Assessing the productivity of the Estonian deverbal suffix *-mine* in five registers of Estonian

Maarja-Liisa Pilvik
University of Tartu

Abstract

This article takes a usage-based quantitative approach to assess the morphological productivity of the deverbal action nominal suffix *-mine* (e. g. *magamine* 'sleeping') in Estonian newspaper texts, fiction, scientific texts, spoken regional dialects, and spoken spontaneous common language. While it is possible to derive an action noun with *-mine* from every verb, its productivity in different registers varies according to the aspects highlighted by the different measures: realized, expanding, and potential productivity. In addition to these measures, the ratio of the verb stems realized as verbs, and the stems realized as *-mine* nouns is compared to detect stems which are attracted to the nominalized structure more than would be expected by their overall occurrence in the corpora. The results of this study indicate that even derivation suffixes with 'absolute' productivity vary in terms of their contribution to the growth rate of the vocabulary of a given register.

Keywords: corpus linguistics, derivation, morphological productivity, nominalization, Estonian

1 Introduction

The suffix *-mine* (e. g. *kakle-mine* 'fighting') is considered to be the most regular and productive means for deverbal nominalization in Estonian (Erelt et al. 1993; Kasik 2009; Erelt 2017). Regularity is generally understood to refer to the formal, syntagmatic aspect of the derivation process, i. e. the fact that it is possible to derive a noun from every verb stem in exactly the same way using this suffix. *-mine* always attaches to the strong stem of the verb and therefore can be easily formed from the 1st infinitive form (*hüp**p**a-ma* jump-1INF 'to jump' → *hüp**p**a-mine* jump-NM.NOM 'jump-ing' vs. *hü**p**a-ta* jump-2INF 'to jump'). While the lexical category changes, the semantics generally does not, and the relationship between the derivative and

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

the underlying verb is transparent. Productivity, in turn, is defined somewhat differently in morphological approaches to word formation, depending also on the status of the language as an object of study, and may also include the notion of regularity (e. g. Pinker 1999; in Estonian, Kasik 2009; 2013). According to Rainer (1987, cited in Gaeta & Ricca 2015), as many as six meanings can be associated with the notion of morphological productivity alone:

1. the number of words formed with a certain word-formation rule;
2. the number of novel words created with a certain word-formation rule;
3. the possibility of creating new words with a certain word-formation rule;
4. the probability of creating new words with a certain word-formation rule;
5. the number of possible words formed with a certain word-formation rule;
6. the relation between occurring and possible words formed with a certain word-formation rule.

Even more extensive lists have been made with regards to the meanings of 'productive' in linguistics in general (cf. Barðdal 2008: 10–11 for a list of 19 different usages). Regardless of the exact definition, the notion of morphological productivity is most often used when distinguishing between morphological categories with a more or less fixed membership and those with a growing membership. While the suffix *-mine* surely belongs to the latter class due to the regularity of the word-formation rule, its productivity would vary with regards to at least some of the 6 aforementioned definitions. The different units of assessment, such as *number*, *probability*, and even the abstract *possibility*, imply that the productivity of a category can be understood as either something fixed, relational, or purely potential/theoretical.

In this article, I follow the general understanding in usage-based linguistics, that productivity is the probability of a grammatical pattern being used to create novel structures (e. g. Bybee 2001; Baayen 2003). It has been shown that a productive morphological pattern is characterized by a large number of *hapax legomena*, i. e. lexemes/types in the category occurring only once in a given corpus since the productivity of the word-formation rule also guarantees the understanding of new entities. The distribution of lexemes of a non-productive category, however, is completely different: there are many high-frequency

lexemes, fewer rare and very few (if any) lexemes which occur only once (Baayen & Renouf 1996; Baayen 2009). The notions of *type* and *hapax legomena* are utilized also in more precise measures developed to highlight different aspects of productivity in quantitative approaches to morphological productivity (e. g. Baayen 1992; 1993; Baayen & Renouf 1996).

It has also been demonstrated that in addition to a number of syntactic and lexical phenomena (Biber 1995), also certain derivative affixes may be more characteristic of certain types of texts, although as Plag et al. (1999) note, nominalization does not appear to play a crucial role in register differentiation. Nevertheless, it has been noted that in Estonian, derivatives with the suffix *-mine* are especially characteristic of written scientific, legal, and administrative texts (Kerge 2003), but have spread also to journalistic texts (Kasik 2006).

Although the derivational patterns (incl. *-mine* derivation) and their productivity in Estonian have been extensively described (e. g. Vare 1994; Kasik 2009), frequencies and their distributions in different registers have not been empirically and systematically accounted for.[1] This is probably because productivity has been associated with language capacity rather than language use (similarly to e. g. Anshen & Aronoff 1999 and Dressler 2003), and the role of frequency has only been seen in the context of lexicalization or semantic specialization of derivatives (Kasik 2011). However, by defining productivity through the concept of probability, the productivity of a derivative pattern can be considered scalarized, and usage frequencies could be operationalized not only to identify lexicalized linguistic units, but also to study the variation in the productivity of certain derivational patterns across different types of texts.

The purpose of the current study is to approach morphological productivity from a quantitative, usage-based perspective and to evaluate the productivity of the suffix *-mine* in different Estonian sub-languages and registers represented in the Balanced Corpus of Estonian (written newspaper texts, fiction, and scientific texts), the Corpus of Estonian Dialects (oral traditional dialects), and the Phonetic Corpus of Estonian Spontaneous Speech (oral common language). To examine the differences between the five registers, I use the realized, expanded, and potential productivity (e. g. Baayen

[1] Krista Kerge's studies (2002; 2003), where the quantitative analysis of the complexity of text types implicitly also included productivity as the ratio between token count of *-mine* nouns and the number of tokens in the corpus, may be considered as exceptions here. Kerge also studied the change of this ratio diachronically, providing an interesting insight into the effect of language policies and ideological orientation on the use of *-mine* nouns.

1992; 2009; Plag et al. 1999) of *-mine* derivational pattern. Based on e. g. Plag et al. (1999: 218), I hypothesize that derivation, which is said to contribute the most to the growth of vocabulary, is much less productive in spoken registers than in written registers. I will also study the ratios of verbal stems realized as verbs and *-mine* nominalizations in the different corpora to unveil the (dis)preferences of the suffix *-mine* with regards to the stems to which it attaches.

In what follows, I will first describe *-mine* nominalization in the light of its formation and main functions. Then, I will discuss the aforementioned measures that have been proposed for assessing morphological productivity. This will lead to presenting my data and comparative findings for the Estonian suffix *-mine*. Finally, I will discuss the implications of these findings to the general understanding of the productivity of *-mine* in Estonian and the pitfalls of the quantitative corpus-based approach taken in this article.

2 Form and functions of *-mine* nominalization

The suffix *-mine* is the most common means for deverbal predicate nominalization in Estonian (Erelt et al 1993; Erelt 2017) and translates roughly as *-ing* in English. Syntactically, it is used to create a secondary predication for using a clause as an argument or modifier in another clause. The agent (1) or the patient (2) of the underlying predicate is usually expressed as a genitive modifier in the resulting noun phrase. The patient can also form a compound with the action nominal (3), in which case the patient specifies a type of action expressed by the verb.

