47–56.

Lai, Siwei, Kang Liu, Shizhu He, and Jun Zhao. 2016. How to Generate a Good Word Embedding. *IEEE Intelligent Systems* 31(6): 5–14.

Landauer, Thomas K and Susan T Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104(2): 211.

Lapesa, Gabriella and Stefan Evert. 2017. Large-scale Evaluation of Dependency-based DSMs: Are They Worth the Effort? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. 394–400.

Laporte, Éric. 2018. Choosing Features for Classifying Multiword Expressions. *Multiword expressions: Insights from a multi-lingual perspective* 1: 143.

Lavagnino, Elisa and Jungyeul Park. 2010. Conceptual Structure of Automatically Extracted Multi-Word Terms from Domain Specific Corpora: A Case Study for Italian. In *Proceedings of the 2nd Workshop on Cognitive Aspects of the Lexicon*. 48–55.

Lenci, Alessandro. 2008. Distributional Semantics in Linguistic and Cognitive Research. *Italian journal of Linguistics* 20(1): 1–31.

Levy, Omer and Yoav Goldberg. 2014. Neural Word Embedding as Implicit Matrix Factorization. In *Advances in Neural Information Processing Systems*. 2177–2185.

Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. Improving Distributional Similarity With Lessons Learned from Word Embeddings. *Transactions of the Association for Computational Linguistics* 3: 211–225.

Li, Jiwei and Dan Jurafsky. 2015. Do Multi-Sense Embeddings Improve Natural Language Understanding? *arXiv Preprint ArXiv:1506.01070* .

Li, Linlin and Caroline Sporleder. 2009. A Cohesion Graph Based Approach for Unsupervised Recognition of Literal and Non-literal Use of Multiword Expressions. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, 75–83.

Lin, Dekang. 1999. Automatic Identification of Non-compositional Phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*. Association for Computational Linguistics, 317–324.

Liu, Yi and Yuan F Zheng. 2006. FS_SFS: A Novel Feature Selection Method for Support Vector Machines. *Pattern Recognition* 39(7): 1333–1345.

Louw, Bill. 1993. Irony in the Text or Insincerity in the Writer? the Diagnostic Potential of Semantic Prosodies. *Text and Technology: In honour of John Sinclair* 240: 251.

Magerman, David M. 1995. Statistical Decision-Tree Models for Parsing. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 276–283.

Mahmood, Ali Mirza, Naganjaneyulu Satuluri, and Mrithyumjaya Rao Kuppa. 2011. An Overview of Recent and Traditional Decision Tree Classifiers in Machine Learning. *International Journal of Research and Reviews in Ad Hoc Networks* 1(1): 2011.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

McCarthy, Diana, Bill Keller, and John Carroll. 2003. Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Association for Computational Linguistics, 73–80.

McCarthy, Diana, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding Predominant Word Senses in Untagged Text. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, page 279.

McDonald, Scott. 2000. *Environmental Determinants of Lexical Processing Effort*. Ph.D. thesis, University of Edinburgh. College of Science and Engineering. School of Informatics.

McDonald, Scott and Chris Brew. 2004. A Distributional Model of Semantic Context Effects in Lexical Processing. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, page 17.

Mel'čuk, Igor A. 1998. Collocations and Lexical Functions. *Phraseology. Theory, Analysis, and Applications* 23–53.

Mel'čuk, Igor A. and Alain Polguere. 1987. A Formal Lexicon in the Meaning-Text Theory (or How to Do Lexica with Words). *Computational Linguistics* 13(3-4): 261–275.

Mihkla, Karl. 1964. *Eesti keele süntaks I. Prooviartikleid lihtlause süntaksi alalt [Estonian Syntax I]*. Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut, Tallinn.

Mihkla, Karl, Lehte Rannut, Elli Riikoja, and Aino Admann. 1974. *Eesti keele lauseõpetuse põhijooned I. Lihtlause [The Basic Characteristics of Estonian Syntax I. Simple Sentence]*. Tallinn: Valgus.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *arXiv Preprint ArXiv:1301.3781* .

