ESUKA – JEFUL 2021, 12–1: 189–258

COMPARING THE PRODUCTIVITY OF ESTONIAN DEVERBAL SUFFIXES *-MINE*, *-US*, AND *-JA* IN FIVE REGISTERS: A QUANTITATIVE USAGE-BASED APPROACH

Maarja-Liisa Pilvik
University of Tartu, EE
maarja-liisa.pilvik@ut.ee

Abstract. This article provides an empirical, usage-based account of the different aspects of morphological productivity of three Estonian deverbal suffixes: *-mine*, *-us*, and *-ja*, in five different registers. The fundamental quantitative measures developed by Baayen (1989, 1992, 1993) and his colleagues are applied to relatively small corpus samples in order to test how well these measures conform to the linguist's intuition about the productivity of the derivation patterns under different communicative settings. The results suggest that while the sample size does affect the reliability of the results, the measures prove a useful approximation of productivity in different registers, even for samples with low token counts.

Keywords: nominalization, morphology, productivity, register variation, corpus linguistics, Estonian

DOI: https://doi.org/10.12697/jeful.2021.12.1.06

1. Introduction

In a morphologically rich language such as Estonian, there are several derivation suffixes which enable the creation of nouns from other word classes. In fact, nominal derivation (nominalization) is the domain with the highest number of suffixes in Estonian (Kasik 2004). Nouns can be derived from verbs, adjectives, adverbs, other nouns or even pronouns (Kasik 2004: 12). However, not all suffixes are equally productive in terms of their availability for coining novel word forms. In this article, I focus on nouns which are derived from verbs and compare the productivity of three deverbal nominalization suffixes, namely *-mine* (1), *-us* (2) and *-ja* (3).

(1) *valitse-**mine*** ‘ruling; governing; controlling’

(2) *valits-**us*** ‘(a) government; ruling; governing; controlling’

(3) *valitse-**ja*** ‘ruler; governer; controller’

The suffixes *-mine* (typically used for deriving action nouns) and *-ja* (typically used for deriving agent nouns) can, in principle, attach to any verb stem. Since their formation is systematic, and the relationship between the base and the suffix is semantically transparent, those suffixes are considered highly productive. The *-us* pattern, on the other hand, can be used for deriving nouns from multiple word classes, has several morphophonological constraints and even just in its deverbal form gives rise to a semantically diverse group of nouns. For this reason, its productivity has been regarded as more restricted. There are also other suffixes which can derive action nouns (e.g. *otsi-**ng*** ‘search(ing)’, *ränn-**e*** ‘migration’, *rünna-**k*** ‘attack(ing)’, *pes-**u*** ‘wash(ing)’, *vul-**in*** ‘gurgle’) or agent nouns (e.g. *valv-**ur*** ‘guard’, *mõrv-**ar*** ‘murderer’, *oma-**nik*** ‘owner’, *joo-**dik*** ‘drunk’, *õpi-**lane*** ‘student’, *näri-**line*** ‘rodent’, *asu-**kas*** ‘inhabitant’), but their productivity is more limited (Kasik 2015: 185–186, 198–199).

Word formation in Estonian has been considered a complex research subject, which is why no single dominant word-formation theory or methodology has emerged for studying the relevant phenomena (Kasik, Vare & Kerge 2002: 51). Nevertheless, the field itself seems well saturated with thorough descriptions of Estonian derivational patterns, their structural restrictions, functions and semantics (e.g. Neetar 1990, Vare 1994, Kasik 1975, 2004, 2014, 2015, Kerge 2002, 2003, and Saari 1997, among others). Henn Saari (1997: 286–287) lists frequency, productivity, and activity of derivations as one area in need of further investigation. However, the role of frequency in the assessments of morphological productivity has still not been systematically and empirically accounted for, despite the fact that relatively large and diverse text corpora have been available for some decades now. Some studies (Kerge 2002, 2003) have operationalized the frequency of *-mine* nominalizations in measures of textual complexity in different registers of written Estonian, thereby inadvertently assessing at least one aspect of the pattern’s productivity. However, most of the research

about Estonian derivational morphology has so far largely neglected usage, as manifested through the frequency of existing formations, and considered productivity rather as a property of unintentional, intuitive linguistic competence (see Kasik 2011: 64–65). This conception of productivity as characteristic only of unintentional word-formation is also shared by widely-cited researchers like Schultink (1961), van Marle (1985) or even Bauer (2001: 35–36, 56–58), who sets *societal productivity* (unconscious shared understanding of a pattern's productivity as manifested through e.g. dictionary entries representing some common vocabulary) above *individual productivity* (may entail coining new forms intentionally, consciously, and often playfully). The attestation of intentionally created novel forms, which are found e.g. only in literary prose, headlines, technical texts, or in an individual's speech, would not be considered as contributing to a pattern's productivity (Bauer 2001: 57–58).

This study seeks to offer a complementary, usage-based perspective on word-formation in Estonian. It disregards such a distinction between intentional and unconscious word-formation and describes aspects of morphological productivity in Estonian as a function of relative usage frequencies in different registers. Indeed, it has been demonstrated that morphological productivity is closely correlated with frequency (e.g. Baayen 1992, Hay & Baayen 2002). An unproductive derivation pattern is usually characterized by many types with high token frequency and few types with low token frequency – types being the unique forms in a text and tokens the particular realizations of types – while productive patterns exhibit the opposite tendency. This results from the empirical evidence that "the formation of abstract mental representations is encouraged by varied types but counteracted by automation of high-frequency types" (De Smet 2020: 251). The number of types occurring only once in a given text or corpus (the so-called hapax legomena) becomes particularly important in assessing morphological productivity as a way for estimating the degree to which a category can be extended (Baayen & Renouf 1996, Baayen 2009). When a derivation pattern is productive, it is highly activated in memory in order to guarantee the adequate categorization of previously unencountered or simply very rare formations, whereas an unproductive pattern is not (Baayen 1992). This is why also intentional and creative formations are informative, as

they indicate the speaker's confidence that the formation is understood as intended due to the category's productivity.

The defining aspect of usage-based approaches is that grammar is regarded as a product of usage, not a mere repository to be accessed in language use (Perek 2015: 6). Therefore, derivation patterns and their productivity as parts of the speaker's or a community's linguistic knowledge are regarded as a potentially changing system, shaped by general cognitive processes such as categorization, prototypicality, extension etc. (Perek 2015, Diessel 2019, Lemmens 2019). Consequently, I view productivity as a graded and potentially changing phenomenon conditioned by various syntagmatic, semantic, paradigmatic and contextual factors, one of which is also register. It has not only been demonstrated that certain syntactic and lexical phenomena tend to occur more frequently in certain types of texts (e.g. Biber 1995), but also that such preferences emerge for derivative affixes as well (Plag, Dalton-Puffer & Baayen 1999). While *-mine* derivation in Estonian has been associated mostly with formal written texts (e.g. Kerge 2003, Kasik 2006), *-us* and *-ja* have not been investigated in the same manner. Furthermore, the use of derivation patterns in spoken language has received close to no attention. Pilvik (2019) assesses the morphological productivity of *-mine* in five types of discourse of Estonian – newspaper texts, scientific texts, fiction, contemporary spoken common language and spoken regional dialects – using the productivity measures introduced by Baayen and his colleagues already in the 1990s (e.g. Baayen 1992, 1993, Baayen & Renouf 1996): realized productivity, expanding productivity, and potential productivity. The results revealed, among other things, that while *-mine* nouns as a category do contribute the most to the size of the vocabulary of written registers, especially of newspapers and academic texts, spoken registers exhibit the most lexical variation, i.e. are most likely to expand the derivation schema for creating novel formations. In Pilvik (2019), I argued that the reasons for such differences largely rest in the functions in which *-mine* nominalization is used in different registers.

In this article, I tackle the same measures and types of discourse (hereinafter referred to as registers) but add additional perspective in two ways: I compare the behaviour of three different nominalization patterns (*-mine*, *-us* and *-ja*) and contrast Baayen's original approach with two modifications to that approach. These modifications were

proposed to overcome the tendency to overestimate the productivity of less frequent affixes. The analysis will focus on the following questions: whether the productivity as captured by different quantitative measures is conditioned by the register, how the suffixes rank within the registers in terms of the rate at which their respective categories are extended, and how well this ranking conforms to the linguist's intuitions about the productivity of the suffixes when assessed on relatively small corpus samples. Additionally, I explore the correlation between base and derivation frequency to see whether there are systematic correspondences between the two, which would suggest high regularity, or whether it reveals lexemes with idiosyncratic meanings, which would imply an increase in semantic uncertainty and a decrease in the semantic regularity of the morphological pattern. Finally, I use relative frequencies to examine the relationship between *-mine* and *-us* formations derived from the same base to address the potential rivalry of the two action noun suffixes.

The article is structured as follows: first, in section 2, I describe the three derivation patterns in terms of their structural and semantic properties as well as their pragmatic[1] and paradigmatic status; then, in section 3, I will give an overview of what I mean by productivity in this study and which measures can be used for assessing its quantitative output; in section 4, I present the data used for this comparative study; finally, I will present the results in section 5. In section 6, I will discuss the implications of these results as well as the shortcomings of the applied methodology. Section 7 concludes the article by highlighting the most important findings.

2. Overview of the derivation patterns

The suffixes under investigation are *-mine*, *-us* and *-ja*, which are considered the three most frequent patterns for noun derivation in Estonian (Kasik 2015: 281). Two of them (*-mine* and *-ja*) are considered highly productive, and two (*-mine* and *-us*) can be considered rival

1 In this article, I use the word *pragmatic* not strictly in its linguistic sense, but in its more general meaning to refer to the practical, utilitarian aspects behind making linguistic choices.

forms for expressing similar meanings in some contexts. As with all word-formation processes (Baayen 1992: 109–110), the use of these three derivation patterns is subject to various kinds of restrictions conditioned by syntagmatic, paradigmatic, semantic or contextual factors.

-mine is considered the most regular and general deverbal suffix for deriving action nouns in Estonian. This means that it is possible to derive a noun from every verb in exactly the same way, without any morphophonological restrictions on the stem or the form of the suffix itself (e.g. *asenda-ma* 'replace' → *asenda-mine* 'replacing', *esita-ma* 'perform' → *esita-mine* 'performing'). Because of the category's high generalizability and some shared features with non-finite verb forms (see e.g. Neetar 1988, Pilvik 2017), *-mine* forms were considered to belong to the verb's inflectional paradigm up until the year 1933 (Kasik 2015: 69).

From a syntactic perspective, some lexical restrictions to *-mine* nominalization have been mentioned, whether they arise from the assumption that only normal clauses (with nominative subject) can be nominalized (Erelt et al. 1993: 269) or arise from the semantics of the verbal predicates (Kasik 1975). Therefore, while it would be syntagmatically possible to derive nouns from verbs such as *piisama* 'suffice' (Erelt et al. 1993: 269) or *huvitama* 'interest', they are usually not nominalized due to the non-canonical argument marking in clauses in which those verbs occur (4–5).

(4) *Mu-lle* *piisa-b* *paari-st* *nädala-st* *puhkuse-st*
I-ALL suffice-3SG couple-ELA week-ELA vacation-ELA
'A couple of weeks off (work) is enough for me.'

(5) *Te-da* *huvita-b* *joonista-mine*
he-PRT interest-3SG draw-NMLZ
'He is interested in drawing.'

