

MAARJA-LIISA PILVIK

Action nouns in a constructional network:  
A corpus-based investigation of  
the productivity and functions  
of the deverbal suffix *-mine*  
in five different registers of Estonian





**MAARJA-LIISA PILVIK**

Action nouns in a constructional network:  
A corpus-based investigation of  
the productivity and functions  
of the deverbal suffix *-mine*  
in five different registers of Estonian



University of Tartu, Institute of Estonian and General Linguistics

Dissertation accepted for the defence of the degree of Doctor of Philosophy on September 21<sup>st</sup>, 2021 by the Committee of the Institute of Estonian and General Linguistics, Faculty of Arts and Humanities, University of Tartu

Supervisors: Professor Liina Lindström, PhD (University of Tartu)  
Petar Kehayov, PhD (Ludwig Maximilian University of Munich)

Opponent: Professor Jurgis Pakerys, PhD (Vilnius University)

Commencement: November 18<sup>th</sup>, 2021 at 16.15 at Ülikooli 18-204, Tartu or in Zoom

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



European Union  
European Regional  
Development Fund



Investing  
in your future

ISSN 1406-1325

ISBN 978-9949-03-736-0 (print)

ISBN 978-9949-03-737-7 (pdf)

Copyright: Maarja-Liisa Pilvik, 2021

University of Tartu Press  
[www.tyk.ee](http://www.tyk.ee)

## ACKNOWLEDGEMENTS

Having done both my BA and MA theses about different topics, I must confess, I never imagined that starting my PhD studies on yet another completely uncharted territory for myself would be an undertaking that lasts for 8 years. During this time, I have been on multiple paths, taken several detours, and been lost more times than I would care to admit. My fascination with *-mine* nominalizations originated from their potential to encompass a charming bundle of syntactic issues. However, having been (and in many respects, still continuing to be) a young and naive researcher, it took me by a slight surprise that these syntactic issues turned out to be closely intertwined with morphological, semantic, as well as pragmatic and functional aspects of language use, and the sudden explosion of potential angles for analysis that accompanied this realization was quite overwhelming. Still, while my journey with the *-mine* nominalizations has been anything but a textbook example of successful completion of PhD studies, I have finally reached a stopover, humble and grateful for everything I have had the chance to discover and experience during my studies thus far.

First and foremost, I am ever so grateful to my supervisors, Liina and Petar, who have let me wander on my own while providing their support and feedback in the most understanding and non-invasive way. They are wonderful people whose sophistication, broad knowledge of various languages, and sincere interest in the constant advances made in different fields of linguistics is something I greatly admire. I deeply value Petar's helpful comments, shared references, and encouraging words, especially during the final stages of putting this thesis together. I cannot even begin to list the aspects in which I have received support from Liina over all these years and having her as my supervisor and a friend has been a great privilege and a source of security. I am convinced that without her, my liberty to explore different research topics and methods as well as my opportunities for teaching or traveling during my studies would have been immensely more limited.

I want to thank all the anonymous reviewers, who have given me useful feedback about my articles and helped me rethink what I at different stages thought I knew. My special thanks go to Ann Veismann, Heete Sahkai, and Jurgis Pakerys, whose comments, suggestions, and questions were kind, yet very helpful for revising and improving several aspects of the final draft of this thesis. I am also grateful to Joshua Wilbur, David Ogren, and the people at Keelekord for helping me turn my sentence monsters into actual readable sentences in English, often with very short notice, and Ann Siiman for very tactfully editing my Estonian sentence monsters.

I owe much of my chances to be involved in research projects, study abroad, participate at international conferences and events, and discuss my research with researchers at the top of their field to the support of the Archimedes Foundation, the Estonian Ministry of Education and Research, and the European Union (through European Social Fund and European Regional Development Fund). I

would also like to thank the Estonian branch of the Alfred Kordelin Foundation for their grant which I received at the very beginning of my PhD studies.

I am glad I did not have to attempt all this without (academic) friends. I am grateful to the people with whom I've had the chance to teach different courses and who, in the process, have also taught me a lot. I also want to thank all the people with whom I've collaborated in various research projects and who have really expanded my horizons. Room 430 in the Jakobi 2 building has hosted several people over the years who already were or have become my good friends. Thank you, Kristel, for all the inspiration, support, and help with staying on the right side of the fragile border between burning out and keeping busy. My dear colleagues and comrades in the persons of Ann, Triin, Eleri, Helen, and Mari have helped reflect my often incoherent thoughts and at other times (in other places), clear my head from those thoughts completely. Of course, the list of people both in and outside of academia who have given me useful advice over the years, who have been supportive, and from whom I've drawn inspiration is long. Most of all, however, I want to thank Karoliina, whose kindness and endless (even if ungrounded) optimism towards my capabilities has provided much-needed ballast, who has given me courage, brought comfort, and made sure I understood that the world outside the office doors is (just as) exciting and worth discovering.

I want to use this opportunity here to thank Merike, Arvo, Malle, and Tõnu for everything they've done for me over a long period. I've considered them my extended family and in different ways, they've played an important role in the chances I've had in life. But finally and above all, I thank my mother Svea to whom I can attribute everything good and decent that is in me. She is the one who has always encouraged me to be curious and whose subtle guidance has given me the perseverance to keep on with strenuous but rewarding endeavors. It is from her that I've learned what it takes to be independent. For this and so much more, I will always be in debt.

On the 26th of September 2021 in Tartu

# CONTENTS

ACKNOWLEDGEMENTS .....	5
LIST OF PUBLICATIONS.....	8
ABBREVIATIONS.....	9
1 INTRODUCTION.....	10
2 THEORETICAL BACKGROUND AND CORE CONCEPTS.....	18
2.1 Theoretical frameworks .....	18
2.1.1 Usage-based linguistics.....	18
2.1.2 Usage-based construction grammar .....	20
2.2 Productivity.....	24
2.2.1 Aspects of productivity .....	24
2.2.2 Measuring productivity .....	28
2.3 Register variation .....	31
3 DATA AND METHODS .....	34
3.1 The corpora .....	34
3.2 Data extraction and analysis methods.....	36
4 OVERVIEW OF THE PUBLICATIONS .....	40
5 RESULTS AND DISCUSSION.....	45
5.1 Morphological productivity as a function of register.....	45
5.2 Ranking the suffixes .....	50
5.3 Paradigmatic support for productivity .....	54
5.4 Syntactic constructions as co-determinants of morphological productivity.....	58
5.5 Limitations and further research .....	66
6 CONCLUSION .....	70
REFERENCES.....	72
SUMMARY IN ESTONIAN .....	78
PUBLICATIONS .....	87
CURRICULUM VITAE .....	247
ELULOOKIRJELDUS.....	249

## LIST OF PUBLICATIONS

**Publication 1 (P1):** Pilvik, Maarja-Liisa 2019. Assessing the productivity of the Estonian deverbil suffix *-mine* in five registers of Estonian. *SKY Journal of Linguistics* 32. 75–103.

**Publication 2 (P2):** Pilvik, Maarja-Liisa 2021. Comparing the productivity of Estonian deverbil suffixes *-mine*, *-us*, and *-ja* in five registers: A quantitative usage-based approach. *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEFUL* 12(1). 189–258.  
DOI: 10.12697/jeful.2021.12.1.06.

**Publication 3 (P3):** Pilvik, Maarja-Liisa 2017. Deverbil *-mine* action nominals in the Estonian Dialect Corpus. *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEFUL* 8(2). 295–326. DOI: 10.12697/jeful.2017.8.2.10.

**Publication 4 (P4):** Pilvik, Maarja-Liisa 2016. *olema* + *Vmine* konstruktsioonid eesti murretes. *Keel ja Kirjandus* 6. 429–446.

## ABBREVIATIONS

1INF – 1 <sup>st</sup> infinitive ( <i>-da</i> )	GEN – genitive case
1SG – 1 <sup>st</sup> person singular	ILL – illative case
1PL – 1 <sup>st</sup> person plural	INE – inessive case
2INF – 2 <sup>nd</sup> infinitive ( <i>-ma</i> )	IPS – impersonal passive mood
2SG – 2 <sup>nd</sup> person singular	NEG – negation marker
3PL – 3 <sup>rd</sup> person plural	NEWS – newspaper texts
3SG – 3 <sup>rd</sup> person singular	NMLZ – nominalization
ADE – adessive case	NOM – nominative case
ALL – allative case	PAR – partitive case
APP – active past participle	PCL – particle
APRP – active present participle	PL – plural
CND – conditional mood	PPP – passive past participle
CNG – connegative	PST – past tense
COM – comitative case	SCI – scientific texts
DIA – spoken dialects	SP – spoken spontaneous language
ELA – elative case	TR – translative case
FICT – fiction texts	

# 1 INTRODUCTION

In this thesis, I explore and describe, from a usage-based perspective, the productivity and functions of deverbal action noun constructions with the suffix *-mine* as emerging from naturally occurring written and spoken data. The suffix *-mine* is considered the default suffix for deverbal nominalization in the Estonian language, enabling the expression of a process or a state in the position of a noun (Erelt 2017: 831–840). This highly general and regular derivation pattern represents a borderline case between inflection and derivation, and therefore, its functions and properties have been compared to both those of nouns and those of non-finite verb forms (Vare 1994, Erelt 2017, Neetar 1988). Up until the year 1933, the suffix was considered to belong to the verb’s inflectional paradigm (Kasik 2015: 69, Erelt 2017: 831), but in later descriptions of Estonian, *-mine* has been categorized as a derivational suffix and the distributional properties of *-mine* nouns have been described as mostly nominal (e.g., Erelt et al. 1993: 269–271, Erelt 2014: 236, Erelt 2017: 831–840). Indeed, *-mine* nouns are more often than not structurally and syntactically very similar to non-derived referential nouns: they can have full case and number paradigms, they can be modified by adjectives (1) and demonstratives, and the NPs which they head can easily function in a sentence as typical subjects, objects, predicatives, attributes, and complements of adpositional phrases. The possibility of expressing a subject or object argument in the NP links the nominalization to the corresponding verb’s argument structure, but the form of the arguments is again conditioned by the primarily nominal syntactic distribution of *-mine* nouns. Either argument can be typically expressed as a prenominal genitive modifier (2), representing the nominal double-possessive type in Maria Koptjevskaja-Tamm’s (1993, 2003) typological classification of action nominal constructions, but when both arguments are explicitly expressed, the subject is usually demoted to an oblique (ex. 3, the ergative-possessive type) or the patient is incorporated into a compound (ex. 4, the possessive-incorporating type).

- (1) *Sest see pole õige ristimine, kui kolm  
because this be.NEG.CNG proper christen:NMLZ when three  
tilka vett sülelapsele pähe pillatakse. (FICT)<sup>1</sup>  
drop.PAR water:PAR child:ALL head.ILL drop:IPS  
‘Because this is not a proper christening when three drops of water are dropped on  
a child’s head.’*
- (2) *Augusta närveerib Lioneli kadumise pärast,  
Augusta be\_nervous:3SG Lionel:GEN disappear:NMLZ.GEN for  
kuid ei räägi Lakenile põhjust. (NEWS)  
but NEG tell.CNG Laken:ALL reason:PAR  
‘Augusta is nervous about Lionel’s disappearance, but doesn’t tell Laken why.’*

---

<sup>1</sup> The source corpora, if available and relevant, are indicated in the brackets by the abbreviation of the corresponding register. The corpora for the registers are introduced in Section 3.1.

- (3) *Ka teised küsitlused ennustavad enam kui 300 kohta*  
 also other:PL poll:PL predict:3PL more than 300 seat:GEN  
*saamist LDP poolt. (SCI)*  
 get:NMLZ:PAR LDP by  
 ‘Other polls also predict that the LDP will receive more than 300 seats.’
- (4) *Jah, see on nagu sinu majaehitamine,*  
 yes this be.3SG like you:GEN house:GEN+build:NMLZ  
*alguses teed palgikorra maas valmis*  
 beginning:INE make:2SG log-arrangement:GEN ground:INE ready  
*ja siis tõstad üles seina peale. (FICT)*  
 and then lift:2SG up wall:GEN onto  
 ‘Yes, this is like you building the house, at first, you arrange the logs on the ground and then you lift them up on the walls.’ (lit. ‘your building of the house’)

Despite this apparent categorial ‘nouniness’, *-mine* nouns become relevant on various levels of language description involving the interface of semantics and morphosyntax, and demonstrate a versatility which is also characteristic of nominalizations cross-linguistically (see e.g., Noonan 1997, Yap et al. 2011). For example, like derivatives in many other languages, *-mine* nouns can acquire specialized meanings (e.g., that of a result or an object of an action, Kasik 2015) which are more or less systematic semantic extensions through the processes of metonymy and metaphor and often involve changes in the noun phrase structure, such as loss of arguments (5) or the use of plural cases (6) (cf. Kerge 2001: 42–43); case-marked *-mine* action nouns can be used as converbs (7, 8), which is also a cross-linguistically attested phenomenon (Koptjevskaja-Tamm 1993: 44, Ylikoski 2003: 199, 203); and they emerge as central participants in syntactic structures whose form and meaning cannot be straightforwardly derived from those of their components (9, 10). In some varieties of Estonian, the deverbal *-mine* forms can express their ‘verbiness’ even by copying the verbal argument marking (11).

- (5) *Kärneri sõnul ei tohi panga lisatav intress*  
 Kärner:GEN word:PL.ADE NEG may:CNG bank:GEN add:APRP interest  
*olla kõrgem sellest, mis oli kirjas*  
 be:1INF higher this:ELA what be:PST.3SG writing:INE  
*fondile tehtud pakkumises. (NEWS)*  
 fund:ALL make:PPP offer:NMLZ:INE  
 ‘According to Kärner, the interest added by the bank may not be higher than what was stated in the offer made to the fund.’
- (6) *Ta pea oli liiga väike, et sinna mingeid*  
 he:GEN head be:PST.3SG too small that there some:PL:PAR  
*teadmisi talletada. (FICT)*  
 know:NMLZ:PL:PAR store:1INF  
 ‘His head was too small to store any facts/knowledge.’

- (7) *līdolite* *nāgemisel* *vallandusid* *pisarad* *ja* *kisakoor*  
 idol:PL.GEN see:NMLZ:ADE well\_up:PST:3PL tear:PL and scream+choir  
*oli* *meeletu.* (NEWS)  
 be:PST.3SG insane  
 ‘When the idols were seen, tears welled up and the screaming was insane.’
- (8) *Tegime* *seda* *Mardi* *teadmisel.* (Sahkai 2011)  
 do:PST:1PL this:PAR Mart.GEN know:NMLZ:ADE  
 ‘We did this with Mart’s knowing.’
- (9) *Maja* *on* *lammutamisel.* (Sahkai 2011)  
 house be.3SG demolish:NMLZ:ADE  
 ‘The house is being demolished.’ (lit. ‘The house is at demolishing.’)
- (10) *see* *küsimus* *tahab* *lahendamist* (SP)  
 this question want:3SG solve:NMLZ:PAR  
 ‘this issue needs to be solved’ (lit. ‘the question wants solving’)
- (11) *põle* *sull* *seda* *väl’lä* *saagimest* *ega* *kedagi* (DIA)<sup>2</sup>  
 be.NEG.CNG you:ADE it:PAR out saw:NMLZ:PAR NEG anything  
 ‘you do not have to cut it out or anything’ (lit. ‘you do not have that out cutting or anything’)

Such functional diversity is without a doubt supported and maintained by the high productivity of the *-mine* derivation pattern, providing an extensive lexical base. Productivity is a concept extensively used and debated in linguistics, especially with regard to morphology (see Bauer 2001, Baayen 2009, Gaeta & Ricca 2015). On a more abstract level, productivity can be understood as the potential of a linguistic construction to be used for coining new members. Most scholars agree that this availability or readiness for expanding the vocabulary is a matter of degree and results from a complex interplay between various semantic, structural, pragmatic, and paradigmatic constraints. However, moving beyond intuition in accessing and evaluating the different aspects of linguistic productivity in practice requires more concrete tools. An attempt to operationalize corpus frequencies in measuring and comparing the productivity of the *-mine* suffix to that of other deverbal suffixes across different registers is made in this thesis. Some of the aspects that arise from the quantitative output of productivity are also examined in a more contextualized manner. In particular, the effect of the different functions of word-formation on morphological productivity as well as the relationship between morphological and syntactic productivity will be discussed.

---

<sup>2</sup> The expected form for ‘it’ would be the genitive *selle* (*põle sull selle väl’lä saagimest*). Here, however, the partitive *seda* copies the argument marking in the corresponding impersonal finite clause (*seda ei saeta välja* it:PAR NEG saw:IPS.NEG out ‘it is not cut out’) or in the equivalent modal construction (*põle sull vaja seda väl’lä saagida* be.NEG.CNG you:ADE need it:PAR out saw:1INF ‘you do not have to cut it out’).

Naturally, the highly ambivalent behavior of *-mine* nouns illustrated above has driven research with different foci and different premises. The *-mine* suffix is included in many works describing the structure of nominalized NPs (Erelt et al. 1993, Erelt 2017) and the relationship between the structure and meaning of various derivational suffixes (e.g., Vare 1993, 1994, Erelt et al. 1995, Kasik 1994, Kasik 2004, Kasik 2015, Saari 1997)<sup>3</sup>. In broad terms, the research specifically concerning the category of *-mine* nouns (or action nouns more broadly) in Estonian can be divided into three types: studies which focus on the syntactic or grammatical aspects of nominalization (i.e., applying the nominalization suffix to use a clause in the position of a noun, Kasik 1968, 1975, Vare 1991, Kerge 2001), studies which focus on the lexical function of derivation (i.e., applying the nominalization suffix to create new terms, Kerge 2002, 2003), and studies covering its textual or stylistic functions such as abstracting, generalizing, or thickening the text (Kasik 2006), although the different types are often intertwined. These three main strands of research converge with the two main functions of word-formation mentioned in the literature: labeling / lexical enrichment and syntactic recategorization / transposition (e.g., Kastovsky 1986, Bauer 2004). Lexical enrichment entails creating new words for naming or labeling concepts, and transposition refers to changing the word class of a lexeme in order to transfer the same meaning to a different function in a sentence. The latter is often closely linked to anaphoric referencing, i.e., using derived words to refer to previous events in discourse, which, in turn, often serves to create stylistic variation or text cohesion.

The functions for which *-mine* nouns are created have also been shown to depend on register. As most studies have been done based on written language, the functions considered have been those prevailing in namely written domains of language use. For example, it has been shown that legal texts mostly make use of *-mine* nouns in the transpositional function (Kerge 2002: 38); the same is true for journalistic texts (Kasik 2006, Kerge 2003), although the lexical function of creating abstract concepts is also mentioned (Kerge 2002); academic and scientific texts have been found to use *-mine* nouns mostly in the creation of terminology (Kerge 2002, 2003).

In addition to the aforementioned lines of research which focus on different aspects of word-formation, there are also individual studies which seek to describe the *-mine* nouns in certain syntactic settings, where the part-of-speech properties of *-mine* nouns and phrase structure rules prove insufficient for characterizing the constructions as a whole. Helmi Neetar, who has described verbal nouns in Estonian dialects (1990), has also written an article about action noun constructions in Estonian dialects (1988). In the article, she lists several constructions consisting of a finite verb and an action noun (mainly with the suffix *-mine*), the latter often interchangeable with a non-finite verb form in dialects (12, 13), and presents these as a ‘syntactic problem’.

---

<sup>3</sup> For an overview of the development of research on Estonian word-formation throughout the 20<sup>th</sup> century see Kasik et al. 2002.

- (12) *lehm tahab joota / jootmist*  
 cow want:3SG give\_water:1INF give\_water:NMLZ:PAR  
 ‘the cow needs to be given water’ (lit. ‘the cow wants to give water / water-giving’)
- (13) *sinna ei maksa teil minna / minemine*  
 there NEG pay.CNG you.PL:ADE go:1INF go:NMLZ  
 ‘you should not go there’ (lit. ‘it doesn’t pay for you to go there’)

Decades later, Heete Sahkai (2011) convincingly demonstrates the possibilities of adopting a constructionalist approach for describing action nouns (including *-mine* nouns) in phrasal expressions that cannot be analyzed as regular noun phrases. The constructions she analyzes are the Nominalization-Based Auxiliary Construction (*Maja on lammutamisel* ‘The house is being demolished’) and the Adessive Manner and Cause Adverbial Construction (*Tegime seda Mardi teadmisel* ‘We did this with Mart’s knowing’) which, in turn, inherits several formal and semantic properties from the Genitive Agent Phrase Construction (e.g., *Projekt on Mardi kirjutatud* ‘The project is written by Mart’). Sahkai demonstrates that such constructions are stable and at least partially productive, meaning that they license several action nouns either from a specific semantic class or from a larger pool of candidates. She also points out that more similar constructions are likely to be found among expressions where the nouns are used in adverbial, attributive, or predicative function (Sahkai 2011: 165).

The constructionalist perspective is also somewhat presented in the more recent complete description of Estonian word-formation (Kasik 2015), where Reet Kasik considers complex words as morphological constructions. In a very usage-based spirit, Kasik also recognizes the limits of using dictionaries as a single source for describing the functions and productivity of different word-formation patterns and acknowledges the role of electronic text corpora, social media, and multiple other sources of actual language use (including introspection). She also gives constructional analogy a prominent role in active word-formation, even when the morphological rule itself cannot be considered productive (Kasik 2015: 16–17). Despite all this, the methodology adopted in the complete description does not differ much from Kasik’s earlier structuralist and functionalist works: rather than taking corpora as independent, representative collections of language with their own internal dynamics, also reflected in the frequency with which certain constructions are used, the research is simply informed by the usage events attested in the corpora, and carefully selected corpus examples serve to validate the linguist’s general impressions. This should by no means be interpreted as a shortcoming of Kasik’s work since such thorough descriptive works informed by naturally occurring examples of language use form the essential qualitative foundation against which quantitative approaches can be tested through individual case studies. Still, it remains a fact that corpus

frequencies have so far largely been neglected or very broadly presented in most of the research about Estonian derivational morphology<sup>4</sup>.

Despite ample attention to word-formation and derivation in Estonian linguistics, there are thus two very important aspects which have not been addressed in existing research. First, the degree to which different morphological categories hold the potential to coin new members has not been empirically tested and directly compared across several categories. While *-mine* has been considered the most productive deverbal suffix in Estonian, it is not clear whether this productivity is absolute or conditioned by factors such as register and function, and how other deverbal suffixes rank with respect to that reference standard. Second, studies have focused only on certain registers of written language<sup>5</sup>, even though it has been argued that speech has the potential to express a more richly developed and more fully revealed grammatical system (Halliday 1994: xxiii), thereby also establishing the most fertile ground for linguistic change (Nevalainen & Raumolin-Brunberg 2003: 28). Both gaps can be justified by the lack of access to large collections of (spoken) data or to the methods with which such large collections could be analyzed. However, this is at least in part also a reflection of certain attitudes and dispositions according to which productive word-formation is a property of intuitive linguistic competence (Kasik 2011), and the role of usage frequency becomes prominent only in the context of lexicalization or semantic specialization of derivatives (Kasik 2011, Kasik 2015: 56, 340).

The current thesis seeks to start filling these gaps and complement the descriptions of *-mine* nouns by adopting a usage-based perspective and making more extensive use of available corpus data, including assigning meaning to the (relative) frequencies with which different patterns are attested. The seven main gaps along with the corresponding research questions are outlined as follows.

**Goal 1:** Provide observations from spoken spontaneous language in the discussion of Estonian derivational morphology which so far has excessively been dominated by reasoning based on standard written language. Speech is an inherently different medium than written language and this has consequences for both morphological productivity, the syntactic combinations into which deverbal nouns are likely to enter, as well as to the spectrum of functions which deverbal nouns can realize. This goal motivates the first, most general research question (**RQ1**): **Is the use of common derivational morphology a differentiating feature between written and spoken registers?**

---

<sup>4</sup> Krista Kerge's studies (2002, 2003) can be considered a promising exception, where the quantitative analysis of text complexity (among which *-mine* nominalization is also analyzed) is combined with a qualitative interpretation of the functions of word-formation patterns in different decades and the lexicalization of individual word forms. However, these studies are mostly based on simple token frequencies, which have been shown to be inflated by a few types occurring very frequently. For assessing productivity in this thesis, the focus is rather on the other end of the frequency spectrum, i.e., on words which rarely occur in the samples.

<sup>5</sup> Some studies (Hennoste 2000a, 2000b, Keevalik 2003, and the references mentioned therein) have, in passing, also dealt with aspects related to morphological derivation in spoken Estonian, but their focus has been more on non-standard suffixes and the derived words whose use is a differentiating lexical feature of the colloquial register.

**Goal 2:** Use quantitative measures derived directly from corpus frequencies to highlight the multiple facets of morphological productivity and analyze its relation to register and function. Defining productivity through the scalar concept of probability entails operationalizing type, token and *hapax legomenon* frequencies and facilitates the study of productivity across different types of texts. This goal motivates the second research question (RQ2): **Can corpus-based quantitative measures of morphological productivity tell us something about the constraints and functions of deverbal suffixes in different registers of Estonian?**

**Goal 3:** Compare the morphological productivity of *-mine* with that of two other frequent deverbal suffixes – *-ja* and *-us* – across different registers in Estonian. The existing quantitative measures of morphological productivity provide means for ranking multiple affixes and for characterizing the use of the suffixes in different communicative settings. This goal motivates the third research question (RQ3): **How do the three most frequent deverbal suffixes rank in terms of their productivity in different registers?**

**Goal 4:** Discuss the effect of sample size and methodology on the interpretation of the quantitative productivity measures. Various methods exist for assessing and comparing morphological productivity, but their (sometimes conflicting) implications should be carefully contextualized and anchored in the underlying data. This goal motivates the fourth research question (RQ4): **Do the basic productivity measures correspond to general linguistic intuitions when they are based on relatively small samples?**

**Goal 5:** Examine the relative frequencies of bases and suffixes as well as the relative frequencies of rival suffixes *-mine* and *-us* in order to capture regular semantic correspondences. Derivatives with lower relative frequencies than their bases have been shown to contribute more to the productivity of a pattern. Rival suffixes, in turn, should exhibit a complementary distribution with regard to bases to which they attach. This goal motivates the fifth research question (RQ5): **Can relative frequencies help detect semantic regularity?**

**Goal 6:** Examine the versatile properties of *-mine* nouns as a fluid category between prototypically nominal and verbal domains. Spoken dialectal data, in particular, is investigated in more detail as it shows richer variation and sheds light on features lost or hidden in contemporary (written) data. This goal motivates the sixth research question (RQ6): **Which combinatorial features of *-mine* nouns emerge from older colloquial data and what implications does this have for the quantitative assessment of morphological productivity?**

**Goal 7:** Introduce a periphrastic action noun construction thus far not discussed in the literature on Estonian action nouns – the impersonal-modal *olema* ‘be’ + *Vmine<sub>NOM/PAR</sub>* construction. The use of *-mine* action nouns in periphrastic constructions differs from their use as regular and referential nouns, while simultaneously both relying on and contributing to the productivity of the derivation pattern. This goal motivates the last research question (RQ7): **What are the properties and potential development paths of a periphrastic action noun construction and what does it tell us about the relationship between morphological and syntactic productivity?**

Through combining morpheme-based (formal) and schema-based (semantic) perspectives, this study seeks to exemplify the role of derivational morphology in actual language use and to find possibilities for describing word-formation patterns as categories between morphology and syntax in a more holistic manner. Namely, constructional approaches to morphology and syntax are utilized to bring the different aspects of nominalization closer together. The analysis conducted in this thesis is mainly synchronic and descriptive on the premise that the corpora used as a primary data source are representative collections of texts fixing the state of language from one particular point in time or period of time. The diachronic development of word meanings and constructions is not the focus of the articles with the exception of P4, where constructional relations and potential development paths are also discussed. However, since diachronic changes are often reflected in synchronic variation (Östman & Trousdale 2013), strictly separating the two aspects is neither possible nor desirable.

The thesis comprises 4 articles and the current overview section. While the main motivation for writing this thesis lies in syntax (this is visible in the order in which the articles were originally published), the ordering of the articles in this thesis follows a structure which presents the analysis and argumentation from more general to more specific, i.e., from general word-formation tendencies in different registers to one specific action nominal construction in one particular register. I hope to show, however, that there is an inherent, natural gradience to these windows into the behavior of *-mine* nouns. In the first case studies (P1 and P2), the morphological productivity of *-mine* is analyzed both across different registers and in comparison with two other frequent deverbal suffixes (*-us* and *-ja*) in order to establish its relative prominence over other derivational means and to argue for the need to account for register-specific variation in descriptions of word-formation. In the third article (P3), one specific register, namely spoken regional dialects, is chosen to map the properties and functions of *-mine* derivatives in data thus far largely excluded from contemporary usage-based studies on Estonian derivation. This includes also describing the forms in a wider syntactic context. Finally, in the case study described in P4, dialect data is examined in even more detail by presenting and analyzing one previously under-described action noun construction with alternating semantics. Parallels from other Finnic languages support the argument that the construction has gradually lost modality and has expanded to the passive-impersonal domain in Estonian.

Before proceeding, a terminological remark has to be made. In the four publications and throughout this overview section, the derived nouns and the relevant morphological processes are referred to using several different terms. Therefore, *suffix*, *derivation pattern*, *morphological construction*, *morphological category*, and *morphological process* all have been used to refer to a schematic form-meaning pairing used for the creation or categorization of words. In turn, *action noun*, *action nominal*, *-mine noun*, *-mine derivative* and *-mine formation* all refer to the specific instantiation of that schematic linguistic unit. *Nominalization* is used in both functions.