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. 2013b. Exploiting Similarities Among Languages for Machine Translation. *arXiv Preprint ArXiv:1309.4168* .

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*. MIT Press.

Moon, Rosamund. 1998. *Fixed Expressions and Idioms in English: A Corpus-based Approach*. Oxford University Press.

Moore, Joshua, Christopher JC Burges, Erin Renshaw, and Wen-tau Yih. 2013. Animacy Detection With Voting Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 55–60.

Muischnek, Kadri. 2006. *Verbi ja noomeni püsiühendid eesti keeles [Fixed Expressions Consisting of Verbs and Nouns in Estonian]*. Ph.D. thesis, University of Tartu.

Muischnek, Kadri, Kaili Müürisep, and Tiina Puolakainen. 2013. Estonian Particle Verbs and Their Syntactic Analysis. In *Human Language Technologies a a Challenge for Computer Science and Linguistics: 6Th Language & Technology Conference Proceedings*. 338–342.

Muischnek, Kadri, Kaili Müürisep, and Tiina Puolakainen. 2017. Parsing and Beyond: Tools and Resources for Estonian. *Acta Linguistica Academica* 64(3): 347–367.

Muischnek, Kadri, Kaili Müürisep, Tiina Puolakainen, and Krista Liin. 2016. Parsing Estonian: Tools and Resources. In *Second International Workshop on Computational Linguistics for Uralic Languages*.

Müller, Andreas C. and Sarah Guido. 2016. Introduction to Machine Learning With Python: A Guide for Data Scientists.

Muru, Monika. 2018. Esinemissageduse mõju ühendverbide tähenduse moodustumisel [The Influence of the Frequency on The Compositionality of Estonian Particle Verbs]. *Bachelor Thesis. Manuscript in the Estonian Language Department of the University of Tartu* .

Muuk, Elmar. 1938. Verbide ja verbaalnoomenite kokkukirjutamisest [The Orthography of Verbs and Verbal Nouns]. *Eesti Keel* 1: 4–17.

Müürisep, Kaili. 2000. *Eesti keele arvutigrammatika: süntaks [Computer Grammar of Estonian: Syntax]*. Ph.D. thesis, Tartu Ülikool.

Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2015. Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space. *arXiv Preprint ArXiv:1504.06654* .

Nematzadeh, Aida, Afsaneh Fazly, and Suzanne Stevenson. 2013. Child Acquisition of Multiword Verbs: A Computational Investigation. In *Cognitive Aspects of Computational Language Acquisition*, Springer, 235–256.

Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language* 70(3): 491–538.

Orasan, Constantin and Richard Evans. 2001. Learning to Identify Animate References. In *Proceedings of the 2001 Workshop on Computational Natural Language Learning*. Association for Computational Linguistics, page 16.

Orasmaa, Siim, Timo Petmanson, Alexander Tkachenko, Sven Laur, and Heiki-Jaan Kaalep. 2016. EstNLTK-NLP Toolkit for Estonian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*.

Padó, Sebastian and Mirella Lapata. 2007. Dependency-Based Construction of Semantic Space Models. *Computational Linguistics* 33(2): 161–199.

Pearce, Darren. 2001. Synonymy in Collocation Extraction. In *Proceedings of the Workshop on WordNet and other Lexical Resources, Second Meeting of the North American Chapter of the Association for Computational Linguistics*. 41–46.

Pearce, Darren. 2002. A Comparative Evaluation of Collocation Extraction Techniques. In *Proceedings of the Third International Conference on Language Resources and Evaluation*.

Pecina, Pavel. 2008. A Machine Learning Approach to Multiword Expression Extraction. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*. Citeseer, 54–61.

Pedersen, Ted. 1996. Fishing for Exactness. In *Proceedings of the South-Central SAS Users Group Conference*. Austin, TX.

Pelevina, Maria, Nikolay Arefiev, Chris Biemann, and Alexander Panchenko. 2016. Making Sense of Word Embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*. Association for Computational Linguistics, Berlin, Germany, 174–183. `http://anthology.aclweb.org/W16-1620`.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. `http://www.aclweb.org/anthology/D14-1162`.