Modal predicates, such as *saama* 'can', *pidama* 'must; have to', *võima* 'can; may', *näima* 'seem', *tunduma* 'seem', *paistma* 'seem' are also said not to allow nominalization (Kasik 1975: 33), or at least to be very marginal among *-mine* nouns (Kasik 2015: 267). This does not mean that the predicates cannot be nominalized in their non-modal meanings (e.g. *saama* 'get', *pidama* 'keep', *paistma* 'shine').

From the perspective of word-formation, *-mine* derivation is also sometimes referred to as grammatical derivation (Kasik 2015: 267), since it does not add any semantic properties to the stem of the base verb, but simply changes the word class. Therefore, *magama* 'sleep' and *magamine* 'sleeping' both refer to the process of sleeping, even though the construal mechanism behind the two words is different (see e.g. Langacker 1987: 22). Some *-mine* nouns can also be fully lexicalized (e.g. *majapidamine* 'household') or semantically extended through metonymical and metaphorical links, e.g. for referring to the result of the process expressed by the verb (*nõudmine* 'demand', *pakkumine* 'offer', *teadmine* 'knowledge') or the event where the process happens (*esinemine* 'performance', *valimised* 'elections', *kogunemine* 'gathering'). *-mine* can also be used unproductively in some adjectives (e.g. *välimine* 'outermost'), but such formations are excluded from this study.

-us is another frequent deverbal nominalization suffix, but this derivation pattern is considerably more complex both in terms of morphophonology and semantics. The semantic relationship between the meaning of the verbal base and the meaning of the derivation can be general and transparent as with *-mine* nouns (*asenda-ma* 'replace' → *asend-us* 'replacing; replacement', *esita-ma* 'perform' → *esit-us* 'performing; performance'). However, in most cases where *-us* and *-mine* nouns share similar meanings, they cannot be considered synonymous, and the main difference often lies in aspect: *-us* usually expresses actions and states as atemporal phenomena or temporally bounded, resultative actions, while *-mine* creates a more processual and temporally unbounded reading (Erelt et al. 1995: 484, Kasik 2015: 187). An *-us* noun can also express other meanings, for example the result (*muutus* 'change'), instrument (*kaunistus* 'decoration'), object (*annetus* 'donation'), subject (*kerjus* 'beggar'), collective subject (*valitsus* 'government') or location of an action (*peatus* 'stop; station'). Compared to *-mine* nouns, there are also more *-us* nouns which can have synchronically non-processual and idiosyncratic meanings (e.g. *katus* 'roof' from *katma* 'cover', *mõistus* 'mind' from *mõistma* 'understand') (Erelt et al. 1995: 484–485).

In addition to a wide semantic scope, there are also morphophonological restrictions to the *-us* construction, making it less general than the *-mine* suffix. Although the suffix *-us* is usually attached to the consonant stem of the verb, with *u* replacing the stem vowel, there are several

conditions in which *-us* nouns cannot be derived, according to Erelt et al. (1995: 486–487). For example, the derivation pattern is said to be unsuitable when

a) the verb stems are derived from nouns through zero-derivation, and the stem itself does not express any action or state (e.g. *värv* 'paint' → *värvima* '(to) paint', *rohi* 'grass' → *rohima* '(to) weed');
b) there is morphological homonymy with the 3rd person singular past tense verb form (e.g. *ebaõnnestu-ma* 'fail' → *ebaõnnestu-s* '(he) failed');
c) the verb stem ends with *-ne* (e.g. *elavne-ma* 'brisk, enliven') or *-ise* (e.g. *helise-ma* '(to) ring, (to) sound');
d) the verb stem is a loan stem (e.g. *kritiseeri-ma* 'criticize');
e) *etc.* (see Erelt et al. 1995: 486–487).

To further complicate the issue, these restrictions do not always cover the whole group of structurally similar stems and can combine with some additional issues. For example, the suffix can also be used for deriving nouns from other nouns (*sõber* 'friend' → *sõprus* 'friendship'), adverbs (*palju* 'many, much' → *paljus* 'abundance, plurality') or adjectives[2] (*tark* 'wise' → *tarkus* 'wisdom'), in which case it can sometimes be unclear whether the base for the *-us* noun is an adjective or a verb derived from that adjective (e.g. *hull* 'mad; a mad person', *hulluma* 'go mad', *hullus* 'madness'). When there are both causative and reflexive stems for a verb (e.g. *erut**a**ma* 'excite' and *erut**u**ma* 'get excited', *reost**a**ma* 'pollute' and *reost**u**ma* 'get polluted'), it is unclear from the word form alone which of those functions as the base of the *-us* noun (e.g. *erut**u**s* 'excitement', *reost**u**s* 'pollution') (Erelt et al. 1995: 486–487, Kasik 2004: 94–95). However, if one were to consider restriction *b* above, the causative stem would become a more likely base here. Otherwise, the 3rd person singular imperfective forms of the reflexive stems would be homonymic with the derived nouns, e.g. *erutu-s* '(he) got excited', *reostu-s* '(it) got polluted'.

-us can also attach to the markers of a verb's non-finite forms, for example the 2nd infinitive (*tule-**m**a* 'come' → *tule-**m**-us* 'result'), in which case the derived noun expresses a result (*uurima* 'study' →

2 Interestingly, for deadjectival derivation, no structural restrictions apply and there are only a few semantic constraints (Vare 1994: 10).

uurimus '(a) study', *tüdima* 'get bored' → *tüdimus* 'boredom') or a property (*väsima* 'get tired' → *väsimus* 'tiredness, fatigue') (Kasik 2015: 271). It can also attach to the marker of the active present participle (*läiki-**v*** 'shiny' → *läiki-**v**-us* 'shinyness'), the passive present participle (*loe-**tav*** 'readable' → *loe-**tav**-us* 'readability'), the passive past participle (*hari-**tud*** 'educated, scholarly' → *hari-**t**-us* 'scholarliness'), or, on rare occasions, to the active past participle (*igane-**nud*** 'outdated, obsolete' → *igane-**n**-us* 'outdatedness, obsoleteness' or the lexicalized *jää-**nud*** 'remained; stayed' → *jää-**n**-us* 'remnant'). *-us* nouns derived from participles are semantically closer to nouns derived from adjectives (Erelt et al. 1995: 487–489), although derivations ending in *-tus* could often also be interpreted as the result of a process or action, e.g. *treenitud* 'fit, trained' → *treenitus* 'fitness, the property of being fit' (Kasik 2015: 193).

The structure of the base word has been shown to be highly relevant in morphological productivity (Aronoff 1976, Baayen & Renouf 1996). According to Kasik (2015: 281, 283), *-us* derivation is more productive from derived stems (e.g. *vallu-**ta**-ma* 'conquer', *kirje-**lda**-ma* 'describe', *vest-**le**-ma* 'chat, converse', *laie-**nda**-ma* 'extend'). However, this, along with the possibility of deriving *-us* nouns from participles, has also given ground to a rather extensive suffix allomorphy, where the border between base and suffix morphemes is not always clear. For example, ***-dus*** nouns from a monosyllabic base (*loodus* 'nature', *saadus* 'product') could be analysed as either formed from the passive past participle (*loo-dud* 'created' → *loo-d-us* 'nature; (lit.) what has been created') or with an allomorph *-dus*, though only the latter interpretation would be possible for polysyllabic bases (*hari-dus* 'education'). The allomorph ***-tus*** in derivations which lack a corresponding base ending with *-ta-*, such as *noomi-tus* ('reprimand') (**noomi-ta-*), is claimed to result from analogy with formations such as *vallu-ta-ma* 'conquer' → *vallu-t-us* 'conquest; conquering' (Kasik 2015: 282), but these forms could just as well be considered to be derived from participles and aquired additional meanings (*noomi-tud* 'reprimanded' → *noomi-t-us* 'the state of being reprimanded; reprimand'). Many ***-ndus*** and ***-lus*** nouns are also baseless in the sense that their potential *-nda-* or *-le-* bases are never used (e.g. **korjanda-* → *korjandus* 'fundraising; collection', **nõudle-* → *nõudlus* 'demand'), while their semantic relationship

between the underived base (e.g. *korja-* 'collect', *nõud-* 'demand') is rather transparent.

Finally, the ***-ja*** suffix is used to create nouns referring to the agent, or more precisely, the internal subject of the activity, since the entities need not be agents in the strict sense. For example, *seisja* 'the one who stands' and *kaotaja* 'the one who loses' code the participant in or the experiencer of a state or process (Kasik 2015: 197). The formation is syntagmatically completely systematic with one exception: when the stems are in the first degree of quantity and end with an *e*, the stem vowel is replaced with an *i* (e.g. *tul**e**-ma* 'come' → *tul**i**-ja* 'the one who comes', *näg**e**-ma* 'see' → *näg**i**-ja* 'the one who sees'). Only the stem *mine-* 'go' does not conform to this transformation (hence, *min**e**-ma* 'go' → *min**e**-ja* 'the one who goes') (Kasik 2004: 110).

-ja nouns presuppose a potentially active participant and therefore are only marginally formed from stems which occur in impersonal verbs (e.g. *müristama* 'thunder'), do not code the agent as the most active participant (e.g. *meeldima* 'like') or express spontaneous, unfacilitated events (e.g. *külmenema* 'become colder') (Erelt et al. 1995: 480, Kasik 2004: 109, 2015: 197). Erelt et al. (1995: 480) also doubt the usage of *-ja* nouns derived from reflexive verbs, such as *selguma* 'become apparent', *ummistuma* 'become clogged' etc. Although all *-ja* nouns can be used in the general sense of someone doing something, the derivation pattern has also semantically expanded to include more specialized meanings, such as professions (e.g. *ehitaja* 'builder', *laulja* 'singer', *lapsehoidja* 'babysitter') or instruments (e.g. *kruvikeeraja* 'screwdriver'). In the latter interpretation, the actor is no longer conceptualized as an animate entity (Kasik 2004: 109–110). The property of *-ja* to express multiple semantic roles (e.g. that of an agent, experiencer, stimulus, instrument or even a theme) is a cross-linguistically attested phenomenon with agent noun suffixes (see Booij & Lieber 2004, Denistia & Baayen 2019).

So far, the descriptions of the derivation patterns have mostly dealt with their syntagmatic and semantic properties, but some things could also be said about their paradigmatic and contextual/pragmatic behaviour. First, it has been proposed that synonymous affixes tend to select their base words from complementary domains (van Marle 1985). Although *-mine* and *-us* cannot be considered fully synonymous, they do cover overlapping semantic domains. Therefore, it is expected that in the case of rival forms with similar readings, the more frequent bases

used for derivation with one of the suffixes will have a low frequency among the formations with the other suffix. In the case of similar relative frequencies for the same stem, there should be a clear semantic opposition between *-mine* and *-us* nouns. With regard to the productivity of rival affixes, it has been observed for English deadjectival nominalization suffixes *-ness* and *-ity*, that although *-ness* is generally considered far more productive, some of the bases to which both *-ity* and *-ness* can attach clearly prefer the less productive *-ity*. It is worth seeing if this also applies for *-mine* and *-us*.