(1) *isa* *peksa-b* → *isa* *peks-mine*
father beat-3SG father.GEN beat-NM.NOM
'father is beating' 'father's beating'

(2) *isa* *peks-t-akse* → *isa* *peks-mine*
father.PART beat-IPS-PRS father.GEN beat-NM.NOM
'father is being beaten' 'the beating of the father'

(3) *isa+peks-mine*
father+beat-NM.NOM
'fatherbeating'

Formally, the resulting forms are considered to belong to the noun word class, since they have a full number and case paradigm, they can take genitive, pronominal or adjectival attributes (including those not in the structure of the underlying predicate), and among other things, can function syntactically also as subjects, objects and predicatives of the clause. These NPs also exhibit a number of verbal properties, such as being able to retain the underlying verb's adverbial modifiers and predicatives or sometimes even govern the object (cf. Pilvik 2016; 2017). Due to this categorial ambiguity, the "productively" formed action nominals are sometimes distinguished from the lexicalized, aspectually bounded, or semantically specialized verbal nouns (Kasik 2009; Erelt 2017: 831–832). Indeed, *-mine* nominalizations do not always refer to the aspectually unbounded processes, but also to single events as well as the results or objects of these events (Kasik 2009). Those structures have been said to be idiomatized, i. e. semantically opaque, and do not strictly fit in the action nominal category (Erelt 2017: 832). The properties which are usually linked to the lexicalized meanings are pluralization (*vali-mise-d* 'elections' vs. 'electings'), absence of the predicate's obligatory arguments (*tead-mine* 'knowledge' vs. 'knowing'), and incorporating a specifier (*söögi-tege-mine* 'cooking' vs. 'foodmaking') (Kerge 2001: 42–43). However, the mere presence of one or several of these attributes does not suffice to distinguish between the processual and referential *-mine* nouns, nor is there enough reason for differentiating between the suffix producing processual nominalizations and the one responsible for just the idiomatic structures. Instead, defining processuality as a gradient and highly context-related property of *-mine* nouns seems more appropriate to account for all the possible aspectual and referential readings in addition to the strict reading of action (Kerge 2001: 40; Pilvik 2017: 304). Also, while productivity characterizes the whole derivational pattern, idiomatization only concerns specific derivatives, which can acquire additional semantic properties in addition to their categorial meaning (Kasik 2009). This implies that the more lexicalized units could be detected based on their type frequencies as it is well-known that opaque forms tend to be high-frequency words.

Research on *-mine* nominalization in Estonian seems to fall largely into three: studies which focus on the syntactic function of the nominalization (i. e. applying the nominalization suffix to use a clause in the position of a noun, Kasik 1968; 1975; Kerge 2001), studies which focus on its lexical function (i. e. applying the nominalization suffix to create new terms, Kerge 2002; 2003), and studies covering its textual functions such as abstracting,

generalizing, or thickening the text (Kasik 1995; 2006). All functions seem to operationalize a productive rule, but they do that for different purposes. The functions for which *-mine* nouns are created have been shown to depend on the register. As most studies have been done based on written language, the functions considered have been the ones prevailing in namely written domains of language use. Thus, legal texts mostly make use of *-mine* nouns for the syntactic purpose as well as for thickening the text (Kerge 2002: 38); same functions apply in journalistic texts (Kasik 2006), although the need for abstract concepts is also mentioned (Kerge 2002); academic and scientific texts mostly use *-mine* nouns for the need of specific terminology (Kerge 2002; 2003). Not much has been said about spoken Estonian, but one can assume that the very conditions under which spoken and written language are produced also affect the functions in which nominalization is used. A study on dialect data, for example, demonstrated the discourse-related functions of *-mine*, such as anaphoric referencing, relaxing the processing load, or providing time for utterance planning (cf. Pilvik 2017).

In addition to functions, different registers also tend to be used to communicate different kinds of topics (Baayen 2009), which affects the distribution of lexemes affected by suffixation. The effects of the functional and topical bias have to be taken into account when interpreting productivity measures described in the next subsection.

3 Measuring productivity

In a probabilistic approach to morphology, productivity is not affected only by syntagmatic regularity, but also by schemas, constructions, and generalizations. According to Baayen (2009: 901), "morphological productivity can be understood as resulting from a great many factors such as the individual language user's experience with the words of her language, her phenomenal memory capacities, her conversational skills, her command of the stylistic registers available in her language community, her knowledge of other languages, her communicative needs, her personal language habits and those of the people with which she interacts".

Several corpus-based measures have been developed in order to seek evidence for intuitions about morphological productivity of certain patterns, catch different aspects of productivity, and to provide comparable statistics across multiple genres and registers. I exploit the three measures discussed in

e. g. Baayen (1993), Baayen & Renouf (1996), and Baayen (2009) – realized productivity, expanding productivity, and potential productivity – since these have been extensively used in corpus studies concerning the morphological productivity of affixes in a wide range of languages. Corpus-based quantitative studies on derivational productivity have focused mainly on the differences between several derivational affixes in Indo-European languages, such as English, Dutch, and Italian, both within one register and across multiple registers (e. g. Baayen 1993; Plag 1999; Plag et al. 1999; Gaeta & Ricca 2006). Using these measures to study the variation in productivity for one suffix across registers has been less common. For Finnic languages, the measures have been applied in e. g. experimental studies in Finnish morphology (Mäkisalo 2000; Järvikivi 2003; Nikolaev & Niemi 2008).

Realized productivity C of a pattern refers to *the contribution* of a category (e. g. *-mine* derivatives) to the size of the whole vocabulary. It is also known as the *extent of use* (Baayen 1993) and is estimated by the type count (*V*) of the category. Since the productivity measures are compared across registers with slightly differing token counts in this article, *realized productivity C* here refers to the quotient of the number of types (*V*) in a given category and the total number of tokens (*N*) in a corpus. Realized productivity is therefore a relational measure in this article, instead of an absolute count. The higher the value, the more a given pattern occurs in the vocabulary and the more productive it can be considered.

*Expanding productivity P** of a pattern indicates *the expansion rate* of a given category. This is also known as the *hapax-conditioned degree of productivity* and is calculated by dividing the number of hapax legomena in a given category (*n1*) by the number of hapaxes in the whole corpus (*N1*). The higher the value, the more a given category attracts new structures (compared to other possible categories) and the more productive it can be considered. In probabilistic terms, expanding productivity *P** shows the conditional probability that a word belongs to a certain category, given that this word occurs in the corpus only once.

Finally, *potential productivity P* reflects *the growth rate of the vocabulary* in a given category. It is also called the *category-conditioned degree of productivity* and is calculated by dividing the number of hapax legomena in a given category (*n1*) by the number of tokens in this category (*Nc*). The higher the value, the more lexically diverse the category is and the more productive it can be considered. In probabilistic terms, potential productivity shows the conditional probability, that a word occurs in the corpus only once, given

Table 1. Three measures of morphological productivity

Measure	Calculation	Reflects
Realized productivity	$C = V/N$	Contribution of the category to the size of all vocabulary.
Expanding productivity	$P* = n1/N1$	Attraction of novel structures by the category.
Potential productivity	$P = n1/Nc$	The potential of creating new structures within the category.

that it belongs to a certain category. This measure is designed to reflect the intuition that it is easier to think of neologisms with one suffix (e. g. *-mine*) than it is with another (e. g. another Estonian deverbal suffix *-us*) (Baayen 2009). Table 1 presents the summary of the three measures.