## 2 THEORETICAL BACKGROUND AND CORE CONCEPTS

### 2.1 Theoretical frameworks

I have not written the publications comprising this thesis in the spirit of one specific theoretical framework and have often borrowed bits and pieces from different approaches in order to describe the different aspects involving *-mine* nouns in Estonian. The concepts and terms used in the articles (e.g., the ones referring to participant roles in syntactic composition) are mostly coherent with Basic Linguistic Theory in the sense of Dixon (1997) and Dryer (2006). However, the main principles of analysis overarching this thesis undoubtedly come from a usage-based approach to studying linguistic variation, both in methodological as well as substantive terms. In addition, as this thesis presupposes the existence of a complex network of interrelated constructions as form-meaning pairings, I also adopt the common conceptions from the usage-based construction grammar framework.

#### 2.1.1 Usage-based linguistics

From a usage-based perspective, speaker's linguistic knowledge is structured through language use and grammar arises from complex recurrent regularities in human communication (e.g., Langacker 1987, Bybee 2010, 2013). The defining aspect of this approach is that grammar is regarded as the product of usage, and is not a mere repository to be accessed in language use (Perek 2015: 6). It is a dynamic and constantly changing system, consisting of fluid categories, symbolic conventions, and flexible, probabilistic constraints, reorganized under the pressure of general cognitive processes such as categorization, prototypicality, extension, etc. (Lemmens 2019, Perek 2015, Diessel 2019). The symbolic conventions constitute abstract schemas, which are extracted and gradually generalized from specific instances. In contrast to generative approaches, in which linguistic units either have to be produced by abstract rules in grammar or be stored in the lexicon (what Langacker (1987: 26) calls the 'rule-list fallacy'), a usage-based approach allows for the simultaneous storing of both the higher-level generalizations and lower-level units in the grammar (Perek 2015: 6). Although in the context of morphology at least, many formal linguists also admit that acts of innovation can change the way language potential is used, the thing that changes for them is the norm, not the system (Bauer 2001: 28). From a usage-based perspective, however, such a distinction is generally considered irrelevant, since there is no adequate access to the system separate from the norm (Lemmens 2019).

This possibility to constitute linguistic expressions at different levels of abstraction and the notion of fluid categories are particularly beneficial for studying *-mine* nouns due to their ambivalent part-of-speech properties (nominal vs. verbal), the gradual transition from purely transpositional to semantically

idiosyncratic instances among them, the partial functional overlap with non-finites and other action nouns, etc. This is not to say that discrete categories in grammar do not exist or that there is no distinction between syntax, morphology, and the lexicon. Rather, the usage-based approach enables handling those fuzzy and overlapping areas in language where *-mine* nouns also seem to belong without unnecessary tension arising from rigid categories and the gradient phenomena they describe.

From a methodological perspective, the name *usage-based* suggests that one should study data reflecting naturally occurring language use in order to make theoretical generalizations about the language system. While *usage-based* and *corpus-based* are not synonymous and studying language use is certainly not restricted to studying usage events<sup>6</sup> represented in a corpus (Lemmens 2019), structured corpora are the most accessible source of actual language use to a linguist interested in linguistic variation. This is especially true when investigating highly variable non-standard spoken data such as Estonian dialects, for which the recordings in the corpus and in the archives are currently the only source for accessing data produced in (semi-)natural settings for all 10 traditional language varieties.

While usage-based linguistics cannot be regarded as a grammatical model by itself, its basic views and principles are more compatible with some theories than others. In this thesis (particularly in this overview section), I have mainly utilized the constructionalist models which have adopted the usage-based approach (e.g., Bybee 1995, Croft 2001, Croft & Cruse 2004, Diessel 2019). The basic assumptions and tools within that direction of construction grammar are described in more detail below. By comparing linguistic productivity in different registers and acknowledging the somewhat diverging roles of *-mine* nouns in spoken and written registers, I also share common ground with other functionalist and/or corpus-based approaches (e.g., Biber & Conrad 2019) which accentuate register awareness in descriptions of grammar.

All the usage-based approaches have common tenets in their rejection of traditional discrete categorization and the prominent role of frequency, among other things. According to Diessel, ‘functional and cognitive linguistics have always emphasized that linguistic structure is motivated by semantic and pragmatic aspects of communication and discourse, but in the recent literature, the focus of analysis has shifted from communication and meaning to frequency and processing’ (Diessel 2019: 30). This means that a more prominent role has been attributed to cognitive processes responsible for mental storage, representation, and activation, guiding the speakers’ choice of linguistic means. These processes are also reflected in usage frequency. The frequency with which a linguistic unit is used has been shown to play a major role in linguistic representations and processing. Perek (2015: 7) presents an excellent list of relevant references about studies which have found *inter alia* that frequent words are more resistant to

---

<sup>6</sup> In Langacker’s terms, *usage event* is a symbolic expression which a speaker uses for a specific purpose in a specific context (Langacker 1987: 494).

morphosyntactic change and stored in the mental lexicon as a whole, that frequent sequences of words are processed more easily, and that the co-occurrence of verbs and their frequent complements also facilitates processing. A particularly important observation in the context of this thesis is that the tendency of complex words to be stored and parsed as whole words rather than combinations of stem and affix has been shown to correlate with the frequency of their base form (Hay 2001, Hay & Baayen 2002). The importance of base frequency in detecting lexicalized *-mine* forms has also been suggested by Kerge (2003: 24–25), but it is notable that it is not the absolute frequencies but the relative frequencies between bases and derivatives that should matter for semantic decompositionality. Lexical processing is therefore determined not solely by the frequency of the complex word itself but also by the frequency with which its components appear by themselves and in other words. This paradigmatic interdependence is also addressed in this thesis.

### **2.1.2 Usage-based construction grammar**

Construction grammar (CxG) is a family of theoretical models which has its roots in the Saussurean idea of conventionalized form-meaning pairings. Constructions are symbolic units (or signs in the Saussurean sense) consisting of syntactic/phonological form and semantic/pragmatic content, and they are considered the basic units of language (Goldberg 2006). As such, the notion of construction covers a wide range of such pairings with different degrees of internal complexity and abstractness, i.e., complex words, idioms, and argument structures can all be considered constructions. Sometimes, the notion of construction is extended to also include simple lexemes or bound morphemes (e.g., Croft & Cruse 2004). Constructionalist approaches therefore presuppose a continuum between syntax and lexicon, instead of forcing a fixed division between the two (Goldberg 2003: 223), and provide the means for a uniform analysis of both ‘core’ and ‘peripheral’ linguistic features (Hoffmann & Trousdale 2013) as all linguistic structure is essentially emergent and fluid. The difference between constructions representing ‘normal syntax’ in CxG and more formal rules of syntax (e.g., phrase structure rules) is in the respective bottom-up and top-down manner in which the two approaches handle linguistic generalizations (Hilpert 2014: 68–69). The bottom-up approach means deriving more general and abstract patterns from recurrent low-level linguistic sequences, whereas the top-down approach generates low-level sequences based on some predetermined relationships between categories and rules.

Constructions as abstract representations, known as constructional schemas in Cognitive Grammar (Langacker 1991: 17–20), emerge from recurrent experience with lexical expressions with overlapping properties. Constructions are organized in a complex network. In this network, hierarchical inheritance relations capture formal and functional similarities between more abstract and more specific constructions whereas specific constructional characteristics are not projected

upwards in the hierarchy (Hilpert 2014: 57, 59). Low-level constructions/schemas are generalized over more specific lexical sequences (e.g., *-mine* nouns, idioms) and can even be lexically fully specified (e.g., *by and large*). High-level constructions/schemas (e.g., the RESULTATIVE construction)<sup>7</sup>, in turn, are abstract and generalized over highly variable outputs. In addition to taxonomical/hierarchical links (also known as instance links) connecting low-level constructions with higher-level constructions, the constructional network also accommodates polysemy links and metaphorical links (for extending the prototypical meaning of a construction to slightly different senses) as well as subpart links (for associating formally or semantically related constructions with the same level of abstraction, not seen as instances of one another) (Goldberg 1995: 75–81). A linguistic item can therefore be linked to more than one constructional schema and receive motivation from multiple sources (Booij 2019: 6, Hilpert 2014: 63).

Usage-based construction grammar differs from more formal CxG (widely adopted in computational linguistics and lexicography, for example) in that it generally eschews complex formalizations (Hoffmann & Trousdale 2013) and that it is output-oriented, i.e., it considers low-level constructions as primary over the high-level constructions (Bybee 1995, Croft & Cruse 2004). Consequently, in contrast with the so-called complete inheritance models of CxG, the usage-based approach assumes that the same information can be represented across different levels of abstraction, not just once in the highest-level schema (Hilpert 2014: 66–68). While the general line of research within usage-based CxG has used linguistic output in attempts to propose the more entrenched high-level abstractions, often serving to make crosslinguistic generalizations (Goldberg 2013), the low-level schemas and lexically specified units themselves are also of great importance for the language user. As Diessel (2019: 16–17) points out, recent studies have shown that for processing, the role of low-level and item-specific constructions (rather than abstract schemas) can be of much more relevance than theoretically assumed (see also Pirrelli et al. 2020). Low-level constructions are usually mentioned in the context of language acquisition where they ‘model the speaker’s knowledge of individual words and their combinatorics’ (Hilpert 2014: 95). Low-level generalizations have been shown to be relevant for *-mine* nouns with regard to the syntactic constructions with non-compositional meanings (Sahkai 2011), but they are definitely important for the analysis of morphological productivity as well.

---

<sup>7</sup> The constructional status of highly general, ‘semantically null’ constructions (e.g., the SUBJECT-PREDICATE construction or the MODIFIER-HEAD construction) has been debated within the CxG (cf. Croft 2001, Goldberg 2006, Fillmore et al. 2012, Hilpert 2014) as such generalizations seem to fail to meet the requirement of constructions being combinations of form and meaning. Postulating them as semantically empty constructions would not differ significantly from the dictionary-and-grammar model of linguistic knowledge (Hilpert 2014: 52). Croft (2001: 63–131) and Hilpert (2014: 68–69) point out that this forces constructionalists to also reconsider the cognitive status of highly abstract categories such as word classes or syntactic categories.

Like syntactic constructions, complex words are also instantiations of constructional schemas (e.g., Croft 2001, Goldberg 2006, Booij 2010). Word-formation patterns are seen as partially schematic generalizations across related words, which then license the individual lexemes. Schemas express predictable properties of existing complex words and indicate how new ones can be created. Again, the difference between word-formation *rules* and morphological constructions mainly lies in the top-down *vs.* bottom-up direction in which the relevant processes take place. Rather than composing the meaning of words from individual morphemes with specific meanings, speakers are expected to form generalizations about the meanings associated with particular forms and their components based on recurring experience with them. The application of CxG solely to the analysis of words has emerged within the field under the name construction morphology, abbreviated as CxM (Booij 2010, Booij 2019). In the framework of CxM, all lexically specified instantiations of a construction are considered ‘lexical constructions’. They can also constitute a subschema, given that their conventionalized form-meaning pairing differs from that of the more general schema (Hilpert 2014: 80). For example, *suitsetaja* ‘smoker’ or *laterdaja* ‘blabbermouth’ can receive an additional semantic component, ‘one who V-s’ + ‘habitually’, and therefore constitute a subschema of the more general schema only carrying the meaning ‘one who V-s’. The cross-linguistically attested tendency for derivation patterns to also realize semantically specialized derivatives, therefore, could be analyzed either as the hierarchical instantiation of individual subschemas (as with *suitsetaja*) or as extensions of the prototypical meaning of the construction through some metonymic or metaphorical links. For example, *nõudmine* ‘demand’ can either mean the act of demanding or the object/result of demanding. The latter would constitute a metonymic extension of the more general schema. *Liikumine* ‘movement’, in turn, can represent a metaphoric extension of the more prototypical meaning of the process of moving when it refers to a group of people with a common ideology or an organized action of such a group. The difference between the schematic and prototype extensions, of course, is not always clear as will be shown in Section 5.2. An important characteristic of morphological constructions is the fact that they are also heavily interlinked through subpart relations, forming paradigms around some common element (Hilpert 2014: 83), e.g., what is traditionally called the *base*. From the CxG viewpoint, however, grammar (including morphology) is not derivational. So while using the terminology of *base* and *derived nouns*, *derivatives*, or *deverbal nouns* implies that the grammatical status of derivatives is secondary to that of the verbs, in the constructionalist framework, all members in a paradigm belong to a network of constructions on different levels of abstraction (Bybee 1995, Diessel 2019). This assumption is also held in this thesis, despite the particular terminology used for referring to the suffixed words and the relevant morphological constructions. The bidirectional relationship between verbs and derivatives can also be seen in the fact that similarly to the possibility of deriving verbs from regular nouns (e.g., *kohv* ‘coffee’ → *kohvitama* ‘have coffee’), the use of a

novel derivative can also precede the use of its base in verb forms (e.g., *ööklubistamine* ‘turning something into a nightclub’ → *ööklubistama* ‘turn something into a nightclub’, Kasik 2015: 189). If the knowledge of paradigms such as the one containing the words *kohv* ‘coffee’, *kohvitama* ‘have coffee’, *kohvitaja* ‘coffee drinker’, *kohvitamine* ‘having coffee’ is entrenched enough, it would not pose a problem for a speaker to infer, upon hearing a word such as *ööklubistamine* ‘turning something into a nightclub’, that there also might be words such as *ööklubi* ‘night club’, *ööklubistama* ‘turn something into a nightclub’ or *ööklubistaja* ‘the one turning something into a nightclub’ (see Hilpert 2014: 84). It has been shown that paradigmatic relations (i.e., the support a word form gets from other words belonging to the same ‘family’) play a crucial role in characterizing lexical representations and determining lexical processing (Hay & Baayen 2005).

Finally, while all CxG theories claim that constructions form a network (Goldberg 2013), an actual holistic model of the grammar network is described in Diessel (2019). He proposes that a speaker’s knowledge of a linguistic sign (a morpheme, word, phrase, or larger structures – usually referred to as a construction in CxG<sup>8</sup>) involves symbolic, sequential, and taxonomical relations, which are responsible for retrieving the meaning of the sign, the automatization of recurrent sequences, and anchoring the sign to some level of abstraction, respectively. The signs in turn are interconnected in a higher-level network through lexical, constructional, and filler-slot relations (Diessel 2019: 11–22). Lexical relations connect lexemes with similar or contrastive forms and meanings, constructional relations link constructions on the same level of abstraction, and filler-slot relations are responsible for connecting lexemes or phrases with slots in constructional schemas (Diessel 2019: 13). On both levels of analysis, the interaction between the different relations determines how the expressions are represented in the network. The network of associations which gets activated with a linguistic element varies in different contexts and speech situations, accessing or creating meaning ‘on-the-fly’ rather than storing it in a fixed lexicon. On a morphological level, for example, the strength of sequential links within a complex word is determined by the degree to which upcoming elements (e.g., the suffix *-mine*) are automatized and predictable, while lexical links are established by categorization and analogy (Diessel 2019: 73). Through filler-slot relations, this complex word, in turn, becomes part of other constructions (e.g., a noun phrase, a converb, or a periphrastic construction), where the same processes apply. Sequential predictability, lexical categorization, and analogical extension become particularly important in assessing the productivity of different constructions, although their effect on productivity is inversely proportional.

---

<sup>8</sup> Diessel (2019) does not consider bound morphemes and monomorphemic lexemes as constructions in the same sense as grammatical patterns, which involve at least two meaningful elements (e.g., morphemes, words, phrases), because he claims that simple lexemes and constructions are learned and processed in very different ways. However, he does treat both notions as linguistic signs.

For the purposes of this study, the described constructional network has to be able to account for at least

- the prototypical meanings of *-mine* nouns,
- the ‘specialized’ meanings of *-mine* nouns,
- the rivalry and complementarity of different deverbal nouns,
- the syntactic behavior of *-mine* nouns according to their word-class properties,
- the syntactic behavior of *-mine* nouns in periphrastic constructions and catenative constructions,
- the syntactic behavior of *-mine* nouns in converb constructions.

I will try to show that with a little goodwill, this can be done. In addition, I will argue that the productivity of suffixes could be better assessed in the context of constructional schemas and that a speaker’s knowledge of the communicative settings in which language is produced (the register) has an effect on which schemas and associations become more activated.

## 2.2 Productivity

### 2.2.1 Aspects of productivity

The notion of linguistic productivity is a multi-faceted complex theoretical construction which is not easily defined, although many attempts have been made (see Bauer 2001: 11–32, Bauer 2005: 315–334, Barðdal 2008: 9–54). It is generally used to distinguish constructions with more or less fixed membership from those which are easily extendable to subsume new lexical items. However, most scholars appear to accept that productivity is in nature a scalar concept instead of a categorical one (Bauer 2001: 125 ff.). Productivity has been investigated primarily in the context of inflectional and derivational morphology (for compounding, see e.g., Shen & Baayen 2020), while studies which set their focus explicitly on syntactic productivity have been less common (cf. Barðdal 2008, Zeldes 2012). The notion is still – even if inadvertently – highly relevant in studies about alternation in idiomatic constructions, argument structure constructions, etc., where the construction’s attractiveness to new or existing lexical items is investigated. In this study, morphological productivity of the *-mine* construction is explicitly assessed and measured in different registers, while the syntactic productivity of the constructions in which *-mine* nouns occur is also addressed, albeit less directly and rather in a qualitative manner.

Barðdal (2008) summarizes the three characteristics associated with productivity in the literature: generality, regularity, and extensibility. Generality refers to the structural and semantic openness of the construction (i.e., the ability to be instantiated by a wide range of lexical items or other constructions); regularity reflects the construction’s semantic transparency (i.e., the ability to be instantiated by items which share a large part of the construction’s core meaning); extensibility, in turn, gauges the likelihood with which the construction attracts new items or functions. These three aspects are related through links of entailment

and co-occurrence, which is one of the reasons why the concept of productivity is so complex in linguistics (Barðdal 2008: 24).

One aspect of productivity often emerging from the literature is the fact that it is defined as pertaining only to unintentional, spontaneous extension of constructions, without requiring conscious and deliberate creativity (e.g., Schultink 1961, Bauer 2001: 57–58). From this perspective, isolated instances of individual creative coinages of new words, for example, do not indicate productivity and should not be examined as novel formations in empirical studies of productivity (Bauer 2001: 62–71). This pronounced distinction between *creativity* and *productivity* seems to be motivated mostly by cases where the meaning of the novel word lacks predictability and where new instances are coined by analogy rather than by merit of productivity. Essentially, this boils down to the question of whether productivity belongs to competence or performance. Characterizing linguistic creativity as a rule-changing process as opposed to a rule-based one leaves creativity outside the scope of competence. However, in addition to the irrelevance of competence vs. performance to the usage-based theories, Barðdal (2008: 91) argues that the difference between full productivity and low-level analogical extension is a matter of quantity, not quality, since there is no reason to assume an ontological difference between the extension of constructions with different degrees of abstraction. In this line of reasoning, a construction which is instantiated only by few items, but these items are semantically highly similar, can also be considered productive. From a more practical perspective, drawing a line between spontaneously and intentionally coined instances of constructions would prove impossible in empirical studies, especially when relying only on corpus data. This difficulty of operationalizing unintentionality is also acknowledged by Bauer (2001: 68). Hence, from a usage-based and constructionalist perspective, all low-level instantiations of a higher-level construction are equally valuable in the empirical analysis of productivity.

The concept of gradient or scalar productivity makes it possible to compare the productivity of different categories. Classical tools which are used for explaining diverging degrees of productivity are phonological, morphological, syntactic, semantic, lexical, and pragmatic restrictions or constraints on base words (Bauer 2001: 126–143). For example, the suffix *-mine* in Estonian as a borderline case between derivation and inflection can be considered highly productive since it is possible to form a *-mine* noun from any verb with no phonological or morphological constraints. From the perspective of deriving nouns from verbs, some lexical or syntactic constraints have been mentioned (Erelt et al. 1993) for bases with non-canonical argument marking (e.g., *mulle piisab heast tervisest* I:ALL be\_enough:3SG good:ELA health:ELA ‘good health is enough for me’), making the actual occurrence of lexical constructions such as *piisamine* impossible or at least highly marked. With regard to semantic restrictions, modal verbs such as *saama* ‘can’, *pidama* ‘must; have to’, *võima* ‘can; may’, *näima* ‘seem’, *tunduma* ‘seem’, *paistma* ‘seem’ have been considered marginal among

the bases for *-mine* nouns (Kasik 1975: 33, Kasik 2015: 267)<sup>9</sup>, but no other semantic restrictions exist. Compared to the *-mine* construction, another deverbal nominalization suffix, *-us*, has restrictions on all levels, despite covering overlapping semantic domains with *-mine* (see P2 for examples), and is therefore regarded as considerably less productive. Still, even these somewhat soft lexical restrictions for *-mine* might not apply in all constructions in which *-mine* nouns occur. The acceptability of actually occurring sentences<sup>10</sup> such as in examples 14, 15, and 16 should be tested and compared to that of more conventionalized *-mine* forms.

(14) (Lääne Elu, 2019)

<i>Lisaks</i>	<i>vanade</i>	<i>elementide</i>	<i>korrastamisele</i>	<i>võiks</i>		
addition:TR	old:PL.GEN	element:PL.GEN	fix:NMLZ:ALL	could:CND		
<i>vahendite</i>	<i>piisamisel</i>		<i>juurde</i>	<i>ehitada</i>	<i>ka</i>	<i>mõne</i>
fund:PL.GEN	be_sufficient:NMLZ:ADE	more	build:1INF	also	some.GEN	
<i>uue</i>	<i>ja</i>	<i>huvitava</i>	<i>elemendi,</i>	<i>mida</i>	<i>staadioni</i>	
new.GEN	and	interesting.GEN	element.GEN	that:PAR	stadium.GEN	
<i>skate-pargis</i>	<i>pole.</i>					
skate+park:INE	be.NEG.CNG					

‘In addition to fixing the old elements, some new and interesting elements which are not available in the stadium skate-park could be built if the funds are sufficient.’

(15) (Postimees, 2009)

<i>Teine</i>	<i>on</i>	<i>aga</i>	<i>ühiskondlik</i>	<i>lepe,</i>	<i>mis</i>	<i>ütleb,</i>	<i>et</i>	<i>sa</i>
other	be.3SG	PCL	social	contract	what	say:3SG	that	you
<i>võid</i>	<i>tahta</i>	<i>küll,</i>	<i>aga</i>	<i>sa</i>	<i>pead</i>	<i>kontrollima</i>	<i>ka</i>	<i>sedä,</i>
may:2SG	want	PCL	but	you	must:2SG	control:2INF	also	that:PAR
<i>mida</i>	<i>sa</i>	<i>võid,</i>	<i>mis</i>	<i>on</i>	<i>sinu</i>	<i>võimise</i>	<i>piir.</i>	
what:PAR	you	can:2SG	what	be.3SG	you.GEN	can:NMLZ:GEN	limit	

‘The other [feeling] is a social contract which states that you may well want to, but you also have to consider what you actually can, what are the limits of your abilities.’

(16) (foorum.vvklubi.ee, 2019)

<i>Suurim</i>	<i>teepeale</i>	<i>jääda</i>	<i>võimist</i>	<i>põhjustav</i>	<i>tõrge</i>
greatest	road.GEN+on	stay:1INF	may:NMLZ:PAR	cause:APRP	failure
<i>on</i>	<i>siiani</i>	<i>olnud</i>	<i>geneka</i>	<i>tuksimine</i>	
be.3SG	thus_far	be:APP	generator.GEN	malfunction	

‘The greatest failure causing [us] to potentially stay in the middle of the road has so far been the malfunction of the generator.’

<sup>9</sup> This does not mean that the same verbs in their non-modal meaning (e.g., *saama* ‘get’, *pidama* ‘keep’) cannot be nominalized.

<sup>10</sup> Examples 14, 15 and 16 are obtained from the Estonian National Corpus 2019 via Sketch Engine. Specific sources (newspapers, journals or websites) are presented in the brackets above the examples.

While the first example actually seems quite natural<sup>11</sup>, the second one might be classified under the notion of linguistic creativity, and the third example might come off as a bit clumsy, but it is difficult to see why the use of any of these forms should not contribute to and rely on the productivity of the more general *-mine* construction, even if indirectly. In the same manner, the acceptability of otherwise perfectly ‘normal’ *-mine* nouns can and will vary in different constructions. Such examples clearly advocate and call for a non-transformational, constructional account of derived nouns.

Pragmatic constraints on productivity are associated with the use of constructions in specific contexts and can apply both to all elements of schematic constructions (e.g., the bases and the suffixes) as well as to the lexical constructions (Bauer 2001, Plag 2018). This restriction plays an important role in comparing data from different registers because while there is no reason to assume that the generalized phonological, morphological, or semantic constraints would differ much under various communicative settings, the need for specific concepts and constructions is definitely different. The pragmatic constraints as I see them in this thesis seem to be associated with what Bauer (2001) refers to as *profitability*: the actual utilization of a morphological process to meet the naming needs of a particular speech community.

Finally, Bauer (2001: 136–138) also lists *blocking* as a potential constraint on productivity. This concerns mainly competing affixes (e.g., *-mine* and *-us*), where the existence of one form may prevent the use of the other (either due to synonymy or homonymy) (Aronoff 1976: 43). Blocking does not always work systematically and often seems to be determined by individual cases (‘type blocking’). Suffixal rivalry between *-mine* and *-us* is in this spirit also briefly examined in P2, although since it is impossible to claim the nonexistence of some form simply because it does not occur in the corpus samples, blocking is interpreted more leniently and rivalry is assessed by comparing relative frequencies between derivatives formed from the same base.

Processing and producing linguistic constructions thus entails online resolution of the various (and possibly conflicting) constraints in an optimal way by combining cognitive processing skills, entrenched memory traces, pragmatic expectations, and paradigmatic relations with the generalized structural and semantic requirements of the linguistic input (Pirrelli et al. 2020: 3). The quantitative output of the complex interplay between all of the potential restrictions, in turn, can be measured in usage, as suggested by e.g., Aronoff (1976, 1983), Baayen (1989), Baayen & Lieber (1991), and many others after them.

---

<sup>11</sup> From a transformational perspective, such examples could be an indication of a change in the verb’s argument structure, where the partitive subject (*vahendeid piisab* fund:PL:PAR be\_sufficient:3SG ‘the funds are sufficient’) is replaced with a canonical subject in the nominative (*vahendid piisavad* fund:PL:NOM be\_sufficient:3PL ‘the funds suffice’).

## 2.2.2 Measuring productivity

One of the first to offer a quantitatively oriented definition of productivity was Bolinger (1948: 18) who defined productivity as ‘the statistically determinable readiness with which an element enters into new combinations’. This definition describes productivity as a probabilistic phenomenon and this is how the concept is mostly also used in this thesis, although from a usage-based perspective, perhaps a more speaker-oriented definition would be more appropriate (see Hilpert 2014: 79) as the probability of a construction being extended is associated with how speakers produce and process those constructions.

Three basic notions of frequency have to be distinguished in the context of measuring productivity: type frequency, token frequency, and the frequency of words occurring only once (known as *hapax legomena*). Here, *type* is a class of objects (e.g., a lemma or a schematic construction), *token* is an instantiation of a type (e.g., a word form or a specific expression), and *hapax legomenon* is a type occurring only once in a corpus. In addition to providing the basis for quantification, these notions are situated against a larger cognitive background and closely linked to the acquisition of new structures.

**Type frequency** is the number of different instantiations of a construction found in a corpus and in the context of morphology, type frequency therefore approximates ‘the size of the vocabulary’ (Cowie & Dalton-Puffer 2002: 416). Productive constructions are usually said to be instantiated by many types (e.g., more different words), while only a few types occur for unproductive constructions. In the context of constructions consisting of two or more slots for individual lexemes, the type frequency of a construction is the number of the different lexically specified configurations. Type frequency is considered central to schema extraction (Diessel 2019: 16) and extension (Barðdal 2008) and it is a fairly intuitive measure of linguistic productivity, but it is an imperfect one, as vigorously pointed out in literature (e.g., Bauer 2001: 144, Baayen 1993, Keune et al. 2006, Säily 2011). Type counts alone are only useful for assessing past productivity, since a construction may have become unproductive over time and while it may exhibit a high number of types, it may no longer be able to form new ones (e.g., the English *-ment*). Another widely cited point is Aronoff’s (1976: 36) criticism that comparing type frequencies of different word-formation processes ‘isn’t fair’ since the number of types a pattern can be instantiated by depends on the number of available bases.

**Token frequency** refers to the total number of all (including repeated) instantiations of a construction found in a corpus. While the term *token* is more widely used in studies about complex words, it is easily also extended to the study of larger constructions (see Barðdal 2008). On the one hand, token frequency can be interpreted as the frequency with which instantiations of a particular category (e.g., *-mine* construction) occur in a given corpus. In that respect, token frequency is also responsible for entrenchment of grammatical structures: the more a speaker is confronted with an expression, the more the sequential relations become automatized and processed as a construction (Diessel 2019, Lemmens 2019).

However, the token count of a schematic construction is often inflated by few very common types and therefore is by itself not a good predictor of productivity. On the other hand, token frequency also matters for individual instantiations of more schematic constructions: the higher the token frequency for a particular type, the more closely the meaning of that lexical construction becomes associated with the meaning of the whole construction and the less need there is for the support from the more abstract schema in processing and production. Token frequency of individual types therefore has a strengthening effect on lexical information (Diessel 2019: 16), but the contribution of high-frequency types to the schematization of the more abstract construction is low. Types with high token frequency have also been associated with higher degrees of lexicalization and semantic specialization (Bauer 2001: 147).

Finally, in corpus-based studies of linguistic productivity, specific instantiations of a construction occurring only once in a corpus are thought to best demonstrate the productivity of a schematic construction because they are most likely to include neologisms (Baayen & Renouf 1996). Consequently, **the number of hapax legomena** is taken as an indication of the construction's extendability. In order for this number to be informative, however, it has to be anchored to some other frequency (e.g., to the total number of tokens or to the number of *hapax legomena* in the whole corpus). The notion of a neologism, of course, is somewhat problematic here because not all words occurring once are necessarily neologisms and not all neologisms occur only once. This issue will be returned to in more length in Section 5.5.