As for the contextual aspect, it makes sense to talk about the pragmatic usefulness of the formations in different situational or stylistic contexts. As stressed in e.g. Baayen and Lieber (1991: 818), new words are produced out of some pragmatic necessity. So, one might ask whether the pragmatic demand for new agent nouns is equal to that for new action nouns in a given register, for example. In addition to competing with several other frequent categories (pronouns, proper names, common nouns) for the same functional slot in a sentence, *-ja* nouns are not always neutral. Their use often encodes a negative judgement of a person's character or behaviour as a constant property (e.g. *moraalitseja* 'someone who's always preaching', *märatseja* 'someone who's always raving', *õelutseja* 'someone who's always being mean'; Erelt et al. 1995: 481). It is therefore expected that the usefulness of the concepts associated with such formations will be more limited in more formal registers of language and less limited in less formal registers. Furthermore, the pragmatic usefulness of different bases in different registers will affect the pragmatic usefulness of those bases' derivations.

Another distinction in pragmatic usefulness is linked to a pattern's different functions in different registers. According to Bauer (2001: 208–209), there are fewer pragmatic constraints on processes which are used for transpositional purposes (i.e. in a syntactic or anaphoric function). This is why they are more productive than the processes used for lexical innovation (e.g. for the creation of new terminology). One pattern can also perform both of these functions, and the higher pragmatic demand for one or the other function might depend on the register. For example, *-mine* has been shown to be more productive in less formal registers of Estonian (fiction and spoken language), where it is less likely to be used in the lexical function (Pilvik 2019).

The interplay of the conditions discussed in this section, as well as many more, has consequences for the productivity of the word formation patterns. 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." While much of the information concerning this individual variation is lost when analysing corpus data, there are ways of assessing the quantitative output of this complex interplay between those different circumstances. These methods are discussed in the next section.

3. Measuring morphological productivity

The term productivity is used to refer to several different, though related phenomena in linguistics. According to Barðdal (2008: 9–24), three main concepts arise in the literature: productivity as GENERALITY, productivity as REGULARITY, and productivity as EXTENSIBILITY, each of which is closely linked to frequency and to the others. The **generality** aspect appears to deal with the schematic openness of the pattern: a highly general construction does not impose (semantic or structural) restrictions onto the members that can instantiate it. A **regular** construction, in turn, gives rise to formations which are semantically transparent and compositional. In morphological terms, the suffix bears a clear meaning and modifies its base in a predictable way for every formation. **Extensibility**, in turn, can be seen as a property of a construction which characterizes its ability to include novel items or develop new functions. Productivity can, therefore, be associated with the speakers' knowledge of constructions in terms of the members which instantiate these constructions, the general meaning(s) of the constructional schemas themselves, their ability to categorize expressions never encountered before and to extend the constructions to create new expressions.

In this article, I take a quantitative, usage-based perspective and focus mainly on the property of extensibility in measuring and comparing the morphological productivity of three deverbal nominalization

patterns. In other words, I define productivity here broadly as the frequency-based probability that a morphological pattern will be used to create novel formations (e.g. Baayen 2003, Bybee 2007), although generality and regularity play a role in whether or how much a pattern can be extended. Comparing the productivity measures across different registers also entails contrasting the pragmatic usefulness and functions of derivations in different communication settings. While productivity is also something often attributed to either affixes, morphological processes, rules, words, or even complete modules of grammar (see an overview in Bauer 2001: 12–15), I do not wish to tackle the theoretical premises behind each of those approaches in this article. I take productivity to be a property of a *suffix*, a *derivation pattern*, a *morphological construction*, a *morphological category* or a *morphological process* – terms which are used more or less interchangeably in this article. The inclusion of the term *construction* refers to the fact that similarly to syntactic constructions, complex words are also seen as instantiations of constructional schemas (see e.g. Booij 2010).

3.1. Realized productivity, potential productivity, expanding productivity

Baayen (1992: 110–111) has posed four requirements for a quantitative measure of morphological productivity: 1) it has to enable ranking word formation processes in a way which would generally correspond to linguistic intuitions; 2) it has to express the availability of the category for producing new formations; 3) it should be sensitive to formally or semantically idiosyncratic properties of some formations in that such formations have a decreasing effect on the measure; 4) it should support the empirical fact that type frequencies alone are not sufficient as productivity measures. With regard to the last desideratum, the raw type frequency can be taken to reflect the pragmatic usefulness of a morphological process, but it does not say anything about the rate at which the category is extended to include new formations (Baayen 1992: 117, 123). It is therefore insufficient as a single measure of productivity.

One way of assessing productivity as a probability is by using the methodology largely developed and applied by Baayen and his collaborators (Baayen 1989, 1992, 1993, 2001, 2003, Baayen & Renouf 1996, Plag, Dalton-Puffer & Baayen 1999). This approach makes use

of text corpora as a source for actual language use and relies largely on three principal counts: the number of tokens, the number of types and the number of hapax legomena (tokens/types occurring only once in a given corpus) of a given morphological category (Plag, Dalton-Puffer & Baayen 1999). Hapaxes are taken to be examples of novel or rare formations not likely to be included in a large dictionary and provide evidence of the category's extensibility.

First, for assessing the productivity of a pattern in the past, one could use the so-called ***realized productivity***, which is reflected in the number of types (*V*) with a given suffix occurring in the corpus with a given number of tokens (*N*). This measure is perhaps the most straightforward and most closely related to the observations made in the literature about the *-mine* derivation being more characteristic of formal written Estonian. As *-mine* nouns are extensively used in e.g. journalistic and legal works for thickening the text (Kerge 2003, Kasik 2006) and in scientific texts for creating new terms (Kerge 2002, 2003), their overall high proportion among all words compared to spoken data is not surprising. The higher the number of types, the higher the proportion of different members from a given morphological category among all words in the corpus, and the more useful the concepts created with this construction have been. Baayen (1993) has also called this measure the *extent of use*. However, realized productivity is perhaps also the least informative measure, since it only accounts for the observed, already existing contribution of a morphological category to the size of the whole vocabulary synchronically. While a suffix may have become unproductive over time, it could still exhibit a high number of types, despite not being able to form new ones (Bauer 2001: 144, Keune, van Hout & Baayen 2006).

A second, central measure of productivity is the so-called ***potential productivity* (*P*)**, which relates the number of hapaxes formed with a given suffix (*n1*) to the number of all tokens carrying this suffix in the corpus (*Nc*) (6).

(6) $P = n1/Nc$[3]

3 In this article, *Nc* is used to refer to the token count limited by the relevant morphological category (e.g. all *-mine* nouns) and *N* is reserved for denoting the total number of tokens in a sample.

Potential productivity reflects the speed at which new lexical items are added to a category. In other words, the more productive a pattern, the higher the value of *P* and the more likely we are to encounter a new, previously unseen type with the next added token of this category. In turn, when a category is fully unproductive, then it has exhausted all its potential members, and all new tokens will represent a type already encountered (Baayen 2009). Even though we are still using already existing formations, this probabilistic measure allows the estimation of a pattern's productivity in a more broader sense, because it explicitly addresses the extensibility of a morphological category, i.e. its ability to produce new items. The presence of idiosyncratic formations with a given affix (i.e. lexicalized, semantically opaque items) tends to reduce potential productivity, since such formations are very unlikely to occur as hapaxes and instead typically have high token frequencies (Baayen 1992: 117). When *P* is compared against type count *V* in a corpus with *N* tokens, it becomes possible to assess the ***global productivity*** (Baayen 1992: 123–124, 2001: 203–205) of different affixes. This provides a multidimensional view of morphological productivity and enables us to assess whether the category's potential to expand has actually been realized in practice.

Finally[4], I will consider the so-called ***expanding productivity*** **(*P**)**, which is calculated by dividing the number of hapaxes occurring with a given suffix (*n1*) by the total number of hapaxes in the corpus (*N1*) (7). It shows the likelihood that any new word should belong to the category with a given suffix. The higher the value of expanding productivity, the more attractive the category is for expressing a previously unencountered concept. Therefore, this measure complements realized productivity in that it also tries to estimate the future utility of the morphological category in question.

(7) P* = n1/N1

4 These three measures, although most extensively used and the most intuitive, do not at all exhaust the whole range of possible empirical measures and techniques for assessing the productivity of a process (see an excellent overview in Zeldes 2012: 48–95).

3.2. A variable-corpus approach

Some criticisms of hapax-based measures and Baayen's procedures have been raised. For example, hapaxes in smaller corpora can represent rare events rather than true neologisms; it is unclear whether complex words should be counted as forming the same type or individual types or whether synchronically lexicalized items should be counted as instantiations of the word-formation pattern etc. (Plag 1999: 28–29, Gaeta & Ricca 2006: 59–60). The main critique posed by Gaeta and Ricca (2003, 2006), however, concerns the comparison of multiple affixes. They assert that affixes with considerably different token frequencies in a given corpus should not be compared on the basis of the potential productivity *P*, because *P* is not a constant but a decreasing function of token count (Gaeta & Ricca 2006: 62). In other words, the more tokens of a given suffix we encounter, the smaller the likelihood of seeing a previously unseen type (a hapax) (Baayen 1992: 114, Gaeta & Ricca 2006: 62). In Figure 1, taken from Gaeta & Ricca (2003: 96), the Italian nominalization suffixes *-mente*, *-mento*, *-(t)ura*, and *-nza* are compared in terms of their potential productivity *P* (the *y*-axis) and the total number of tokens with a given suffix (the *x*-axis).

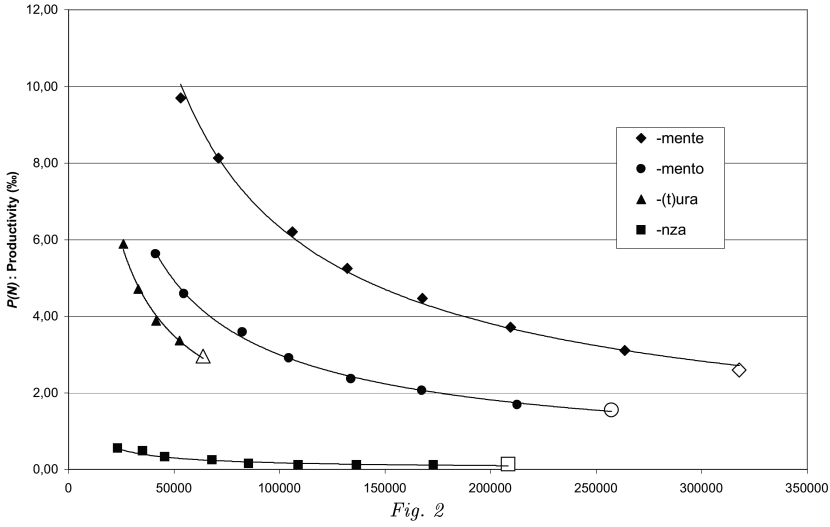

Figure 1. Comparison of the potential productivity of Italian nominalization suffixes according to their token counts in the corpus (published in Gaeta & Ricca 2003: 96).

Due to the decreasing nature of the function, comparing the productivity measures at the endpoint of the sampling process means overestimating the values of *P* for the remarkably less frequent affixes (Gaeta & Ricca 2006: 62–63). For example, the potential productivity of *-(t)ura* in Figure 1 would be estimated the highest, despite the fact that the corpus overall contains significantly fewer words with that suffix than words with the other 3 suffixes. If we examined the productivity of all 4 suffixes at the point where approximately 7000 tokens (N^5) have been sampled from the corpus for each affix, the suffix *-(t)ura* would rank only third after *-mente* and*-mento*. Baayen (1992: 119) himself also notes that *P* "does not hand us the means for obtaining a measure of productivity that has a fixed value irrespective of sample size".