It is well-known that words with less compositional meaning tend to have higher usage frequencies, but a binary division between unproductive lexicalized elements (which are stored in lexicon) and productive schemas (which are stored in grammar) is usually not appropriate, especially with *-mine* nominalization which produces structures on the whole scale from non-referential to fully referential (cf. Pilvik 2017). It has also been shown with experiments that all complex words (not only irregular or "lexicalized" words) leave traces in lexical memory (Hay & Baayen 2005). In order to compare the likelihood of encountering nouns with a more or less fixed meaning in different registers, I have calculated what I call here *NV-score* for each individual verb stem in the samples of the five different subcorpora. NV-score of a verb stem is the difference between the proportion of that stem among the *-mine* nouns and the proportion of that stem among verbs. When this difference is 0, the likelihood of this stem occurring in the corpus as a verb is equal to the likelihood of it occurring as a *-mine* noun. When the difference is > 0, the stem is less likely to occur in a nominalized structure than would be expected given its probability to occur as a verb in the corpus. When the difference is < 0, the stem is more likely to occur in a nominalized structure than would be expected. The reason for doing this instead of simply looking for the most frequent types in each corpus is to eliminate the effect of the above-mentioned topical bias in different registers. The type-based NV-score provides additional perspective to straightforward productivity measures: if

some verbs are more likely to be nominalized with *-mine* than would be expected by their overall frequency in the corpus, then we can assume the corresponding nominalized structures are possibly also more entrenched in the lexicon. The admittedly simple reasoning behind this assumption relies on the regularity of the word-formation pattern: if all verbs can be nominalized with *-mine*, then *-mine* nouns created for syntactic functions should show similar distribution to verbs with regards to the lexical stems from which they are formed. In turn, *-mine* nouns created for lexical or textual purposes are independent of the verbs used to discuss certain topics in a given register. The stems rarely nominalized are expected to exhibit semantic properties, which make them less likely to be construed as an entitized referent instead of a situational one, or belong to the class of verbs often used in non-normal clauses such as existential, possessive, exeriencer, or resultative clause.[2] Formally, however, no lexical restrictions exist for *-mine* nominalization.

In recent decades, a wealth of researchers have turned from corpus data to experimental studies in order to provide even stronger support for the gradual and probabilistic view of morphology (cf. an overview in Hay & Baayen 2005). In these approaches, the morpheme as a theoretical construct is backgrounded, and instead, full words are viewed as the basic units in the lexicon. It is paradigmatic relations, instead of a set of lexical entries and rules, which help characterize lexical representations and determine lexical processing. In the context of *-mine* nouns, this would mean accounting for the support a derived noun gets from other words occurring in the same morphological family (*laul-mine* sing-NM 'singing', *laul-ma* sing-1INF 'to sing', *laul-da* sing-2INF 'to sing', *laul-ja* sing-AN 'singer' etc.) to explain e. g. reaction times and precision in lexical decision or word naming tasks. This highly intriguing perspective remains outside the scope of this article, which at this stage limits itself only to corpus data. However, considering the frequency of individual verb stems (covering all the verbs' inflectional forms) in different corpora is also a step towards higher paradigmatic awareness when assessing the degree of lexicalization.

[2] Erelt et al. (1993: 269) make the claim that verbs which are never used in Estonian normal clauses (e. g. *piisama* 'to suffice, to be enough' in *Pildi tegemiseks piisab ka tavalisest kaamerast* 'An ordinary camera is enough to take a photo') can not be nominalized. While I do not hold this to be true, I do expect non-normal clauses to be nominalized less frequently.

4 Data

The data for this study comes from 3 different corpora, representing altogether five different registers of Estonian. As the aim of this article is to provide a global view of the variation in probabilistic productivity of the suffix *-mine*, the subdivision into registers is taken to be predefined by the corpora and a more detailed view of the different text types (e. g. sports commentaries against interviews) is currently ignored.

The Balanced Corpus of Estonian[3] (BCE) comprises 5 mln tokens from newspapers, 5 mln tokens from fiction, and 5 mln tokens from scientific texts. The newspapers subcorpus (NEWS) dates from 1995 to 2007 and contains both daily and weekly newspapers; fiction subcorpus (FICT) contains excerpts from Estonian literature, both prose and poetry, from the period between 1987 and 2011; the scientific subcorpus (SCI) holds Ph. D. dissertations and articles from scientific journals, covering a wide range of disciplines, and dates from 1995 to 2006. BCE is a subpart of the big Mixed Corpus of Estonian and is compiled first and foremost to enable the comparison of the genres and text classes of written Estonian. The corpus also has a morphologically annotated version, where each token has received automatic morphological analysis.

The Corpus of Estonian Dialects[4] (CED) comprises about 900,000 tokens of manually morphologically annotated transcriptions of recordings from the 1960s–1970s, covering all 10 Estonian traditional dialect areas (DIA). The recordings are unstructured interviews, where elderly informants speak about their everyday life, childhood, customs, traditions, past events, and work. The transcripts follow the conventions of the dialect corpus's simplified transcription.[5] However, as both the original transcriptions (following Finno-Ugric transcription rules) as well as morphological annotation have been done manually and neither strictly follows the orthographic rules of written Estonian, there are also some idiosyncracies. For example, compound words are written using both separate writing (*ära viimine* 'taking away') and solid writing with the transcription sign + (*ära+viimine* 'taking away').

The Phonetic Corpus of Estonian Spontaneous Speech[6] (PCESS) is the

[3] https://www.cl.ut.ee/korpused/grammatikakorpus/ (accessed 2018-01-01).
[4] https://www.keel.ut.ee/et/keelekogud/murdekorpus/ (accessed 2015-09-29).
[5] In addition to phoneme realizations, simplified transcription also marks short and long pauses, liaisons, quantity alternations and compound words; https://www.keel.ut.ee/sites/default/files/www_ut/emk_teejuht2015.pdf (accessed 2019-03-01).
[6] https://www.keel.ut.ee/en/languages-resourceslanguages-resources/

Table 2. Token counts, type/token ratios, and growth rates of the (sub)corpora[7]

(Sub)corpus	Corpus Size in tokens	Type/token ratio (TTR)	Growth rate
NEWS	4,675,823	0.045	0.024
FICT	4,953,823	0.033	0.018
SCI	4,798,966	0.048	0.025
DIA	890,788	0.028	0.013
SP	338,807	0.052	0.027

smallest corpus in this study, comprising only around 340,000 tokens (SP). The corpus contains 60 hours of speech recordings, both dialogues between familiar people and monologues, recorded between the years 2006 and 2016. Most recordings are done in a studio, some have also been done during fieldwork. The average length of the recordings is 30 minutes. As the name of the corpus suggests, the purpose of its compilation was to provide data on the phonetic traits of spontaneous speech (e. g. speech sounds, sound structures, syllables, assimilation, voice quality), but its accessibility and the abundance of different segmentation layers (incl. morphological annotation) makes this corpus a valuable resource also for non-phoneticians.

Table 2 shows the size of the corpora in morphologically annotated tokens[8] and the type/token ratios (TTR), which are considered a good indication of the range of vocabulary used in the corpus under consideration. A high TTR indicates more lexical variation and a low TTR relatively little lexical variation (Baker et al. 2006: 162). The growth rate in the last column is calculated by dividing the number of hapaxes by the total number of tokens and translates as the probability of the occurrence of a type not encountered before in the corpus.