In order to anchor the type, token, and *hapax legomenon* counts and to illustrate different aspects of a construction's productivity, several corpus-based measures were introduced by Baayen and his colleagues in the 1990s (e.g., Baayen & Lieber 1991, Baayen 1992, Baayen 1993, Baayen & Renouf 1996) and later used in a wide range of studies on morphological productivity. The three central measures also used in this thesis are *realized productivity*, *potential productivity*, and *expanding productivity*.<sup>12</sup> These measures of productivity have widely been used in studies of morphological productivity, but they could, in principle, also be extended to analyze larger syntactic units (e.g., Zeldes 2012), especially in cases where the grammatical construction's potential to attract new lexical fillers is assessed in a lexically partially specified constructional idiom. However, since the main focus of corpus-based analyses of syntactic and idiomatic constructions seems to be more on the need to establish new, independent constructions instead of comparing the productivity of different constructions or suggesting that the construction under investigation is indeed simply the sum of its parts, collocation methods (e.g., Stefanowitsch 2013) seem to be more popular within that line of research. In any case, what these productivity measures collectively capture is essentially the degree of constructional schematicity, generality, and pragmatic usefulness of a construction.

---

<sup>12</sup> Of course, these three measures do not exhaust by far all the possibilities for measuring productivity. For a thorough overview, see e.g., Zeldes (2012: 48–95).

The **realized productivity** ( $V(C,N)$ ) of a construction ( $C$ ) corresponds to a simple type count ( $V$ ) in a given corpus with  $N$  tokens<sup>13</sup>. Baayen (1993) has also called this measure the *extent of use* of a particular construction. As mentioned, this is a fairly intuitive measure and also corresponds well to observations made in the literature about the *-mine* derivation being very productive in formal written Estonian (Kerge 2003, Kasik 2006). A construction with high realized productivity (i.e., with many types) is likely to have very few restrictions and it is likely pragmatically useful for expressing certain concepts of states-of-affairs. However, it captures only past productivity and is insufficient for measuring how well a construction can be extended to be instantiated by new items. As an example, the English nominalization suffix *-ment* is often cited as a suffix which is frequently attested in different words but not used in the coining of new forms (e.g., Aronoff 1983, Bauer 2001).

The central measure in studies of (morphological) productivity is **potential productivity** ( $P$ ), which relates the number of *hapax legomena* in a given category ( $V(I,C,N)$ ) to the total number of tokens in that category ( $N(C)$ ) (17). The measure has also been called the *category-conditioned degree of productivity* (Baayen 2009: 902). The higher the potential productivity, the higher the proportion of *hapax legomena* within that category, and – when translated into terms of probability – the more likely we are to encounter a new, previously unseen type with the next token added to that category. Potential productivity therefore reflects the speed with which language users extend the construction with novel instantiations. The more restrictions a construction has, the less frequently it can be instantiated by a type not seen before. Since types with high token frequency are more likely lexicalized or semantically specialized, the proportion of such types within the category is also negatively correlated with the category's potential productivity (Baayen 1992: 117).

$$(17) P = V(I,C,N)/N(C)$$

The third measure utilized in this thesis is **expanding productivity** ( $P^*$ ), which relates the number of *hapax legomena* in a given category ( $V(I,C,N)$ ) to the total number of *hapax legomena* in the corpus ( $V(I,N)$ ) (18). When comparing the productivity of two constructions from the same corpus, the differences between their respective  $P^*$  values therefore correspond to the differences between the absolute number of *hapax legomena*. Expanding productivity is a completely *hapax*-based measure and has as such also been called the *hapax-conditioned degree of productivity* (Baayen 2009). It is a rough indication of a category's attractiveness for new items. In other words, it expresses the probability that this particular construction will be chosen to express the next novel token added to the corpus (Baayen 1993: 192–193). Expanding productivity complements realized

---

<sup>13</sup> For reasons I cannot recall, I decided to denote this productivity measure with the symbol  $C$  in P1. The conventional annotation in Baayen (1992, 1993, 2009), however, seems significantly more motivated and should be preferred.

productivity in that it (albeit indirectly) allows one to estimate the future usefulness of a construction.

$$(18) P^* = V(I, C, N) / V(I, N)$$

These measures are not without their faults and cannot be taken as absolutes (cf. Plag 1999: 28–29, Gaeta & Ricca 2006, Zeldes 2012). The main critique concerns the measure  $P$  when comparing multiple affixes; this originates from the fact that  $P$  is a decreasing function of token count (Gaeta & Ricca 2006: 62). This means that when sequentially scanning through the corpus, the more instantiations of a given construction we encounter, the smaller the likelihood of the next token representing a new type never seen before. This puts two affixes with very different token counts in a given corpus in an unequal position when their extendability is assessed at the endpoint of this scanning because the more frequent construction has ‘used up’ much more potential types than the less frequent one. Gaeta & Ricca (2003, 2006) propose an improvement of Baayen’s approach – the variable-corpus approach – where  $P$  is estimated at an equal number of tokens for all affixes compared. This entails splitting the corpus into multiple chunks of progressively increasing size and estimating the productivity of the more frequent suffixes not at their total token count in a given corpus but at the token count which corresponds to the total token count of the least frequent affix. This means that suffixes get estimated in differently sized corpus samples. Plag et al. (1999) have also used a method that splits the corpus into different subcorpora, but their method was based on averaging the (estimated) productivity values in the different subcorpora. In P2 and in Section 5.2, I will suggest however that these two approaches do not offer significant advantages over Baayen’s original approach.

## 2.3 Register variation

Before turning to the data and methods section of this thesis, the central notion of register has to be defined according to the way in which it is used in this study. I will also briefly explain the motivation for considering the analysis of different registers in the first place.

Regarding the terminology, there is a lot of confusion among using the terms ‘genre’, ‘text type’ and ‘register’, partly because their definitions often overlap (see the relevant overview and discussion in e.g., Melissourgou & Frantzi 2017). In this thesis, I follow Biber (2010) and use the term *register* to refer to those text varieties which are differentiated by the different functions which core linguistic features (or constructions) serve. Therefore, I assume that the use of complex words, similarly to other parts of grammar, simultaneously contributes to and is affected by the nature of the types of texts they occur in, due to distinctions in the topic, the speakers’ communicative purposes, the production circumstances, and the relationship between the speaker/writer and the hearer/reader. The differences in functions also have consequences for productivity and the assessment of

naturally occurring data. In addition to a wide range of linguistic phenomena, it has been shown that the productivity of word-formation patterns also varies across registers (Plag et al. 1999) and differences in the results reflect the slightly different conventions and communicative needs of text varieties. Morphological productivity has also been shown to have the potential to differentiate text types or registers (see Baayen 1994, who uses principal component analysis in order to categorize different types of texts based on morphological productivity).

For Estonian, it has indeed been shown that the different functions in which *-mine* derivatives have been used correlate with registers, although functions are understood here mainly from the perspective of text linguistics and refer to the more general functions of word-formation. For example, in legal texts and journalistic texts, *-mine* nouns are said to be constructed mostly in the syntactic function as well as for making the text denser (Kerge 2002, 2003, Kasik 2006) or for creating abstract terms (Kerge 2002); academic and scientific texts, in turn, mainly operationalize the *-mine* derivation with the lexical function, in other words for creating specific terminology (Kerge 2002). Kerge (2003) has also shown that the functions of both the derivation patterns as well as individual words within one register have changed over time. These functions (lexical vs. transpositional) for *-mine* nouns are also considered in P1 and somewhat in P2, although only from a synchronic perspective. The functions of complex words are associated with certain constructions (e.g., syntactic functions), which in turn can be captured from corpora via some formal properties such as the different use of case constructions. This will be illustrated in Section 5.4. Thus far, however, only written registers have been analyzed in more detail, eluding the discussion of the very different natures of written and spoken modes of communication and their effects on the functions and productivity of complex words.

One of the main goals of this thesis is to include observations about spoken data in the discussion of *-mine* nouns. While spoken dialects receive the most attention throughout this thesis because they constitute the most variable register among those investigated, in the first two case studies, the use of *-mine* nouns in dialects is compared to spoken common language (or *colloquial Estonian*, see Keevallik 2003) and three written registers. In the context of complex words, I hypothesized that the use of *-mine* nouns (and nouns with suffixes *-us* and *-ja*) in spoken registers is more ‘regular’ than in the written registers, meaning that they occur in contexts which more often evoke the prototypical, schematic *-mine* construction, paradigmatically closely linked to the corresponding finite clause constructions (the transpositional function). Consequently, the construction should exhibit higher potential and expanding productivity due to the lower proportion of semantically specialized subconstructions and metonymic extensions. Differences in productivity should also reflect differences in the pragmatic usefulness of the constructions in different functions. I also hypothesized that the syntactic constructions in which *-mine* nouns occur would differ between spoken and written registers (e.g., the Adessive Manner and Cause Construction (Sahkai 2011) is expected to occur less often in spoken data). Although the productivity and use of the syntactic constructions in different registers are not directly

measured in this thesis, the constructional differences are assessed roughly based on the case distribution of the *-mine* nouns and on the initial analysis of a smaller, contextually anchored sample of 1000 sentences per register.

Due to my strong desire to avoid complicated, highly fragmented, and unbalanced corpus structures with already small samples, I take the registers analyzed in this thesis as predefined by the corpora used (described in the next section). However, a more detailed analysis of text types could certainly provide many more insights into the results obtained in the case studies and remains an undertaking for further research.

### 3 DATA AND METHODS

This thesis makes use of corpus data as the primary data source and therefore, focuses on inference from past production rather than direct observations of online processing. In the case studies, I rely largely on corpus-based frequency analysis, despite having to conclude at times that for rare linguistic events such as syntactic action noun constructions, the size of the corpus (sample) sets limits to the possibility of postulating well-grounded generalizations based on the individual usage events.

With regard to studying constructional productivity and constructional alternations, corpus data proves extremely useful because it provides the researcher with counts and distributions of different usage events instantiating a schematic construction. In addition, relative corpus frequencies have been shown to correlate well with the cognitive ease with which a construction is semantically decomposable (Hay 2002, Hay & Baayen 2005, Hay & Plag 2004), and they provide means for ranking different suffixes along the productivity cline (e.g., Plag et al. 1999). Naturally, using corpus data also suffers from known limitations, such as lack of control over the production, the non-representativeness of the data, and the non-independence of data points. These limitations and their potential effects on the results of the case studies are discussed and elaborated on in Section 5.5.

I extracted naturalistic language data from three corpora of Estonian: the Balanced Corpus of Estonian (BCE), the Phonetic Corpus of Estonian Spontaneous Speech (PCESS), and the Corpus of Estonian Dialects (CED). I used both qualitative and quantitative methods in order to answer the research questions presented in the introduction. The corpora represent the five registers analyzed in the case studies: scientific and academic writing (SCI), journalistic texts (NEWS), fiction (FICT), spoken spontaneous common language (SP), and spoken regional dialects (DIA). In the next two subsections, I give an overview of the corpora used, the process of data collection, and the methods used in the analysis.

#### 3.1 The corpora

The Balanced Corpus of Estonian<sup>14</sup> (BCE) contains approximately 15 million tokens which represent fiction, newspapers, and scientific texts. The fiction subcorpus comprises 5 million tokens in excerpts from Estonian literature from the period 1987–2011. The journalistic subcorpus contains 5 million tokens from texts published in the newspapers *Eesti Ekspress*, *Maaleht*, *Postimees*, *Päevaleht*, and *SL Õhtuleht* in 1995–2007. The first two are published weekly, the latter three are daily newspapers. The 5 million tokens of the scientific subcorpus are divided equally between PhD dissertations from various fields and scientific journals

---

<sup>14</sup> <https://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=en>  
(Accessed on 23/02/2021)

(e.g., Eesti Arst, Arvutitehnika ja Andmetöötlus, etc.) published in 1995–2006. While the files containing dissertations and fiction are authored by individual writers, the files in newspapers and journals contain many individual articles with multiple authors. For this study, the morphologically annotated version of the BCE was used.

The Phonetic Corpus of Estonian Spontaneous Speech<sup>15</sup> (PCESS) comprises around 96 hours of high-quality recordings of unedited, spontaneous speech, both monologues as well as dialogues. The recordings were done between 2006 and 2016; most were recorded in a studio, but some were collected during fieldwork. Each transcribed file corresponds to a single speaker's part in a given session, each of which is approximately 30 minutes long. Some speakers were recorded multiple times. Transcriptions include automatic morphological annotation which was done using Filosoft's Vabamorf<sup>16</sup> tagger trained on written Estonian. The version of the corpus used in P1 contained around 339,000 tokens, and the one used in P2 around 426,500 tokens (both counts exclude fillers (e.g., *mm*, *ee*) since these are not morphologically tagged).

The Corpus of Estonian Dialects<sup>17</sup> (CED) represents areally varying spoken language from a considerably earlier time period than the previous two corpora. The recordings from all 10 traditional dialect areas were mostly made in the 1960s and 1970s, but range in total from 1938 to 2010. The recordings are unstructured interviews with mostly elderly local people and contain long monologue passages about informants' everyday life, past events, customs, and work. At the time of data collection (January 2015), the corpus contained about 890,000 morphologically annotated tokens. For P3 and P4, additional files without morphological annotations were analyzed, yielding a total of ~950,000 tokens. The differences between the 10 dialects can be quite significant, which could make it difficult to justify grouping them as one register in P1 and P2. Compared to other corpora, the dialect data is definitely lexically more diverse which should be kept in mind when interpreting the number of *hapax legomena* in quantitative productivity measures. However, the analyses in P3 and P4 did not reveal any significant differences between the dialects regarding the ways in which *-mine* action nouns were used in the corpus. Furthermore, in the productivity analyses, standardized lemmas instead of text tokens in special transcription were used for type and token counts, which slightly alleviates at least the effect of morphophonological differences.

In addition to covering a diverse range of language functions, these corpora were chosen for analysis due to the fact that they were morphologically analyzed and enabled queries targeting the suffix of the lemma and part-of-speech tag, thus making it possible to extract all the case forms of *-mine*, *-us*, and *-ja* nouns at once. This also enabled automatic elimination of many of the cases which merely

---

<sup>15</sup> <https://www.keel.ut.ee/en/languages-resources/languages-resources/phonetic-corpus-estonian-spontaneous-speech> (Accessed on 23/02/2021)

<sup>16</sup> <https://github.com/Filosoft/vabamorf> (Accessed on 23/02/2021)

<sup>17</sup> <https://www.keel.ut.ee/et/keelekogud/murdekorporus> (Accessed on 23/02/2021)

contained the character sequences *mine, us* or *ja* (*minema* ‘go’, *tummine* ‘thick’, *kuus* ‘six’, *jalg* ‘leg’) or which were formed with the unproductive relational adjective suffix *-mine* (*alumine* ‘bottom; lower’, *äärmine* ‘outermost’). The use of derivatives as modifiers in compound phrases (e.g., *käitumis+mustrid* ‘behavioral patterns’) was disregarded. The corpora were also available as source files, which made it possible to specify both the query and the desired output structure. While the largest and the most comprehensive corpus for written Estonian today is the Estonian National Corpus 2019 (Kallas & Koppel 2020), which also contains genres not included in this study (such as the new media or legal texts), using the BCE as a source for written Estonian allows for a more controlled comparison of the productivity of the deverbal *-mine* suffix in different registers of written language, especially for making comparisons with the two smaller spoken corpora.

Needless to say, the text varieties in the corpora do not uniformly define the five registers under investigation. Consider the obvious differences between sports commentaries and interviews, the terminology used in popular science and the specific jargon of a scientific field, the aesthetic choices made by fiction writers, etc. As these are rather matters of genres and text types (see Biber 2010) and would require a separate detailed investigation, the aspects of internal variation within the registers are ignored in this study. Among the registers, however, at least four global dimensions of variation can be identified as presented in Table 1.

**Table 1.** Dimensions of variation in five registers of Estonian. SCI – scientific texts, NEWS – newspaper texts, FICT – fiction, SP – contemporary spontaneous speech, DIA – regional spoken dialects (Pilvik 2021)

	SCI	NEWS	FICT	SP	DIA
Written (vs. spoken)	+	+	+	–	–
Edited (vs. spontaneous)	+	+	+	–	–
Formal (vs. informal)	+	+	+/-	–	–
Common (vs. local)	+	+	+/-	+	–

### 3.2 Data extraction and analysis methods

The articles in this thesis combine qualitative and quantitative methods. While three of the four articles simply rely on absolute and relative frequencies to describe the distribution, functions, and productivity of *-mine* nouns, P2 also utilizes inferential statistics when modeling the productivity of different suffixes. As a large part of data extraction in P1 and P2 is immediately relevant to data analysis, these aspects are described together in this subsection.

For case studies in P3 and P4, all clauses containing deverbal *-mine* nouns and their immediate contexts were extracted from the ~950,000 tokens in the CED based on lemma and part-of-speech tags. The dataset was manually cleaned and analyzed first and foremost with regard to the syntactic properties of the *-mine* nouns to explore the more ‘nouny’ and ‘verby’ properties of the action nouns (P3)

as well as to detect the instantiations of the *olema* ‘be’ + *Vmine*<sub>NOM/PAR</sub> periphrastic construction (P4). In order to adequately describe the distribution of *-mine* forms across all 10 dialect areas, the proportion of *-mine* nouns among all tokens from a dialect was used instead of the absolute frequency.

For the case studies in P1 and P2, which examined the productivity of the *-mine* construction in the five registers, extracting comparable data first required taking equally sized samples from the corresponding (sub)corpora, since any type- and *hapax*-based measures are very sensitive to the size of the text or corpus (Hardie & McEnery 2006: 139). The longer the text, the more types have already been seen and the lower the likelihood of encountering a new, previously unseen type. In both case studies (P1 and P2), the samples were extracted with reference to the size of the smallest corpus, the PCESS. The sampling included complete files and stopped when the total token count in a sample reached more than 335,000 tokens in P1 and 426,000 tokens in P2 (this condition was checked prior to sampling another file). The samples for SP included all files available in the PCESS at the times of data collection. As the sampling process for other corpora was random, studies in P1 and P2 were conducted on different corpus samples. This enables one to compare and validate the results obtained in the two case studies for the suffix *-mine*.

In P1, which described only the productivity of *-mine* in the five registers, all occurrences of noun lemmas ending with *-mine* were extracted from the samples and the lists were manually cleaned of foreign words (e.g., *examine*) and false hits. Manual checking of the corpus data is crucial for using *hapax legomena* (Gaeta & Ricca 2015: 847) since the group of words occurring only once in the corpus has the highest likelihood of containing typos and other idiosyncrasies. In order to make the data from different corpora really comparable, only simple lemmas of derivatives were used in the two case studies. This is a practical consequence of the different annotation principles and practices in the corpora. The main issue here concerns the marking of compounds and phrasal verbs. In the dialect corpus, the elements of compound words are usually separated with a ‘+’ (e.g., *ära+viimine* ‘taking away’, *asja+ajamine* ‘conducting business, handling matters’), whereas in the BCE and in the PCESS, ‘\_’ is used (*ära\_viimine*, *asja\_ajamine*). However, this annotation is not systematic and relies firstly on human transcribers in CED and PCESS and secondly on human annotators in the CED and on the automatic morphological analyzer’s ability to recognize complex words in the BCE and PCESS. Therefore, expressions such as *ära viimine* can also occur. Furthermore, phrasal verbs (e.g., *asju ajama* ‘conduct business; handle matters’) and particle verbs (e.g., *ära viima* ‘take away’) are not marked as compounds in the corpora, making it impossible to reliably compare the relative frequencies of derivatives and verb forms sharing the same base in order to detect formations which are less likely to be semantically decomposed in lexical processing (see Aedmaa 2019 for a comparison of automatic methods for detecting Estonian particle verbs). For this reason, all compound words were stripped to their final component and the lists were manually corrected for *-mine* nouns. The type, token, and *hapax legomenon* counts were then used to calculate the values

for realized, potential, and expanding productivity in all five registers at their maximal sample size (around 335,000 tokens). Since different registers were compared with slightly different sample sizes ranging from 335,325 tokens in DIA to 398,796 tokens in FICT, the realized productivity (the number of different types in the sample) was normalized to the sample size.

Since in P1 only one single suffix was analyzed, the issues regarding productivity measures (especially potential productivity  $P$ ) described in Section 2.2.2 were not relevant. However, for comparing multiple affixes in P2, these issues had to be addressed. This meant taking further steps in data extraction and preparation as well. For a more rational use of computational resources, first lists of all tokens, types, and *hapax legomena*, all verb lemmas as well as individual lists of tokens, types, and *hapax legomena* formed with the suffixes *-mine*, *-us*, and *-ja* were compiled from each file in the samples. The verb lists and the suffix lists were cleansed of false hits (spelling errors, false analyses, foreign words) and unwanted derivatives (derivatives from other word classes and non-finite verb stems, e.g., *tule-m-us* ‘result’). Next, in order to take into account the vocabulary growth curves, the data from all files in each register was split into 21 chunks of progressively increasing size, adapting the procedure described in Gaeta & Ricca (2006). Each following subcorpus therefore included the previous subcorpus as well as an additional 21,300 tokens, until the 21<sup>st</sup> subcorpus chunk contained the whole sample (~426,000 tokens). This also meant that data from longer texts got split into several consecutive chunks. The three productivity measures were calculated for each chunk (e.g., at 21,300 tokens, at 42,600 tokens, at 63,900 tokens, etc.). This meant that instead of a single value for each productivity measure as in P1, I produced productivity curves with 21 nodes. However, since the samples (and smaller chunks) consist of unrelated texts of differing lengths and do not constitute one continuous discourse, choosing one specific ordering of the sample files creates a structural and topical bias towards that specific configuration of textual sequence. In order to alleviate this effect, the above-mentioned ‘chunking’ procedure was repeated on 100 random permutations of the sample files and the final productivity measures along with the 95% confidence intervals were obtained by averaging over those 100 permutations. Doing this helps to relieve the effect of discourse structure imposed by the specific ordering of the files on specific chunks, while at the same time not assuming that words appear randomly and independently in texts (cf. Baayen 1996). With the actual productivity curves where potential productivity  $P$  was also anchored in the average number of tokens included in a particular chunk/subcorpus, it became possible to compare Baayen’s original approach for assessing  $P$  for different suffixes at the endpoint of the sampling process with the variable-corpus approach used in Gaeta & Ricca (2006) and the averaging approach in Plag et al. (1999). While Gaeta & Ricca’s method for obtaining productivity values was power regression, which enabled them to predict the productivity value for a suffix at some fixed number of suffix tokens, I used generalized additive models (GAM, Wood 2017) instead as they are better suited for modeling nonlinear, non-monotonic relationships with less risk of overfitting. Given that we can calculate

e.g., 21 productivity values for a suffix X, based on the number of suffix tokens and the number of *hapax legomena* among them in each of the 21 chunks/subcorpora, we can then use statistical models to estimate that suffix's productivity at any given number of suffix tokens in the range of the 21 observed values. That given number allows for a comparison of multiple suffixes at a common number of suffix tokens.

In addition to comparing the different aspects of productivity, both P1 and P2 explored the relationship between base and derivative frequencies to see whether this provides a systematic way for detecting a construction's degree of decompositionality. In P1, I used a simple relative score (the 'NV-score') which subtracted the relative frequency of a base among all verbs from its relative frequency among all *-mine* derivatives. When this score was  $> 0$ , the base was more attracted to the nominalized structure than expected by its occurrence in verb forms; when the score was  $< 0$ , it was less attracted to the nominalized structure. Any score close to 0 was interpreted as the derivatives with the corresponding base expressing a fairly regular action noun schema. The measure was based on absolute frequencies and highlighted only the extremes among the lexical constructions. In P2, relative log frequencies of bases and derivatives were compared instead of absolute frequencies because it has been shown that people tend to perceive frequency in a logarithmic manner (Hay & Baayen 2002): differences between lower frequencies are perceived as more substantial than differences between higher frequencies. The strength and statistical significance of the association between bases and derivatives were measured using Spearman's nonparametric correlation coefficient ( $r_s$ ), and relative frequency effects were visualized using GAM curves. Finally, in P2, I also explored suffixal rivalry between *-mine* and *-us* nouns by comparing the relative frequencies of derivatives sharing the same lexical base.

This overview section also contains some additional analyses which have not been published but were conducted on data from the same corpora. They serve first and foremost to exemplify the distribution of *-mine* nouns and their functions across registers (not only in dialects) as well as to provide an overview of case distributions of the three investigated suffixes. However, the observations from the additional data will remain unelaborated and do not constitute the core of this thesis.

Data manipulation, analysis, and visualization were conducted in R (R Core Team 2020). The scripts for data collection and analyses were most complex for P2 and these have been published online along with input and output datasets<sup>18</sup>. The process for P1 was similar and can be derived from the published scripts. Data analysis for P3 and P4 was largely qualitative and the datasets can be extracted from the Corpus of Estonian Dialects by a simple noun lemma query [*\*mine*].

---

<sup>18</sup> [https://www.github.com/MPilvik/morphological\\_productivity](https://www.github.com/MPilvik/morphological_productivity)

## 4 OVERVIEW OF THE PUBLICATIONS

**P1. Pilvik, Maarja-Liisa. 2019. Assessing the productivity of the Estonian deverbal suffix *-mine* in five registers of Estonian. *SKY Journal of Linguistics* 32. 75–103.**

In this first article, I analyzed the productivity of the deverbal suffix *-mine* in five different registers of Estonian: scientific texts, newspaper texts, fiction, contemporary spoken language, and spoken regional dialects. The different aspects of productivity were captured through operationalizing the three measures of morphological productivity described in Section 2.2.1 – realized productivity, potential productivity, and expanding productivity. The size of the analyzed samples was approximately 335,000 tokens per register. A further attempt was made to capture those *-mine* formations which have a higher likelihood of being semantically more entrenched. This was done using a very simple NV-score which expresses the difference between the relative frequency of a verb stem among the *-mine* nouns and the relative frequency of the same stem among verbs. This simple measure relies on the assumption that semantically nonregular derivatives generally have a high usage frequency and therefore, the stems with higher NV-scores are more likely to occur in nominalized structures than would be expected by their overall use in verb forms.

Even though *-mine* is often described as the most productive suffix in Estonian or the suffix with absolute productivity (Kasik 1979: 11), the analysis revealed that the quantitative output of different semantic, paradigmatic, and pragmatic constraints on the productivity of *-mine* does indeed vary across registers. Each of the three measures used, however, provided a slightly different perspective on this variation. Realized productivity, which reflects the past contribution of *mine*-suffixed words to the size of the vocabulary as a whole, conformed most with the general observations on productivity made in the literature and was the highest in the more formal written registers (newspapers and scientific writing) and the lowest in dialects and spoken contemporary Estonian. However, expanding productivity, which expresses the potential of *-mine* nouns to further expand the vocabulary as a whole, was the highest in fiction and contemporary spoken Estonian, while this estimate for scientific texts was undeniably the lowest. Finally, potential productivity, which shows the future growth potential of the *-mine* category itself, also clearly favored fiction and the spoken registers (especially dialects) as the registers with most innovation potential. These results were associated with the proportion of syntactic and lexical functions of *-mine* in the corresponding registers, with syntactic functions showing more potential for category expansion. The implications from the NV-scores for the lexical processing of individual lexemes highlighted the formations with more ambivalent semantics (e.g., *kohtumine* ‘meeting’, *valimised* ‘elections’) as the stems more attracted to the nominalized form.

**P2. Pilvik, Maarja-Liisa. 2021. Comparing the productivity of Estonian deverbal suffixes *-mine*, *-us*, and *-ja* in five registers: A quantitative usage-based approach. *Eesti ja soome-ugri keeleteaduse ajakiri / Journal of Estonian and Finno-Ugric Linguistics* 12(1). 189–258.**

The second article further built upon P1 by examining the same three productivity measures and five registers while adding some additional perspectives. In addition to comparing different registers in terms of how the productivity of a suffix varies, I also used the quantitative measures to rank the three most frequent suffixes for noun derivation in Estonian (Kasik 2015: 281): *-mine*, *-us*, and *-ja*. The *-mine* and *-ja* suffixes are parallel in their generality and regularity, while *-mine* and *-us* as action noun suffixes cover overlapping semantic domains. However, all the suffixes differ in their structural, semantic, paradigmatic, and pragmatic constraints. Since the ranking of different affixes in terms of the rate with which they expand their category is sensitive to the number of existing usage events, I compared three approaches in terms of how well their implications conformed to the linguist's intuition. I also operationalized relative usage frequency for assessing the regularity of the three derivation patterns and for capturing the derivatives which would more likely be accessed as a whole word and those which are more easily parsed into components in lexical processing. Instead of NV-scores as in P1, I used log frequencies of the verbal base and derivative and Spearman's correlation coefficient. Finally, I used log derivative frequencies to explore the potential rivalry between the two action noun suffixes *-us* and *-mine*, hypothesizing that in the case of similar functions or meanings, any two derivatives formed from the same base but with these two suffixes, respectively, have diverging frequencies, and that the action noun readings for both suffixes are more readily available in the lower frequency range, where the likelihood of the speaker utilizing a productive word-formation rule is higher.

The results confirmed that the derivation patterns demonstrate varying degrees of productivity in different registers. Again, the pragmatic usefulness of the words formed with the different suffixes correlates well with the dimension of written *vs.* oral mode of communication, but this can also be attributed to the different distribution of word classes in different registers more generally. However, it is the contrast between formal and informal language use which determines the possibility to enrich the lexicon with new formations. Fiction emerged as the register which seems to facilitate the most productive use of the derivation patterns in all aspects of productivity. The ranking of the affixes mostly agrees with linguistic intuition and correlates with the number of restrictions a derivation pattern has: the overwhelmingly most productive suffix is *-mine* as the borderline case between inflection and derivation. It is followed by *-ja*, and finally, by *-us*. However, the latter two are argued to be constrained by different aspects: either structural conditions and semantic vagueness of the category (*-us*) or the semantic conditions and low pragmatic usefulness / paradigmatic competition (*-ja*). The samples used in P2 had around 80,000 more tokens than the ones used in P1, but the sample sizes were still very modest due to the small amount of corpus data

available for spoken Estonian. Therefore, the results of the different approaches should be interpreted with caution since in small samples, the productivity of the suffixes may be over- or underestimated.