In order to overcome this limitation, Gaeta and Ricca (2003, 2006) propose a so-called variable-corpus approach for evaluating *P* for different affixes at equal values of *Nc*. This means extracting an equal number of tokens for each affix but consequently from corpora of different sizes. For this purpose, the authors first divided a global newspaper corpus of 75,000,000 tokens into 36 chunks of progressively increasing size and then extracted all the occurrences of the suffixes under investigation from the complete list of word-forms. Then, they calculated the relevant type, token, and hapax counts for each chunk. The potential productivity values *P* for each subsample were fitted against the number of tokens for each affix. This made possible to obtain the values of *P* at a fixed value of *Nc* (for example, at the number of tokens of the least frequent suffix) for each affix by using power regression (Gaeta & Ricca 2006: 64–66).

Gaeta and Ricca's approach results in a measure which is actually closely linked to the notion of expanding productivity. In practice, *P** often displays a strong correlation to the measure they propose in their work (Gaeta & Ricca 2006: 62, Baayen 2009). Their approach is also similar to the one taken in Plag, Dalton-Puffer & Baayen (1999), where the suffixes were compared based on the average values of *V* and *P* which, in turn, were calculated based on the expected[6] number of types occurring at twenty equally spaced intervals.

5 In the context of this article, *Nc*.

6 *Expected* number of types is "the number of different types one may expect to count on average for a great many different orderings of the text fragments in a given subcorpus"

In this article, I bring an additional perspective to the analysis of variation in morphological productivity by also studying the variation that occurs across registers. I use Baayen's simple approach to compare different registers in terms of how productively each deverbal pattern appears to be used in them, and Gaeta and Ricca's variable-corpus approach to compare the productivity of the suffixes *-mine*, *-us*, and *-ja* themselves. For this purpose, I use Generalized Additive Modeling (GAM, Wood 2017) to interpolate the productivity values at a fixed token count. GAMs are better suited for modeling nonlinear, nonmonotonic relationships with less risk of overfitting. Later in this article, I will also examine the effect of averaging the productivity values (as done in Plag, Dalton-Puffer & Baayen (1999)) and compare that with the variable-corpus approach and Baayen's original approach.

In addition to comparing the productivity values, I also make use of the actual lexemes in order to provide more insight into the nature of the frequent and the less frequent derivations in the data. First, I explore the relationship between base and derivation frequency: if base and derivation frequencies correspond, there is reason to believe that the derivation pattern is used regularly, that there is a clear semantic parallel between the base and the derivation, and that processing or producing items in a morphological category is more likely to also involve lexical procedural knowledge. Second, I compare the distribution of *-mine* nouns with that of *-us* nouns formed from the same bases to see if this may shed light on the potential rivalry of the two suffixes.

4. Data

The data used for this study is extracted from three corpora, representing five different registers of Estonian: The Balanced Corpus of Estonian (BCE)[7], The Corpus of Estonian Dialects (CED)[8], and The Phonetic Corpus of Estonian Spontaneous Speech (PCESS)[9].

(Plag, Dalton-Puffer & Baayen 1999: 217). For this estimation, they used binomial interpolation, which makes strong assumptions about the randomness and independence of words occurring in running texts.

7 https://www.cl.ut.ee/korpused/grammatikakorpus/ (Accessed 01.01.2018.)

8 https://www.keel.ut.ee/keelekogud/murdekorpus (Accessed 29.09.2015.)

9 https://www.keel.ut.ee/en/languages-resourceslanguages-resources/phonetic-corpus-estonian-spontaneous-speech (Accessed 31.11.2019.)

The **BCE** contains 15 million tokens of written Estonian, subdivided into three equally-sized subcorpora. The media subcorpus comprises of 5 million tokens from daily and weekly newspapers published between 1995 and 2007; the fiction subcorpus consists of 5 million tokens of excerpts from Estonian literature dating from 1987 to 2011; the subcorpus of scientific texts contains 5 million tokens from dissertations and articles published in scientific journals between the years 1995 and 2006. For this study, I used a version of the corpus with automatic morphological annotation.

The **CED** contains approximately 900,000 tokens of morphologically annotated dialect transcriptions. The recordings of unstructured interviews, where elderly informants talk about their childhood, everyday life, past customs and events, date to the 1960s and 70s and cover all Estonian traditional dialect areas, although the amount of data available from different regions may differ substantially. The transcriptions do not follow standard spelling conventions, but instead use a special simplified version of the Finno-Ugric phonetic transcription[10]. Both transcribing and annotating have been done manually, guaranteeing the high quality of the data, but at the same time, resulting in some inconsistencies in the spelling and annotating conventions (e.g. *ära viimine* 'taking away' and *ära+viimine*[11] 'taking away').

The **PCESS** is the smallest corpus used in this study, comprising only about 426,000 morphologically annotated tokens[12]. However, compared to the version used in Pilvik (2019), this version contains over 80,000 more tokens. The recordings in the corpus are made between 2006 and 2019 and cover dialogues and monologues. Although the main application of this corpus rests in the analysis of the phonetic traits of spontaneous speech, it is a valuable resource for a variety of different tasks.

Obviously, the text types contained in the corpora are in no way uniform: not in form, communicative function or situational settings. For example, sports commentaries differ significantly from interviews

10 https://www.keel.ut.ee/sites/default/files/www_ut/emk_teejuht2015.pdf (Accessed 01.03.2019.)

11 The plus sign marks the boundary between compound lexemes.

12 The tokens have been automatically analyzed with Filosoft's Vabamorf morphological analyzer (https://github.com/Filosoft/vabamorf, accessed 19.01.2021), which is trained on written standard Estonian.

in all three of those aspects; popular science texts in journals use less specific vocabulary than dissertations, etc. PCESS also covers a variety of speech situations, ranging from public lectures to conversations held between two familiar people. The CED comprises speech from 10 regions which are traditionally viewed as diverging in many aspects of language use (although an analysis of *-mine* nominalization, for example, did not reveal significant differences between dialects, see Pilvik 2016). However, since a more fine-grained analysis of the role of different derivation patterns in specific text types remains an undertaking too detailed for the scope of this article, the registers are currently taken as predefined by the corpora. Among the registers, 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.

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

The productivity measures used in this study are formally different ratios of type and token counts which are interpreted as conditional probabilities. However, as mentioned in the previous section, the likelihood of encountering a previously unseen type in a corpus is a decreasing function of the corpus size: the longer the text, the more types have already occurred and the lower the likelihood that a new token representing a new type should appear (Hardie and McEnery 2006: 139, Gaeta and Ricca 2006). Therefore, in order to be able to compare the productivity measures across registers, it was necessary to extract random samples with equal numbers of tokens[13]. The size of the samples was determined by the number of tokens in the smallest corpus, the PCESS. The token counts were extracted from randomly chosen complete files from each subcorpus, and the file sampling stopped when

13 All the analyses in this article are conducted with lemmatized tokens.

the total token count in the corpus sample reached higher than 426,000 (this condition was checked prior to sampling another file). Table 2 presents the size of the corpora, the size of the samples, the type/token ratios in the samples (demonstrating the degree of lexical variation in the sample) and the growth rate of the samples (demonstrating the probability of encountering a type previously unseen in the sample).

Table 2. Token counts, TTRs and growth rates of the five registers. Punctuation is excluded.

Register	Tokens in corpus	Tokens in total sample	Type/token ratio	Growth rate
NEWS	4,675,823	437,003	0.073	0.034
SCI	4,798,966	427,987	0.063	0.028
FICT	4,953,609	439,917	0.045	0.020
SP	426,516	426,516	0.035	0.016
DIA	890,788	427,012	0.025	0.010

We can observe that written registers exhibit considerably higher lexical variation than spoken registers. They are also more likely to make use of rare words and expand their vocabulary at a higher rate. In the context of contemporary corpus linguistics, these samples are extremely small, considering that most studies base their analyses on tens of millions of words. However, since we currently lack access to larger automatically analyzable corpora on spoken Estonian, I must rely on Baayen (1992, 1993) who got interesting results even with a relatively small corpus such as the Dutch Eindhoven Corpus, which at that time contained approximately 600,000 words of written text.

From each sampled file, I compiled lists of all tokens, types and hapaxes as well as individual lists of tokens, types and hapaxes formed with each suffix. The layer of morphological annotation enabled me to extract the lists of all *-mine*, *-us* and *-ja* nouns using the part-of-speech tag and the suffix of the lemma. Derivations functioning as first parts of compound words (e.g. *korrut**us**+tabel* 'multiplication table') were not included. The lists were manually cleaned from false hits (spelling errors, false analyses, foreign words) and unwanted formations (derivations from other word classes and non-finite verb stems). *-us* nouns which could be derived from participles are not included in the current

analysis, since including them would require the inclusion of multiple *-us* derivations from same base (e.g. *kohusta-* → *kohust-us* 'obligation; obliging', *kohusta-t-us* 'being obliged', *kohusta-v-us* 'being obliging'), which in turn would make the comparison with *-mine* nouns (*kohusta-mine* 'obliging') more difficult. However, nouns ending with *-dus* are included, as these are more naturally interpreted as formed with an allomorph *-dus*. Other allomorphs, such as *ol-**l**us* 'matter, substance' (from *ole-ma* 'be') or those in baseless forms such as *korja-**nd**us* 'fundraising, collection' are also included in the sample. In the latter case, the corresponding underived stems (e.g. *korja-* 'collect') have been considered as the bases. The reasons for including the allomorphs are three: one, such derivations are complementary to simple *-us* nouns (allomorphs cannot usually derive nouns from the same bases as can the simple *-us* suffix, and vice versa[14]); two, they are occasional; three, they are not expected to influence the analysis as a semantically or functionally distinct subgroup of *-us* nouns. The fact that they cannot be interpreted as action nouns cannot be a sufficient criterion for disregarding their contribution to the productivity of *-us*, since the action noun reading is not necessarily present for *-us* nouns either.

Finally, all verb tokens were extracted and stripped of grammatical markers as the possible bases for the derivations. To match a derived noun to its corresponding verbal base (especially in the case of *-us* nouns), I consulted Kasik (2015) and the electronic database of Estonian Word Families[15] at the Institute of the Estonian Language.

Following the procedural approach in Gaeta and Ricca (2003, 2006) and modifying it to fit the considerably smaller samples and five different registers, the data from all sampled files was split into 21 chunks of progressively increasing size, with each chunk/subcorpus including 21,300 more tokens than the previous one. This yielded altogether 105 subcorpora, with 21 for each register. This means that the data from longer files was split into several consecutive subcorpora. In each subcorpus, productivity measures for each suffix were calculated based on

14 On rare occasions, the suffix *-lus* can attach to stems of *-us* forms (compare *and-ma* 'give' → *ann-us* 'dose' and *aru and-ma* 'report' → *aru+and-**lus*** 'reporting; report'), even though the corresponding base verb (e.g. **aru andlema* or **aruandlema*) is not used. As such parallelisms are very infrequent, this is not expected to affect the overall results.

15 http://www.eki.ee/dict/sp/ (Accessed 15.12.2019.)

the token, type and hapax frequencies at that particular sampling point. Since the 21 subcorpora are cumulative and each subcorpus includes the smaller subcorpora that come before it, the subcorpora do not represent independent samples and the productivity measures are affected by their respective values in the previous samples. As the samples consist of unrelated texts of differing lengths and do not constitute one continuous discourse, choosing only one ordering of the files in the samples would be biased towards that specific configuration of textual sequence. In order to alleviate the effect of the order in which the files are sampled, the above-mentioned procedure was repeated on 100 random permutations of the sample files. Then, for each subcorpus size, productivity values were averaged over the 100 different orderings of files[16] and 95% confidence intervals were calculated[17] in order to demonstrate the range of values which is likely to include the actual productivity value in a subcorpus of that particular size. In doing so, the effect of the order in which the individual files are sampled decreases, since longer texts are broken into parts and divided into different subcorpora. Therefore, the discourse structure of texts written by single individuals does not affect the occurrence of the derived forms as strongly as it otherwise might. However, this approach does not assume that words appear randomly and independently in texts (cf. Baayen 1996).