Simple type/token ratio (TTR) is very sensitive to the size of the text or

phonetic-corpus-estonian-spontaneous-speech (accessed 2018-09-02).

[7] Types are counted according to their specified part-of-speech to decrease the chance that homonymic forms, which represent different linguistic categories, will be counted as one type. While *palk* 'log' and *palk* 'payment' will still remain one type, *või* 'or' and *või* 'butter' will not. This, however, can artificially increase the type count where manual annotation has been involved.

[8] Punctuation is excluded.

Table 3. Token counts, type/token ratios, and growth rates of subcorpora

Sampled corpus	Sample size in tokens	Type/token ratio (TTR)	Growth rate
NEWS	336,984	0.123	0.068
FICT	398,796	0.075	0.041
SCI	337,696	0.100	0.052
DIA	335,325	0.041	0.019
SP	338,807	0.052	0.027

corpus. The longer the text, the more types have already been encountered and the lower the likelihood of any given token representing a new type. Therefore, it is crucial to provide texts of equal sizes when comparing their TTR (Hardie & McEnery 2006: 139). Productivity measures described above also rely on type frequencies. To enable comparison between different registers, I took a random sample from each corpus, with reference to the size of the smallest corpus, the PCESS.[9] The sampling of the files in each corpus stopped when the token count in that sample reached higher than 335,000 (condition was checked prior to sampling another file). While these samples can be considered rather small, they are still able to shed some light on the variation of productivity between the registers.[10] The sample sizes, their TTRs and growth rates are presented in Table 3.

As the comparison of Table 2 and Table 3 reveals, the TTRs and growth rates are indeed considerably higher in the samples than in the full-sized corpora (except for the spoken spontaneous language, where sample size equals corpus size) and sampling, therefore, is necessary for adequate comparisons. Table 3 shows that both TTR and growth rate are higher in the written registers than in the spoken registers. This means that written language

[9] A similar approach, although a more refined one, was taken in Plag et al. (1999), where sub-corpora were compared for the largest range of token sizes that they had in common. They also took into account the shape of vocabulary growth curve and assessed the number of types and productivity measures by calculating the average of each measure at twenty equally spaced intervals. In this article, however, the aspects of productivity are assessed only at the final values of the sample sizes.

[10] I rely on e. g. Plag et al. (1999: 215), who have written that "What exactly counts as sufficiently large is not easy to determine, but even relatively small corpora like the Dutch Eindhoven Corpus (600,000 words of written text) seem to yield interesting results".

corpora are lexically more diverse and also expanding their vocabulary at a higher rate.

From the samples, I exhaustively extracted all occurrences of noun lemmas formed with the suffix *-mine*. Extracting *-mine* nouns was a seemingly effortless task, since there are no morphonological alternations (*valmista-ma* 'to make' → *valmista-mine* 'making', *arutle-ma* 'to discuss' → *arutle-mine* 'discussing', *uuri-ma* 'to study' → *uuri-mine* 'studying', *hari-ma* 'to educate' → *hari-mine* 'educating'). The morphologically annotated corpora, thus, enabled to simply look for lemmas which ended with *-mine* and were marked for being nouns.[11] The lists were manually cleaned from foreign words (e. g. *examine*, *phenylpropanolamine*). The extraction of deverbal *-us* nouns, for example, would be somewhat more difficult, because *-us* can also be used to derive nouns from adjectives (e. g. *haige* 'ill, sick' → *haig-us* 'illness, sickness') and the derivation is subject to both stem and suffix alternations (e. g. *valmist**a**-ma* → *valmist-us* 'making', *arutl**e**-ma* → *arutl-us* 'discussion', *uuri-ma* → *uuri-**m**us* '(a) study', *hari-ma* → *hari-**d**us* 'education'). However, going through the word lists and manually correcting the mistakes also for the suffix *-mine* was crucial for the measures to be informative. Especially since with rare words, the automatic morphological analyzer does not (yet) have the capacity to always understand these structures the same way that a human interpreter would. Therefore, it occasionally proposes analyses which are not appropriate and skew the statistics.

Finally, I collected the stems of all verbs in the samples in order to see, whether the distribution of the nominalized verb lemmas corresponds to the overall distribution of the verbs used in certain registers.

5 Results

Corpus data allow examining, how the morphological productivity of the suffix *-mine* varies across registers. Each of the three measures discussed provides a slightly different perspective on this variation. As the magnitudes of the measures are essentially proportions and can be interpreted as percentages, they vary according to the frequency of the phenomenon in question. Individual derivational affixes form rather specific grammatical categories and cannot, therefore, account for a very large proportion of the

[11] The suffix *-mine* can also non-productively derive adjectives (e. g. *alumine* 'lower; lowermost', *pealmine* 'upper; uppermost').

Table 4. Realized productivity for the suffix *-mine* in 5 registers

Register	Realized productivity
NEWS	0.0031
FICT	0.0019
SCI	0.0030
SP	0.0012
DIA	0.0007

overall words. Additionally, the NV-score will help detect the verbs which are considerably more or less likely to be nominalized than would be expected by their average presence in the corpora.

The productivity measures are calculated based on the samples of approximately 335,000 tokens per corpus, which is very small, compared to the corpora used to study e. g. English derivational suffixes. However, as productivity in spoken registers has so far largely been neglected in Estonian linguistics, the results based on even relatively small datasets are extremely valuable.

5.1 Realized productivity of *-mine*

In this article, realized productivity means finding the ratio between the number of types of *-mine* nouns and the total number of tokens in the corpus. The measure reflects the contribution of the *-mine* nominalization category to the size of the whole vocabulary in the corpus, or in other words, the productivity of a morphological process in the past. The values for realized productivity are presented in Table 4.

The lowest realized productivity for the suffix *-mine* is in the dialect corpus (0.0007), the highest in news and scientific writing. This means that out of the five registers, *-mine* nouns as a grammatical category contribute the most to the size of vocabulary in these written text types where *-mine* nouns most likely function as means for abstraction and generalization, thickening the text, or providing new terms. The formation rule seems to be the least productive in dialect data with respect to the category's contribution as a whole. These results are in accordance with what is generally claimed for derivational patterns: they are much less productive in spoken

registers than in written ones. However, it seems that written fiction is somewhat closer to spoken common language. Both registers are used to communicate less specific purposes than newspapers or scientific texts. Therefore, realized productivity is a measure which depicts the different functions of *-mine* derivation and is closely linked to the written-oral dimension of communication. The differences between the registers seem quite small, even when considering that *-mine* derivation is not the most frequent phenomenon among morphological categories. When translated into percentages, the difference between the register with the highest realized productivity (NEWS) and one with the lowest (DIA) is only 0.24 percentage points. However, the absolute type frequencies of 1060 and 232, respectively, do speak of a significant difference.

5.2 Expanding productivity of *-mine*

Expanding productivity or the *hapax-conditioned degree of productivity* is found when dividing the number of hapax legomena among *-mine* derivates by the number of hapaxes in the whole corpus. The measure shows the rate with which a category gives rise to novel structures and is a more refined one compared to the previous measure. The values for realized productivity are presented in Table 5.