Pereira, Lis, Elga Strafella, Kevin Duh, and Yuji Matsumoto. 2014. Identifying Collocations using Cross-lingual Association Measures. In *Proceedings of the 10th Workshop on Multiword Expressions (MWE)*. 109–113.

Pichotta, Karl and John DeNero. 2013. Identifying Phrasal Verbs using Many Bilingual Corpora. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 636–646.

Podesva, Robert J. and Devyani Sharma. 2014. *Research Methods in Linguistics*. Cambridge University Press.

Pudil, Pavel and Jana Novovičová. 1998. Novel Methods for Feature Subset Selection With Respect to Problem Knowledge. In *Feature Extraction, Construction and Selection*, Springer, 101–116.

Ramisch, Carlos. 2014. *Multiword Expressions Acquisition: A Generic and Open Framework*. Springer.

Ramisch, Carlos, Silvio Ricardo Cordeiro, Agata Savary, Veronika Vincze, Verginica Barbu Mititelu, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Carla Parra Escartín, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*. Association for Computational Linguistics, Santa Fe, New Mexico, USA, 222–240. `https://www.aclweb.org/anthology/W18-4925`.

Ramisch, Carlos, Vitor De Araujo, and Aline Villavicencio. 2012. A Broad Evaluation of Techniques for Automatic Acquisition of Multiword Expressions. In *Proceedings of ACL 2012 Student Research Workshop*. Association for Computational Linguistics, 1–6.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. 2008. An Evaluation of Methods for the Extraction of Multiword Expressions. In *Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008)*. 50–53.

Rätsep, Huno. 1978. *Eesti keele lihtlausete tüübid [Patterns of Estonian Simple Sentences]*. Valgus, Tallinn.

Reddy, Siva, Ioannis Klapaftis, Diana McCarthy, and Suresh Manandhar. 2011a. Dynamic and Static Prototype Vectors for Semantic Composition. In *Proceedings of 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, 705–713.

Reddy, Siva, Diana McCarthy, and Suresh Manandhar. 2011b. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*. Asian Federation of Natural Language Processing, 210–218.

Rey, Denise and Markus Neuhäuser. 2011. Wilcoxon-signed-rank test. *International Encyclopedia of Statistical Science* 1658–1659.

Richter, Frank and Manfred Sailer. 2009. Phraseological Clauses in Constructional HPSG. In *Proceedings of the 16th International Conference on Head-Driven Phrase Structure Grammar*. 297–317.

Roweis, Sam T. and Lawrence K. Saul. 2000. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290(5500): 2323–2326.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, Springer, 1–15.

Sahlgren, Magnus. 2008. The Distributional Hypothesis. *Italian Journal of Disability Studies* 20: 33–53.

Sahlgren, Magnus and Alessandro Lenci. 2016. The Effects of Data Size and Frequency Range on Distributional Semantic Models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 975–980.

Sailer, Manfred. 2003. *Combinatorial Semantics and Idiomatic Expressions in Head-driven Phrase Structure Grammar*. Ph.D. thesis, Universität Tübingen.

Sailer, Manfred and Stella (eds.) Markantonatou. 2018. *Multiword Expressions: Insights from a Multi-lingual Perspective (Phraseology and Multiword Expressions)*, volume 1. Language Science Press.

Salehi, Bahar and Paul Cook. 2013. Predicting the Compositionality of Multiword Expressions Using Translations in Multiple Languages. In ** SEM@ NAACL-HLT*. 266–275.

Salehi, Bahar, Paul Cook, and Timothy Baldwin. 2014. Using Distributional Similarity of Multi-way Translations to Predict Multiword Expression Compositionality. In *EACL*. 472–481.

Salehi, Bahar, Paul Cook, and Timothy Baldwin. 2015. A Word Embedding Approach to Predicting the Compositionality of Multiword Expressions. In *HLT-NAACL*. 977–983.