By limiting the data to include only deverbal formations which can be identified as nouns (instead of adjectives, such as *alumine* 'lowermost') and considering different allomorphs of *-us* (e.g. *-dus* as in *haridus* 'education'), I move from a strictly morpheme-based perspective towards a more schema-based perspective (see Fonteyn & Hartmann 2016). However, I do not differentiate between the eventive and non-eventive readings of action nouns and consider lexicalized formations equal to all other instantiations of a morphological construction. While not focusing on word formation from only a synchronic perspective could admittedly be problematic, a finer analysis of the derivation patterns would be highly time-consuming, due to the necessity of including

16 In the largest subcorpus with total number of tokens (e.g. 437,003 tokens in NEWS), the token, type, and hapax counts as well as the productivity measures were identical in all 100 permutations.

17 This was done using the function *groupwiseMean* from R (R Core Team 2020) package *rcompanion* (Mangiafico 2020), which provides means for calculating adjusted bootstrap confidence intervals for non-normally distributed data.

contextual information. In addition, such an analysis could also be, at least to some extent, subjective, due to the semantic and formal ambivalence of some forms. In addition to the practical predicaments, making a strict distinction between semantically opaque and transparent formations would also be theoretically problematic, since word meanings are not fixed, but are subject to continuous change (see Baayen et al. 2019). Likewise, the use of morphotactically transparent but semantically opaque items may also contribute to the activation of a derivation pattern for native speakers. Therefore, including lexicalized items is also the preferred option on psycholinguistic grounds (Gaeta & Ricca 2006: 79).

Table 3 presents the raw type, token and hapax counts of all derivation patterns in the five registers at the maximum corpus size (these counts are the same in all 100 permutations, because permutations do not allow replacements and all files must eventually be used). Only the outermost derivation cycle is considered, and only simple lemmatized words are counted. This has negative consequences for the number of hapaxes, but since the spelling conventions differ between the corpora, this helps to control for the amount of manual work needed to check the word lists for errors (misprints, incorrect or inconsistent analyses) which would artifically increase the hapax count and make it more difficult to compare different registers.

Table 3. Raw frequency counts of tokens, types and hapaxes occurring with the three suffixes across five registers (assessed at the endpoint of the sampling process).

	-mine			***-us***			***-ja***		
Register	tokens	types	hapaxes	tokens	types	hapaxes	tokens	types	hapaxes
SCI	15,052	1114	360	13,987	323	61	2346	218	79
NEWS	7417	1155	477	8464	436	95	4135	442	172
FICT	2570	771	409	3418	451	158	1472	392	207
SP	1894	481	240	2566	286	88	947	173	73
DIA	952	271	128	642	133	54	717	138	59

5. Results

5.1. Productivity of *-mine*, *-us* and *-ja* in five different registers

First, I will analyze the 3 productivity measures at 21 different corpus sizes for each individual suffix in order to highlight the differences between registers. Then, the productivity of the three suffixes is calculated at an equal number of suffix tokens to emphasize the differences between suffixes. Since the only measure based on the token count of the suffixed nouns (*Nc*) is the potential productivity *P*, the variable-corpus approach will compare the suffix patterns only with regard to that measure, i.e. their extensibility to include new items. The results of this approach are compared with those of Baayen's original procedure and the process of averaging the productivity over multiple samples as done in Plag, Dalton-Puffer & Baayen (1999).

5.1.1. Differences between registers

Differences in the productivity curves between the registers can be assessed based on the way in which the measures for *-mine*, *-us* and *-ja* change when the corpus size increases. While the curves depicted in Figures 2, 3 and 6 do not enable a straightforward comparison between the suffixes (due to the reasons mentioned in section 3.2), they do highlight the ways in which the five registers analysed in this study differ in terms of the 3 proposed productivity measures. The *y*-axes in the figures do not follow the same scale, but situate the productivity curves of each suffix such that they are relative to their individual maximums and minimums.

Figure 2 represents the **realized productivity** curves. Those curves are relatively linear and monotonic, which means that the type count for all suffixes increases gradually as more tokens are sampled from the corpus. This process occurs at a higher rate in written registers. The confidence intervals are very narrow and hardly visible, meaning that the depicted average values from 100 permutations of the samples are very close to the number of types expected to occur in a new sample of the same size and design. Only *-us* and *-ja* in fiction show slightly more variation.

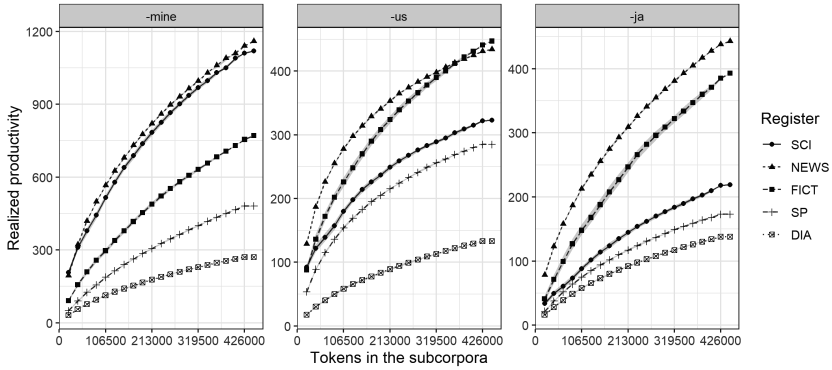

Figure 2. Realized productivity of *-mine*, *-us* and *-ja* in cumulative samples of five registers. The symbols and black lines represent the mean values of 100 random permutations of the samples. 95% confidence intervals are shown with light grey shading.

A high realized productivity value is a reflection of the suffix's profitability in the past. The written registers seem to make more use of derived nouns than do the spoken ones, which is in accordance with what is generally claimed about derivation patterns (Plag, Dalton-Puffer & Baayen 1999). While this is perhaps not surprising, this has had little empirical support in Estonian linguistics up until now. The two regular suffixes, *-mine* (e.g. *valitsemine* 'ruling; governing; controlling') and *-ja* (e.g. *valitseja* 'ruler; governer; controller'), are similar in that they both show high realized productivity values for NEWS and low values for the two spoken registers, SP and DIA. However, *-mine* is also highly productive in SCI, while *-ja* becomes more productive in FICT. This is a reflection of the relative pragmatic usefulness of a pattern in different written registers: on the one hand, fiction writing frequently uses *-ja* for occupations, just as newspaper writing does (e.g. *näitleja* 'actor', *uurija* 'detective, investigator'), but the pattern is also actively applied to refer to participants of temporary actions (*helistaja* 'caller', *lugeja* 'reader', *sõitja* 'passenger; rider') or states (e.g. *vaikija* 'the one who is quiet', *sööja* 'the one who eats', *soovija* 'wisher'), and to characterize someone through their negative behaviour (e.g. *sittuja* 'shitter', *karjuja* 'yeller', *pugeja* 'brown-nose'). The latter are very infrequent functions in the scientific register, where most *-ja* derivations are occupations, fixed roles of participants in specific procedures (*kaebaja* 'prosecutor', *kõneleja* 'speaker') or scientific abstract terms and instruments (e.g.

näitaja ‘indicator’, *mõõtja* ‘meter’). Such vocabulary, in turn, is not expected to expand very rapidly or comprise a highly variable lexical base.

The more restricted *-us* derivation pattern (*valitsus* ‘government; governing; ruling; controlling’) is similar to *-ja* in that it also contributes most to the vocabulary of NEWS and FICT, but it also appears to be relatively more profitable in contemporary spoken spontaneous Estonian (SP). The fact that scientific texts and spoken spontaneous language use a similar amount of unique *-us* derivations is unexpected considering the very different nature of those registers, but may make more sense when one takes into account that the spoken corpus also contains public lectures, and that even conversations between two familiar people may include people working in a university. Since this parallel between the two registers is not apparent with other suffixes, one might hypothesize that the *-us* derivation boosts its pragmatic usefulness in the SP sample by realizing more semantically irregular derivations or special terms than the other two suffixes do. However, Table 3 shows that *-us* is responsible for more hapaxes in SP than is *-ja*. This suggests two possibilities. The first is that both *-us* and *-ja* occur in SP in more semantically specialized nouns. The second – and more likely – is that *-us* is pragmatically more useful in SP because it can indeed expand the vocabulary of SP more rapidly than that of the other registers. Whatever the case, the absolute type frequencies (or realized productivity curves) for *-mine* reach higher in all registers than those of the other two suffixes, making *-mine* a definite leader in terms of how often it has been used to express different concepts or relations.

Realized productivity only deals with past productivity. It is not concerned with the question of whether new types are also synchronically derived by speakers using the relevant morphological process (Zeldes 2012: 50). While the differences in realized productivity are a reflection of the extent to which the suitable base words for each derivation pattern *have been used*, the differences in what can be considered a central measure in studies on morphological productivity, **the potential productivity** *P*, relate to the extent to which the remaining available base words *can be used* to create new words (Baayen 1992: 124), i.e. to the likelihood of forming more types than are actually attested in the corpus. The curves in Figure 3 represent *P* as a decreasing function of corpus size: the more tokens that have been sampled, the smaller the

chance that the next noun formed with a given suffix will be a type not encountered before. This holds for all suffixes and all registers, although the curves for *-mine* appear to be less steep than those of the other two suffixes. The latter implies that it takes more time (i.e. more running text) to reach a situation where new types of *-mine* nouns start occurring infrequently. The confidence intervals for *P* are more visible, especially in the smaller subcorpora, indicating more variation in the productivity estimates. The curves and the confidence intervals for *-ja* in the two spoken registers overlap, suggesting that the two registers behave very similarly in terms of how productively *-ja* can be used, given the number of tokens already processed.

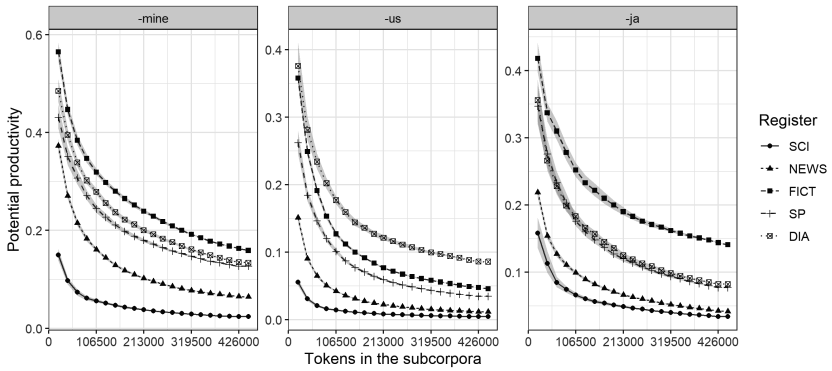

Figure 3. Potential productivity of *-mine*, *-us* and *-ja* in cumulative samples of five registers. The symbols and black lines represent the mean values of 100 random permutations of the samples. 95% confidence intervals are shown with light grey shading.