Expanding productivity shows the highest value again in a written register, but this time, it is fiction which exhibits high productivity of *-mine*, similarly to spoken common language, whereas scientific writing fell to the bottom of the list. A high expanding productivity means that if we were to add one extra token into our corpus, expressing some new concept, a *-mine* noun would be a plausible grammatical category for realizing that new structure. Consequently, this category will often require lexical procedural knowledge in both comprehension and production. Differences between registers imply that the relative contribution of *-mine* to the growth rate of the vocabulary in both written and spoken common language (FICT and SP, respectively) is an increasing function of the sample size with a far greater slope than, for example, in scientific writing. This is somewhat counterintuitive at first sight, since the terminological function of *-mine* in scientific texts should make it highly attractive to novel concepts. However, expanding productivity is a hapax-based measure. One can assume that terms in scientific writing are usually not created for single use, but are meant for denoting a specialized concept which is important to the subject of the text. Therefore, finding

Table 5. Expanding productivity for the suffix *-mine* in 5 registers

Register	Expanding Productivity
NEWS	0.0202
FICT	0.0274
SCI	0.0186
SP	0.0236
DIA	0.0204

specialized terms among hapaxes in scientific writing is less likely.

5.3 Potential productivity of *-mine*

The third and final measure, potential productivity, shows the growth rate of the vocabulary among *-mine* nouns and is found by dividing the number of hapaxes found with *-mine* by the total number of tokens with *-mine*. It reflects the likelihood of forming more types than are actually attested in the corpus. Potential productivity values for the five subcorpora are presented in Table 6.

The potential productivity measures show the most extensive variability amongst the three measures. Here, dialects exhibit the highest productivity for *-mine* and the differences between the registers are slightly clearer than for the other measures. In dialects, a considerable proportion (19%) of *-mine* nouns occur only once, indicating that the potential to encounter a *-mine* noun not attested before in the corpus is over nine times higher in dialects than it is in scientific texts and nearly three times higher than in newspapers. Fiction, once again, leans towards the spoken registers in terms of the more productive use of the suffix *-mine* than other written registers, making it a likely intermediate category between written and spoken genres. This measure is strongly affected by lexical diversity. As the dialect corpus contains data from 10 traditional dialect areas, the large proportion of hapaxes reflects lexical variability between these areas. A high number of types coined only once in spoken common language and fiction, in turn, reflects creativity in language use (slang, intentional wordplay etc.). Another interpretation of the measure involves stating the higher likelihood of using lexical procedural knowledge to formulate a preverbal concept (e. g. an action of some sort) in spoken dialect interviews than in newspaper texts or scientific writing. As this

Table 6. Potential productivity for the suffix *-mine* in 5 registers

Register	Potential productivity
NEWS	0.0764
FICT	0.1746
SCI	0.0260
SP	0.1358
DIA	0.1934

measure is restricted by one category, it is also more restricted as a measure for productivity with not being able to account for the potential for coining new structures with other means than the suffix *-mine*. Potential productivity is also said to be negatively affected by the amount of opaque words in a category (Baayen 2009), because of their tendency to be high-frequency words. Therefore, the low value for potential productivity in newspapers and scientific texts can be considered an indication of the presence of more *-mine* nouns with specialized meanings. Figure 1 summarizes the magnitude of different productivity measures in the five registers.

When distributed across the same scale, it can be observed that compared to potential productivity, the other two productivity measures differ less in magnitude across the five registers. Therefore, in comparing the productivity of *-mine* in these registers, the extent to which the regularity of the derivation pattern can be used to create *-mine* nouns in lexically diverse settings (irrespective of other available categories) seems to make the most difference. However, as the other two measures are relative frequencies based on considerably larger denominators, this bears no statistical significance.

5.4 NV-scores for the verbal stems

NV-scores reflect the degree to which certain verbal stems are attracted to the nominalized structure, given their overall frequencies in the different subcorpora. The overall distribution of the scores in the corpora is presented in Figure 2.

Figure 2 shows that most stems receive a score close to 0, while few stems rank considerably higher or lower. This is in accordance with the overall distribution of words in corpora, meaning that there are a lot of low-frequency

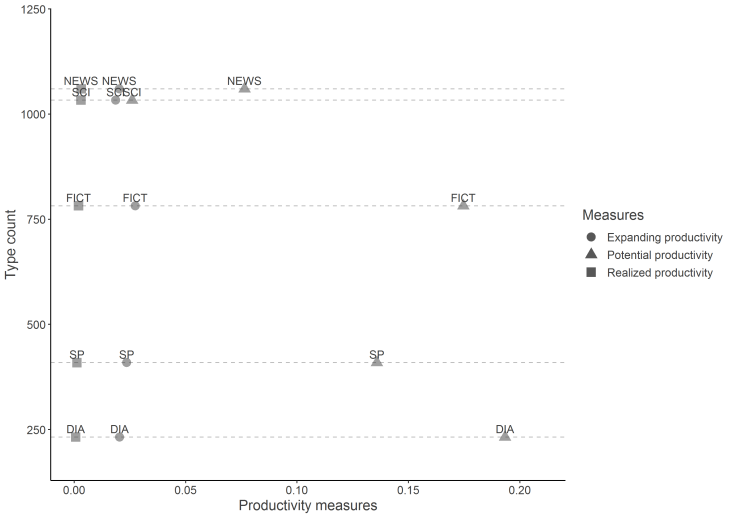

Figure 1. Comparison of three different productivity measures for the suffix *-mine*

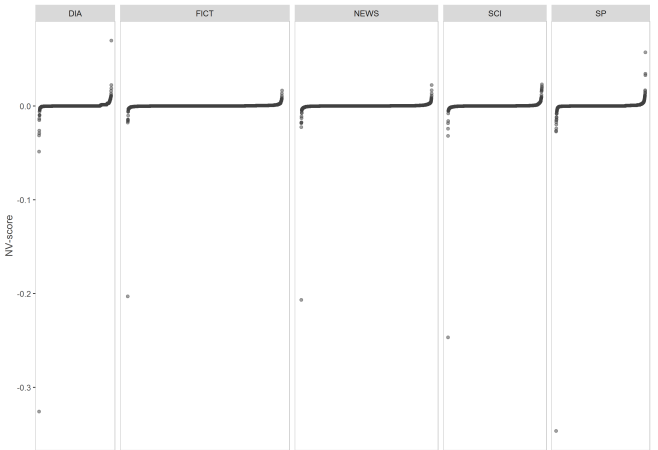

Figure 2. Distribution of NV-scores in the five registers

Table 7. Verb stems with highest NV-scores

NEWS	Score	FICT	Score	SCI	Score	DIA	Score	SP	Score
kohtu- 'meet'	0.022	*tege-* 'do'	0.017	*käitu-* 'behave'	0.023	*tege-* 'do'	0.070	*tege-* 'do'	0.057
vali- 'elect'	0.017	*liiku-* 'move'	0.013	*tead-* 'know'	0.021	*söö-* 'eat'	0.022	*liiku-* 'move'	0.034
tege- 'do'	0.013	*tead-* 'know'	0.011	*aruta-* 'discuss'	0.020	*ela-* 'live'	0.019	*luge-* 'read'	0.033
kasuta- 'use'	0.011	*ela-* 'live'	0.009	*kasuta-* 'use'	0.018	*aja-* 'drive; handle smth; cause smth'	0.016	*pakku-* 'offer'	0.017
pakku- 'offer'	0.010	*kohtu-* 'meet'	0.008	*kohalda-* 'apply; customize'	0.017	*peks-* 'beat'	0.013	*varieeru-* 'vary'	0.016

words, whose frequency divided by a larger number approaches 0, and few high-frequency words, which give slightly higher scores. The width of each facet reflects the overall number of different stems in the corresponding subcorpus. The five stems with the highest scores in each corpus are presented in Table 7.