While base frequency did not appear to be a very strong or reliable predictor of semantic decompositionality for individual lexical items, I argued that the stronger the correlation, the stronger the implication for the use of a more general construction in a particular register. With regard to the comparison of the suffixes *-mine* and *-us*, the semantic rivalry could indeed more likely be found among lexical constructions in the low-frequency range, but the relative frequencies of individual decontextualized lexemes do not provide a useful tool for semantic analysis.

**P3. Pilvik, Maarja-Liisa. 2017. Deverbal *-mine* action nominals in the Estonian Dialect Corpus. *Eesti ja soome-ugri keeleteaduse ajakiri / Journal of Estonian and Finno-Ugric Linguistics* 8(2). 295–326.**

The third article zoomed in on dialectal data, moved beyond morphological constructions, and examined the cross-linguistically attested ambiguity of action nominals as an intermediate category between verbs and nouns (see Koptjevskaja-Tamm 1993: 263–366, Ylikoski 2003, Comrie 1976, Haspelmath 1995). Unedited spontaneous language with an areal dimension is highly suitable for this since it demonstrates variation in many aspects which might be rare in written standard language. The study drew on dialect corpus examples to highlight how the *-mine* nominalization is used as an ambivalent category in communication. Next to the contexts in which the prototypical word-class properties of Estonian nouns are sufficient for describing the distributional properties of *-mine* nouns, many examples of periphrastic, catenative, and converbal constructions with their own form-meaning conventions were presented where the processual reading of the nouns becomes more salient. More specifically, the analyzed dialect data contained usage events of the following constructions:

- impersonal-modal Cx: *olema* ‘be’ +  $V_{mine_{NOM/PAR}}$ ;
- passive ingressive Cx: *minema* ‘go’ +  $V_{mine_{ALL}}$ ;
- active ingressive Cx: *minema* ‘go’ +  $V_{mine_{TR}}$ ;
- active progressive Cx: *olema* ‘be’ +  $V_{mine_{ADE}}/V_{mine_{GEN}}$  *peal* ‘on’;
- passive modal Cx: *tahtma* ‘want’ +  $V_{mine_{PAR}}$  (+ *saada* ‘get’);
- emphatic Cx: V1 + V1*mine*;
- colorative Cx: V1 + V2*mine\_{PAR}*;
- various converbal constructions.

This list does not exhaust all the possible *-mine* action nominal constructions and contains only the constructions attested in the relatively small dialect dataset consisting of 1928 observations of *-mine* nouns. However, the use of these constructions gives reason to believe that the *-mine* morphological category is indeed also quite flexible in the speakers’ minds, and the network of associations which

is activated is calibrated to match the different contexts and speech situations. It is then in this specific environment where the ‘nouny’ or ‘verby’ properties of the action nouns are determined for the speaker. The article also draws attention to the additional properties and functions of the *-mine* nominalization in spoken discourse where in addition to being a means for anaphoric referencing, *-mine* nouns seem to be used for relaxing the processing load or for providing time for utterance planning.

**P4. Pilvik, Maarja-Liisa. 2016. *olema* + *Vmine* konstruktsioonid eesti murretes. *Keel ja Kirjandus* 6. 429–446.**

In the fourth case study, I analyzed one rather frequent linguistic construction which nonetheless has thus far not been addressed in studies concerning Estonian *mine*-nominalizations or action noun constructions, namely the impersonal-modal *olema* ‘be’ + *Vmine*<sub>NOM/PAR</sub> construction (e.g., ex. 19).

- (19) *polnud*      *tal*      *enesel*      *vaevanagemist*  
 be.NEG:APP   she:ADE   own:ADE   pain.GEN+see:NMLZ:PAR  
 ‘she did not have to work hard herself’ (lit. ‘she did not have any pain-seeing herself’)

Again, the Corpus of Estonian Dialects was chosen as the primary data source, since it provides substantially more observations of this construction than written corpora<sup>19</sup> and also reveals more variation within the construction. A total of 364 attestations of the construction were analyzed, which is 19% of all observed *-mine* nouns when averaged across all 10 dialect areas. Between the dialects, however, no significant differences in the relative usage frequencies could be detected.

The construction itself consists of the copula *olema* ‘be’ and an action noun (typically a *-mine* noun, but other action nouns are also acceptable) in either the partitive or the nominative case. An optional adessive argument can be added to express the demoted agent or experiencer. Structurally, the construction therefore resembles an existential clause or an experiencer clause which states the (non)-existence of a process or a state. However, the fact that it is the action nominal which conditions other elements in the clause gives the construction independent syntactic status, much like a complex predicate. The functions of the construction are also different from that of a typical existential or experiencer clause: while most of the constructions have an impersonal meaning, some also express non-epistemic modality (e.g., ex. 20).

---

<sup>19</sup> For example, in samples consisting of 1000 randomly drawn observations of *-mine* nouns from all five registers, this construction occurs 225 times in dialectal data, 32 times in spoken contemporary data, and 16 times in fiction. It does not occur in newspapers and the one observation from scientific corpus is a direct quotation from a historical chronicle.

- (20) *pole koju kandmist seda rämps*  
 be.NEG.CNG to\_home carry:NMLZ:PAR this:PAR trash.PAR  
 ‘one does not have to carry that trash home’ (lit. ‘one does not have carrying that trash home’)

Distinguishing between impersonal and modal readings, however, is not a trivial task, and is often a matter of interpretation. The construction is likely to have developed from a more archaic Finnic modal construction (Grünthal 1941: 177–179), the relicts of which can also be seen in the dialect data, namely in the possibility for the action noun to inherit the verb’s argument structure (e.g., ex. 21).

- (21) *enne koitu oli vedamine neid välja*  
 before dawn.PAR be:PST.3SG carry:NMLZ they:PL:PAR out  
 ‘they had to be taken outside before dawn’

The construction in question can be considered a productive one since the action noun slot can be filled by various lexemes and there is only one lexeme which seems to occur in the construction more frequently than expected – *tegemine* ‘doing, making’. This lexeme, however, is the most frequent *-mine* noun in the dialect data outside this specific construction as well but is used in a slightly different manner in the *olema* ‘be’ + *Vmine* construction.

## 5 RESULTS AND DISCUSSION

Before turning to the discussion of the results, I will again present the seven research questions introduced at the beginning of this thesis.

**RQ1:** Is the use of common derivational morphology a differentiating feature between written and spoken registers?

**RQ2:** Can corpus-based quantitative measures of morphological productivity tell us something about the constraints and functions of deverbal suffixes in different registers of Estonian?

**RQ3:** How do the three most frequent deverbal suffixes rank in terms of their productivity in different registers?

**RQ4:** Do the basic productivity measures correspond to general linguistic intuitions when they are based on relatively small samples?

**RQ5:** Can relative frequencies help detect semantic regularity?

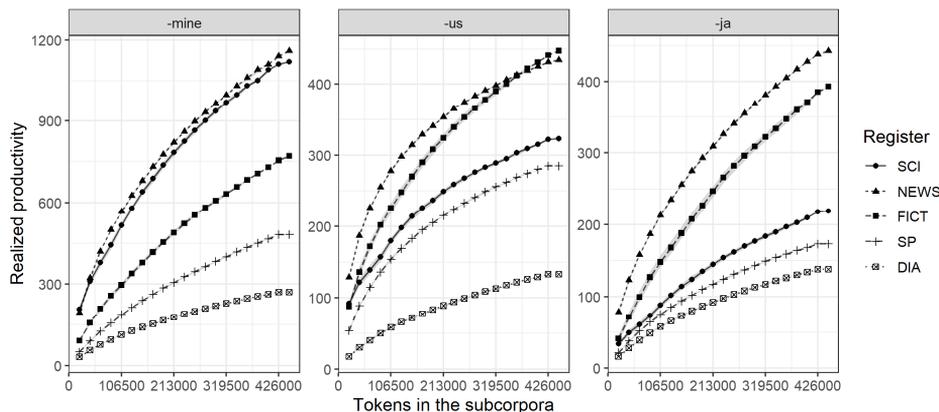
**RQ6:** Which combinatorial features of *-mine* nouns emerge from older colloquial data and what implications does this have for the quantitative assessment of morphological productivity?

**RQ7:** What are the properties and potential development paths of a periphrastic action noun construction and what does it tell us about the relationship between morphological and syntactic productivity?

Answers to the proposed questions are presented and discussed in the following 4 subsections, which summarize the different aspects more relevant to some research questions than to others. Subsection 5.5 discusses the limitations of the current work and suggests potential avenues for further research.

### 5.1 Morphological productivity as a function of register

The corpus analyses of the derived words in P1 and P2 suggested that the aspects of morphological productivity as gauged by the realized, potential, and expanding productivity measures introduced by Baayen and his colleagues do indeed vary across different registers of Estonian. The main distinction between written and spoken registers, relevant to **RQ1**, was most apparent when attention was set to past/realized productivity, i.e., to the contribution the suffixes make to the size of the existing vocabulary as a whole in the samples. The lexical growth curves from P2, which compared all three suffixes in five registers, are presented in Figure 1.

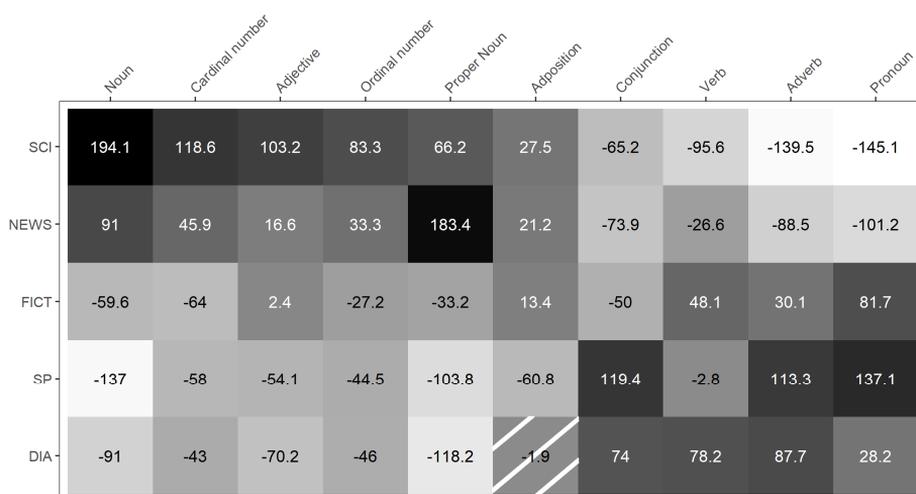


**Figure 1.** Realized productivity ( $V(N)$ ) of *-mine*, *-us*, and *-ja* in cumulative samples of five registers. The symbols and black lines represent the mean values of 100 random permutations of the 426,000 token samples. 95% confidence intervals are shown with light grey shading. (Pilvik 2021).

In Figure 1, the  $x$ -axis presents the number of tokens in the 21 subcorpora of cumulatively increasing size and the  $y$ -axis presents the mean number of types formed with a suffix (= the realized productivity of a suffix) in a given subcorpus. The derived nouns clearly contribute the most to the overall vocabulary of the written registers (journalistic texts, newspaper texts, and fiction, especially the first two), and the differences become more significant as the size of the corpus increases. For example, when the difference between the realized productivity of *-mine* in NEWS and DIA at 106,500 tokens is around 470 types, then at 426,000 tokens, this difference is almost 900 types. One implication of this could therefore be that written registers exhibit a more productive use of the morphological constructions in question. The reasons for this may lie in the higher pragmatic need for the resulting concepts and relations as well as the more general traditional functions of word-formation. For example, the high realized productivity values of *-mine* in journalistic and scientific texts could be attributed to the necessity to use the means of nominalization for abstraction and creating new terms, and perhaps also for stylistic reasons (see Kasik 2011: 70), whereas such a need is less obvious in the spoken registers, which are more likely to use derived nouns for syntactic recategorization and anaphoric referencing in online communication. The distribution of *-ja* and *-us* nouns in the different registers could also be analyzed with regard to active word-formation *vs.* using automatized lexical sequences, but they also relate to the pragmatic usefulness of the relevant concepts in different registers (see P2 for detailed discussion). Therefore, with regard to **RQ2**, we might conclude that the functions for which derived words are created in different registers are indeed reflected in the visible differences between realized productivity curves.

However, given that realized productivity is based on a simple type count, we must ask whether this high productivity in written registers can be associated with

the morphological class of derived nouns or with the class of nouns and their functions more generally. Indeed, in P2, the differences between the distribution of word classes in the used samples were clearly visible: nouns as a word class are used considerably less in the spoken registers than in written ones, whereas verbs and adverbs are used more. Figure 2 illustrates this functional labor division between word classes in different registers in more detail<sup>20</sup>. The cells are colored according to the standardized Pearson residuals from a Chi-squared test of independence performed on the registers and part-of-speech tags in corpus samples used in P2. The darker the color of the cell and the larger the value of the residual, the stronger the attraction between the categories. Conversely, the lighter the color and the smaller the value of the residual, the stronger the repulsion between the categories. The striped cell indicates a combination where the residuals are  $< 2$  or  $> -2$ , meaning that this particular combination of categories does not have a significant contribution to explaining the significant association between registers and their use of word classes.



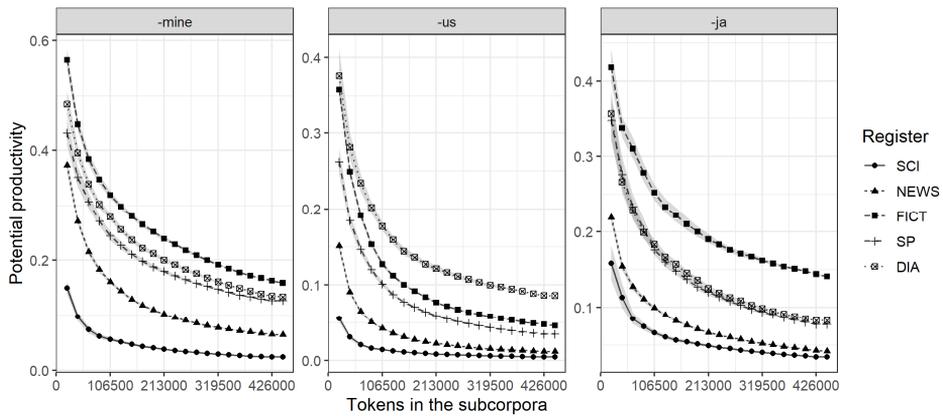
**Figure 2.** Standardized Pearson residuals from Chi-squared test of independence performed on the registers and part-of-speech tags ( $X^2(36, N = 2,074,937) = 250,973, p < .01$ ).

It is quite remarkable how much of the differences between registers can be explained by looking at simple distributions of word classes. It is apparent that scientific and journalistic texts clearly cluster together based on their higher use of nouns, cardinal numbers, adjectives, ordinal numbers, proper nouns, and appositions, whereas the spoken registers make significantly more use of the core discourse mediators – conjunctions, adverbs and pronouns (fillers are not shown in this analysis). Fiction emerges as an intermediate register between the two clusters while sharing a remarkable similarity with the spoken registers. It is important to be aware of the specifics of the distributions of words and word classes across

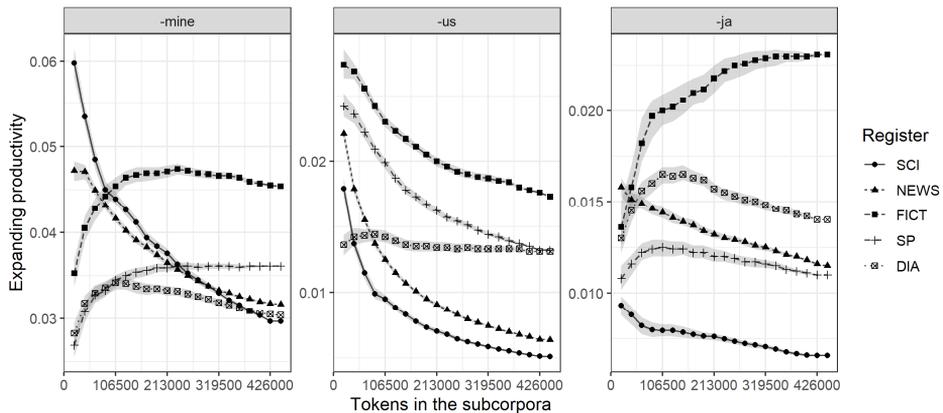
<sup>20</sup> Only the word classes which are used in all corpora are analyzed.

corpora because otherwise, a lot of information could be lost (Racz et al. 2016). Having fewer instances of particular word classes in certain registers, therefore, seems to be a hitherto not very well studied co-determinant of productivity. Coming back to research question **RQ2**, the realized productivity values can still be related to the pragmatic need for the different functions of word-formation, but they are also influenced by the general need for nouns in different registers.

The other two productivity measures, which are less dependent on the overall contribution of nouns as a word class to each specific register, make diametrically different implications with regard to the productivity of the complex words in the five registers. Potential productivity curves from P2 are shown in Figure 3 and expanding productivity curves in Figure 4.



**Figure 3.** Potential productivity of *-mine*, *-us*, and *-ja* in cumulative samples of five registers. The symbols and black lines represent the mean values of 100 random permutations of the 426,000 token samples. 95% confidence intervals are shown with light grey shading (Pilvik 2021).



**Figure 4.** Expanding productivity of *-mine*, *-us*, and *-ja* in cumulative samples of five registers. The symbols and black lines represent the mean values of 100 random permutations of the 426,000 token samples. 95% confidence intervals are shown with light grey shading (Pilvik 2021).

Instead of reflecting the category's past contribution to the vocabulary, these measures lend themselves to be interpreted as conditional probabilities with which the categories can be extended with new items. Potential productivity as a central measure in quantitative studies of morphological productivity expresses the probability that the next instantiation of a morphological construction (e.g., a *-mine* noun) added to the corpus represents a previously unseen type. Expanding productivity, in turn, expresses the probability that the next previously unseen type added to the corpus instantiates the morphological construction in question (e.g., a *-mine* construction). These probabilities are not so much determined by the distinction between written and oral modes of communication (**RQ1**), but instead by the dimension of formality *vs.* informality. This is clearly visible for potential productivity curves in Figure 3, where SCI and NEWS curves are situated at lower values on the *y*-axis throughout the sampling process, i.e., with respect to the *x*-axis. Fiction and the spoken registers appear to facilitate more productive and active word-formation in that sense than the written registers, which exhaust the list of their effective bases at a generally much higher rate (the rate with which new *hapax legomena* are found, shown on the *y*-axis, decreases more rapidly as more tokens are sampled). Since this appears to also apply to the *-mine* construction, which has a theoretically infinite list of available bases, the key to explaining the different expanding potential a morphological process demonstrates has to lie in the pragmatic factors and the functions for which the process is used in different registers (Bauer 2001: 208–209). According to Bauer (2001: 208–209), there are fewer pragmatic constraints on processes which are used for transpositional purposes; for this reason, they are more productive than the processes used for lexical innovation. I do not have the necessary analyses to claim, for example, that *-us* would be used more for lexical innovation than *-mine*, although it is possible that for transitive *-ta*-ending bases, the potential to easily create a word with an object or a result reading (e.g., *süüdista-* ‘accuse’ → *süüdist-us* ‘accusation’) is often utilized for this purpose. The tendency cited in Bauer (2001) might, however, explain the differences in productivity for the same suffix across registers. For example, *-mine* or *-us* nouns can exhibit more transpositional functions in the informal registers, while the lexical functions prevail in the more formal ones.

For expanding productivity, it is worth looking at the general trend of the curves in Figure 4: the trend is clearly negative in SCI and NEWS for all suffixes, indicating that these registers might benefit more from an existing lexicon of derived nouns rather than active word-formation. Since potential productivity (and, indirectly, expanding productivity as well) is negatively affected by the proportion of frequently used and consequently highly automatized lexical sequences among the constructions, it seems reasonable to suggest that the amount of such sequences is lower in the informal registers and complex words are more often produced on-the-fly (**RQ2**). This, admittedly, is a somewhat bold generalization to make based on corpus data alone and should be further tested with other methods. It must also be noted that the exact ranking of the registers according to potential and expanding productivity of *-mine* construction differed slightly in P1

and P2, which prompts us to be cautious not to interpret the quantitative results in a very absolute manner, especially with small samples (**RQ4**).

The comparison of quantitative productivity measures for a single suffix in different registers might also evoke a different interpretation of the results, namely one where a construction (e.g., *-mine*) would be considered equally productive in all registers, but this productivity would be manifested as a high number of types when the usage frequency of the construction is high in a register and as a high number of *hapax legomena* when the usage frequency of the construction is low<sup>21</sup>. In this interpretation of productivity, (structural and semantic) generality and (semantic) regularity would be given higher prominence over functional, pragmatic, and paradigmatic factors, which would be responsible for distorting the somehow inherent productivity of a pattern in usage. From a usage-based perspective, however, it seems that the functions in which the derivatives are (not) used or for which they are (not) created in a particular register as well as the number of competing linguistic means for a construction should be considered important co-determinants of productivity, rather than something that simply happens to potentially productive or unproductive patterns in usage. In this thesis, I adopt the view that productivity should not be reduced to some abstract potential but instead be regarded as an empirically derived and constantly changing property of linguistic constructions. Therefore, a suffix can be considered more productive in one register than another because while its structural restrictions are not expected to change in different registers, the semantic, functional, pragmatic, and paradigmatic restrictions might. And while it is true that the more types we go through, the less likely we are to encounter a type never used before, the (functional, pragmatic, paradigmatic) mechanisms behind the high or low number of types instantiating a construction in some register are also constantly shaping its productivity. As I see it, the quantitative productivity measures are meant to gauge these different aspects and complement each other. However, I readily agree that the tendencies found in this thesis should be tested against larger corpora, and validated with other methods (see Section 5.5).

## 5.2 Ranking the suffixes

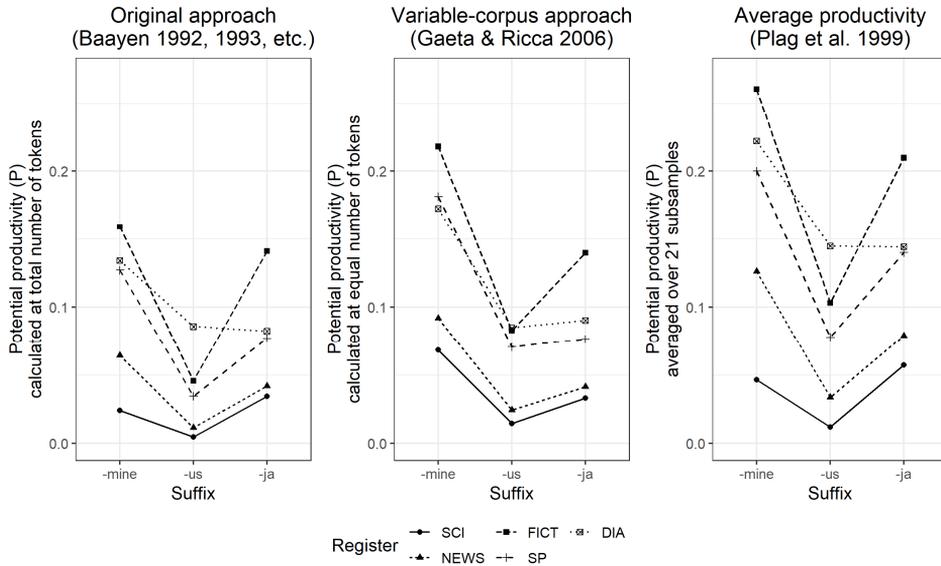
When generalizing across all three productivity measures and all three suffixes analyzed, fiction seems to be the central breeding ground of productive language use and lexical innovation based on the samples used in the studies. However, not all constructions behave uniformly with regard to their productivity in different registers. The use of a morphological construction may display vast differences in productivity across registers, different from that of other constructions.

Different approaches utilizing the potential productivity measure for ranking different morphological processes were compared in P2 (Figure 5). Since potential productivity is the only measure of the three measures discussed in this thesis

---

<sup>21</sup> I thank one of the reviewers of this thesis for pointing this out.

which is related to the number of usage events of a given construction in a corpus and can thus give different results based on how the usage events are counted, mainly the relevant methodology and results concerning that measure will be discussed here further. Results for ranking the suffixes according to their expanding productivity values will also be mentioned.

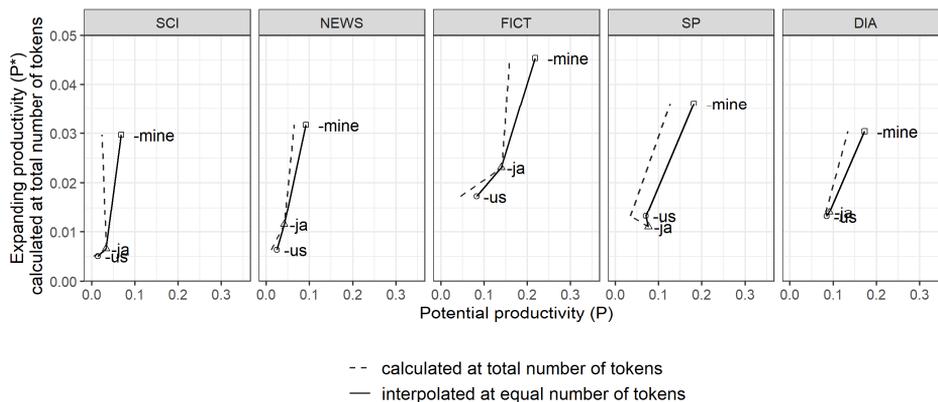


**Figure 5.** Comparison of the values for potential productivity  $P$  obtained with three different approaches (Pilvik 2021).

Overall, *-mine* construction emerged as the most productive suffix in all registers, followed by *-ja*, and *-us* (RQ3), which is not a surprising result by itself, considering the complexity of the combination of restrictions each pattern is subjected to (RQ2). The suffix *-us*, which has several morphophonological constraints with regard to the bases with which it can combine, which mostly prefers derived bases ending with *-ta* or *-da* (see P2), and has several allomorphs (e.g., *-dus*, *-lus*, *-ndus*) is also semantically quite vague as a construction and in its most regular use as a deverbal suffix forming action nouns in rivalry with the more general *-mine* pattern. This should be taken as a natural co-determinant of a pattern's low productivity: due to formal and semantic fuzziness, speakers would rather stick to known formations than create new ones. The suffix *-ja* represents a highly general and regular construction, which nevertheless is semantically inclined to attach to bases which presuppose a potentially active participant (the internal subject of the activity). The *-ja* suffix as an agent noun suffix also has many paradigmatic competitors in contexts where the denotation of the related process is not relevant, such as pronouns, proper nouns, and common nouns, making it often pragmatically a less useful construction. The productivity of the constructions is thus governed by the interplay between these quite different

forces which guide speakers' choices and consequently influence the reinforcement of particular constructional schemas but also by the importance of specific constraints in different registers. Therefore, the reason why the *-ja* category is extended at a low rate in scientific texts might be its typical use in denoting fixed roles of participants in specific procedures (e.g., *kaebaja* 'prosecutor', *kõneleja* 'speaker') or scientific abstract terms and instruments (e.g., *näitaja* 'indicator', *mõõdja* 'meter') – categories which are not likely to be rapidly extended. On the other hand, the fact that it is less productive than *-mine* in fiction might stem from the overall lower pragmatic usefulness of the concepts which instantiate the general schema.

Although for some registers, the variable-corpus approach did provide a more intuitive ranking of the suffixes than Baayen's approach (the latter, for example, ranked *-ja* the most productive suffix in the scientific register), I argue that in small samples at least, there is no real reason to prefer the variable-corpus approach of Gaeta & Ricca (2006) to Baayen's original approach (**RQ4**). While the original approach has been criticized for overrating the productivity of less frequent constructions (which also might have happened to *-ja* in the scientific register), Gaeta & Ricca's approach can essentially suffer from the same problem when samples are small and constructions occur with very different frequencies. For example, in the scientific sample used in P2, there were 15,052 *-mine* tokens, 13,987 *-us* tokens, but only 2346 *-ja* tokens. In this case, Baayen's approach seems statistically more valid since it anchors the measures to the same corpus and thereby also accounts for the information that the suffixes are not equally profitable. Gaeta & Ricca's variable-corpus approach assesses the potential productivity of more frequent suffixes at completely random points on their growth curve, which results in the different status of the next tokens added to the corpus for the different morphological constructions. The expanding productivity measure  $P^*$  could also be used, since it is closely correlated with Gaeta & Ricca's modified version of  $P$ , while still being anchored in the same corpus. The comparison of Baayen's expanding productivity  $P^*$  with his original potential productivity  $P$  (dashed line) and with the interpolated  $P$  of Gaeta & Ricca (solid line) in P2 also shows that the measures give very similar rankings (Figure 6) but Baayen's potential productivity  $P$  is systematically estimated lower for the more frequent suffixes *-mine* and *-us*.



**Figure 6.** Comparison of expanding productivity  $P^*$  and potential productivity  $P$ . The lines connecting the suffixes do not imply continuity but are there for visualization purposes (Pilvik 2021).