Salle, Alexandre, Marco Idiart, and Aline Villavicencio. 2016. Matrix Factorization using Window Sampling and Negative Sampling for Improved Word Representations. *arXiv Preprint ArXiv:1606.00819* .

Savary, Agata, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, and Ivelina Stoyanova. 2017. The parseme shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Association for Computational Linguistics, 31–47.

Schone, Patrick and Daniel Jurafsky. 2001. Is Knowledge-Free Induction of Multiword Unit Dictionary Headwords a Solved Problem. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*. 100–108.

Schulte im Walde, Sabine, Anna Hätty, and Stefan Bott. 2016. The Role of Modifier and head Properties in Predicting the Compositionality of English and German Noun-Noun Compounds: A Vector-space Perspective. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. 148–158.

Schulte im Walde, Sabine, Stefan Müller, and Stephen Roller. 2013. Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*. 255–265.

Schütze, Hinrich. 1998. Automatic Word Sense Discrimination. *Computational Linguistics* 24(1): 97–123.

Scornet, Erwan, Gérard Biau, and Jean-Philippe Vert. 2015. Consistency of Random Forests. *The Annals of Statistics* 43(4): 1716–1741.

Seretan, Violeta. 2008. *Collocation Extraction Based on Syntactic Parsing*. Ph.D. thesis, University of Geneva.

Shutova, Ekaterina, Simone Teufel, and Anna Korhonen. 2013. Statistical Metaphor Processing. *Computational Linguistics* 39(2): 301–353.

Siyanova-Chanturia, Anna. 2013. Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings. *The Mental Lexicon* 8(2): 245–268.

Smadja, Frank. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics* 19(1): 143–177.

Snider, Neal and Inbal Arnon. 2012. A unified lexicon and grammar? Compositional and non-compositional phrases in the lexicon. *Frequency Effects in Language* 127–163.

Stevens, Stanley Smith. 1946. On the Theory of Scales of Measurement. *Science* 103: 677–680.

Taslimipoor, Shiva, Omid Rohanian, Ruslan Mitkov, and Afsaneh Fazly. 2018. Identification of multiword expressions: A fresh look at modelling and evaluation. In *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop*. Language Science Press, volume 2, page 299.

Tauli, Valter. 1973. *Standard Estonian Grammar. Volume 1: Phonology, Morphology, Wordformation*, volume 8 of *Studia Uralica Et Altaica Upsaliensia*. Almqvist & Wiksell, Uppsala.

Tauli, Valter. 1983. *Standard Estonian Grammar. Volume 2: Syntax*, volume 14 of *Studia Uralica Et Altaica Upsaliensia*. Almqvist & Wiksell, Uppsala.

Tian, Fei, Hanjun Dai, Jiang Bian, Bin Gao, Rui Zhang, Enhong Chen, and Tie-Yan Liu. 2014. A Probabilistic Model for Learning Multi-Prototype Word Embeddings. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 151–160.

Tsvetkov, Yulia, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor Detection With Cross-lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 248–258.

Tsvetkov, Yulia and Shuly Wintner. 2014. Identification of Multiword Expressions by Combining Multiple Linguistic Information Sources. *Computational Linguistics* 40(2): 449–468.

Turney, Peter D. and Michael L. Littman. 2003. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems (TOIS)* 21(4): 315–346.

Turney, Peter D., Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and Metaphorical Sense Identification Through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 680–690.

Turney, Peter D and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research* 37: 141–188.

Uchiyama, Kiyoko, Timothy Baldwin, and Shun Ishizaki. 2005. Disambiguating Japanese Compound Verbs. *Computer Speech & Language* 19(4): 497–512.

Uiboaed, Kristel. 2010. Statistilised meetodid murdekorpuse ühendverbide tuvastamisel [Statistical Methods for Phrasal Verb Detection in Estonian Dialects]. *Eesti Rakenduslingvistika Ühingu Aastaraamat* 6: 307–326.