The stems which receive the highest NV-scores and are thus strongly attracted to the nominalized structure tend to occur in more specialized meanings and constructions. For example, *kohtumine* 'meeting', *valimine* 'election', *liikumine* 'movement', *teadmine* 'knowledge', *pakkumine* 'offer' are more frequently used as referring to events, or the objects or results of the process, rather than to the processes themselves. A closer examination of the observations in the data reveals that these are the nouns also more actively used in the plural and without the predicate's arguments, fulfilling at least two of the criteria for higher degree of lexicalization. The curious case of *tege-* 'do', in turn, can be explained by its use in special constructions. In written language, *tegemine* is often used in the predicative construction *olema* 'be' + *tege-*$_{\text{NM}_{\text{PART}}}$ + $_{\text{PRED}_{\text{COM}}}$ (examples 4 and 5).

(4) *Tege-mis-t* *ol-i* *nalja-ga.*
do-NM-PART be-IPF.3SG joke-COM
'It was a joke.'

(5) *Tema näol on tege-mis-t varga-ga.*
she.GEN POSTP be.3SG do-NM-PART thief-COM
'She is a thief.'

In spoken registers, however, this construction is rare and instead, *tege-* occurs in the so-called *busy*-construction A_{ADE} + *olema* 'be' + *tege-*NM_{PART} (+ X_{COM}), illustrated by example (6), where the agent expressed as an adessive experiencer is doing something keeping her busy. The optional argument in comitative can be interpreted as the patient affected by the agent's activity. This construction is probably specialized from a general modal construction A_{ADE} + *olema* 'be' + V-$NM_{NOM/PART}$ in example (7) (cf. Pilvik 2016).

(6) *Ema-l on (looma-de-ga) tege-mis-t.*
mother-ADE be.3SG animal-PL-COM do-NM-PART
'Mother is busy (with the animals)'

(7) *Anne-l ol-i kõvasti õppi-mis-t, et teis-te-le järele jõu-da.*
Anne-ADE be-IPF.3SG hard study-NM-PART that other-PL-ADE ADV catch_up-2INF
'Anne had to study hard to catch up with the others.'

The productivity of *-mine* nouns in different syntactic constructions (cf. Neetar 1988; Sahkai 2011; Pilvik 2016; 2017) as well as the productivity of the constructions themselves definitely calls for a closer inspection. However, this additional perspective does not fit in the scope of this article.

The five stems less likely to occur in *-mine* nominalizations are presented in Table 8. The table contains stems which occur proportionally considerably more as verbs than they do as *-mine* nominalizations. Not surprisingly, these are the most frequent verb stems which are multifunctional, often used as grammatical verbs and semantically generic, but also the verbs typically used in non-normal clauses (cf. § 3). While all of these stems do also occur in nominalized structures, where *-mine* nouns are processual and most likely used for syntactic purposes, their overall high frequency as verbs keeps them at the low end of the NV-scores. An exception here is the modal verb stem *või-* 'can; may', since modal verbs are unlikely candidates for nominalization when they don't also carry a non-modal meaning (e. g. *saa-* 'get; can').

Table 8. Verb stems with lowest NV-scores

NEWS	Score	FICT	Score	SCI	Score	DIA	Score	SP	Score
ole- 'be; have'	-0.207	*ole-* 'be; have'	-0.203	*ole-* 'be; have'	-0.247	*ole-* 'be; have'	-0.326	*ole-* 'be; have'	-0.346
saa- 'get; can'	-0.022	*tule-* 'come'	-0.018	*või-* 'can; may'	-0.032	*saa-* 'get; can'	-0.048	*mine-* 'go'	-0.027
tule- 'come'	-0.018	*mine-* 'go'	-0.016	*pida-* 'must; keep'	-0.024	*mine-* 'go'	-0.031	*tule-* 'come'	-0.026
või- 'can; may'	-0.018	*saa-* 'get; can'	-0.015	*saa-* 'get; can'	-0.018	*tule-* 'come'	-0.030	*saa-* 'get; can'	-0.024
ütle- 'say'	-0.017	*või-* 'can'	-0.015	*tule-* 'come'	-0.016	*pane-* 'put'	-0.026	*ütle-* 'say'	-0.020

6 Discussion

The aim of this study is to examine the productivity of the suffix *-mine* in different registers of Estonian and use available corpus data in order to provide concrete measures gauging different aspects of productivity. This means going further than simply presenting token frequencies or posing intuitive statements based on the number of grammatical restrictions for a given morphological pattern. Accounting for the necessity of equally-sized corpora for comparisons as well as looking into the lexical distribution of the corpora allows making more targeted and specific claims about both the functional as well as structural characteristics of each register.

Findings from this study clearly show that the three measures of productivity used enable describing the productivity of *-mine* in more detail and stress important aspects of register-specific variation. Out of the three measures, the realized productivity meant for gauging the contribution of the *-mine* category to the size of the whole vocabulary showed the least differences between the five registers. The main distinction emerged between the written and oral registers: *-mine* in the three written registers exhibited higher realized productivity than in the two oral ones. This measure could also most clearly be linked to the traditional functions of *-mine* nouns in different registers: high productivity in newspapers and scientific writing can be attributed to the necessity to use nominalization as means for abstraction and

creating new terms. The low realized productivity in spoken registers could mean there are less specific functions for this category in orally communicated language.

The second, hapax-based measure (expanding productivity) set a clear distinction between scientific writing and common language. The low expanding productivity in scientific writing could be explained by the need to repeatedly use created terms, instead of creating novel structures for single use. A high expanding productivity in spoken common language and fiction, in turn, means that the relative contribution of the suffix *-mine* to the growth rate of the whole vocabulary in these registers is higher and *-mine* nominalization is a likely category to attract novel concepts.

The third, category-internal measure (potential productivity) highlighted dialects as the most likely register for forming more types with the suffix *-mine* than are actually attested in the corpus, followed by fiction and spoken spontaneous speech. In other words, in CED, PCESS and the fiction subcorpus of BCE, the lexical variability in the category of *-mine* nouns appeared to be considerably higher than in the subcorpora of newspapers and scientific texts. This can be explained by dialectal vocabulary from 10 distinct dialect areas and creative language use in spoken language and fiction. As potential productivity is sensitive to opaque words, the lower potential productivity values in newspapers and scientific texts can also be a sign of a higher use of lexicalized and idiomatized *-mine* nouns in these registers. Examining the rate with which new types are formed with the suffix *-mine* in the corpora, thus, provides a clue for measuring the availability of this word-formation rule in different registers. The speed at which the lexical inventory of the types formed with *-mine* is enriched is considerably higher in spoken registers and thus, more linked with the syntactic, rather than lexical or textual functions of *-mine* derivation.