When comparing multiple suffix constructions, which can be instantiated by many lexical items with diverging degrees of automatization and decompositionality, it becomes important to explicate what has been considered to be a morphological construction when measuring and comparing productivity. As we have already seen (and will see in the next subsection), general schemas can be instantiated by novel structures based on familiar patterns, but such instantiations can range from being completely automatized to borderline rule-bending cases. When claiming something about a morphological construction’s productivity, the meaning of the construction should thus be either very general (e.g., ‘entity having some relationship to the base’, see Bauer 2001: 200) or a prototypical meaning should be established (e.g., ‘an action noun’ or ‘an agent noun’), which then serves as the central node for various extensions through polysemy links. Affixes can be highly non-specific in terms of identifying the relevant metonymy and can underspecify the relationship involved (Janda 2010). The realizations of a construction which somehow deviate from the semantics of the more general (or prototypical) derivation pattern (e.g., occupation names such as *õpetaja* ‘teacher’, *õmbleja* ‘seamstress’) have been treated as idiomatic (Kasik 2004: 25) or lexicalized (Kasik 2011: 68) instances, defined as derivatives which have lost their semantic transparency. However, systematic groups of such words seem to be better analyzed as subschemas of the more general pattern since they appear to be partially productive by themselves, i.e., they can be instantiated by new items without necessarily having to wait for the specialized meaning to develop in usage. Which of the two aforementioned approaches is more adequate is unfortunately not completely clear to me. The (arguably somewhat trivial) solution seems to be considering all relevant formations (e.g., *-mine* nouns) as the result of the same very general process if we restrict our analysis to the morpheme-level, and to distinguish between a prototype and more peripheral uses if a schema-based, semantic perspective is taken. The latter concerns the analysis of the productivity of semantically more specific subconstructions (e.g., constructions denoting results or objects instead of processes among *-mine* or *-us* nouns) or the productivity of the

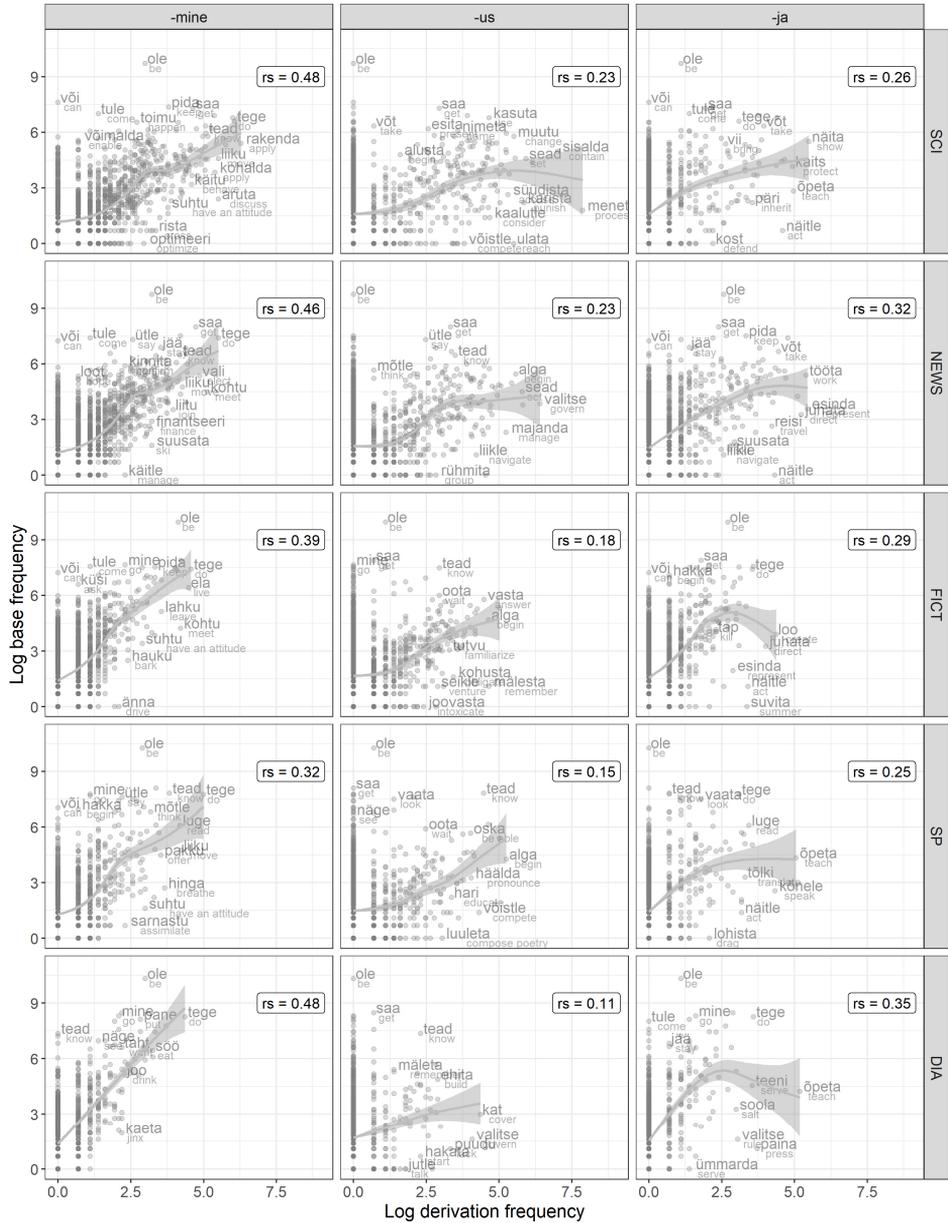
larger syntactic constructions in which the complex words occur. The different uses can show different grammatical behavior. For example, the use of *-ja* nouns denoting people differs certainly from *-ja* nouns denoting instruments as the latter are more likely to occur in the comitative and consequently are often used in different syntactic constructions. In the first two case studies of this thesis, which deal with measuring productivity (P1 and P2), ‘identity of form was taken as a guide to identity of morphological structure’ (Bauer 2001: 199–200), whereas in the two more qualitatively-oriented studies based on dialect data (P3 and P4), a more schema-based analysis was conducted, although restricted to constructions with *-mine* nouns.

### 5.3 Paradigmatic support for productivity

As shown in many studies, the productivity of a construction is closely related to semantic compositionality. There is a negative correlation between the strength of what Diessel (2019: 75) calls sequential links and lexical connections. Frequent complex words become more automatized and are processed faster. Consequently, the number of linguistic decisions that producing and understanding texts requires is also reduced and the sequential relations become more activated at the expense of the general schema. This might result in the weakening of lexical connections, i.e., an item’s associations with other items on the same level of abstraction, which are used in categorization. The association of one derived lexeme with other lexemes belonging to the same morphological category, for example, becomes weaker. In more concrete terms, the more semantically opaque, idiomatic, or simply very frequent types there are within a morphological category, the less productive the general schema; this is because the lexical sequences themselves are stored in and retrieved from memory as whole chunks and do not significantly contribute to the productivity of the general schema. Studies (e.g., Hay 2001, Hay & Baayen 2002, Blumenthal-Dramé 2012) have shown that the tendency of complex words to be stored and parsed as whole words rather than combinations of stem and affix correlates with the frequency of their base form: if the complex word is more frequent than its base, then it is more likely to be accessed as a whole, whereas when the base is more frequent, the complex word is perceived as more compositional. To answer **RQ5**, these paradigmatic relationships were also tested in this thesis (P1, P2) and the idea was further extended to explore the potential ‘blocking’ between *-mine* and *-us* types (P2). It must be acknowledged, however, that both analyses were conducted in a global manner without thoroughly analyzing individual lexemes or the semantic groups they form. The limitations on the semantic analysis were imposed by the methodological decision to analyze the decontextualized simple base stems of derivatives. Therefore, any conclusions will also be quite general and only suggest potential avenues for further research.

The comparison of base and derivative frequencies in P2 (Figure 7) suggested that there is indeed a moderate to strong positive monotonic correlation between the log of base frequency and the log of derivative frequency, which implies high regularity and strong support for the derived nouns being from bases realized as

verb forms. However, this correlation only applies for the *-mine* construction, where the lexical relations are stronger. For the other two suffixes, the correlation was weaker and not necessarily monotonic, especially with *-us* nouns. Lexical *-us* and *-ja* constructions therefore might facilitate a higher degree of automatized processing, while *-mine* activates the general schema.



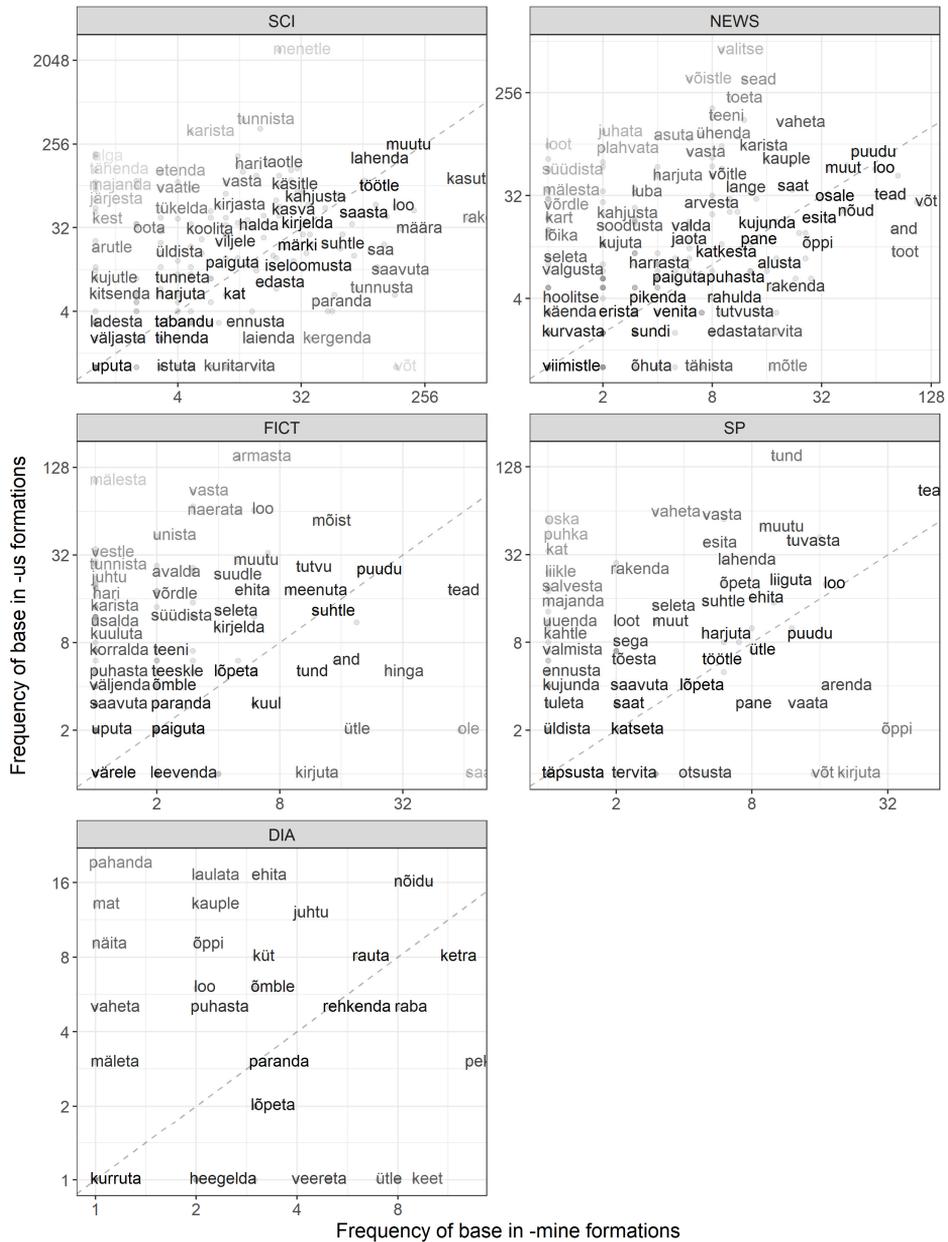
**Figure 7.** The relation between log base frequency and log derivative frequency for the suffixes *-mine*, *-us*, and *-ja* in five registers (Pilvik 2021).

Furthermore, while the derivatives whose base frequency was higher than the derivative frequency appeared to be semantically more regular and less specific (e.g., *olemine* ‘being’, *saamine* ‘getting’, *tulemine* ‘coming’) and the ones whose base frequency was lower were more likely to express metonymic or metaphorical extensions of the general schema (e.g., *kohtumine* ‘meeting’, *suhtumine* ‘attitude’, *valimine* ‘election’), this was not absolute and semantically extended complex nouns could be found on both sides of the trend line. It is thus difficult to claim anything substantial about semantic decompositionality based on small samples of corpus data. More refined measures or a more detailed analysis of the various subschemas could prove beneficial along with experimental techniques such as lexical naming or lexical decision tasks. However, as global indicators for semantic regularity of the constructions, the correlations worked quite well. The bases most attracted to the *-mine* construction and the ones most repulsed by it were also captured with NV-scores in P1. While there were also topical differences between registers, one of the most likely bases to be nominalized across all five registers was *tegemine* ‘doing, making’. It should be the case, therefore, that this type does not particularly contribute to or benefit from the productivity of the *-mine* construction. Its frequent use, however, originates from different sources in the different registers, as will be shown in Section 5.4.

As mentioned in Section 2.1.2, morphological constructions are interconnected through an excessive amount of subpart relations (Hilpert 2014: 83), which in Diessel’s (2019) approach are described using the terms of *lexical* and *constructional links*. Therefore, the effect and nature of correlations between constructions belonging to the same paradigm or ‘word family’ were further explored in the context of the potential type blocking between *-us* and *-mine* nouns, which often cover overlapping semantic domains. Type blocking was interpreted in a non-exclusive and gradient form through relative token frequencies. The derivatives whose bases occurred at least once with both suffixes were analyzed in the five registers (Figure 8).

In the less formal registers, the frequency of types formed with the less productive *-us* was clearly higher than the frequency of the corresponding *-mine* types. This is a robust demonstration of the complementary relationship between rival affixes, where the existence of one form may constrain the use of the other. Since the *-us* construction is structurally very selective toward the bases it licenses, it is actually not striking that there are more *-us* tokens than *-mine* tokens from the same bases: those bases populate the *-us* category more densely than the *-mine* category. This effect is at first glance less visible in the formal registers SCI and NEWS, but this is really an artifact of using the logarithmic scale and the fact that the token frequencies are considerably higher in the more formal registers. While most of the derivative pairs seem to prototypically fulfill different functions, similar semantics are more likely to be detected in the lower frequency spectrum and in higher-frequency pairs with a similar relative frequency. This is, on the one hand, contrary to the previously proposed hypothesis about suffixes being used in the same functions to demonstrate different relative frequencies but, on the other hand, directly associated with the tendencies of low-frequency types

to exhibit regular uses of a morphological construction, while high-frequency types tend to acquire specialized meanings or functions. To answer **RQ5**: relative frequencies are a useful global heuristic for examining the general semantic regularities of patterns, but they do not prove very beneficial in a closer semantic analysis of rival *-mine* and *-us* types when stripped of all contextual information.



**Figure 8.** Comparison of the frequencies of *-us* and *-mine* nouns derived from the same base (Pilvik 2021).

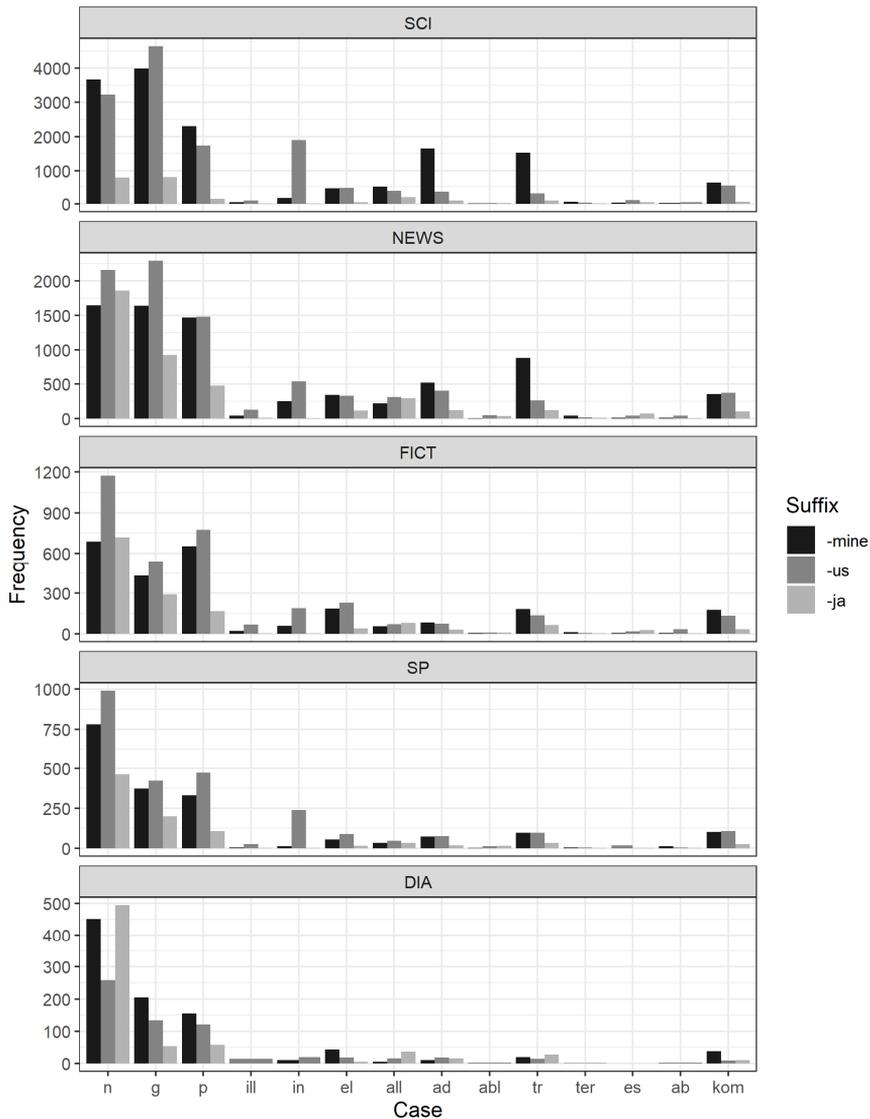
## 5.4 Syntactic constructions as co-determinants of morphological productivity

The main reason why *-mine*, *-us*, and *-ja* nouns are categorized as nouns is the fact that they are inflected for case and number. Much like the productivity of the suffixes in different registers is affected by the overall distribution of word classes in those registers, the cases in which the suffixed nouns are frequently used – indicating the core functions of the nouns belonging to a particular morphological category – also influence the outcomes in quantitative assessments of their morphological productivity through the activation of different filler-slot relations. When many usage events of a morphological construction occur in a particular case, which in turn correlates with a specific morphosyntactic environment, the productivity of a morphological construction is controlled by the productivity of the construction in which the complex word is used. When that particular construction represents an open schema, licensing a wide range of lexical constructions, the effect on the productivity of the morphological construction would be less drastic than when that schema constitutes a construction in which the slot of the deverbal noun is lexically fixed. For example, when assessing the pragmatic usefulness of particular concepts formed with some morphological construction, we would actually be assessing the pragmatic usefulness of the syntactic construction. The case distribution of complex nouns would also have consequences for word processing effects. In paradigms in which the members are distributed uniformly the processing time for the members would be comparable. However, when some members occur much more frequently than the rest, then the low-frequency forms would be processed more slowly since their high-frequency competitors would be harder to eliminate (Baayen et al. 2011, Marzi et al. 2020: 282–283).

Figure 9 shows the distribution of case forms of the three derivation suffixes analyzed in P2 according to the register. It is evident that the case distribution of the forms is not uniform either across suffixes or across registers. As expected, the nominative, the genitive, and the partitive are generally more frequent also for the nouns containing the three suffixes, but even here, there are remarkable differences. For example, the *-ja* construction in dialects occurs surprisingly frequently in the nominative case construction compared to other registers<sup>22</sup>. The case distribution of *-us* in scientific texts shows a very frequent use of the inessive; *-mine*, in turn, has a tendency to occur in the adessive and the translative case in the more formal written registers, etc.

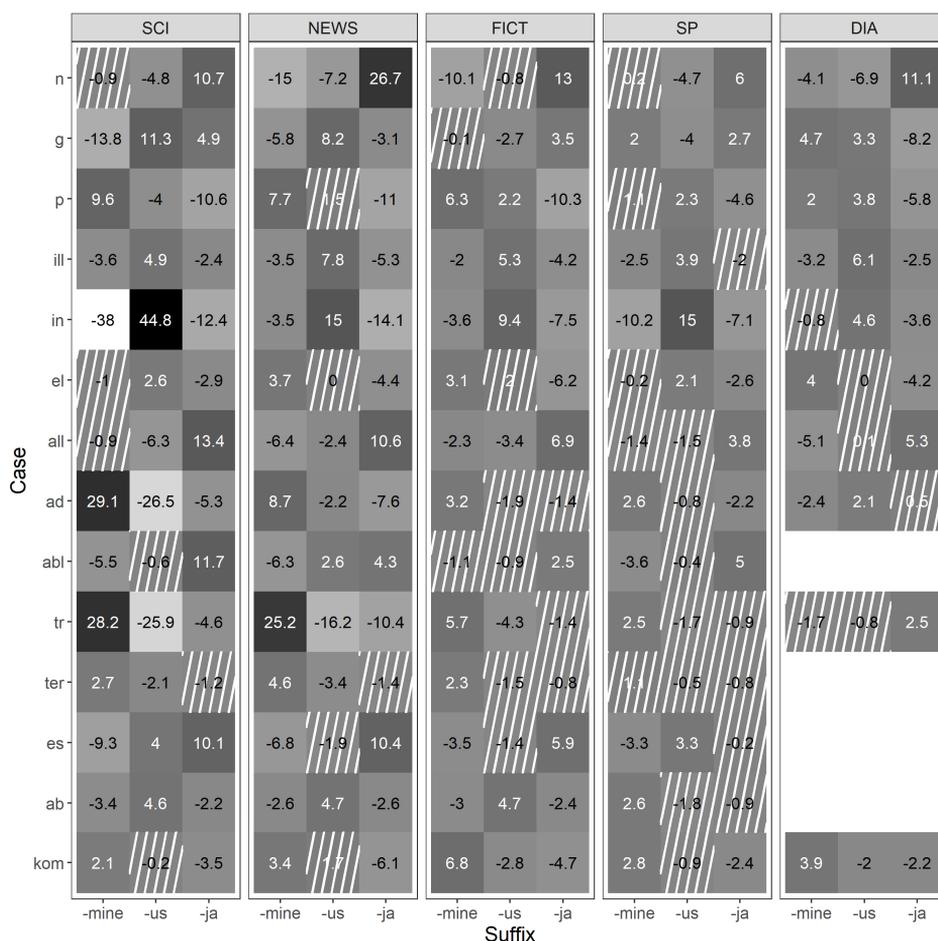
---

<sup>22</sup> Where the process of assigning morphological categories is largely automatic, this might also be affected by the morphological analyzer's ability to distinguish between different case forms in context (e.g. the nominative *kirjutaja* 'writer' and the genitive *kirjutaja* 'writer's').



**Figure 9.** Case distribution of the nouns formed with the suffixes *-mine*, *-us* and *-ja*.

These tendencies are confirmed by statistical analyses where the distribution of case forms is compared across the three suffixes within each register. Again, the Chi-squared test of independence was used and Figure 10 presents the standardized Pearson residuals. The striped cells indicate combinations where the residuals are  $< 2$  or  $> -2$ , meaning that these particular combinations of categories do not have a significant contribution to explaining the significant association between suffixes and their case forms. The empty rows in DIA refer to the case constructions which were too rare with the three complex noun constructions to be included in the analysis (namely ablative, terminative, essive, and abessive).



**Figure 10.** Standardized Pearson residuals from Chi-squared test of independence performed on the suffix and case constructions in the five registers. SCI:  $X^2(26, N = 31255) = 4341.2, p < .01$ ; NEWS:  $X^2(26, N = 19947) = 2100.3, p < .01$ ; FICT:  $X^2(26, N = 7501) = 608.15, p < .01$ ; SP:  $X^2(26, N = 5413) = 381.75, p < .01$ ; DIA:  $X^2(18, N = 2255) = 290.74, p < .01$ .

The test confirms, for example, that in scientific texts, *-mine* is used in the adessive and translative case considerably more often than expected, considering the case distributions of the other two suffixes. These cases are closely related to specific morphosyntactic constructions which require the regular meaning of an action noun. For example, based on an analysis of 1000 random concordances of *-mine* nouns from the scientific subcorpus, the adessive *-mine* nouns are mostly used in converb constructions (22), in the converbal Adessive Manner and Cause Construction as well as in the passive auxiliary construction described in Sakhai (2011).

- (22) *viimasest olen lähtunud esitatud muusikalise draama*  
 latter:ELA be:1SG follow:APP present:PPP musical.GEN drama.GEN  
*kujundamisel* (SCI)  
 design:NMLZ:ADE  
 ‘I have followed the latter when designing the musical drama presented’

All these constructions require using the prototypical action noun schema instead of metonymic or metaphorical extensions which do not entail a processual reading of the noun. Although in these constructions, *-us* nouns (and other action nouns) could also be used in a regular and processual manner (e.g., *draama kujundusel* drama.GEN design:NMLZ:ADE ‘when designing the drama’), they are far less likely to be used in the adessive than *-mine* nouns. If many *-us* types were used a lot in these constructions, this would result in higher productivity ratings of *-us*. Based on earlier observations (e.g., Sahkai 2011, Kasik 2015), however, the converbal constructions with *-us* are expected to be generalized over a limited number of low-level schemas, meaning that the speakers readily accept and use only some *-us* lexemes in such constructions. The converbal constructions as well as other syntactic constructions in which the *-mine* nouns occur were analyzed qualitatively in P3 using dialect data, which shows extensive variation with regard to the syntactic functions of action nouns. Concerning **RQ6**, I showed that in addition to the action nouns’ nominal syntactic distribution, a good amount of their behavior is better described in the context of larger syntactic units, where the *-mine* noun can function as the semantic and syntactic core of the clause (23, 24).

- (23) *eks ta tuulamesele läks* (DIA)  
 PCL it winnow:NMLZ:ALL go:PST.3SG  
 ‘it [the grain] was going to be winnowed’ (lit. ‘it went to winnowing’)

- (24) *siis läheb ta üles kiskumiseks* (DIA)  
 then go:3SG it.GEN up tear:NMLZ:TR  
 ‘then it will be torn up’ (lit. ‘then goes to its up-tearing’)

Such constructions simultaneously rely on and contribute to the productivity of the *-mine* morphological construction when they relate to other, verbal constructions on the same level of abstraction (e.g., 25, 26) through constructional/subpart links since they presuppose that the action noun is semantically coherent with other action nouns and also with the lexical meaning(s) of the base.

- (25) *eks teda hakati tuulama*  
 PCL it:PAR begin:IPS:PST winnow:2INF  
 ‘it [the grain] was going to be winnowed’

- (26) *siis kisutakse ta üles / siis hakatakse teda üles kiskuma*  
 then tear:IPS it up then begin:IPS it:PAR up tear:2INF  
 ‘then it will be torn up’

At the same time, the use of different syntactic constructions can also have a negative effect on the quantitative assessment of morphological productivity when the productivity of the syntactic construction itself is low, but token frequency is high, as I will show in the upcoming examples. The interaction between syntactic and morphological productivity in cases where *-mine* nouns are used in ‘regular’ NPs is considerably less obvious since the higher-level syntactic constructions (for example, the adpositional phrase or the subject-predicate construction) do not usually presuppose strong lexical associations between the NPs which instantiate them.

I do not have comparable contextualized data for the use of the suffixes *-us* and *-ja* in the different registers nor is it feasible in this thesis to give a description of all the constructions in which such complex nouns occur. Therefore, the case distributions above are mainly used to advocate for the need to take case constructions into consideration in quantitative models of productivity. Although models of productivity would ideally also account for the paradigmatic relations as well as a wider morphosyntactic context, this modeling has not been attempted in this thesis due to a large amount of manual analysis required for such purposes, and small corpus samples. In such small samples, creating additional categories would yield dispersed and highly unbalanced datasets, problematic for most statistical analyses.

The importance of looking into different morphosyntactic constructions when analyzing productivity also becomes apparent when looking at the more frequent lexical constructions. One of the most frequent *-mine* nouns was *tegemine* ‘doing, making’, which by itself, seems to denote a very general, semantically unspecified process. Therefore, we can consider these tokens as instantiations whose reading is similar to the use of the corresponding verb forms (e.g., ex. 27), highlighting the transpositional function of word-formation.

(27) *autoteede paremaks tegemisest räägitakse meil kogu aeg* (NEWS)  
road:PL.GEN better:TR make:NMLZ:PAR speak:IPS we:ADE all time  
‘There is constant talk about making the roads better’

The regularity of *tegemine* should be supported especially by the frequent use of its base stem *tege-* in different inflectional verb forms. However, based on the contextually anchored samples of 1000 sentences, most of its uses in written registers and contemporary spoken language concern the instantiations of the predicative construction *olema* ‘be’ + *tegemist* ‘making’ +  $X_{\text{COM}}$  (e.g., ex. 28).

(28) *tegemist oli põrandaaluse kihlveokontoriga* (NEWS)  
make:NMLZ:PAR be:PST.3SG underground.GEN betting.GEN+office:COM  
‘It was an underground betting office’

The predicative in the comitative case identifies the subject complement, which can be inferred from the context (28) or explicitly expressed with an oblique phrase (29, 30).

(29) *nende puhul on tegemist puhtalt riistvara*  
 they:PL.GEN case:ADE be.3SG make:NMLZ:PAR purely hardware.PAR  
*teenindavate käskudega (SCI)*  
 servicing:PL.GEN command:PL:COM  
 ‘These are purely hardware servicing commands.’

(30) *siis on antud katkendi valiku näol juba*  
 then be.3SG given passage.GEN choice.GEN face:ADE already  
*tegemist tõlgendamise (SCI)*  
 make:NMLZ:PAR interpret:NMLZ:COM  
 ‘then the choice of a given passage is already an interpretation’

In this construction, the action noun *tegemist* can be replaced with a shorter deverbal action noun *tegu*. However, it would seem incorrect to claim that the action nouns themselves are lexicalized or instantiate some subschema of the more general action noun construction. Rather, it is the use of the whole construction that requires the action noun to be interpreted as semantically highly abstract and as forming a complex predicate with the copula. Interestingly, this predicative construction does not occur at all in dialect data, where the construction responsible for a little less than half of all the uses of *tegemine* is the impersonal-modal construction *olema* ‘be’ + *Vmine*<sub>NOM/PAR</sub> described in P3 and P4. While a subschema of this construction (ex. 31) can formally share similarities with the predicative construction, it has a completely different meaning.