The spoken and written registers seem to have switched their status based on this productivity measure: the more formal written registers (NEWS and SCI) exhibit the lowest potential productivity for all suffixes throughout the whole sampling process, while the spoken registers (SP and DIA) align with the less formal written register FICT in using a more diverse lexicon of bases for the three suffixes. In dialects, this can partly be attributed to the fact that the data comes from 10 different dialect areas, each of which occasionally makes use of distinct vocabulary not used in others. A high potential productivity in spoken common language and fiction, in turn, indicates a higher possibility of accommodating creative language use, which can be manifested in e.g.

slang or intentional wordplay. While realized productivity only showed that derivations have been more profitable in written registers, potential productivity also accentuates the structural and semantic generality of the patterns: in written registers, the derivation patterns have exhausted more of their potential bases (high realized productivity) and are therefore less likely to expand with new types, especially when the suffix has many restrictions on the set of bases to which it can attach. However, when talking about exhausting the bases for a theoretically infinite repository, as is the case for *-mine* derivation, it becomes apparent that the pragmatic purpose and the topical distribution of a given register limit this repository of available bases in practice.

The potential productivity *P* is sometimes plotted against the number of types *V* (the realized productivity) to assess the so-called **global productivity** of one or more affixes (Baayen 1992: 123–124, 2001: 203–205): two affixes with the same potential productivity can vary greatly in the extent of their use. When this assumption is adapted to compare registers instead of affixes, we can hypothesize that a suffix with more or less the same potential to combine with different bases in two registers (e.g. *-us* in SP and FICT or *-mine* and *-ja* in SP and DIA) will not be equally profitable in those registers. In principle, Figure 2 would be plotted against Figure 3 to capture this. However, in order to avoid 3D-graphs that are difficult to interpret on a 2-dimensional plane, the dimension of the increasing corpus size is disregarded here and *P* is plotted against *V* for each suffix close to the endpoint of the sampling process, namely at a corpus size of 426,000 tokens (Figure 4).

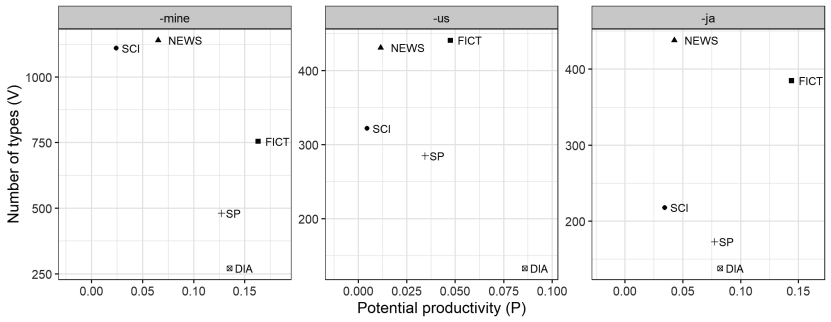

Figure 4. The global productivity of *-mine*, *-us* and *-ja* in five registers (averaged over 100 permutations at a corpus size of 426,000 tokens).

A **high potential productivity** along with a **high number of types**, which seems to be the case for all the suffixes in FICT (when considering the relative measures for each suffix individually), indicates simultaneously high lexical diversity and a relatively high proportion of novel structures among the derivations with the corresponding suffix. This would be the ideal position for a productive derivation pattern, since it presupposes that the derivation pattern, generalized over its many instantiations, can effectively be used for categorizing new, previously unencountered formations. Therefore, fiction seems to be the register where all suffixes are used most productively in the global sense: the suffixes are pragmatically useful for creating multiple different concepts, many of which are used only once, implying that the semantic relationship between the base and the meaning of the suffix is transparent enough for the hapaxes to be understood. As the value for potential productivity is negatively affected by the proportion of semantically idiosyncratic and opaque formations, fiction is also the register where such formations are less likely to occur.

A **high potential productivity** and a (relatively) **low number of types** characterizes all three suffixes in the spoken registers (except for *-us* in SP) and suggests that while deverbal derivation with those suffixes has not in general been pragmatically very useful, many new formations could be expected for larger corpora. The low number of types in spoken registers also results from the fact that while formal written registers are known to be very "nominal" (Kerge 2003), spoken discourse hosts a considerable amount of speech particles, which are not attested or are used in a very limited manner in written registers. Therefore, at any given sampling point, there are fewer nouns in the spoken registers, because other word classes (e.g. speech particles, interjections) take up a considerable proportion of the total tokens. The exact proportion of such words in the corpora used in this study is difficult to assess. The CED has a special tag for speech particles, but since the corpus relies on manual annotation, such particles have been classified unsystematically. For instance, some annotators have assigned particle status to some very frequent unstressed adverbs, such as *siis* 'then' and *nüüd* 'now', while others have always tagged them as adverbs. In PCESS, the morphological annotation has been done with tools trained on written data. This means several things: there is no special tag for particles; *siis* and *nüüd* are always tagged as adverbs; some particles are

analyzed as interjections (e.g. *aa* 'oh'); and some fillers which would be tagged as particles in the CED (*ee*, *mm*) are not fed as input to the morphological analyzer at all. In any case, when we compare the word classes tagged in the samples from all 5 subcorpora (Figure 5)[18], it is evident that the overall proportion of nouns (including derived nouns) is considerably smaller in the spoken registers.

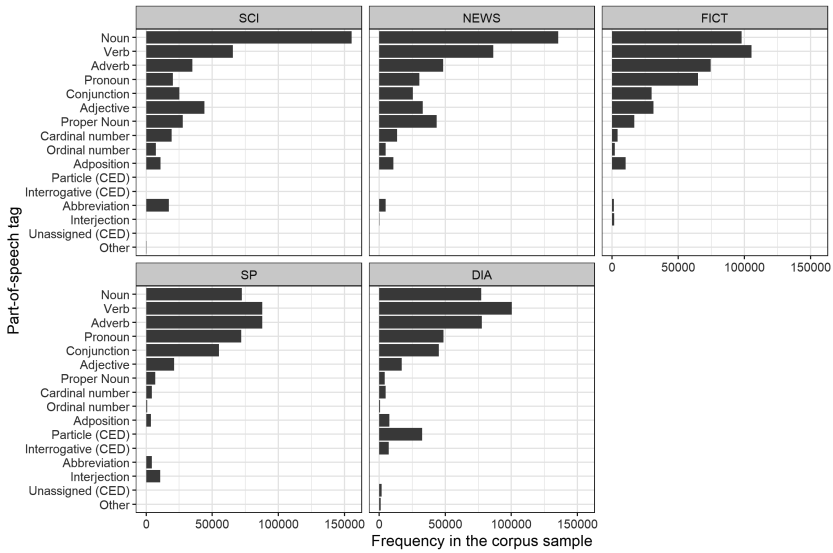

Figure 5. Word class distribution in the five samples. The difference between the distribution of part-of-speech tags is statistically significant (X^2 (60, N = 2,158,435) = 494,582, $p < .01$).

On the one hand, this distinction between written and spoken registers inevitably distorts the comparison of productivity measures across different registers, especially the category-internal measure *P*, since it estimates the growth rate of the derivational categories in spoken registers based on a considerably smaller token count. This would essentially result in problems similar to those associated with the straightforward comparison of suffixes with different token counts, such

18 The more detailed tagging systems of individual corpora are generalized to be comparable across all the registers. For example, the verb and the auxiliary in the CED are subsumed under a general class *Verb*, and adjectives tagged for different degrees of comparison (comparative, superlative) are converged to a general class *Adjective*.

as overestimating the productivity of the suffixes in registers where the token count is low (see Section 3.2). On the other hand, this distinction in the distribution of word classes is naturally characteristic of the types of texts that occur in those registers. It is also an inherent property of the different modes of communication and the dynamics that shape productivity.

Therefore, the lower number of derived noun types in spoken data has to be accepted as relevant information about how discourse is structured differently in speech and writing. It can be used to explain why the spoken registers appear to show higher likelihood of forming new types with the deverbal nominalization categories but at the same time actually profit less from the concepts related to those derivational patterns.

Low potential productivity and a **high number of types** as for all three suffixes in journalistic texts (NEWS) and *-mine* (perhaps also *-us*) in scientific texts (SCI) in Figure 4 suggests that while those derivation patterns are extensively used in the more formal written registers, there is a relatively low chance that those derivations should be either unintentional rare formations or intentional neologisms, i.e. single instances of creative language use. Instead, given that the formally or semantically idiosyncratic properties of some formations may reduce potential productivity (Baayen 1992), there is reason to suggest that in NEWS and SCI, speakers more frequently use vocabulary that is in some sense specialized or derivations that are simply conventionalized instead of profiting from the relevant morphological processes for the creation of regular and semantically transparent derivations.

Low potential productivity and a **low number of types** means that the derivation pattern is not pragmatically very useful in a particular register and has also exhausted most of its potential members. This might be the case for *-ja* in SCI. While *-ja* is restricted mostly in terms of its semantic generality, the overall limited pragmatic usefulness of actively formed agent nouns in scientific texts prevents a rapid extension of the base domain to which this morphological process applies.

The third core measure I seek to assess is the completely hapax-based **expanding productivity**. This measure reflects the attractiveness of a morphological category for expressing *any* novel concept. In other words, it assesses the probability that any previously unencountered new word added to the corpus will be a *-mine*, *-us* or a *-ja* noun and thereby measures how much the suffixes contribute to the growth rate of

the vocabulary as a whole (see Table 2). Expanding productivity should gauge both semantic and structural generality as well as the future pragmatic usefulness of the derivation patterns.

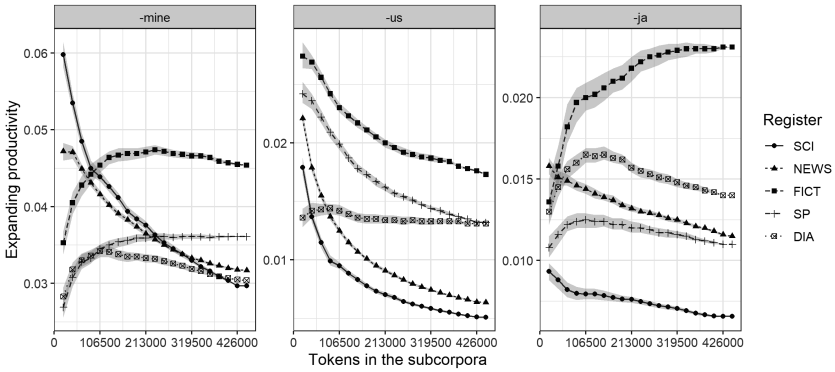

Figure 6. Expanding productivity of *-mine*, *-us* and *-ja* in cumulative samples of five registers. The symbols and black lines represent the mean values of 100 random permutations of the samples. 95% confidence intervals are shown with light grey shading.

The curves in Figure 6 represent a less monotonic function between *P** and corpus size than appears in the other two measures of productivity. Only for *-us* does the general trend appear to be negative, which means that in most registers (except for DIA), the likelihood that any new concept added to the corpus is expressed as an *-us* noun decreases as the corpus size increases. This tendency is expected, since *-us* is the suffix which has the most structural restrictions. Therefore, the list of available verbal bases for *-us* becomes shorter as more and more bases have already been used, while the list of available bases for the general suffixes *-mine* and *-ja* is, in principle, infinite: any new verb which is created can be used as the base for *-mine* nouns, and any new verb which entails an actor of some sort can be used as the base for *-ja* nouns. The negative trend is also visible for the other two suffixes in the two formal written registers, suggesting again that these registers might benefit more from an existing lexicon of derived nouns. For the less formal registers, however, the contribution of *-mine* and *-us* to the growth rate of the vocabulary does not decrease as the sample grows. In fact, it has been argued that for productive affixes, there should be a positive correlation between expanding productivity and sample size (Baayen 1994).