The role of lexicalized and idiomatized nouns in written registers was partly confirmed for newspaper texts based on NV-scores, which compared the ratios of verbal stems realized as verbs with the ratios of those stems realized as *-mine* nouns in the corpora. In newspapers, the stems most attracted to the nominalized construction were ones which are easily pluralized and used as referring to events or results of processes rather than to the processes themselves (*kohtumine* ‘meeting’, *valimine* ‘election’, *pakkumine* ‘offer’). However, the scores for these stems were high also in fiction and spoken spontaneous language, but not in scientific writing. In the latter, the stems which were far more likely to be nominalized

than used as verbs were the ones used in describing objects of study and scientific procedures (*käitumine* 'behaviour', *arutamine* 'discussion', *kasutamine* 'using', *kohaldamine* 'customization'). In dialects, the verbal stems most attracted to *-mine* nominalization were words which described everyday activities (*söömine* 'eating', *elamine* 'living', *peksmine* 'beating'). A common stem amongst all registers was *tege-* ('do; make') which occurs most often in specific grammatical constructions, namely the predicative construction in written language (*Tegemist on pettusega* 'This is fraudery') and the *busy*-construction in spoken language (*Emal oli tegemist* 'Mother was busy').

These results have to be accompanied by mentioning the possible pitfalls of this study. First, as many corpus studies, this study also suffers from known limitations, starting with the lack of control over the production to the non-representativeness of the subcorpora. The different registers in this study have been taken as predefined by the corpus compilation in order to provide sufficient data for quantitative analysis. However, it is clear that none of the five registers discussed are completely homogenous (or on the contrary, sufficiently heterogenous) in terms of the content and types of text they represent. Even scientific and academic writing have different, discipline- and publication-specific writing traditions, let alone media texts or spoken spontaneous language. Second, the measures used here do not account for the fact that the observations in the corpora are not independent, i. e. same writers/speakers can contribute multiple observations. Speakers of a linguistic community have individual preferences towards the grammatical means which the language provides and in order to reflect the preferences of the whole community, each speaker would ideally contribute only one observation or passage of equal length with other passages. This is often difficult to achieve with corpus studies due to the way in which the corpora have been compiled. Third, the methodology used in this article assesses the productivity of the suffix *-mine* only at the endpoint of sampling all the corpus files. This means that the non-linear growth curve of vocabulary is not accounted for. While this is somewhat less important in examining the productivity of one single suffix, this would have more elaborate consequences when comparing the productivity of different suffixes, since the productivity of less frequent suffixes could be overestimated.

Keeping all this in mind is necessary when drawing conclusions based on the results. However, the issues mentioned above do not lessen the importance of looking for usage-based evidence for morphological productivity of certain

derivational patterns. For quantitative studies of morphological productivity, corpus studies are shown to be more adequate than using dictionaries, and considerably faster than doing experimental studies, often providing a necessary starting point for the latter. The statistical productivity measures have several applications in e. g. authorship recognition, modeling processing constraints, and dictionary compilation.

There are plenty of alleys for further research. The continuation of the current study involves the comparison of the productivity measures for *-mine* with those of other deverbal action nominal suffixes, mainly *-us*, which is said to be less productive and semantically more complex. A clear issue with this is going to be the difficulty of obtaining appropriate data due to the morphonological alternations. Based on the high degree of regularity, the productivity of *-mine* (*kõndimine* 'walking') could be compared with the productivity of the agent noun suffix *-ja* (*kõndija* 'walker'). The variation in different registers should not be omitted from either study. The diachronic aspect, which has been neglected in this article, would shed light on the dynamics of change in the productivity of *-mine* in different registers, given that appropriate data can be found for also spoken language. Finally, the increasing amount of experimental research done in linguistics to complement corpus studies should not be disregarded in the study of morphological productivity of Estonian derivational affixes. Studies involving e. g. eyetracking experiments would significantly enrich our knowledge about the constraints involved in processing and producing examples of (non-)productive morphological categories.

7 Summary

This article described a study which assessed the morphological productivity of the Estonian deverbal suffix *-mine*, which is said to be the most regular and productive means for nominalizing a verb in Estonian. By examining corpus data representing five different registers of Estonian (written newspapers, fiction, scientific texts, spoken spontaneous common language, and spoken regional dialects) it was shown that the productivity of this suffix varies across the registers, depending on the specific aspects highlighted by 3 different productivity measures: realized productivity, expanding productivity, and potential productivity. Realized productivity, reflecting the contribution of *mine*-suffixed types to the size of the whole vocabulary, was highest in

written registers, especially in newspapers and scientific texts, while lowest in dialects. This is in accordance with the general claims about derivational patterns being much less productive in spoken registers than in written ones and is likely linked to the textual functions of *-mine* nominalization. Expanding productivity, which reflects the probability that it is indeed the suffix *-mine* that is used to grow the vocabulary by a type never encountered before, was highest in fiction and lowest in scientific texts, indicating that *-mine* is not a likely category for creating novel concepts in scientific texts, contrary to what intuition of a native speaker would imply. Finally, the third measure – potential productivity – characterizes *-mine* as being most productive in dialects and spoken common language. This means that spoken registers are lexically more diverse in the category of *-mine* nouns than would be expected by their overall lexical growth rate. In addition to these three measures, which showed that even a suffix with 'absolute' productivity exhibits variation in terms of its contribution to the vocabulary used in specific registers, the proportions of verb stems realized as verbs and as nominalized structures was compared in order to detect the stems with a higher likelihood of being more lexicalized. The stems attracted to the nominalized form were indeed the ones which have acquired additional meanings (e. g. *kohtumine* 'meeting', *valimised* 'elections') or are used in frequent grammaticalized constructions. While the samples used in this study were relatively small, the results provide a solid basis for more elaborate usage-based examination of derivational morphology in Estonian.

Acknowledgements

The study was supported by the European Union through the European Regional Development Fund (The Centre of Excellence in Estonian Studies).

Abbreviations

A	agent argument of transitive verb
ADV	adverbial
AN	agent noun
ADE	adessive case
GEN	genitive case
COM	comitative case
IPF	imperfect tense
IPS	impersonal mood
NM	nominalization (with the suffix *-mine*)
NOM	nominative case
PART	partitive case
POSTP	postposition
1INF	1st infinitive (the *-ma* infinitive)
2INF	2nd infinitive (the *-da* infinitive)
PPP	passive past participle
PL	plural
PRED	predicative
PRS	present tense
SG	singular
V	verb stem

References

Anshen, Frank & Aronoff, Mark. 1999. Using dictionaries to study the mental lexicon. *Brain and Language* 68(1–2). 16–26.

Baayen, R. Harald. 1992. Quantitative aspects of morphological productivity. In Booji, Geert E. & van Marle, Jaap (eds.), *Yearbook of morphology 1991*, 109–149. Dordrecht: Kluwer Academic Publishers.

— 1993. On frequency, transparency, and productivity. In Booji, Geert E. & van Marle, Jaap (eds.), *Yearbook of morphology 1992*, 181–208. Dordrecht: Kluwer Academic Publishers.

— 2003. Probabilistic approaches to morphology. In Bod, Rens & Hay, Jennifer & Jannedy, Stefanie (eds.), *Probabilistic linguistics*, 229–287. Cambridge, MA: The MIT Press.

— 2009. Corpus linguistics in morphology: Morphological productivity. In Lüdeling, Anke & Kytö, Merja (eds.), *Corpus linguistics: An international handbook*, 900–919. Berlin: Mouton de Gruyter.