(31) *no eks loomadega ole tegemest (DIA)*  
 PCL PCL animal:PL:COM be.3SG make:NMLZ:PAR  
 ‘Well, animals do require some work / animals keep you busy’

The aforementioned construction, which is relevant to **RQ6** and **RQ7**, is analyzed in P3 and P4 as the most frequent construction which cannot be considered as a combination of more higher-level constructions. The function of the construction is mainly to emphasize process over agent either by omitting the agent or by forcing it to the position of an adessive experiencer or a genitive modifier. Some such constructions also have a modal meaning which entails a judgment of obligation, permission, or necessity concerning the process referred to by the action noun. Perhaps one of the reasons why this construction has not been analyzed in more detail is its relative ambiguity, which makes it difficult to distinguish it from the more general EXISTENTIAL and EXPERIENCER constructions as well as from the high-level SUBJECT-PREDICATE construction. Indeed, the usage events which are considered to be licensed by this particular constructional schema in this thesis form a continuum from clear cases with modal readings (32) through mostly impersonal readings (33) to cases where their status as impersonal-modal constructions is questionable and conditioned by prosodic cues and previous discourse (34).

- (32) *põle kodo kanmest seda rimpsu* (DIA)  
 be.NEG.CNG to\_home carry:NMLZ:PAR this:PAR trash.PAR  
 ‘One doesn’t have to carry that trash home’
- (33) *ess ole meil rahagor’jam’ist* (DIA)  
 NEG.PST be.CNG we:ADE money.GEN+collect:NMLZ:PAR  
 ‘We did not collect money / ?we did not have to collect money’
- (34) *Petserimaal (.) säääl olli tökäti tegemine* (DIA)  
 Pechory:ADE there be:PST.3SG tar.GEN make:NMLZ  
 ‘In Pechory, there was tar making / tar was made there / there, the tar was made’

The nominalizations in these constructions are not considered ordinary noun phrases modifying the verb mainly because the use of the action noun is not conditioned by the lexical meaning of the finite verb and it is the action noun which conditions all other elements in the clause. For example, instead of the prenominal genitive modifier, the agent argument can be expressed as an adessive experiencer in all of the above constructions; the patient is usually incorporated in the action noun phrase (33), but interestingly, it can also be expressed as an adessive argument (see ex. 22 and 23 in P4), or in some cases, it can retain its sentential form (32), which suggests a similar functional status as with the non-finites in catenative constructions (e.g., ex. 35).

- (35) *ei pea seda rämpsus koju kandma*  
 NEG have\_to.CNG this:PAR trash:PAR to\_home carry:2INF  
 ‘One doesn’t have to carry that trash home.’

The action nouns in the constructions are not referential, i.e., they cannot be modified with a relative clause, and when adjectival modifiers are used, they refer to circumstances which would be expressed with an adverbial in the corresponding impersonal clause (compare ex. 36 and 37).

- (36) *ja siis oli kange viina vedamine* (DIA)  
 and then be:PST.3SG heavy vodka.GEN import:NMLZ  
 ‘and then vodka was heavily imported’
- (37) *ja siis veeti kangesti viina*  
 and then import:IPS:PST heavily vodka:PAR  
 ‘and then vodka was heavily imported’

The analysis in P4 suggested that the construction is semantically fairly general and regular, although still relatively infrequently used in the dialect data (compared to e.g., various verb + non-finite verb constructions), which is why its productivity was not quantitatively assessed or compared against that of alternative, more general means of expression. Since in periphrastic and catenative constructions the finite element is usually semantically general and its function is to express the grammatical categories of mood and tense (or manner in some cases),

the meaning of the action noun has to correspond to that of the base. We can therefore establish a link between semantic regularity and the use of action nouns in periphrastic and catenative constructions (this was also suggested in Sahkai 2011). In turn, the use of such constructions promotes higher productivity of the general action noun schema if we are not dealing with constructional idioms. While the syntactic constructions themselves can be productive, they can also be instantiated by lexical sequences which become automatized due to higher usage frequency and therefore might acquire a separate status as an individual sub-schema in processing. Some instances of the *olema* ‘be’ + *Vmine*<sub>NOM/PAR</sub> construction have become partially idiomatic, e.g., *A*<sub>ADE</sub> + *olema* ‘be’ + *pistmist/tegemist* ‘sticking/making’ + *X*<sub>COM</sub> (e.g., ex. 38) or *A*<sub>ADE</sub> + *olema* ‘be’ + *tegemist* ‘making’ + *X*<sub>COM</sub> (e.g., ex. 39).

(38) *sel ei ole minuga pistmist*  
 this:ADE NEG be.CNG I:COM stick:NMLZ:PAR  
 ‘this has nothing to do with me’ (lit. ‘this does not have sticking with me’)

(39) *mul on lastega tegemist*  
 I:ADE be.3SG child:PL:COM make:NMLZ:PAR  
 ‘I am busy with the children’ (lit. ‘I have making with the children’)

The action noun slots in the constructions are more or less lexically fixed or at least restricted to a small set of lexemes and the constructions express some kind of a close interaction between A and X.

The individual constructional status of *olema* ‘be’ + *Vmine*<sub>NOM/PAR</sub> is further supported by the existence of similar modal constructions in other Finnic languages. In fact, it is suggested in P4 that the construction may have developed from a proto-Finnic transitive modal construction (Grünthal 1941: 177–179) and has become semantically more vague in Estonian, even licensing intransitive verbs. Due to a lack of diachronic data, this cannot be reliably proved; it is also possible that the two readings of the construction have co-existed and developed in parallel. When browsing the corpus of old literary Estonian<sup>23</sup>, it becomes apparent that both readings of the construction existed long before the time when the dialect corpus recordings were made. For example, both the impersonal (40, 41) and modal constructions (42) can be found in the written texts from the 17<sup>th</sup> century.

(40) (Blume, Christoph, 1666. Geistliche Wochen-Arbeit, pp. 80)  
*ni kauwa kud mul sihn wêel ellamist on*  
 so long as I:ADE here still live:NMLZ:PAR be.3SG  
 ‘as long as I am still living’ (lit. ‘as long as I still have living here’)

<sup>23</sup> <https://vakk.ut.ee/> (Accessed on 26/02/2021)

- (41) (Rossihnius, Joachim, 1632. Catechismus Herrn D. MARTINI LUTHERI, pp. 37)  
*ninck ei olle teṃa man üttekitt kajemist se*  
 and NEG be.CNG he.GEN by not\_any:PAR look:NMLZ:PAR that.GEN  
*Inimesse suhresse pähle.*  
 person.GEN greatness.GEN onto  
 ‘and the person’s greatness does not matter there’ (lit. ‘and there is no looking at the person’s greatness with him / at his place’)
- (42) (Stahl, Heinrich, 1641. Leyen Spiegel, part I, pp. 156)  
*echk temmal küll kuhstöistkümmē pennekohrm teed*  
 although he:ADE PCL sixteen Scandinavian\_mile.GEN road:PAR  
*keimist olli*  
 walk:NMLZ:PAR be:PST.3SG  
 ‘although he had to walk for sixteen Scandinavian miles’ (lit. ‘although he had walking of sixteen Scandinavian miles worth of road’)

To conclude, with respect to **RQ7**, the dual relationship between morphological and syntactic productivity can be examined by analyzing the syntactic constructions in which the morphological constructions occur. This can, among other things, shed light on the different functions in which the same lexical construction is used in different registers, consequently affecting the lexical construction’s contribution to the productivity of the relevant morphological schema.

## 5.5 Limitations and further research

There are some limitations to the studies presented in this thesis. Some of them originate from the practical need to simplify data collection and analysis, some from the assumptions of the quantitative techniques used, or others more generally from the complexity that inevitably emerges when examining derived nouns on different levels of the proposed constructional network.

With regard to the quantitative measures of productivity, the most immediate issue concerns the status of the *hapax legomena*. The *hapax*-based productivity measures attribute a special status to types occurring only once in a given corpus since they are the most likely frequency class to contain neologisms (Baayen & Renouf 1996, Plag et al. 1999). Neologisms, in turn, should reflect the most regular use of a schematic construction since their formation does not receive any support from usage frequency and automatization. Naturally, this assumption simplifies things since not all *hapax legomena* are neologisms and not all neologisms are *hapax legomena*. While the number of *hapax legomena* starts to approximate the number of neologisms in sufficiently large samples, the issue is more serious with small samples (see Säily & Suomela 2009, Säily 2011) where words occurring only once can simply be rare by chance and in fact familiar to the speakers (**RQ4**). For example, *hapax legomena* in the NEWS sample include words such as *küsimine* ‘asking’, *unustamine* ‘forgetting’ and *ennustamine* ‘predicting’, which are known, regular formations unlikely to surprise anyone in discourse and not created out of

a pragmatic need to express a novel concept (cf. Baayen & Lieber 1991, Baayen 1994: 453, Gaeta & Ricca 2006: 68). Using *hapax legomena* to assess the probability that a category is extended with new formations could therefore provide unreliable results. While it is definitely desirable to test the measures on larger samples, provided that such samples can be accessed for spoken Estonian, using *hapax legomena* as substitutes for rare formations has proved a useful and accessible heuristic for comparing morphological constructions, yielding results which correspond to general linguistic intuitions. They make it possible to quantify the potential the morphological category has to include rare formations and as such, these measures are linguistically interpretable, intuitively coherent, consistent, and most importantly, comparable (Zeldes 2012: 68).

For technical and resource-related reasons, only the simple lemmas of the derivatives were analyzed in the first two case studies on morphological productivity. Considering that lexical representations are highly context-sensitive (Marzi et al. 2020), this means disregarding a large amount of information which plays an important role in lexical processing. In addition to neglecting multi-word expressions as different from individual monolexical bases (e.g., *tähele panema* ‘notice’ vs. *kappi panema* ‘put in the closet’, see Kasik 2015: 17), I have also ignored the fact that derivation often occurs in combination with compounding (Vare 1994: 28). This is a robust approach often resorted to in quantitative studies due to the amount of manual work and subjective assessments involved in a more detailed analysis. Ignoring compounding, however, also has a theoretical motivation since accounting for them in the studies as distinct categories would have made it difficult to assess whether the used measures reflect the productivity of the derivation pattern, the productivity of compounding, or the salience of the participants of the process incorporated into the noun phrase. Still, in more complex models of productivity, it would be desirable to account for the complete structure of the formations as the specific lexical sequences have been shown to be highly relevant for morphological processing (see e.g., Diessel 2019). This means that, as shown, also including case information and, consequently, syntactic functions is expected to influence the assessment of morphological productivity and the paradigmatic relations among word families. It is quite clear that in addition to keeping to the conventions of the register, speakers have preferences both with regard to which low-level morphological constructions they choose in a desired syntactic setting as well as to which syntactic construction they pick for expressing a word in a sentence (see Barðdal 2008). The size of the word family itself could be used as a predictor of productivity.

In this thesis, I have considered the semantic aspect of morphological and syntactic constructions only from a synchronic perspective, but according to the usage-based model, constructions are not stable and are constantly changing both in terms of form and meaning (see Baayen & Renouf 1996). Word-formation trends and functions have also changed in Estonian over time (see Kerge 2002, 2003), both as a natural development within a linguistic community as well as resulting from the ideals of language planning. For example, in Kerge (2003) it is shown that at the beginning of the 20th century, the syntactic function of *-mine*

nouns prevailed over the lexical functions and many derivatives were used which have now been replaced with shorter patterns (e.g., *liiklemine* was used for *liiklus* ‘traffic’). This tendency is still visible in the spoken dialect data from the 1960s where forms such as *juhtumine*, *söömine*, *arvamine* were used instead of *juhus* ‘coincidence’, *söök* ‘food’, *arvamus* ‘opinion’, making the comparison of dialect data with contemporary data from a synchronic perspective somehow unfair. However, including older spoken data enables us to see that aspect of variation that is largely lost in contemporary language, which is why the contribution of dialect data to this thesis is very valuable.

Several critiques have been put forward concerning both corpus studies and usage-based studies which largely rely on frequency data. These objections have involved e.g., the inadequate size and data sparsity of corpora, the underrepresentativeness of the texts, the lack of negative evidence (i.e., impossible forms/constructions), oversimplifying, etc. (see Lemmens 2019). While most linguists working with corpus data readily agree with all the shortcomings of corpora, they are also highly aware of the limitations imposed by using other data sources (introspection, conversational data, experimental data, etc.). Obviously, frequency data in a corpus, based on collected observations, can be used only as indirect evidence of lexical processing and the cognitive processes involved. The methods used have been shown to quite adequately reflect speakers’ knowledge. However, they do not necessarily reflect human performance, i.e., how speakers and listeners use this knowledge. The analysis of corpus data is crucial for putting forward hypotheses, which can then be checked in experimental psycholinguistic studies. There are indeed multiple further avenues which could improve the understanding of productivity in a complex network of morphological and syntactic constructions.

For example, it is possible to analyze equal-sized samples instead of samples of cumulatively progressive size in order to compare the variance of the productivity measures for different suffixes in different registers and establish whether there is a statistically significant difference (Zeldes 2012). There are also more sophisticated models for analyzing word frequency distributions such as the LNRE (Large Number of Rare Events) models (Baayen 2001, Evert 2004, Evert & Baroni 2007), which are suited for describing frequency distributions where many types are rare and only a few types are frequent. More advanced statistical techniques (regression modeling and GAMs (Wood 2017), learning models, such as Naive Discriminative Learning (Baayen 2011, Baayen et al. 2011) or Random Forests (Hothorn et al. 2006)) would allow for factoring in more variables in studying the productivity of one or more patterns. For example, productivity has been shown to also be influenced by socio-geographical forces (Keune et al. 2006, Säily & Suomela 2009, Säily 2011), and the effect of individual preferences has also been demonstrated (De Smet 2020). These more advanced statistical techniques would require much more data than what was available for my studies in order to make adequate inferences. A complex statistical analysis on small samples does not provide the benefits it does when applied to large homogenous corpora. Distributional models and artificial neural networks, such as Word2Vec (Mikolov

et al. 2013) could be used for accounting for the semantic similarity between different words, assessing the (de)compositionality of a given derivation pattern, or detecting polysemous derivatives, given a sufficiently large amount of high-quality data. Distributional models (also known as vector-space models) rely on the distributional hypothesis, according to which semantic similarity can be detected based on the context in which the words occur. One potential problem with distributional models, besides the requirement for large volumes of high-quality input data, is that many usage events for a lexeme are required in order for its semantic representation to be adequate, especially when similarities are required for multiple word senses rather than one prominent or generalized sense alone. As shown in P2, however, suffixal rivalry is more likely to be found among formations occurring with low frequency, and low-frequency formations are also more likely to contribute to the productivity of more schematic constructions. Therefore, the advantages and applications of distributional models should be approached with caution. Combining corpus data from different registers with experimental techniques would enable us to focus not only on past, passive production but also on active production and lexical processing. With regard to e.g., acceptability ratings, native speaker intuitions can be considered reliable when asked whether a complex word is a possible word in a language but they should be approached with caution when the question is about the familiarity of the words or the likelihood of the word being a lexical innovation (Baayen & Renouf 1996: 75). Finally, there are interesting advances in neurophysiology which suggest that different word classes are processed in distinct cerebral areas of the brain (see Marangolo & Papagno 2020). This could have interesting implications for the analysis of the more verb- or noun-like categorization of deverbal nominalizations in different syntactic environments and/or with different lexical specifications.

## 6 CONCLUSION

This thesis was born out of an interest for periphrastic constructions headed by *-mine* action nouns and grew into a more general interest concerning the productivity and functions of *-mine* nouns in spoken and written discourse. In the course of conducting the four case studies with very different perspectives and varying degrees of detail, I felt a growing level of discomfort with having to constantly navigate and choose between theories, concepts, and tools of morphology, syntax, and lexicology, while essentially describing one and the same (admittedly complex) linguistic phenomenon. This thesis is an attempt to join the implications from the case studies under one constructionalist framework.

The thesis consists of an overview section and four publications which deal with *-mine* morphological construction, its functions, and its productivity from morphological and syntactic perspectives. By combining qualitative and quantitative analyses, the thesis provides a usage-based account of the role of derivational morphology not only in written registers but also in spoken registers, which have so far largely been neglected in studies of Estonian morphology in general. The main conclusions of the thesis with regard to the seven proposed research questions are summarized in the following.

**First**, the analyses revealed that morphological productivity as reflected in relative corpus frequencies of more productive derivation suffixes can indeed be used to distinguish between written and spoken registers, but mainly in the context of past, realized productivity, whereas the dimensions of formality and informality emerge as more salient features with regard to the potential to expand a morphological construction to include new lexical items. **Second**, the reason why written registers exhibit higher realized productivity – manifested as the number of types in a given category in a given corpus – may lie in the higher pragmatic need for the concepts or processes expressed with deverbal nouns, but also in the fact that written registers are more nominal as a whole. The diverging productivity of the suffixes in different registers results from alternating forces between structural-semantic and pragmatic-paradigmatic constraints. **Third**, *-mine* is clearly the most productive suffix by any measure of productivity, but the magnitude by which its productivity surpasses that of the other two suffixes analyzed (*-us* and *-ja*) differs across registers. **Fourth**, the attested differences between suffixes are also reliant on the approach chosen for the quantitative assessments. For small samples and suffixes with very different token frequencies, all three tested approaches called for certain reservations, whether with regard to their theoretical groundedness or their coherence with linguistic intuitions. Although these quantitative measures prove to be a useful and accessible heuristic, they must be interpreted with caution and not in an unequivocal manner. The *hapax*-based measures in particular are conceptually sensitive to sample sizes. **Fifth**, paradigmatic relations in the constructional network as assessed via relative frequencies between bases and suffixes as well as between rival affixes provide a way to obtain a global overview of the dynamics between different morphological

processes but fail to be very informative on semantic grounds for specific low-level constructions when compared with fully decontextualized words. **Sixth**, older spoken dialect data exhibit rich variation with regard to the combinatorial properties of *-mine* action nouns. These *-mine* nouns are often better described in the context of specific syntactic constructions where they do not behave like regular nouns or noun phrases. The attested syntactic constructions seem to facilitate a regular use of the morphological pattern, thereby suggesting that syntactic constructions are indeed co-determinants of morphological productivity, as words are rarely coined in a decontextualized manner in corpus data. **Finally**, the most frequent action noun construction in the dialect data, the *olema* ‘be’ + *Vmine*<sub>NOM/PAR</sub> construction, which functions as a means for impersonalization and/or expressing speaker stance with regard to the necessity of the process expressed with the action noun, exemplifies the various hierarchical, filler-slot and subpart relations in grammar as a constructional network.

The thesis contributes first and foremost to the study of Estonian word-formation processes by providing little-seen empirical evidence of the quantitative output of productivity and its driving mechanisms in different registers, including spoken data. Although the quantitative measures used in this thesis have become the standard tools in morphological studies, the comparison of productivity in multiple registers is still rare in a cross-linguistic perspective. It is also not often the case that the individual lexemes contributing to or constraining productivity are investigated in more detail in corpus-based quantitative studies about morphological productivity, let alone analyzing the effect of syntactic constructions in which the complex words occur. The attempt to combine the study of action noun constructions on multiple levels can therefore be considered as an original contribution of this thesis.

The thesis also lays a foundation for multiple, potentially very fruitful avenues for further research, and advocates a desperate need for including more empirical data in the study of Estonian derivational morphology.

## REFERENCES

- Aedmaa, Eleri. 2019. *Detecting Compositionality of Estonian Particle Verbs with Statistical and Linguistic Methods*. (Dissertationes linguisticae Universitatis Tartuensis 37.) Tartu: University of Tartu Press.
- Aronoff, Mark. 1976. *Word Formation in Generative Grammar*. Cambridge, MA: The MIT Press.
- Aronoff, Mark. 1983. Potential words, actual words, productivity and frequency. *Preprints of the Plenary Session Papers of the XIIIth International Congress of Linguists. Tokyo*. 163–170.
- Baayen, R. Harald. 1989. *A Corpus-Based Approach to Morphological Productivity. Statistical Analysis and Psycholinguistic Interpretation*. Amsterdam: Vrije Universiteit Amsterdam. Doctoral dissertation.
- Baayen, R. Harald. 1992. A quantitative approach to morphological productivity. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1991*. 109–149. Dordrecht: Kluwer.
- Baayen, R. Harald. 1993. On frequency, transparency and productivity. In Geert E. Booij & Jaap van Marle (eds.), *Yearbook of Morphology 1992*. 181–208. Dordrecht: Kluwer.
- Baayen, R. Harald. 1994. Productivity in language production. *Language and Cognitive Processes* 9(3). 447–469.
- Baayen, R. Harald. 1996. The effect of lexical specialization on the growth curve of the vocabulary. *Computational Linguistics* 22. 455–480.
- Baayen, R. Harald. 2001. *Word Frequency Distributions*. (Text, Speech and Language Technology 18.) Dordrecht: Kluwer.
- Baayen, R. Harald. 2009. Corpus linguistics in morphology: Morphological productivity. In Anke Lüdeling & Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*. 900–919. Berlin: Mouton de Gruyter.
- Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada* 11(2). 295–328.
- Baayen, R. Harald & Rochelle Lieber. 1991. Productivity and English derivation: A corpus-based study. *Linguistics* 29. 801–834.
- Baayen, R. Harald & Antoinette Renouf. 1996. Chronicling The Times: Productive lexical innovations in an English newspaper. *Language* 72(1). 69–96.
- Baayen, R. Harald, Petar Milin, Dusica Filipović Đurđević, Peter Hendrix & Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review* 118(3). 438–481.
- Barðdal, Jóhanna. 2008. *Productivity: Evidence from Case and Argument Structure in Icelandic*. (Constructional Approaches to Language 8.) Amsterdam: John Benjamins.
- Bauer, Laurie. 2001. *Morphological Productivity*. Cambridge: Cambridge University Press.
- Bauer, Laurie. 2004. The function of word-formation and the inflection-derivation distinction. In Henk Aertsen, Mike Hannay & Rod Lyall (eds.), *Words in their Places. A Festschrift for J. Lachlan Mackenzie*. 283–292. Amsterdam: Vrije Universiteit. 2004.
- Bauer, Laurie. 2005. Productivity: Theories. In Pavol Štekauer & Rochelle Lieber (eds.), *Handbook of Word-Formation*. (Studies in Natural Language & Linguistic Theory 64.) 315–334. Dordrecht: Springer.
- Biber, Douglas. 2010. What can a corpus tell us about registers and genres? In Anne O’Keefe & Michael McCarthy (eds.), *The Routledge Handbook of Corpus Linguistics*. 241–254. London: Routledge.

- Biber, Douglas & Susan Conrad. 2019. *Register, Genre, and Style*. 2nd edn. Cambridge: Cambridge University Press.
- Blumenthal-Dramé, Alice. 2012. *Entrenchment in Usage-Based Theories: What Corpus Data Do and Do not Reveal about the Mind*. (Topics in English Linguistics 83.) Berlin: De Gruyter Mouton.
- Bolinger, Dwight. 1948. On defining the morpheme. *Word* 4. 18–23.
- Booij, Geert. 2010. *Construction Morphology*. Oxford: Oxford University Press.
- Booij, Geert. 2019. The construction of words: introduction and overview. In Geert Booij (ed.), *The Construction of Words: Advances in Construction Morphology*. (Studies in Morphology 4.) 3–16. Cham: Springer.
- Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10. 425–455.
- Bybee, Joan. 2010. *Language, Usage, and Cognition*. Cambridge: Cambridge University Press.
- Bybee, Joan. 2013. Usage-based theory and exemplar representations of constructions. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. 49–69. Oxford: Oxford University Press.
- Comrie, Bernard. 1976. The syntax of action nominals: A cross-linguistic study. *Lingua* 40. 177–201.
- Cowie, Claire & Christiane Dalton-Puffer. 2002. Diachronic word-formation and studying changes in productivity over time: Theoretical and methodological considerations. In Javier E. Díaz Vera (ed.), *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics*. 410–437. Amsterdam: Rodopi.
- Croft, William A. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.
- Croft, William A. & D. Alan Cruse. 2004. *Cognitive Linguistics*. Cambridge: Cambridge University Press.
- De Smet, Hendrik. 2020. What predicts productivity? Theory meets individuals. *Cognitive Linguistics* 31(2). 251–278.
- Diessel, Holger. 2019. *The Grammar Network: How Linguistic Structure is Shaped by Language Use*. Cambridge: Cambridge University Press.
- Dixon, Robert M. W. 1997. *The Rise and Fall of Languages*. Cambridge: Cambridge University Press.
- Dryer, Matthew S. 2006. Descriptive theories, explanatory theories, and basic linguistic theory. In Felix Ameka, Alan Dench & Nicholas Evans (eds.), *Catching Language: Issues in Grammar Writing*. 207–234. Berlin: Mouton de Gruyter.
- Erelt, Mati. 2014. *Eesti keele lauseõpetus: Komplekslause*. Tartu: University of Tartu.
- Erelt, Mati. 2017. Sekundaartarinditega laused. In Mati Erelt & Helle Metslang (eds.), *Eesti keele süntaks*. (Eesti keele varamu 3.) 756–840. Tartu: Tartu Ülikooli Kirjastus.
- Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael & Silvi Vare. 1993. *Eesti keele grammatika II. Süntaks. Lisa: Kiri*. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.
- Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, & Silvi Vare. 1995. *Eesti keele grammatika I. Morfoloogia. Sõnamoodustus*. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.
- Evert, Stefan. 2004. A simple LNRE model for random character sequences. *Proceedings of JADT 2004*. 411–422.

- Evert, Stefan & Marco Baroni. 2007. zipfR: Word frequency distributions in R. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Posters and Demonstrations Session*. 29–32.
- Fillmore, Charles J., Russell R. Lee-Goldman & Russell Rhodes. 2012. The FrameNet construction. In Hans C. Boas & Ivan A. Sag (eds.), *Sign-Based Construction Grammar*. 283–299. Stanford: CSLI.
- Gaeta, Livio & Davide Ricca. 2003. Italian prefixes and productivity: a quantitative approach. *Acta Linguistica Hungarica* 50(1–2). 93–112.
- Gaeta, Livio & Davide Ricca. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 44(1). 57–89.
- Gaeta, Livio & Davide Ricca. 2015. Productivity. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-Formation: An International Handbook of the Languages of Europe. Vol. 2*. 842–858. Berlin: De Gruyter.
- Goldberg, Adele E. 1995. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press.
- Goldberg, Adele E. 2003. Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7(5). 219–224.
- Goldberg, Adele E. 2006. *Constructions at Work: The Nature of Generalization in Language*. Oxford: Oxford University Press.
- Goldberg, Adele E. 2013. Constructionist approaches. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. 19–31. Oxford: Oxford University Press.
- Grünthal, Willem. 1941. *Itämerensuomalaisten kielten yksikön nominatiivi objektin edustajana aktiivin yhteydessä: lauseopillinen tutkimuskoe*. Helsinki.
- Halliday, Michael A. K. 1994. *An Introduction to Functional Grammar*. 2nd edn. London: Edward Arnold.
- Haspelmath, Martin. 1995. The converb as a cross-linguistically valid category. In Martin Haspelmath & Ekkehard König (eds.), *Converbs in Cross-Linguistic Perspective*. 1–55. Berlin-Mouton de Gruyter.
- Hardie, Andrew & Tony McEnery. 2006. Statistics. In Keith Brown (ed.), *Encyclopaedia of Language and Linguistics*. 138–146. Oxford: Elsevier.
- Hay, Jennifer. 2001. Lexical frequency in morphology: is everything relative? *Linguistics* 39(6). 1041–1070.
- Hay, Jennifer. 2002. From speech perception to morphology: affix-ordering revisited. *Language* 78(3). 527–555.
- Hay, Jennifer & R. Harald Baayen. 2002. Parsing and productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2001*. 203–235. Dordrecht: Kluwer.
- Hay, Jennifer & R. Harald Baayen. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342–348.
- Hay, Jennifer & Ingo Plag. 2004. What constrains possible suffix combinations? On interaction of grammatical and processing restrictions in derivational morphology. *Natural Language and Linguistic Theory* 22. 565–596.
- Hennoste, Tiit 2000a. Sissejuhatus suulisesse eesti keelde II. Suulise kõne erisõnavara 1. *Akadeemia* 6. 1343–1374.
- Hennoste, Tiit 2000b. Sissejuhatus suulisesse eesti keele III. Suulise kõne erisõnavara 2. *Akadeemia* 7. 1553–1582.
- Hennoste, Tiit. 2000c. Allkeeled. In Tiit Hennoste (ed.), *Eesti keele allkeeled*. (Tartu Ülikooli eesti keele õppetooli toimetised 16.) 9–56. Tartu: Tartu Ülikool.