Although this tendency is not particularly clear in the small sample used in this study, it can be expected that at least in the informal registers, the proportion of *-mine* and *-ja* nouns among hapaxes in the corpora will increase as more tokens are added to the corpus, and P^* will become an increasingly accurate estimate of the number of true neologisms – words which are created out of a pragmatic need and which do not occur in large dictionaries (Baayen 1994). For each hapax in the samples, it is in fiction (FICT) where it is most likely to be a *-mine*, *-us* or *-ja* derivation. This means that compared to other registers, fiction provides an environment in which all the derivational patterns under investigation are attractive categories for forming novel, previously unencountered concepts. As expanding productivity is a completely hapax-based measure and hapaxes are unlikely to represent semantically opaque formations, fiction also appears to facilitate more use of lexical procedural knowledge in the production of deverbal nouns, as well as open up the possibility that a reader might access the forms from their lexical memory. This is true for all suffixes, irrespective of their degree of productivity.

The relative position of the productivity curves in the other four registers depends on the suffix. *-mine* is a somewhat more attractive category for novel formations in spoken spontaneous data (SP) after the sample reaches approximately half of its full size. The fact that formal written registers do not appear to see *-mine* nouns as a very attractive category for expanding their vocabulary might result from the need to repeatedly reuse created forms (e.g. for anaphoric referencing or for terminological purposes), which has a negative effect on the number of hapaxes in the samples. The lower rank of dialects could again be the result of the diverging vocabulary used in different dialects, which causes the overall contribution of *-mine* nouns to the growth rate of the "global" dialectal lexicon to remain relatively low. It must be noted, however, that the results for *-mine* in Figure 6 differ somewhat from those in Pilvik (2019), where the same measure was used to assess the expanding productivity of *-mine* in different samples from the same corpora (at approximately 335,000 tokens). Although the top two registers in Pilvik (2019) were the same – FICT and SP –, dialects actually ranked third, while the low expanding productivity of *-mine* in scientific writing was even more extreme. This encourages a cautious interpretation of the results, since the samples are small and somewhat unstable. In this study, scientific writing (SCI) is also ranked lowest for *-us* and

-ja with regard to expanding productivity. This register is therefore the least likely to express a novel concept through derivational categories. The reason why the suffixes are not particularly attractive in SCI might again be linked to the need to use derived forms more than once, to the structural restrictions and semantic irregularity of *-us* and to the low pragmatic usefulness of *-ja*.

5.1.2. Differences between suffixes: a variable-corpus approach

As mentioned earlier, the main critique of *P* (the ratio of hapaxes to tokens in a given morphological category) is that it does not provide the possibility to adequately compare affixes with different token counts or to rank affixes when both their *P* and *V* are different (Baayen 1992: 124, Gaeta & Ricca 2006). This critique is based on the nondecreasing monotonic relationship between tokens and types in general (see Baayen 1989: 104): if a pattern is even minimally productive, more types will emerge as token count increases. However, the curves for different affixes (and registers) will increase with different slopes (Gaeta & Ricca 2006: 58–59). It is evident from Figure 7, for example, that *-us* in the scientific register (SCI) nearly exhausts its potential for new formations at a sample size of only 5000 *-us* tokens. After that, only a few new types emerge throughout the sampling process, although the suffix token count continues to grow. In other registers, the *-us* suffix token count in the total sample is lower than in SCI, but the rate with which new types occur is higher. This is also expressed in Figure 7 through the type-token ratios (*TTR*), which are calculated close to the endpoint of the sampling process, at 426,000 tokens. *TTR* is a measure usually used to assess lexical variation of a document/text or a whole corpus. The closer the *TTR* value is to 1, the more lexical variation there is (Baker, Hardie & McEnery 2006: 162). The measure can also be used to assess lexical diversity within a category. For example, on average, every 20th *-us* token, every 9th *-ja* token and every 6th *-mine* token in NEWS represents a new type. (There are 5 *-us* types, 11 *-ja* types and 16 *-mine* types per 100 tokens.)

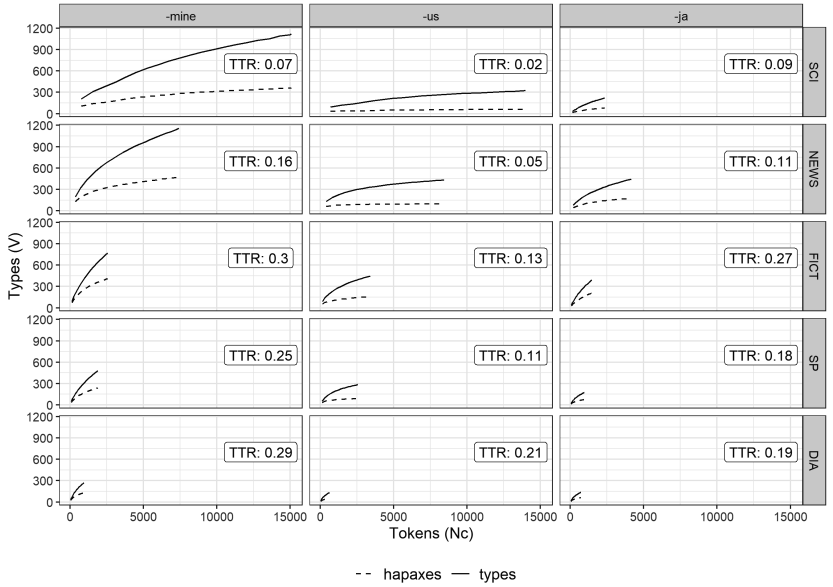

Figure 7. Type counts as a function of the number of tokens in a morphological category.

As the numerator in *TTR* is the type count *V*, this measure is also closely linked to the realized productivity of an affix. However, the more tokens with a given suffix that have been sampled, the smaller the likelihood of coming across new, previously unencountered types and the lower the proportion of unique words among all tokens (Hardie & McEnery 2006: 139, Gaeta & Ricca 2006: 59). Thus, the *TTR* for *-us* in NEWS would be higher if it were calculated at the point where only 5000 tokens were sampled, while the *TTR* for *-ja* in NEWS would likely be lower once we included 3000 more *-ja* tokens in the sample. This makes both *TTR* and *P* sensitive to the number of sampled suffix tokens (*Nc*) when ranking different suffixes, since the values for less frequent suffixes could be overestimated. To overcome this proposed limitation, Gaeta and Ricca (2006) suggest using a variable-corpus approach as a modification of Baayen's procedure, which means calculating the potential productivity value at an equal number of tokens for all affixes.

The potential productivity curves are presented again in Figure 8, but this time, different graphs represent registers instead of suffixes, the curves represent the suffixes instead of the registers, and the *x*-axis expresses the suffix token count instead of the subcorpus size. If *P* were

calculated at the total number of suffix tokens at the endpoint of the curves, the rank order of the suffixes would be *-mine*, *-ja*, and *-us* in all registers except in SCI, where there are so few *-ja* tokens compared to the other two suffixes that *-ja*'s potential productivity is assessed even above that of *-mine*. This is definitely not intuitive, since there is no reason to believe that agent nouns with semantic constraints and low pragmatic usefulness (see Figure 4) have more potential to expand their category in academic and scientific texts than action nouns with almost no semantic restrictions. Therefore, comparing these results with those from the variable-corpus approach would be beneficial to assess the degree of over- or underestimation of *P*. The dashed vertical lines in Figure 8 show the fixed values of *Nc* at which the values for *P* would be estimated and compared (see Table 4 below). In all registers except for dialects, *-ja* is the least frequent suffix. In dialects, *-us* is the least frequent suffix.

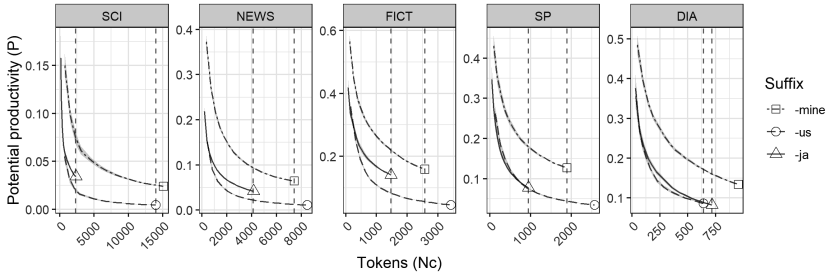

Figure 8. Potential productivity of *-mine*, *-us* and *-ja* in five registers. The vertical dashed lines mark the token counts for which the values of *P* are estimated in the variable-corpus approach.

In order to provide precise productivity measures for comparison, generalized additive models were fitted to the observed data in 100 permutations with the discrete values of *P* as the response and the corresponding suffix token counts *Nc* as the predictor variable. This was done for each suffix in each register: for each suffix, there were altogether 100 times 21 observations per register. This made it possible to interpolate the values of *P* for a fixed token value (Nc_f) in each register and compare the productivity ratios of two suffixes at different points of their curves. Table 4 presents the interpolated values of *P* multiplied by 100 at a given number of tokens Nc_f in each register.

Table 4. The estimated potential productivity values (*P*) at a fixed Nc_f number of tokens, multiplied by 100.

	SCI		NEWS		FICT		SP		DIA	
Nc_f =	**2346**	**13,987**	**4135**	**7417**	**1472**	**2570**	**947**	**1894**	**642**	**717**
-mine	6.83	2.57	9.17	6.33	21.81	15.85	18.14	12.58	17.25	16.07
-us	1.43	0.42	2.42	1.32	8.26	5.69	7.05	4.50	8.49	–
-ja	3.28	–	4.12	–	13.96	–	7.64	–	9.01	8.09

It is estimated, for example, that for 2346 *-mine* formations in the scientific corpus, 6.83% of them are types occurring only once. For the same number of *-us* formations, this figure is only 1.43%. *-ja* ranks between the two with 3.28% of all formations representing hapaxes. At 13,987 suffix tokens, only 2.57% of *-mine* formations and 0.42% of *-us* formations are hapaxes. As there is no data for *-ja* in that frequency range (there are only 2346 *-ja* tokens in the total sample), the second value of *P* can not be interpolated for this suffix. Table 4 suggests that *-mine* is the most productive suffix in all registers at both measuring points; *-ja*, usually the least frequent suffix, ranks second at the first measuring point, and *-us* ranks third. Figure 9, however, illustrates that the magnitude of difference between the values of *P* for two suffixes varies greatly across registers and increases as more tokens are sampled. As the second Nc_f always contains data for only 2 out of the 3 suffixes, the latter situation can be represented by only one ratio.

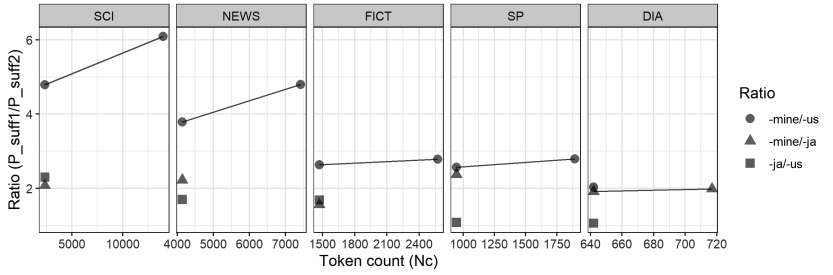

Figure 9. The productivity ratios of suffixes at two fixed token counts (Nc_f).