Baayen, R. Harald & Renouf, Antoinette. 1996. Chronicling The Times: Productive lexical innovations in an English newspaper. *Language* 72(1). 69–96.

Baker, Paul & Hardie, Andrew & McEnery, Tony. 2006. *A glossary of corpus linguistics*. Edinburgh: Edinburgh University Press.

Barðdal, Jóhanna. 2008. *Productivity: Evidence from case and argument structure in Icelandic*. (Constructional approaches to language 8). Amsterdam: John Benjamins.

Biber, Douglas. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press.

Bybee, Joan L. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Dressler, Wolfgang U. 2003. Degress of grammatical productivity in inflectional morphology. *Italian Journal of Linguistics* 15(1). 31–62.

Erelt, Mati. 2017. Sekundaartarindiga laused [Sentences with secondary constructions]. In Erelt, Mati & Metslang, Helle (eds.), *Eesti keele süntaks* [The syntax of Estonian], 756–840. (Eesti keele Varamu III). Tartu: Tartu Ülikooli Kirjastus.

Erelt, Mati & Kasik, Reet & Metslang, Helle & Rajandi, Henno & Ross, Kristiina & Saari, Henn & Tael, Kaja & Vare, Silvi. 1993. *Eesti keele grammatika II: Süntaks. Lisa: Kiri* [Estonian grammar II: Syntax. Appendix: Orthography]. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Gaeta, Livio & Ricca, Davide. 2006. Productivity in Italian word formation: A variable-corpus approach. *Linguistics* 44(1). 57–89.

— 2015. Productivity. In Müller, Peter O. & Ohnheiser, Ingeborg & Olsen, Susan & Rainer, Franz (eds.), *Word-formation: An international handbook of the languages of Europe*, vol. 2, 842–858. (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of linguistics and communication science (HSK) 40/2). Berlin: Walter de Gruyter.

Hardie, Andrew & McEnery, Tony. 2006. Statistics. In Brown, Keith (ed.), *Encyclopaedia of language & linguistics*, 138–146. Oxford: Elsevier.

Hay, Jennifer B. & Baayen, R. Harald. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342–348.

Järvikivi, Juhani. 2003. *Allomorphy and morphological salience in the mental lexicon*. Joensuu: Joensuun yliopisto. (Doctoral dissertation).

Kasik, Reet. 1968. Omastavalise täiendi subjektilisusest ja objektilisusest *mine*-konstruktsioonis [On the subjecthood and objecthood of the genitive attribute in *-mine* construction]. In *Keel ja struktuur: töid strukturaalse ja*

matemaatilise lingvistika alalt 2 [Language and structure: Works from structural and mathematical linguistics 2], 125–139. Tartu: Tartu Riiklik Ülikool.

— 1975. Verbide ja verbaalsubstantiivide tuletusvahekorrad tänapäeva eesti keeles [Derivational relations of verbs and verbal nouns in contemporary Estonian]. In *Keele modelleerimise probleeme 5* [Issues in language modeling 5], 4–162. (Tartu Riikliku Ülikooli toimetised 363). Tartu: Tartu Riiklik Ülikool.

— 1995. Õiguskeele lauseehitus [Syntax of legal texts]. *Õiguskeel* 1. 16–21.

— 2006. Nominalisatsioon meediauudiste tekstimoodustusvõttena [Nominalization as a textual strategy of news]. *Keel ja Kirjandus* 2/2006. 122–134.

— 2009. *Eesti keele sõnatuletus* [Derivation in Estonian]. Third, corrected edn. Tartu: Tartu Ülikooli Kirjastus.

— 2011. Sõnatuletus leksika ja grammatika vahel: *nd*- ja *ndus*-liitelised verbaalnoomenid [Derivation between lexicon and grammar: Verbal nouns with *-nd* and *-ndus* suffixes]. *Emakeele Seltsi aastaraamat* 56 (2010). 63–90.

— 2013. *Komplekssete sõnade struktuur* [Structure of complex words]. (Tartu Ülikooli eesti keele osakonna preprindid 3). Tartu: Tartu Ülikool.

Kerge, Krista. 2001. Action nouns on *-mine* in Estonian grammar and lexicon. In Niemi, Jarkko & Heikkinen, Janne (eds.), *Nordic and Baltic morphology: Papers from a NorFa course, Tartu, June 2000*, 34–45. (Studies in Languages 36). Joensuu: University of Joensuu.

— 2002. Kirjakeele kasutusvaldkondade süntaktiline keerukus [Syntactic complexity of functional domains of literary language]. In Kasik, Reet (ed.), *Tekstid ja taustad: Artikleid tekstianalüüsist* [Texts and backgrounds: Articles on text analysis], 29–46. (Tartu Ülikooli eesti keele õppetooli toimetised). Tartu: Tartu Ülikooli Kirjastus.

— 2003. *Keele variatiivsus ja* mine*-tuletus allkeelte süntaktilise keerukuse tegurina* [Linguistic variativity and *-mine* derivation as a factor in lectal syntactic complexity]. Tallinn: TPÜ kirjastus. (Doctoral dissertation).

Mäkisalo, Jukka. 2000. *Grammar and experimental evidence in Finnish compounds*. Joensuu: Joensuun yliopisto. (Doctoral dissertation).

Neetar, Helmi. 1988. Mõnest teonimekonstruktsioonist Eesti murretes [On some action nominal constructions in Estonian dialects]. *Emakeele Seltsi aastaraamat* 32 (1986). 36–45.

Nikolaev, Alexandre & Niemi, Jussi. 2008. Suomen nominien taivutusjärjestelmän produktiivisuuden indekseistä [On the productivity indices of the inflectional paradigm of Finnish nouns]. *Virittäjä* 112(4). 518–544.

Pilvik, Maarja-Liisa. 2016. *olema* + V*mine* konstruktsioonid eesti murretes [Constructions with *olema* + V*mine* in Estonian dialects]. *Keel ja Kirjandus* 6/2016. 429–446.

— 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.

Pinker, Steven. 1999. *Words and rules: The ingredients of language*. New York: Basic Books.

Plag, Ingo. 1999. *Morphological productivity: Structural constraints in English derivation*. Berlin: Mouton de Gruyter.

Plag, Ingo & Dalton-Puffer, Christiane & Baayen, R. Harald. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3(2). 209–228.

Rainer, Franz. 1987. Produktivitätsbegriffe in der Wortbildungslehre [Productivity notions in word-formation]. In Wolf, Dietrich & Gauger, Hans-Martin & Geckeler, Horst (eds.), *Grammatik und Wortbildung romanischer Sprachen* [Grammar and word formation in the Romance languages], 187–202. Tübingen: Narr.

Sahkai, Heete. 2011. *Teine grammatika: Eesti keele teonimede süntaks konstruktsioonipõhises perspektiivis* [Second grammar: The syntax of Estonian action nominals in a constructionalist perspective]. Tallinn: Tallinna Ülikool. (Doctoral dissertation).

Vare, Silvi. 1994. *Nimi- ja omadussõnatuletus tänapäeva kirjakeeles* [Nominal and adjectival derivation in contemporary literary language]. Tartu: Tartu Ülikool. (Doctoral dissertation).

Contact information:

Maarja-Liisa Pilvik
Institute of Estonian and General Linguistics
University of Tartu
Jakobi 2
51014 Tartu
Estonia
e-mail: maarja-liisa(dot)pilvik(at)ut(dot)ee