- Hilpert, Martin. 2014. *Construction Grammar and its Application to English*. Edinburgh: Edinburgh University Press.
- Hoffmann, Thomas & Graeme Trousdale. 2013. Construction grammar: Introduction. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.
- Hothorn, Torsten, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro & Mark van der Laan. 2006. Survival ensembles. *Biostatistics* 7(3). 355–373.
- Janda, Laura A. 2010. The role of metonymy in Czech word-formation. *Slovo a slovesnost* 71(4). 260–274.
- Kallas, Jelena & Kristina Koppel. 2020. Eesti keele ühendkorpus 2019 / Estonian National Corpus 2019. Eesti Keeleressursside Keskus.
- Kasik, Reet 1968. Omastavalise täiendi subjektilisusest ja objektilisusest *mine*-konstruktsioonis. *Keel ja struktuur* 2. 125–139. Tartu: Tartu Riiklik Ülikool.
- Kasik, Reet. 1975. Verbide ja verbaalsubstantiivide tuletusvahekorrad tänapäeva eesti keeles. *Keele modelleerimise probleeme* 5. 4–162. Tartu: Tartu Riiklik Ülikool.
- Kasik, Reet. 1979. *Eesti keele tuletusõpetus I. Substantiivituletus*. Tartu: Tartu Riiklik Ülikool.
- Kasik, Reet 1994. *Verbid ja verbaalsubstantiivid tänapäeva eesti keeles: tuletusprotsess ja tähendus*. (Dissertationes philologiae estonicae Universitatis Tartuensis 5.) Tartu: Tartu Ülikool.
- Kasik, Reet. 2004. *Eesti keele sõnatuletus*. Teine, parandatud trükk. Tartu: Tartu Ülikooli Kirjastus.
- Kasik, Reet. 2006. Nominalisatsioon meediaaudiste tekstimoodustusvõttena. *Keel ja Kirjandus* 2. 122–134.
- Kasik, Reet. 2011. Sõnatuletus leksika ja grammatika vahel: *nd-* ja *ndus-*liitelised verbaalnoomenid. *Emakeele Seltsi aastaraamat* 56(2010). 63–90.
- Kasik, Reet. 2015. *Sõnamoodustus*. (Eesti keele varamu I). Tartu: Tartu Ülikooli Kirjastus.
- Kasik, Reet, Silvi Vare & Krista Kerge. 2002. Tänapäeva eesti kirjakeele uurimine. Sõnamoodustus. *Emakeele Seltsi aastaraamat* 48(2001). 49–62.
- Kastovsky, Dieter. 1986. The problem of productivity in word-formation. *Linguistics* 24. 585–600.
- Keevallik, Leelo 2003. V. Colloquial Estonian. In Mati Ereht (ed.), *Estonian Language*. (Linguistica Uralica Supplementary Series 1.) 343–378. Tallinn: Estonian Academy Publishers.
- Kerge, Krista. 2002. Kirjakeele kasutusvaldkondade süntaktiline keerukus. In Reet Kasik (ed.), *Tekstid ja taustad: artikleid tekstianalüüsist*. (Tartu Ülikooli eesti keele õppetooli toimetised.) 29–46. Tartu: Tartu Ülikooli Kirjastus.
- Kerge, Krista. 2003. *Keele variatiivsus ja mine-tuletus allkeelte süntaktilise keerukuse tegurina*. Tallinn: TPÜ kirjastus. Ph.D. thesis.
- Kerge, Krista. 2001. Action nouns on *-mine* in Estonian grammar and lexicon. In Jarkko Niemi & Janne Heikkinen (eds.), *Nordic and Baltic morphology: Papers from a NorFA Course, Tartu, June 2000*. (Studies in Languages 36.) 34–45. Joensuu: University of Joensuu.
- Keune, Karen, Roeland van Hout & R. Harald Baayen. 2006. Socio-geographical variation in morphological productivity in spoken Dutch: A comparison of statistical techniques. *Proceedings of JADT 2006*. 571–580.
- Koptjevskaja-Tamm, Maria. 1993. *Nominalizations*. London: Routledge.
- Koptjevskaja-Tamm, Maria. 2003. Action nominal constructions in the languages of Europe. In Frans Plank (ed.), *Noun phrase structure in the languages of Europe*.

- (Empirical Approaches to Language Typology 20(7).) 723–760. Berlin/New York: Mouton de Gruyter.
- Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Vol. 1. Theoretical Prerequisites*. Stanford: Stanford University Press.
- Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar. Vol. 2. Descriptive Application*. Stanford: Stanford University Press.
- Lemmens, Maarten. 2019. In defense of frequency generalizations and usage-based linguistics. An answer to Frederick Newmeyer’s ‘Conversational corpora : when big is beautiful’. *CogniTextes* 19.
- Marangolo, Paola & Costanza Papagno. 2020. Neuroscientific protocols for exploring the mental lexicon: Evidence from aphasia. In Vito Pirrelli, Ingo Plag & Wolfgang U. Dressler (eds.), *Word Knowledge and Word Usage*. (Trends in Linguistics. Studies and Monographs 337.) 127–166. Berlin/Boston: Walter de Gruyter.
- Marzi, Claudia, James P. Blevins, Geert Booij & Vito Pirrelli. 2020. Inflection at the morphology-syntax interface. In Vito Pirrelli, Ingo Plag & Wolfgang U. Dressler (eds.), *Word Knowledge and Word Usage*. (Trends in Linguistics. Studies and Monographs 337.) 228–294. Berlin/Boston: Walter de Gruyter.
- Melissourgou, Maria N. & Katerina T. Frantzi. 2017. Genre identification based on SFL principles: The representation of text types and genres in English language teaching material. *Corpus Pragmatics* 1(4). 373–392.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*. 3111–3119.
- Neetar, Helmi. 1988. Mõnest teonimekonstruktsioonist Eesti murretes. *Emakeele Seltsi aastaraamat* 32(1986). 36–45.
- Neetar, Helmi. 1990. *Deverbaalne nominaaltuletus eesti murretes. I*. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.
- Nevalainen, Terttu & Helena Raumolin-Brunberg. 2003. *Historical Sociolinguistics: Language Change in Tudor and Stuart England*. (Longman Linguistics Library.) London: Pearson Education.
- Noonan, Michael. 1997. Versatile nominalizations. In Joan Bybee, John Haiman & Sandra A. Thompson (eds.), *Essays on Language Function and Language Type: Dedicated to T. Givón*. 373–394. Amsterdam: John Benjamins.
- Perek, Florent. 2015. *Argument Structure in Usage-Based Construction Grammar: Experimental and Corpus-Based Perspectives*. (Constructional Approaches to Language 17.) Amsterdam: John Benjamins.
- Pirrelli, Vito, Ingo Plag & Wolfgang U. Dressler. 2020. Word knowledge in a cross-disciplinary world. In Vito Pirrelli, Ingo Plag & Wolfgang U. Dressler (eds.), *Word Knowledge and Word Usage*. (Trends in Linguistics. Studies and Monographs 337.) 1–22. Berlin/Boston: Walter de Gruyter.
- Plag, Ingo. 1999. *Morphological Productivity: Structural Constraints in English Derivation*. Berlin, New York: Mouton de Gruyter.
- Plag, Ingo. 2018. *Word-Formation in English*. 2nd edn. Cambridge: Cambridge University Press.
- Plag, Ingo, Christiane Dalton-Puffer & R. Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3(2). 209–228.
- R Core Team. 2020. *R: A language and environment for statistical computing* (R Foundation for Statistical Computing). Austria: Vienna. <https://www.R-project.org/>.

- Rácz, Péter, Viktória Papp & Jennifer B. Hay. 2016. Frequency and corpora. In Andrew Hippisley & Gregory Stump (eds.), *The Cambridge Handbook of Morphology*. 685–709. Cambridge: Cambridge University Press.
- Saari, Henn. 1997. *Ein Weg zur Wortgrammatik am Beispiel des Estnischen. Erster Teil*. (Eesti Keele Instituudi toimetised 1.) Tallinn.
- Sahkai, Heete. 2011. *Teine grammatika: Eesti keele teonimede süntaks konstruktiooni-põhises perspektiivis*. Tallinn: Tallinna Ülikool. Ph.D. thesis.
- Schultink, Henk. 1961. Produktiviteit als morfologisch fenomeen. *Forum der Letteren* 2. 110–125.
- Shen, Tian & R. Harald Baayen. 2021. Adjective-noun compounds in Mandarin: A study on productivity. *Corpus Linguistics and Linguistic Theory*.
- Stefanowitsch, Anatol. 2013. Collostructional Analysis. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.
- Säily, Tanja. 2011. Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory* 7(1). 119–141.
- Säily, Tanja & Jukka Suomela. 2009. Comparing type counts: The case of women, men and *-ity* in early English letters. In Antoinette Renouf & Andrew Kehoe (eds.), *Corpus Linguistics: Refinements and Reassessments*. (Language and Computers: Studies in Practical Linguistics 69.) 87–109. Amsterdam: Rodopi.
- Zeldes, Amir. 2012. *Productivity in Argument Selection. From Morphology to Syntax*. (Trends in Linguistics. Studies and Monographs 260.) Berlin: De Gruyter Mouton.
- Vare, Silvi. 1991. Kazuto Matsumura ettekandest Debreceni kongressil ning verbi ja verbaalnoomeni vahekorra. *Keel ja Kirjandus* 7. 408–415.
- Vare, Silvi. 1993. Tuletusliitest uue nurga alt. *Keel ja Kirjandus* 8. 530–536.
- Vare, Silvi. 1994. *Nimi- ja omadussõnatuletus tänapäeva kirjakeeles*. Tartu: Tartu Ülikool. Doctoral dissertation.
- Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R*. 2nd edn. Philadelphia, PA: CRC Press.
- Östman, Jan-Ola & Graeme Trousdale. 2013. Dialects, discourse, and construction grammar. In Thomas Hoffmann & Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*. Oxford: Oxford University Press.
- Yap, Foong Ha, Karen Grunow-Hårsta & Janick Wrona (eds.). 2011. *Nominalization in Asian Languages: Diachronic and Typological Perspectives*. (Typological Studies in Language 96.) Amsterdam: John Benjamins.
- Ylikoski, Jussi. 2003. Defining non-finites: action nominals, converbs and infinitives. *SKY Journal of Linguistics* 16. 185–237.

## SUMMARY IN ESTONIAN

### Teonimed konstruksioonide võrgustikus: korpuspõhine uurimus *mine*-deverbaalsufiksi produktiivsusest ja funktsioonidest viies eesti keele registris

Verbidest nimisõnu tuletavat *mine*-sufiksit peetakse sageli eesti keele produktiivseimaks tuletusliiteks, eeskätt seetõttu, et oma reeglipärasuses asetub *mine*-tuletusmall infleksiooni ja derivatsiooni piirimaile ning lubab teonimesid samal viisil tuletada teoreetiliselt lõpmatust hulgast verbidest. Nõnda peeti 1933. aastani *mine*-liidet suisa verbi infleksioonilise paradigma osaks (Kasik 2015: 69, Erelt 2017: 831). Hilisemates käsitlustes on *-mine* arvatud siiski tuletusliidete hulka ning selle süntaktilisi omadusi ja kombineerumisvõimalusi eesti keele grammatika-kirjeldustes käsitletud pigem nimisõnade klassist ja sellega seotud fraasistruktuurireeglitest lähtuvalt (nt Erelt jt 1993: 269–271, Erelt 2014: 236, Erelt 2017: 831–840). *mine*-liitelistel nimisõnadel on tõepoolest hulk sõnaklassiga seletatavaid omadusi, näiteks võib neil olla täielik käände- ja arvuparadigma, neid võivad laiendada adjektiivid ja demonstratiivid ning need esinevad sageli nimisõnadele tüüpilistes süntaktilistes rollides (subjektina, objektina, predikaatiivina või adpositsioonifraasi laiendina). Ometi on *mine*-tuletiste kirjeldamine tegeliku keelekasutuse põhjal märksa keerukam, kuna sarnaselt paljude teiste keelte nominalisatsioonidega (vt Noonan 1997, Yap jt 2011) on ka *mine*-tuletiste semantilised ja morfosüntaktilised funktsioonid mitmekesised. Näiteks võib tuletiste tähendusvälja laienemine kaasa tuua muutusi nimisõnafraasi struktuuris (argumentide kadu, mitmuse kasutamine, näide 1), teatud käänetes võivad *mine*-tuletised mõtestuda konverbidena (2) ning perifrastilistes ühendites võivad need olla ka lause süntaktiliseks ja semantiliseks tuumaks (3). Seda mitmekesisust toetab kahtlemata justnimelt tuletusmalli produktiivsus, mis annab situatsioonide ja suhete väljendamiseks laia leksikaalse baasi.

- (1) *Ta pea oli liiga väike, et sinna mingeid teadmisi talletada.*
- (2) *Tegime seda Mardi teadmisel.* (Sahkai 2011)
- (3) *Maja on lammutamisel.* (Sahkai 2011)

Eesti keeleteaduses on *mine*-tuletiste uurimused keskendunud põhiliselt sõna-tuletuse kahele põhifunktsioonile: sõnavara rikastamine (nt Kerge 2002, 2003) ja süntaktiline rekategoriseerimine (nt Kasik 1975, 2006, Vare 1991, Kerge 2001). Esimest funktsiooni on peetud omaseks näiteks teadustekstidele (Kerge 2002, 2003), teist juriidilistele tekstidele (Kerge 2002), ajakirjandustekstide puhul on räägitud mõlemast (nt Kerge 2002, 2003, Kasik 2006). Lisaks nendele põhisuundadele on ka uurimusi, mis on kirjeldanud *mine*-tuletisi süntaktilistes kontekstides, kus nende käitumine ei ole seletatav pelgalt tuletise sõnaliigiliste omaduste ja fraasistruktuurireeglite abil, vaid tähenduse saab omistada laiemale konstruksioonimallile, milles *mine*-tuletis esineb (Neetar 1988, Sahkai 2011). Sellistel konstruksioonipõhistel lähenemistel on oluline roll eripäraste süntaktiliste struktuuride (või suisa „süntaktiliste probleemide“) analüüsimisel ja

kirjeldamisel ning need haakuvad hästi ka psühholingvistiliste tõenditega sellest, kuidas kõnelejad keelt töötlevad ja kasutavad.

Siiski on kaks olulist aspekti, mis on senistes uurimustes jäänud tagaplaanile. Esiteks ei ole morfoloogiliste tuletusmallide produktiivsust tuletiste esinemis-sageduste kaudu operatsionaliseeritud ega empiiriliselt võrreldud. Nii ei tea tegelikult, kas *mine*-tuletusliite suur produktiivsus on absoluutne või tingitud teguritest, nagu register, tuletise funktsioon ja süntaktiline kontekst, või kuidas teiste tuletusliidete produktiivsus *mine*-liite seatud standardiga suhestub. Teiseks on tuletisi ja nende produktiivsust kirjeldatud põhiliselt ainult kirjakeele teatud registrites, ehkki suulises keeles võib väljenduda hoopis rikkalikumalt arenenud grammatiline süsteem (Halliday 1994: xxiii). Väitekiri annab panuse nende kahe lünga täitmisse, kasutades keelekorpuste (sagedus)andmeid, ning näitlikustab holistilise ja mitmetasandilise konstruktsioonipõhise lähenemise vajalikkust ja eeliseid *mine*-tuletiste kirjelduses. Ühtekokku püütakse väitekirjas vastata seitsmele uurimisküsimusele:

- K1:** Kas levinud tuletusmorfoloogia kasutamine eristab kirjutatud ja kõneldud eesti keele registreid?
- K2:** Kas korpuspõhised kvantitatiivsed morfoloogilise produktiivsuse mõõdikud viitavad deverbaalsufiksiga seotud tuletusmallide funktsioonidele ja piirangutele eesti keele eri registrites?
- K3:** Kuidas järjestuvad eri registrites produktiivsuse põhjal eesti keele kolm kõige sagedamat deverbaalsufiksit (*-mine*, *-us* ja *-ja*)?
- K4:** Kas põhilised korpuspõhised produktiivsuse mõõdikud haakuvad üldise lingvistilise intuitsiooniga, kui kasutatud on võrdlemisi väikseid korpusvalimeid?
- K5:** Kas suhtelised sagedused aitavad ennustada semantilist reeglipärasust?
- K6:** Milliseid *mine*-tuletiste süntaktilisi kombinatsioone esineb vanemas kõneldud keeles ja millist mõju avaldab see morfoloogilise produktiivsuse kvantitatiivsele hindamisele?
- K7:** Millised on ühe perifrastilise teonimekonstruktsiooni omadused ja võimalikud arenguteed ning mida kõneleb see morfoloogilise ja süntaktilise produktiivsuse vahekorras?

Väitekiri koosneb neljast artiklist ja katuspeatükist, mis esitab artiklite põhitulemused ning seob need ühtsesse raamistikku. Artiklid on järjestatud vastavalt üldistuse astmele, liikudes eri registrite üldistest sõnamoodustustendentsidest konkreetse süntaktilise konstruktsiooni analüüsini ühesainsas registris. Nõnda kasutatakse esimeses kahes artiklis morfoloogilise produktiivsuse uurimustes levinud kvantitatiivseid mõõdikuid, et hinnata *mine*-tuletiste produktiivsust eri registrites (P1, P2) ja võrdluses teiste deverbaalsufiksiga (P2), ning uuritakse, kui hästi aitavad suhtelised sagedused hinnata tuletusmalli semantilist regulaarsust (P1, P2) ja potentsiaalset konkurentsi kattuvaid tähendusvälju tähistavate

sufiksrite vahel (P2). Analüüsiks valitud registrid – teaduskeel, ajakirjanduskeel, ilukirjanduskeel, suuline spontaanne ühiskeel ning suuline murdekeel – on töös defineeritud kasutatavate korpuste kaudu ning registrite hulka on arvatud tervikuna ka murdekeel, mida küll traditsiooniliselt situatiivsetest ehk kasutuskesksetest allkeeltest eristatakse (Hennoste 2000c). Kuna aga murdekeele allikaks olev eesti murrete korpus sisaldab situatiivselt sarnast tüüpi keelekasutust, kus poolstruktureeritud intervjuude vormis meenutatakse möödunud aegade eluolu, töid ja tegemisi, peetakse murdekeele võrdlemist teiste registritega seetõttu selles töös õigustatuks. Kahes viimases artiklis keskendutakse ainult murdekeelele kui kõneldud keele variandile, mis näitab *mine*-tuletiste kasutuses enim varieeruvust ning võimaldab uurida ka jooni, mis tänapäevases ühiskeeles on kadunud või ilmnevad väga harva. Kolmas artikkel (P3) selgitab välja kõneldud murdekeeles avalduvaid *mine*-tuletiste nimisõnalisi ja verbilisi omadusi ning funktsioone, kirjeldades tuletisi ka laiemas süntaktilises kontekstis. Neljas artikkel (P4) vaatlleb murdekeele andmeid veelgi spetsiifilisemalt ning analüüsib üht seni vähekirjeldatud perifrastilist teonimekonstruktsiooni – *olema* + *Vmine*<sub>NOM/PAR</sub> (nt *Seal oli karjumist, Meil on koristamist*) – kui iseseisvat ja mingil määral polüsemset keelekirjeldusüksust.

Väitekiri on kirjutatud kasutuspõhise keelekirjelduse vaimus. Kasutuspõhise lähenemise põhitões on, et keel (sh grammatika) kujuneb kasutuses ning kasutus omakorda vormib keelt (Perek 2015: 6). Sealjuures ei eeldata (ainult) jäikade kategooriate olemasolu, vaid grammatikat nähakse pigem kui dünaamilist, paindlikku, mitmetasandilist ja pidevalt muutuvat süsteemi, mida reguleerivad üldisemad kognitiivsed protsessid (Lemmens 2019, Perek 2015, Diessel 2019). Justnimelt oma dünaamilisuse ja paindlikkuse tõttu sobib kasutuspõhine lähene mine hästi kirjeldama ka sellist kompleksset ja mitmepalgelist protsessi nagu *mine*-tuletus, mis hõlmab sõnaklassi ambivalentsust, ulatuslikku semantilise läbi paistvuse skaalat, funktsionaalset komplementaarsust infinitiivide ja teiste teonimedega jne. Väitekirjas kasutatakse ka kasutuspõhise konstruktsiooni-grammatika (Goldberg 2006, Croft ja Cruse 2004) põhimõtteid ja mõisteid, et kirjeldada *mine*-tuletisi nii sõna- kui ka lausetasandil. Konstruktsiooni all on töös mõeldud niisiis erineva sisemise kompleksuse ja abstraktsuse astmega malle, mis seovad omavahel mingi keelelise vormi ja tähenduse (st nii sõnatuletusmalle kui ka süntaktilisi vormeleid peetakse selle töö kontekstis konstruktsioonideks). Spetsiifilisemad madalama taseme ja abstraktsemad kõrgema taseme konstruktsioonid on seotud võrgustikuks hierarhiliste pärilussuhete kaudu (ingl *hierarchical inheritance links*), ent see konstruktsioonide võrgustik sisaldab ka polüsemia- ja metafooriseoseid (ingl *polysemy links, metaphorical links*) ning alamosaseoseid (ingl *subpart links*), mis ühendavad sama taseme konstruktsioone (Goldberg 1995: 75–81). Sama keeletasandi ja eri keeletasandite konstruktsioonid on omakorda leksikaalsete, konstruktsiooniliste ja nn täitja-pesa-seoste (ingl *filler-slot relations*) kaudu ühendatud kõrgema taseme võrgustikuks (Diessel 2019). Selline mitmetasemeline kompleksne võrgustik aitab kirjeldada *mine*-tuletisi, mis on ise madalama taseme morfoloogilised konstruktsioonid ja võivad kasutussageduse mõjul teatud määral automatiseeruda; mis võivad

polüseemia- ja metafooriseoste kaudu esineda süstemaatiliselt mingis kõrvaltähenduses (nt *nõudmine*, *tahtmine* kui tegevuse objektid või tulemid); mis kuuluvad alamosaseoste kaudu nn sõnaperedesse (*vastutamine*, *vastutus*, *vastutaja*, *vastutama* jne); ning mis võivad täitja-pesa-seoste kaudu esineda süntaktilistes konstruktsioonides, sh sellistes madalama taseme konstruktsioonides, mille kõik omadused ei ole tuletatavad hierarhiliste suhete kaudu kõrgema taseme konstruktsioonidest (nt *Töö tahab tegemist*, *Seal pole enam heinaniitmist*). Süntaktilised *mine*-teonime konstruktsioonid omakorda võivad olla polüseemsed (nt väitekirjas analüüsitud konstruktsioon *olema* + *Vmine*<sub>NOM/PAR</sub> võib olla nii modaalne kui ka lihtsalt impersonaalne), olla alamosaseoste kaudu seotud teiste sarnast tähendust kandvate konstruktsioonidega (nt *Seal ei niideta enam heina*, *Seal ei pea enam heina niitma*) jne.

Kasutuspõhine lähenemine tingib suuresti ka väitekirja aluseks olevate uurimuste materjali ja meetodite valiku. Kuna kasutuspõhisus eeldab, et uuritakse tegelikku keelekasutust ning tegeliku keelekasutuse kõige kättesaadavamad allikad on struktureeritud keelekorpused, on siinsete uurimuste aluseks olevad andmed kogutud eesti murrete korpusest<sup>24</sup> (P1, P2, P3, P4), eesti keele spontaanse kõne foneetilisest korpusest<sup>25</sup> (P1, P2) ning tasakaalus korpusest<sup>26</sup>, mis sisaldab ajalehtede, ilukirjanduse ja teaduskirjanduse tekste (P1, P2). Kasutussagedus, mis on kasutuspõhistes keeleuurimustes üha tähtsamal kohal ning tihedalt seotud lingvistiliste struktuuride kognitiivse töötlusega (Diessel 2019: 30, Perek 2015: 7), on siinsetes töös enam esil esimeses kahes, morfoloogilist produktiivsust käsitlevas uurimuses (P1, P2). Uurimustes on produktiivsust laiemalt mõistetud kui kvantitatiivselt mõõdetavat tõenäosust, et mingit kategooriat (nt *mine*-konstruktsiooni) täiendatakse uute liikmetega, ent produktiivsuse eri aspektide hindamiseks on artiklites P1 ja P2 kasutatud levinud mõõdikuid – realiseerunud produktiivsus, potentsiaalset produktiivsus ja laiendamise produktiivsus (vt nt Baayen ja Lieber 1991, Baayen 1992, Baayen 1993, Baayen ja Renouf 1996) –, mis omakorda operatsionaliseerivad tuletiste sõne-, tüübi- ja *hapax legomena* ehk üks kord esinevate tüüpide sagedusi ning suhestavad need mingi (alam)korpuse sõnade koguarvuga. Lisaks registrite võrdlemisele uurimuses P1 võrreldakse uurimuses P2 *mine*-konstruktsiooni produktiivsust ka *us*- ja *ja*-konstruktsiooni produktiivsusega eri registrites. Erineva kasutussagedusega sufiksiste nende produktiivsuse alusel järjestamiseks kõrvutatakse morfoloogilise produktiivsuse kesksel mõõdikul, potentsiaalset produktiivsust, uurimuses P2 ka kahe alternatiivse lähene-misega. Nii P1 kui ka P2 otsivad tuletusliidete morfoloogilisele produktiivsusele ja tuletusmallide semantilisele (eba)regulaarsusele seletusi ka liidete paradigmaatilistest suhetest teiste samasse sõnaperesse kuuluvate konstruktsioonidega. P1-s võrreldakse leksikaalseid konstruktsioone nn lihtsa NV-skoori abil, mis võrdleb tuletusaluse verbi suhtelist sagedust vastava tuletise suhtelise sagedusega ja võimaldab tuvastada konkreetseid tuletisi, mis on tõenäolisemalt suure

<sup>24</sup> <https://www.keel.ut.ee/et/keelekogud/murdekorpus> (Kasutatud 23.02.2021)

<sup>25</sup> <https://www.keel.ut.ee/en/languages-resourceslanguages-resources/phonetic-corpus-estonian-spontaneous-speech> (Kasutatud 23.02.2021)

<sup>26</sup> <https://www.cl.ut.ee/korpused/grammatikakorpus/index.php?lang=en> (Kasutatud 23.02.2021)

kasutussageduse tõttu automatiseerunud ning mille protsessimisel aktiveerub seetõttu vastav abstraktne konstruktsiooniskeem vähem. P2-s kasutatakse tuletusaluste verbide ja vastavate tuletiste võrdlemiseks logaritmitud sagedusi ning tuletusmalli läbipaistvust hinnatakse Spearmani korrelatsioonikordaja ja üldistatud aditiivsete mudelite abil (Wood 2017). Mõlemad meetodid tulenevad ühesugusest eeldusest: deverbaaltuletised, mille suhteline sagedus on väiksem kui vastava verbi suhteline sagedus, panustavad morfoloogilise konstruktsiooni produktiivsusesse enam, samas kui vastupidisel juhul on tuletis suurema tõenäosusega automatiseerunud ja aktiveerib morfoloogilist konstruktsioonimalli nõrgemalt (vt nt Hay 2001, Hay ja Baayen 2002, Hay ja Plag 2004, Hay ja Baayen 2005). P2 tegeleb põgusalt ka nn tüübablokeerimisega (ingl *type blocking*), kus ühe vormi olemasolu võib takistada teise vormi kasutust (Aronoff 1976: 43, Bauer 2001: 136–138), ning võrdleb samast verbist moodustatud *mine-* ja *us-*tuletiste suhtelisi sagedusi, et uurida, kas ja kuidas avaldub tüübablokeerimine, kui kaks deverbaalsufiksit võivad potentsiaalselt väljendada sama teonime tähendust (nt *kasutamine* ja *kasutus*). Analüüs uurimustes P3 ja P4 põhineb ainult eesti murrete korpuse materjalil, on suuresti kvalitatiivne ning hõlmab *mine-*tuletiste uurimist nende esinemiskontekstis, selleks et kirjeldada võimalikult ulatuslikult *mine-*tuletiste funktsioone ja omadusi (P3), nende levikut 10 murdealal (P4) ning tuvastada ja analüüsida *olema-*verbi ja nominatiivi- või partitiivi-kujulise *mine-*teonime konstruktsioone (P4). Selleks otsiti korpusest *mine-*lõpulise lemma ja sõnaklassi põhjal välja kõik 1928 tuletise esinemisjuhtu ning märgendati iga esinemisjuhu puhul selle kääne, arv, tuletusaluse verbi lemma, laiendite esinemine ja vorm jm tunnused, mis vaatlusi aitaksid klassifitseerida. Väitekirja katuspeatükis on esitatud publikatsioonide tulemuste sidumiseks ning taustaks ka paar lisaanalüüsi, mida ei ole varem publitseeritud, ent nende osa töös on väike. Lisaanalüüsid põhinevad 1000-l igast registrist ja iga sufiksi kohta juhuslikult valitud vaatlusel.

Analüüsi tulemuste põhjal selgus, et vähemalt produktiivsemaks peetud eesti keele deverbaalsufiksitate puhul (*-mine*, *-us*, *-ja*) näitavad kvantitatiivsed produktiivsuse mõõdikud tõepoolest registriti erinevusi (**K1**). Sealjuures oli kirjalikkuse-suulisuse vastandus oluline põhiliselt realiseerunud produktiivsuse puhul, mis hindab seda, kui palju eri tüüpe mingi tuletusmall on minevikus tootnud, ent ei väljenda otseselt seda, kui hõlpsalt saab seda kasutada uute tuletiste loomiseks. Kirjalikes registrites (teaduskirjandus, ajakirjandus, ilukirjandus) oli kõigi kolme sufiksi realiseerunud produktiivsus oluliselt suurem kui suulistes registrites (tänapäevane spontaanne kõne ja arhailisem kõneldud murdekeel). Teiste produktiivsuse mõõdikute puhul aga, mis võimaldavad hinnata morfoloogilise kategooria potentsiaali rikastada sõnavara uute sõnadega, sai suulisusest-kirjalikkusest olulisemaks tekstide formaalsuse-informaalsuse mõõde: formaalsemates registrites (teadus- ja ajakirjandustekstid) oli tõenäosusena mõõdetud potentsiaal sõnavara rikastada oluliselt väiksem kui informaalsetes registrites. Kolme produktiivsuse mõõdiku põhjal kokku tundub ilukirjandus olevat register, mis kasutab tuletusmalle kõige produktiivsemalt.