Van De Cruys, Tim and Begona Villada Moirón. 2007. Semantics-based Multiword Expression Extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*. Association for Computational Linguistics, 25–32.

Veismann, Ann. 2009. *Eesti keele kaas- ja määrsõnade semantika võimalusi [Semantics of Estonian Adpositions and Adverbs]*. Ph.D. thesis, University of Tartu.

Veismann, Ann and Heete Sahkai. 2016. Ühendverbidest läbi prosoodia prisma [Particle Verbs and Prosody]. *Eesti Rakenduslingvistika Ühingu Aastaraamat* 12: 269–285.

Venkatapathy, Sriram and Aravind K Joshi. 2005. Measuring the Relative Compositionality of Verb-noun (VN) Collocations by Integrating Features. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 899–906.

Villada Moirón, Begona. 2005. *Data-driven identification of fixed expressions and their modifiability*. Ph.D. thesis, University of Groningen.

Řehůřek, Radim and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. 45–50.

Wasow, Thomas, Ivan Sag, and Geoffrey Nunberg. 1983. Idioms: An Interim Report. In *Proceedings of the XIIIth International Congress of Linguists*. CIPL Tokyo, 102–115.

Webelhuth, Gert, Sascha Bargmann, and Christopher Götze. 2018. Idioms as evidence for the proper analysis of relative clauses. In *Reconstruction Effects in Relative Clauses*, De Gruyter, 225–262.

Weeber, Marc, Rein Vos, and R. Harald Baayen. 2000. Extracting the Lowest-Frequency Words: Pitfalls and Possibilities. *Computational Linguistics* 26(3): 301–317.

Weinreich, Uriel. 1969. Problems in the Analysis of Idioms. *Substance and Structure of Language* 23–81.

Wendlandt, Laura, Jonathan K. Kummerfeld, and Rada Mihalcea. 2018. Factors Influencing the Surprising Instability of Word Embeddings. *arXiv Preprint ArXiv:1804.09692* `https://arxiv.org/pdf/1804.09692.pdf`.

Wermter, Joachim and Udo Hahn. 2004. Collocation Extraction Based on Modifiability Statistics. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*.

Xu, Weichao, Yunhe Hou, YS Hung, and Yuexian Zou. 2013. A comparative analysis of spearman's rho and kendall's tau in normal and contaminated normal models. *Signal Processing* 93(1): 261–276.

Yazdani, Majid, Meghdad Farahmand, and James Henderson. 2015. Learning Semantic Composition to Detect Non-compositionality of Multiword Expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1733–1742.

Zipf, George Kingsley. 1945. The Meaning-Frequency Relationship of Words. *The Journal of General Psychology* 33(2): 251–256.

CURRICULUM VITAE

Name: Eleri Aedmaa
Citizenship: Estonian
Date of birth: May 24 1989
Telephone: +372 5565 6952
E-mail: aedmaaeleri@gmail.com

Education:

2014–2019	University of Tartu, doctoral studies in Estonian and Finno-Ugric Linguistics
2017 spring	University of Stuttgart, visiting PhD student
2015 fall	Uppsala University, visiting PhD student
2011–2014	University of Tartu, MA in Estonian and Finno-Ugric Linguistics (computational linguistics)
2008–2011	University of Tartu, BA in Estonian and Finno-Ugric Linguistics
2005–2008	Võru Kreutzwald Gymnasium

Professional experience:

01.05.2019–	University of Groningen, Center for Information Technology, ICT process coordinator
2017–2018	University of Tartu, Institute of Estonian and General linguistics, junior research fellow in Estonian and Finno-Ugric Linguistics
2016	University of Tartu, Institute of Estonian and General linguistics, junior research fellow in language processing

Publications:

Aedmaa, Eleri, Maximilian Köper, and Sabine Schulte im Walde. 2018. Combining Abstractness and Language-specific Theoretical Indicators for Detecting Non-Literal Usage of Estonian Particle Verbs. Proceedings of NAACL-HLT 2018: Student Research Workshop: The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 9–16.