Kirjalikes registrites olid realiseerunud produktiivsuse väärtused oluliselt suuremad kui suulistes registrites. See viitab asjaolule, et kirjalikes registrites on

olnud vajadus vastavate tuletistena realiseeritud mõistete järele suurem (K2). Samuti mängivad siin vähemasti *mine*-tuletiste puhul rolli traditsioonilised sõnamoodustuse funktsioonid: näiteks teadus- ja ajakirjandustekstides võib suur realiseerunud produktiivsus viidata abstraherimise ja uute terminite loomise vajadusele, aga ka stilistilistele kaalutlustele (vt Kasik 2011: 70), samas kui suulises keeles tundub nõudlus selliste funktsioonide järele olevat väiksem ning olulisemaks saavad süntaktilise rekategoriseerimise ja anafoorilise viitamise funktsioonid. Teisalt aga tuleb silmas pidada, et vähemalt formaalsemates registrites (teadus- ja ajakirjandustekstid) on ka üldine nimisõnade osakaal tekstides oluliselt suurem, mistõttu näib sõnaklasside jaotus eri registrites olevat morfoloogilise produktiivsuse seni väheuuritud kaastegur. Potentsiaalse produktiivsuse ja laienemise produktiivsuse mõõdikute suuremaid väärtusi omakorda just informaalsetes registrites (ilukirjandus, suuline ühiskeel ja suuline murdekeel) võib seletada samuti pragmaatilise vajaduse ja sõnamoodustuse funktsioonidega: süntaktilises funktsioonis kasutatud sufiksitele on üldiselt vähem pragmaatilisi piiranguid kui leksikaalses funktsioonis kasutatud sufiksitele (Bauer 2001: 208–209) ning süntaktiline funktsioon on omakorda informaalsete registritega loomulikumalt seostatav kui leksikaalne funktsioon. Kuna need kaks mõõdikut on negatiivselt mõjutatud ka (enamasti suure esinemissagedusega) automatiseerunud või leksikaliseerunud tuletiste osakaalust, võib oletada, et informaalsete registrid kasutavad selliseid tuletisi ka vähem.

Kolme deverbaalsufiksi järjestus nende potentsiaalse produktiivsuse põhjal oli küllalt ootuspärane, arvestades tuletusmallide morfosüntaktilisi ja semantilisi piiranguid: *-mine* oli üldiselt kõikides registrites kõige produktiivsem sufiks ning sellele järgnesid *-ja* ja *-us* (K3). Seega on tõenäosus eesti keeles morfoloogilist kategooriat uute tuletistega rikastada kõige suurem just *mine*-konstruktsiooni puhul, samas kui näiteks *us*-konstruktsiooni semantiline ja struktuuriline ähmasus tingib selle, et kõnelejad eelistavad uute tuletiste moodustamisele kasutada pigem olemasolevaid. *ja*-tuletiste produktiivsust pärsivad omakorda semantilised piirangud, hulk paradigmaatilisi konkurente (nt asesõnad, nimisõnad, pärisnimed) ning seega ka väiksem pragmaatiline kasulikkus. Nende piirangute esiletõusmine ja tugevus aga varieerub registriti. Samuti on registriti olulisi erinevusi selles, kui võrd suur on *mine*-sufiksi edu teiste sufiksitate ees.

Kahes esimeses uurimuses (P1 ja P2) kasutatud produktiivsuse mõõdikud ja nende usaldusväärsus sõltuvad siiski ka valimi suurusest (K4), seda eeskätt erineva esinemissagedusega sufiksitate produktiivsuse võrdlemisel. Kuna siinses väitekirjas kasutatud valimid olid olude sunnil korpuslingvistilises mõttes väga väikesed (ulatudes iga registri puhul vaid ligikaudu 426 000 tekstisõnani), peab kvantitatiivse analüüsi tulemuste tõlgendamisel olema ettevaatlik ning suhtuma ka intuiitselt ootuspärastesse tulemustesse teatud reservatsiooniga. Põhiliselt puudutab mõõdikute tõlgendamise probleem asjaolu, et väikestes valimites ei tähista üks kord esinevad sõnad neologisme, vaid sõnad võivad esineda harva ka lihtsalt juhuse tahtel. Näiteks esinevad ajakirjanduskorpuse valimis sõnad *küsimine*, *unustamine* ja *ennustamine* ainult ühe korra, ent on igale eesti keele kõneleajale harjumuspärased ja tuttavad sõnad. Nii muutub keerulisemaks potentsiaalse ja laieneva produktiivsuse mõõdikute otsene tõlgendamine. Ehkki kasutatud

mõõdikud on seega kasulikud, võrdlemisi hõlpsalt rakendatavad ja intuitiivselt mõistetavad morfoloogilise produktiivsuse indikaatorid, on eriti *hapax legomena* põhised mõõdikud valimi suuruse suhtes tundlikud, mida näidati ka uurimuste P1 ja P2 tulemuste võrdlemisel.

Väitekirjas uuriti ka alamosaseoste ehk paradigmaatiliste suhete rolli tuletusmallide kompositsionaalsuse hindamisel (**K5**). Mallide semantilise läbipaistvuse suhestamine tuletiste aluseks olevate verbidega suhteliste sageduste kaudu (P1, P2) andis hea ülevaate sellest, milline on tuletusmallide üldine dünaamika (*mine*-tuletus näib ootuspäraselt olevat oluliselt regulaarsem kui ülejäänud kaks tuletusmalli). Ehkki suhteliste sageduste võrdlus tõi üpris selgelt esile ka üksikud sagedad vormid (nt *tegemine*), mis suure tõenäosusega on kõnelejate jaoks automatiseerunud ega aktiveeri tugevalt vastavat morfoloogilist konstruktsioonimalli, ei osutunud sageduste suhe alati usaldusväärseks semantilise kompositsionaalsuse ennustajaks. Dekontekstualiseeritud tuletiste suhteliste sageduste võrdlus ei sobinud väga hästi ka madalama taseme leksikaalsete konstruktsioonide semantiliseks võrdluseks *mine*- ja *us*-tuletiste potentsiaalse konkurentsi hindamisel (P2), ehkki üldiselt võis osalist sünonüümiat omistada pigem nendele *mine*- ja *us*-tuletistele, mille suhteline sagedus vastava liitega tuletiste seas oli väiksem.

Väitekirja kvalitatiivsem osa keskendus suulise murdekeele analüüsile (P3 ja P4). Arhailisem vanem murdekeel osutus küllalt rikkalikuks allikaks, mille põhjal *mine*-tuletiste omadusi ja süntaktilisi kombineerumisvõimalusi kirjeldada. Ehkki *mine*-tuletistel on ka murdekeele andmete põhjal üksjagu nimisõnadele omaseid jooni, on need sageli paremini kirjeldatavad süntaktiliste konstruktsioonide kaudu, milles need esinevad (**K6**). Sellistes mittekompositsionaalse tähendusega konverbiiladsetes või perifrastilistes konstruktsioonides kasutatakse *mine*-tuletisi kindlates käänetes, ent kuna teonimel on neis enamasti selge protsessuaalne tähendus ning teonimede esinemine konstruktsioonides ei ole vormiliselt ega tähenduslikult piiratud, siis kasutatakse pigem produktiivset tuletusmalli (nt *eks ta tuulamesele läks, siis läheb ta üles kiskumiseks*). Seega on morfoloogilise produktiivsuse olulised kaastegurid sarnaselt sõnaliikide jaotumisega mingis registris ka käändekonstruktsioonid, milles teonimed sagedamini esinevad, ning sedakaudu on omakorda seotud morfoloogiline ja süntaktiline produktiivsus. Morfosüntaktilised konstruktsioonid, milles tuletised esinevad, on olulised ka seetõttu, et need võimaldavad seletada sagedate leksikaalsete morfoloogiliste konstruktsioonide kasutust (**K7**). Nii näiteks esines kõige sagedam tuletis *tegemine* kirjutatud keeles ja tänapäevases spontaanses ühiskeeles põhiliselt predikatiivkonstruktsioonis *olema + tegemist + X<sub>COM</sub>* (nt *tegemist oli põrandaaluse kihlveokontoriga*), samas kui murdetekstides ei kasutatud vastavat konstruktsiooni kordagi. Murdeandmetes prevaleeris *tegemine* aga hoopis konstruktsioonis *olema + V<sub>mine</sub><sub>NOM/PAR</sub>*, mida kirjeldati põhjalikumalt uurimuses P4. Kõnealuse konstruktsiooni komitatiiviga alamkonstruktsioonil (nt *no eks loomadega ole tegemest*) on hulk ühiseid vormilisi jooni eespool kirjeldatud predikatiivkonstruktsiooniga, ehkki konstruktsioonide tähendused on täiesti erinevad. *olema + V<sub>mine</sub><sub>NOM/PAR</sub>* konstruktsiooni näol on tegemist produktiivse konstruktsioonimalliga, mille põhifunktsiooniks võib pidada impersonaliseerimist: konstrukt-

siooniga väidetakse mingi teonimega väljendatud tegevuse toimumist/mitte-toimumist, kusjuures tegija defokuseeritakse ning taandatakse kas genitiivse agentatribuudi või adessiivse kogeja rolli või jäetakse sootuks eksplitsiitselt väljendamata. Osa konstruktsioonidest mõtestub aga ka modaalsena, sisaldades hinnangut tegevuse toimumise kohustatusele, lubatusele või vajadusele. Konstruktsioon pärib niisiis osa oma vormist ja tähendusest eksistentsiaal- ja kogeja-omajalausetelt, ent sel on ka hulk omadusi, mis ei ole taandatavad pelgalt mingi referentsiaalse entiteedi olemasolu väitmisele või omamissuhte väljendamisele. Konstruktsiooni kui eraldiseisva süntaktilise üksuse staatusest kõneleb ka see, et sarnased modaalse tähendusega konstruktsioonid esinevad ka teistes läänemeresoome keeltes (näiteks liivi ja soome keeles). Ka on pakutud, et konstruktsioon on arenenud algläänemeresoome transitiiivsest modaalstruktsioonist, ent olnud aegade jooksul aldis muutustele sõnajärjes (Grünthal 1941: 177–179). Seega on ühelt poolt võimalik, et konstruktsioonil on algul olnud mitte-episteemiline modaalne tähendus, ent aja jooksul on eesti keeles modaalne tähendus ähmastunud ning prevaleerima hakanud konstruktsiooni impersonaalne funktsioon, samas kui liivi ja soome keeles pole säärast arengut toimunud. Teisalt on ka võimalik, et konstruktsiooni kaks tähendust on arenenud paralleelselt.

Seitsme uurimisküsimuse kaudu püüdis siinne väitekiri ühendada morfeemipõhise (vormilise) ja skeemipõhise (täendusliku) perspektiivi, kirjeldamaks tuletusmorfoloogia rolli tegelikus keelekasutuses ning käsitlemaks sõnamoodustusmalle kui morfoloogia ja süntaksi vahele jäävaid nähtusi terviklikumal moel. Väitekiri panustab eelkõige eesti keele sõnamoodustusprotsesside uurimisse, pakkudes empiirilist tõendusmaterjali lingvistilise produktiivsuse kvantitatiivse väljundi ning produktiivsust mõjutavate protsesside kohta eri registrites, sealhulgas kõneldud keeles. Väitekirja originaalseks väärtuseks võib pidada katset ühendada teonimede analüüs keelekirjelduse eri tasanditel ühtseks tervikuks. Väitekiri loob ka aluse paljudele edasistele võimalikele uurimissuundadele. Näiteks võiks eesti keele tuletusmorfoloogia produktiivsust edasi uurida keerukamate statistiliste mudelite abil, mis võimaldavad adekvaatsemalt analüüsida sagedusjaotusi, mis on tugevalt paremale kaldu (vt Baayen 2001, Evert 2004, Evert ja Baroni 2007), või võtta produktiivsust mõjutavate faktoritena lisaks registrile arvesse ka sotsiogeograafilisi tegureid ja kõnelejate individuaalseid eelistusi (nt Keune jt 2006, Säily ja Suomela 2009, De Smet 2020). Distributsioonilise semantika mudelite ning tehisnärvivõrkude abil saaks uurida eri tuletiste tähenduste sarnasusi, hinnata tuletusmalli semantilist (de)kompositsionaalsust, või tuvastada polüseemseid tuletisi. Samuti oleks kasulik kontrollida korpusanalüüsist esile kerkivaid tendentse katseliste meetoditega, et hinnata tuletusmallide produktiivsust ka aktiivses, reaalses toimivas tekstiloomes ning tuletiste automatiseerumist leksikaalse info töötlemisel.

## **PUBLICATIONS**

## CURRICULUM VITAE

**Name:** Maarja-Liisa Pilvik  
**Date of birth:** 26.02.1989  
**Citizenship:** Estonian  
**Address:** Jakobi 2–430, 51005 Tartu  
**E-mail:** maarja-liisa.pilvik@ut.ee

### Education:

2013–2021 University of Tartu, doctoral studies in Estonian and Finno-Ugric Linguistics  
2014 fall KU Leuven, visiting scholar  
2010–2013 University of Tartu, MA in Estonian and Finno-Ugric Linguistics (Finnish language and culture)  
2013 spring University of Turku, visiting scholar  
2010 fall University of Helsinki, Erasmus exchange student  
2007–2010 University of Tartu, BA in Estonian and Finno-Ugric Linguistics (Finnish language and culture)  
2004–2007 Noarootsi Gymnasium  
1995–2004 Kärkla Co-educational Gymnasium

### Professional experience:

2021– University of Tartu, Institute of Estonian and General Linguistics, Specialist of Corpora and Quantitative Linguistics  
2016–2021 University of Tartu, Institute of Estonian and General Linguistics, Junior Research Fellow of Applied Dialectology  
2017–2018 University of Tartu, Institute of Estonian and General Linguistics, Junior Research Fellow of Finno-Ugric Linguistics  
2013 University of Tartu, Institute of Estonian and General Linguistics, Dialect Corpus Specialist

### Publications:

Lindström, Liina, Maarja-Liisa Pilvik, Helen Plado. 2021. Variation in negation in Seto. *Studies in Language* [to appear]. DOI: 10.1075/sl.19063.lin.  
Pilvik, Maarja-Liisa, Liina Lindström, Helen Plado. 2021. Murded, varieerumine ja korpusandmed: eitussõna paiknemine võru ja seto eituslausetes. *Keel ja Kirjandus* 8–9. 771–796. DOI: 10.54013/kk764a7.  
Pilvik, Maarja-Liisa. 2021. Comparing the productivity of Estonian deverbal suffixes *-mine*, *-us*, and *-ja* in five registers: A quantitative usage-based approach. *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEFUL* 12(1). 189–258. DOI: 10.12697/jeful.2021.12.1.06.  
Lindström, Liina, Maarja-Liisa Pilvik. 2020. Keeleteaduse muutumine arvuti-ajastul. *Horisont* 6. 38–40.  
Pilvik, Maarja-Liisa. 2020. Horisondi sõnapilv. *Horisont* 6. 41.

- Rumm, Andra, Maarja-Liisa Pilvik, Liina Lindström, Virve-Anneli Vihman. 2020. Kas noored ropendavad liiga palju? *Horisont* 6. 70–72.
- Lindström, Liina, Maarja-Liisa Pilvik, Mirjam Ruutma, Kristel Uihoaed. 2019. On the use of perfect and pluperfect in Estonian dialects: frequency and language contacts. In Sofia Björklöf & Santra Jantunen (eds.), *Multilingual Finnic – Language contact and change*. (Uralica Helsingensia.) 155–193. Helsinki: Finno-Ugrian Society. DOI: 10.33341/uh.85035.
- Pilvik, Maarja-Liisa, Kadri Muischnek, Gerth Jaanimäe, Liina Lindström, Kersti Lust, Siim Orasmaa, Tõnis Tärna. 2019. Mõistus sai kuulotedu: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine. *Eesti Rakenduslingvistika Ühingu aastaraamat* 15. 139–158. DOI: 10.5128/ERYa15.08.
- Pilvik, Maarja-Liisa. 2019. Assessing the productivity of the Estonian deverbalsuffix *-mine* in five registers of Estonian. *SKY Journal of Linguistics* 32. 75–103.
- Lindström, Liina, Maarja-Liisa Pilvik, Helen Plado. 2018. Nimetamiskonstruktsioonid eesti murretes. Murdeerinevused või suuline süntaks? *Mäetagused* 70. 91–126. DOI: 10.7592/mt2018.70.lindstrom\_pilvik\_plado.
- Lindström, Liina, Maarja-Liisa Pilvik. 2018. Korpuspõhine kvantitatiivne dialektoloogia. *Keel ja Kirjandus* 8–9. 643–662.
- Pilvik, Maarja-Liisa. 2017. Deverbalsuffix *-mine* action nominals in the Estonian Dialect Corpus. *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEFUL* 8(2). 295–326. DOI: 10.12697/jeful.2017.8.2.10.
- Pilvik, Maarja-Liisa. 2016. *olema* + *Vmine* konstruktsioonid eesti murretes. *Keel ja Kirjandus* 6. 429–446.
- Ruutma, Mirjam, Aki-Juhani Kyröläinen, Maarja-Liisa Pilvik, Kristel Uihoaed. 2016. Ambipositsioonide morfosüntaktilise varieerumise kirjeldusi kvantitatiivsete profiilide abil. *Keel ja Kirjandus* 2. 92–113.
- Klavan, Jane, Maarja-Liisa Pilvik, Kristel Uihoaed. 2015. The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of Estonian. *SKY Journal of Linguistics* 28. 187–224.
- Lindström, Liina, Maarja-Liisa Pilvik, Mirjam Ruutma, Kristel Uihoaed. 2015. Mineviku liitaegade kasutusest eesti murretes ja keelekontaktide valguses. *Võro Instituudi toimendusõq* 29. 39–70.
- Pilvik, Maarja-Liisa. 2013. Konverents „Grammatika ja kontekst”. *Keel ja Kirjandus* 10. 788–790.
- Pilvik, Maarja-Liisa. 2013. Tähelepanu akendamine. *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEFUL* 4(1). 95–109. DOI: 10.12697/jeful.2013.4.1.06.
- Pilvik, Maarja-Liisa. 2012. Kolmas soome-eesti kognitiivse lingvistika konverents. *Keel ja Kirjandus* 3. 232–234.

## ELULOOKIRJELDUS

**Nimi:** Maarja-Liisa Pilvik  
**Sünniaeg:** 26.02.1989  
**Kodakondsus:** Eesti  
**Aadress:** Jakobi 2–430, 51005 Tartu  
**E-post:** maarja-liisa.pilvik@ut.ee

### Haridus:

2013–2021 Tartu Ülikool eesti ja soome-ugri keeleteaduse doktoriõpe  
2014 fall KU Leuven, külalisuuriija  
2010–2013 Tartu Ülikool, eesti ja soome-ugri keeleteaduse magistrikraad (soome keel ja kultuur)  
2013 kevad Tartu Ülikool, külalisuuriija  
2010 sügis Helsingi ülikool, Erasmuse vahetusüliõpilane  
2007–2010 Tartu Ülikool, eesti ja soome-ugri keeleteaduse bakalaureusekraad (soome keel ja kultuur)  
2004–2007 Noarootsi Gümnaasium  
1995–2004 Kärdla Ühisgümnaasium

### Töökogemus:

2021– Tartu Ülikool, eesti ja üldkeeleteaduse instituut, korpuste ja kvantitatiivse lingvistika spetsialist  
2016–2021 Tartu Ülikool, eesti ja üldkeeleteaduse instituut, rakendusliku dialektoloogia nooremteadur  
2017–2018 Tartu Ülikool, eesti ja üldkeeleteaduse instituut, soome-ugri keeleteaduse nooremteadur  
2013 Tartu Ülikool, eesti ja üldkeeleteaduse instituut, murdekorpuse spetsialist

### Publikatsioonid:

Lindström, Liina, Maarja-Liisa Pilvik, Helen Plado. 2021. Variation in negation in Seto. *Studies in Language* [ilmummas]. DOI: 10.1075/sl.19063.lin.  
Pilvik, Maarja-Liisa, Liina Lindström, Helen Plado. 2021. Murded, varieerumine ja korpusandmed: eitussõna paiknemine võru ja seto eituslausetes. *Keel ja Kirjandus* 8–9. 771–796. DOI: 10.54013/kk764a7.  
Pilvik, Maarja-Liisa. 2021. Comparing the productivity of Estonian deverbal suffixes *-mine*, *-us*, and *-ja* in five registers: A quantitative usage-based approach. *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEFUL* 12(1). 189–258. DOI: 10.12697/jeful.2021.12.1.06.  
Lindström, Liina, Maarja-Liisa Pilvik. 2020. Keeleteaduse muutumine arvuti-ajastul. *Horisont* 6. 38–40.  
Pilvik, Maarja-Liisa. 2020. Horisondi sõnapilv. *Horisont* 6. 41.  
Rumm, Andra, Maarja-Liisa Pilvik, Liina Lindström, Virve-Anneli Vihman. 2020. Kas noored ropendavad liiga palju? *Horisont* 6. 70–72.

- Lindström, Liina, Maarja-Liisa Pilvik, Mirjam Ruutma, Kristel Uiboaed. 2019. On the use of perfect and pluperfect in Estonian dialects: frequency and language contacts. In Sofia Björklöf, Santra Jantunen (toim.), *Multilingual Finnic – Language contact and change*. (Uralica Helsingiensia.) 155–193. Helsinki: Finno-Ugrian Society. DOI: 10.33341/uh.85035.
- Pilvik, Maarja-Liisa, Kadri Muischnek, Gerth Jaanimäe, Liina Lindström, Kersti Lust, Siim Orasmaa, Tõnis Tärna. 2019. Mõistus sai kuulotedu: 19. sajandi vallakohtuprotokollide tekstidest digitaalse ressursi loomine. *Eesti Rakenduslingvistika Ühingu aastaraamat* 15. 139–158. DOI: 10.5128/ERYa15.08.
- Pilvik, Maarja-Liisa. 2019. Assessing the productivity of the Estonian deverbal suffix *-mine* in five registers of Estonian. *SKY Journal of Linguistics* 32. 75–103.
- Lindström, Liina, Maarja-Liisa Pilvik, Helen Plado. 2018. Nimetamiskonstruktsioonid eesti murretes. Murdeerinevused või suuline süntaks? *Mäetagused* 70. 91–126. DOI: 10.7592/mt2018.70.lindstrom\_pilvik\_plado.
- Lindström, Liina, Maarja-Liisa Pilvik. 2018. Korpuspõhine kvantitatiivne dialektoloogia. *Keel ja Kirjandus* 8–9. 643–662.
- Pilvik, Maarja-Liisa. 2017. Deverbal *-mine* action nominals in the Estonian Dialect Corpus. *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEFUL* 8(2). 295–326. DOI: 10.12697/jeful.2017.8.2.10.
- Pilvik, Maarja-Liisa. 2016. *olema* + *Vmine* konstruktsioonid eesti murretes. *Keel ja Kirjandus* 6. 429–446.
- Ruutma, Mirjam, Aki-Juhani Kyröläinen, Maarja-Liisa Pilvik, Kristel Uiboaed. 2016. Ambipositsioonide morfosüntaktilise varieerumise kirjeldusi kvantitatiivsete profiilide abil. *Keel ja Kirjandus* 2. 92–113.
- Klavan, Jane, Maarja-Liisa Pilvik, Kristel Uiboaed. 2015. The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of Estonian. *SKY Journal of Linguistics* 28. 187–224.
- Lindström, Liina, Maarja-Liisa Pilvik, Mirjam Ruutma, Kristel Uiboaed. 2015. Mineviku liitaegade kasutusest eesti murretes ja keelekontaktide valguses. *Võro Instituudi toimõndusõq* 29. 39–70.
- Pilvik, Maarja-Liisa. 2013. Konverents „Grammatika ja kontekst”. *Keel ja Kirjandus* 10. 788–790.
- Pilvik, Maarja-Liisa. 2013. Tähelepanu akendamine. *Eesti ja soome-ugri keeleteaduse ajakiri ESUKA / Journal of Estonian and Finno-Ugric Linguistics JEFUL* 4(1). 95–109. DOI: 10.12697/jeful.2013.4.1.06.
- Pilvik, Maarja-Liisa. 2012. Kolmas soome-eesti kognitiivse lingvistika konverents. *Keel ja Kirjandus* 3. 232–234.

## DISSERTATIONES PHILOLOGIAE ESTONICAE UNIVERSITATIS TARTUENSIS

1. **Ülle Viks.** Eesti keele klassifikatoorne morfoloogia. Tartu, 1994.
2. **Helmi Neetar.** Deverbaalne nominaaltuletus eesti murretes. Tartu, 1994.
3. **Ülo Valk.** Eesti rahvausu kuradi-kujutelm. Tartu, 1994.
4. **Arvo Eek.** Studies on quantity and stress in Estonian. Tartu, 1994.
5. **Reet Kasik.** Verbid ja verbaalsubstantiivid tänapäeva eesti keeles. Tartu, 1994.
6. **Silvi Vare.** Nimi- ja omadussõnatuletus tänapäeva eesti kirjakeeles. Tartu, 1994.
7. **Heiki-Jaan Kaalep.** Eesti keele ressursside loomine ja kasutamine keele- tehnoloogilises arendustöös. Tartu, 1998.
8. **Renate Pajusalu.** Deiktikud eesti keeles. Tartu, 1999.
9. **Vilja Oja.** Linguistic studies of Estonian colour terminology. Tartu, 2001.
10. **Külli Habicht.** Eesti vanema kirjakeele leksikaalsest ja morfosüntaktilisest arengust ning Heinrich Stahli keele eripärast selle taustal. Tartu, 2001.
11. **Pire Teras.** Lõunaeesti vokaalisüsteem: Võru pikkade vokaalide kvaliteedi muutumine. Tartu, 2003.
12. **Merike Parve.** Väited lõunaeesti murretes. Tartu, 2003.
13. **Toomas Help.** Sõnakeskne keelemudel: Eesti regulaarne ja irregulaarne verb. Tartu, 2004.
14. **Heli Laanekask.** Eesti kirjakeele kujunemine ja kujundamine 16.–19. sajandil. Tartu, 2004.
15. **Peeter Päll.** Võõrnimed eestikeelses tekstis. Tartu, 2005.
16. **Liina Lindström.** Finiitverbi asend lauses. Sõnajärg ja seda mõjutavad tegurid suulises eesti keeles. Tartu, 2005.
17. **Kadri Muischnek.** Verbi ja noomeni püsiühendid eesti keeles. Tartu, 2006.
18. **Kanni Labi.** Eesti regilaulude verbisemantika. Tartu, 2006.
19. **Raili Pool.** Eesti keele teise keelena omandamise seaduspärasusi täis- ja osasihitise näitel. Tartu, 2007.
20. **Sulev Iva.** Võru kirjakeele sõnamuutmissüsteem. Tartu, 2007.
21. **Arvi Tavast.** The translator is human too: a case for instrumentalism in multilingual specialised communication. Tartu, 2008.
22. **Evar Saar.** Võrumaa kohanimede analüüs enamlevinud nimeosade põhjal ja traditsioonilise kogukonna nimesüsteem. Tartu, 2008.
23. **Pille Penjam.** Eesti kirjakeele *da-* ja *ma-*infinitiiviga konstruktsioonid. Tartu, 2008.
24. **Kristiina Praakli.** Esimese põlvkonna Soome eestlaste kakskeelne keelekasutus ja koodikopeerimine. Tartu, 2009.
25. **Mari Mets.** Suhtlusvõrgustikud reaajas: võru kõnekeele varieerumine kahes Võrumaa külas. Tartu, 2010.

26. **Karen Kuldnokk.** Militaarne retoorika. Argumentatsioon ja keeleline mõjutamine Eesti kaitsepoliitilises diskursuses. Tartu, 2011.
27. **Kai Tafenu.** Uue Testamendi tõlkimisest Rootsi ajal: käsikirjad, tõlkijad ja eesti kirjakeel. Tartu, 2011.
28. **Küllil Prillop.** Optimaalsusteoreetiline käsitus eesti keele fonoloogilisest kujunemisest. Tartu, 2011, 261 lk.
29. **Pärtel Lippus.** The acoustic features and perception of the Estonian quantity system. Tartu, 2011, 146 p.
30. **Lya Meister.** Eesti vokaali- ja keuskategooriad vene emakeelega keelejuhtide tajus ja häälduses. Eksperimentaalfoneetiline uurimus. Tartu, 2011, 145 lk.
31. **Kersti Lepajõe.** Kirjand kui tekstiliik. Riigieksamikirjandite tekstuaalsed, retoorilised ja diskursiivsed omadused. Tartu, 2011, 141 lk.
32. **Tiit Hennoste.** Grammatiliste vormide seoseid suhtlustegevustega eesti-keelses suulises vestluses. Tartu, 2013, 236 lk.
33. **Helena Metslang.** Grammatical relations in Estonian: subject, object and beyond. Tartu, 2013, 443 p.
34. **Kristel Uihoaed.** Verbiühendid eesti murretes. Tartu, 2013, 227 lk.
35. **Helen Plado.** Kausaalsuhete adverbiaalused eesti keeles. Tartu, 2013, 244 lk.
36. **Annika Kungas.** Pragmatiliste markerite kujunemine ja funktsioonid eesti keeles *It*-sõnade näitel. Tartu, 2014, 200 lk.
37. **Maarika Teral.** Arvutipõhine eesti keele õpe: vahendid ja hinnangud nende efektiivsusele Tartu ülikooli keelekursuste näitel. Tartu, 2015, 175 lk.
38. **Anni Jürine.** The development of complex postpositions in Estonian: a case of grammaticalization via lexicalization. Tartu, 2016, 315 p.
39. **Kairi Tamuri.** Basic emotions in read Estonian speech: acoustic analysis and modelling. Tartu, 2017, 238 p.
40. **Sven-Erik Soosaar.** Eesti keele kultuursõnavara arengujooni – eesti keel kui piirikeel. Tartu, 2017, 157 lk.
41. **David Ogren.** Object case variation in Estonian *da*-infinitive constructions. Tartu, 2018, 212 p.
42. **Reet Hendrikson.** Kas sõjasõna sünnib sõtta? Erialakeele tõhusus sõjandusterminoloogia näitel. Tartu, 2018, 252 lk.
43. **Eva Velsker.** Ajaväljendid eesti murretes. Tartu, 2018, 216 lk.
44. **Riina Reinsalu.** Juhendavad haldustekstid žanriteoreetilises raamistikus. Tartu, 2019, 168 lk.
45. **Ann Siiman.** Vormikasutuse varieerumine ning selle põhjused osastava ja sisseütleva käände näitel. Tartu, 2019, 209 lk.
46. **Tiina Laansalu.** Kesk-Harjumaa kohanimed kujunemine ja nimesiire. Tartu, 2020, 216 lk.
47. **Mariko Faster.** Hargla kihelkonna vetenimed Eesti-Läti piirialal. Tartu, 2020, 218 lk.