Aedmaa, Eleri (2017). Exploring Compositionality of Estonian Particle Verbs. In *Proceedings of the ESSLLI 2017 Student Session*, 197–208.

Aedmaa, Eleri. 2016. Eesti keele ühendverbide kompositsionaalsuse määramine. *Eesti Rakenduslingvistika Ühingu aastaraamat* 12: 5–23.

Aedmaa, Eleri and Mark Fišel. 2016. Whether to English More Alike Reordered Suits Estonian Input for Statistical Machine Translation Better? In *Proceedings of the Seventh International Conference Baltic HLT 2016*. IOS Press, 77–83.

Aedmaa, Eleri. 2015. Statistilised meetodid ühendverbide tuvastamisel tekstikorpusest. *Eesti Rakenduslingvistika Ühingu aastaraamat* 11: 37–54.

Aedmaa, Eleri. 2014. Statistical methods for Estonian particle verb extraction from text corpus. In *Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014)*, 17–22.

Muischnek, Kadri, Kaili Müürisep, Tiina Puolakainen, Eleri Aedmaa, Riin Kirt, Dage Särg. 2014. Estonian Dependency Treebank and its annotation scheme. In *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13)*, 285–291.

ELULOOKIRJELDUS

Nimi: Eleri Aedmaa
Kodakondsus: Eesti
Sünniaeg: 24. mai 1989
Telefon: +372 5565 6952
E-post: aedmaaeleri@gmail.com

Haridus:

2014–2019	Tartu Ülikool, eesti ja üldkeeleteaduse instituut, doktoriõpe
2017 kevad	Stuttgarti Ülikool, külalisdoktorant
2015 sügis	Uppsala Ülikool, külalisdoktorant
2011–2014	Tartu Ülikool, eesti ja üldkeeleteaduse instituut, MA (arvutilingvistika)
2008–2011	Tartu Ülikool, eesti ja üldkeeleteaduse instituut, BA (eesti keel)
2005–2008	Võru Kreutzwaldi Gümnaasium

Teenistuskäik:

al 01.05.2019	Groningeni Ülikool, Infotehnoloogiakeskus, IKT-koordinaator
2017–2018	Tartu Ülikool, eesti ja üldkeeleteaduse instituut, eesti ja soome-ugri keeleteaduse nooremteadur (0,5)
2016	Tartu Ülikool, eesti ja üldkeeleteaduse instituut, keele automaattöötluse nooremteadur (0,25)

Publikatsioonid:

Aedmaa, Eleri; Köper, Maximilian; Schulte im Walde, Sabine 2018. Combining Abstractness and Language-specific Theoretical Indicators for Detecting Non-Literal Usage of Estonian Particle Verbs. – Proceedings of NAACL-HLT 2018: Student Research Workshop: The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, lk 9–16.

Aedmaa, Eleri 2017. Exploring Compositionality of Estonian Particle Verbs. – Proceedings of the ESSLLI 2017 Student Session, lk 197–208.

Aedmaa, Eleri 2016. Eesti keele ühendverbide kompositsionaalsuse määramine. – Eesti Rakenduslingvistika Ühingu aastaraamat, 12, lk 5–23.

Aedmaa, Eleri; Fišel, Mark 2016. Whether to English More Alike Reordered Suits Estonian Input for Statistical Machine Translation Better? – Proceedings of the Seventh International Conference Baltic HLT 2016. IOS Press, lk 77–83.

Aedmaa, Eleri 2015. Statistilised meetodid ühendverbide tuvastamisel tekstikorpusest. Eesti Rakenduslingvistika Ühingu aastaraamat, nr 11, lk 37–54.

Aedmaa, Eleri 2014. Statistical methods for Estonian particle verb extraction from text corpus. – Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations (CCLCC 2014), lk 17–22.

Muischnek, Kadri; Müürisep, Kaili; Puolakainen, Tiina; Aedmaa, Eleri; Kirt, Riin; Särg, Dage 2014. Estonian Dependency Treebank and its annotation scheme. – Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13), lk 285–291.

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 lk.
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.