The comparison of productivity ratios along with the interpolated values of *P* in different registers provides a fuller account of how available the different suffixes are for new formations. In SCI and NEWS

where the proportion of hapaxes among the derivational suffixes is the lowest in absolute numbers (see Table 4), the relative difference in the productivity of *-mine* and *-us* is the most obvious: the proportion of hapaxes among *-mine* nouns is around 4–5 times greater than the proportion of hapaxes among *-us* nouns and is more than 6 times greater in SCI when the token count increases. In other registers, the ratio of *P* for these two suffixes is also the largest, but the magnitude of the difference is considerably lower. As *P* is sensitive to the proportion of semantically opaque forms within the category, these results, along with the absolute values in Table 4, suggest that in fiction and spoken registers *-mine* and *-us* derivations fulfil more similar functions and are therefore perhaps more often used as general and regular derivation patterns. In SCI and NEWS, there are more lexicalized or idiomatized *-mine* and *-us* nouns, and the *-us* derivation pattern in particular hosts several frequent formations, used repeatedly as fixed terms (e.g. *karistus* 'punishment', *ulatus* 'extent').

As a non-category-internal and completely hapax-based measure, expanding productivity *P** suffers less from the fact that there is a different number of tokens for each suffix: while the number of suffix hapaxes in the numerator is different, the number of corpus hapaxes in the denominator is the same for all suffixes from the same register. Therefore, when comparing suffixes from the same register/corpus, differences between *P** correspond to the differences between the simple number of hapaxes (Gaeta & Ricca 2006: 61). This also holds for potential productivity *P*, which is calculated at a fixed suffix token count, because again, the denominator Nc_f is the same for all suffixes. The difference is that for *P**, the hapaxes are compared in the whole sample for all the suffixes, while for *P* in the variable-corpus approach, the hapaxes are compared in the whole sample for only some suffixes and at a random fixed sampling point for others. We can check if the ranking of the suffixes based on *P** (calculated at the total number of suffix tokens) and *P* (interpolated at an equal number of suffix tokens) is correlated as suggested by Gaeta and Ricca (2006: 62), Baayen (2009) and Zeldes (2012: 65). Figure 10 presents *P* on the *x*-axis and *P** on the *y*-axis. The dashed lines show the corresponding trends based on *P* computed from the total sample as in Baayen's original approach.

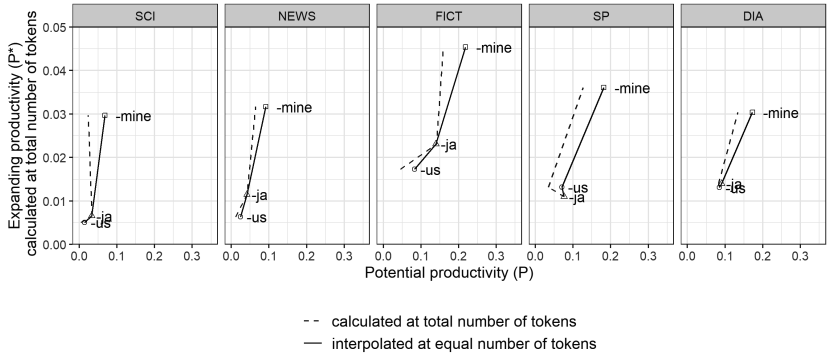

Figure 10. Comparison of expanding productivity *P** and potential productivity *P*. The lines connecting the suffixes do not imply continuity, but are there for visualization purposes.

The ranking of the three suffixes based on the two measures coincides in almost all registers and corresponds to general intuition: *-mine* is by far the most productive suffix both in terms of the growth rate of the category itself (*P*) as well as in terms of its contribution to the growth rate of the whole vocabulary (*P**). *-mine* is followed by *-ja*, which is followed by *-us*. Only in SP is *-us* ranked above *-ja* by its expanding productivity, while their (interpolated) values for potential productivity *P* are very similar (see Table 4). It is also apparent that in the original approach, the *P* of the more frequent suffixes is systematically estimated lower.

Finally, I compare both Baayen's original approach and the variable-corpus approach applied by Gaeta and Ricca (2003, 2006) to the process of averaging the productivity values from multiple subsamples as done in Plag, Dalton-Puffer & Baayen (1999). The results from the three approaches are illustrated in Figure 11.

Figure 11 shows that the approaches differ first and foremost in the absolute estimations of productivity: Baayen's original approach, where *P* is calculated at the total number of tokens, seems to provide the smallest probability for a token with a given suffix to be a hapax, especially when the more informal registers are compared with other approaches. In turn, the average of all the values of *P* in the 21 subsamples (as done in Plag, Dalton-Puffer & Baayen (1999), with the exception that their subsample frequencies were modeled, while mine are simply averaged over 100 permutations) provides the highest and

most divergent estimates. The latter approach seems therefore the most radical, because it suggests that the likelihood of a derived noun occurring only once in the sample is nearly 22 times higher when this noun ends with *-mine* in fiction than when it ends with *-us* in scientific texts. However, considering that *P* is a decreasing function of sample size, these results are not surprising, given that smaller subsamples with larger *P*s factor into the approaches of both Gaeta and Ricca and Plag and his colleagues, while only the total samples are considered in Baayen's original approach.

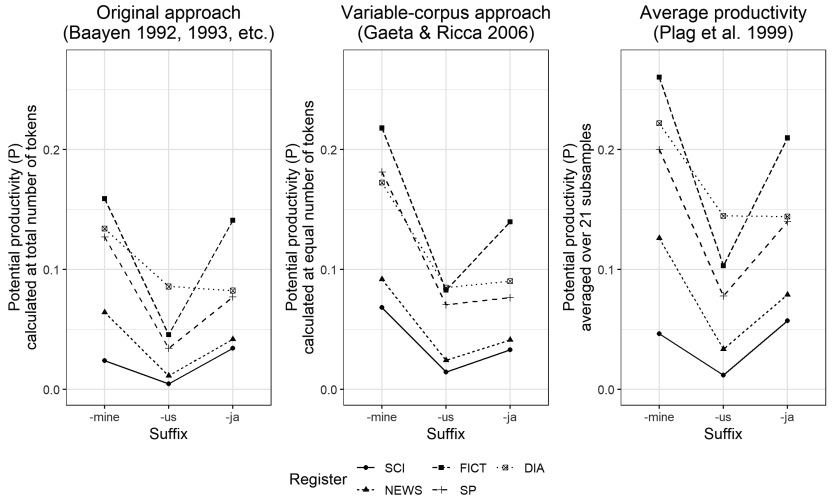

Figure 11. Comparison of the values for potential productivity *P* obtained with three different approaches.

The ranking of the suffixes is quite uniform across registers and approaches: *-mine* is estimated to be the most likely suffix to produce new types, followed by *-ja* and finally *-us*. There are a few exceptions. First, in the original approach and as a result of averaging, the productivity of *-us* and *-ja* is estimated to be equal in dialects (DIA); in the variable-corpus approach, the same thing happens in both spoken registers. As suggested earlier, this means either that *-ja* has lower productivity in dialects than in other registers, or that *-us* has higher productivity in dialects than in other registers. Both could be explained using the pragmatic and structural constraints mentioned in Section 2. The core function of *-ja* would be to identify an actor or an experiencer by

the designated action or state (Kasik 2015: 197). As such, it simply might not be pragmatically very useful in most registers. In most registers, there may be less need for active word formation, and the most common formations may be more or less conventionalized, either as the names of professions, fixed roles in specific processes, or as instruments. In spoken discourse, *-ja* nouns are also actively competing with other referential devices, such as pronouns and proper names. The generally low productivity of *-us*, in terms of both *P* and *P**, is the combination of its structural restrictions and the semantic vagueness of that category (see Baayen 1992: 109), which I describe in Section 2. The diverse meanings that arise from the *-us* category, as well as its extensive use as a deadjectival suffix, might prevent language users from fully exploiting the derivational potential of the suffix for expressing the full range of theoretically possible concepts, especially when there are other, more general patterns available (*-mine*). However, in spoken data and especially dialects, the *-us* pattern can exhibit more productive behaviour thanks to a range of formations which are rare in contemporary written language (e.g. *kaetus* 'jinxing'), so increasing the hapax count. The fact that the *-us* derivation pattern seems to be used more productively in dialects, partly thanks to the rich dialectal lexicon, was also evident in Figure 6, where DIA was the only register in which the expanding productivity *P** of *-us* did not exhibit a negative trend. Not all hapaxes in dialects represent semantically regular examples of *-us*, however. For example, words like *hapatus* 'leaven' and *kastus* 'sauce' occur only once in the dialect sample but designate referential nouns whose meanings have become conventionalized through metonymic association with the state of affairs depicted by the verbal stem.

The second departure from the overall ranking order of the suffixes is that in Baayen's approach, *-ja* is considered the most productive suffix in scientific writing (SCI), even surpassing *-mine*. This, as mentioned earlier, seems somewhat counterintuitive. Although neither of the suffixes have syntagmatic restrictions, *-ja* is semantically less general, because it prefers its base to have a potentially active subject. *-ja* would also seem conceptually considerably less useful in scientific texts unless used in some less regular way, like for referring to instruments (e.g. *lugeja* 'reader', *mõõtja* 'meter, measurer', *saatja* 'transmitter'), certain participant roles (e.g. *kõneleja* 'speaker', *osaleja* 'participant', *kasutaja* 'user') or abstract concepts (e.g. *näitaja* 'indicator', *kordaja*

'coefficient'). These, in turn, would not likely be hapaxes. Therefore, the higher productivity of *-ja* appears to be linked to the original critique of Baayen's procedure that motivated the variable-corpus approach: the productivity of less frequent suffixes is overestimated when calculated based on the total number of tokens (or the productivity of more frequent suffixes is underestimated in comparison). *-ja* also ranks slightly higher than *-mine* in scientific texts in the approach that averages the productivity estimates. As in this case, the productivity values that underly the mean value are also calculated at the number of maximum suffix tokens in each of the 21 subcorpora, this approach might suffer from the exact same problem of over- and underestimation.

The three approaches do differ in terms of which register is estimated to use the derivation patterns most productively. Compared to other methods, the variable-corpus approach makes the clearest distinction between the formal written registers NEWS and SCI and the less formal registers FICT, SP, and DIA, especially with regard to the more frequent suffixes *-mine* and *-us*. Overall, fiction is the register which seems to promote productive use of the more regular and general suffixes, *-mine* and *-ja*, while dialects give stage to expanding the *-us* category with new formations.

With such small samples, it is difficult to say which of the three approaches is the most adequate. Some approaches seem to suffer from overestimating the productivity values while others suffer from the opposite. While Gaeta and Ricca's method is advertized as more desirable, considering the negative correlation between potential productivity and the number of affix tokens, it fails to provide fully reliable results when samples and total token counts are small. Indeed, Gaeta and Ricca (2006: 68) warn us about potentially misleading results when *P* for very frequent affixes is estimated at a relatively low token count (Nc_f), which is the case for *-mine* and *-us* in SCI at least. For example, 2346 tokens of *-mine* are reached when only about 85,000 tokens have been sampled in the corpus of scientific writing. In such a small subset, hapax legomena are not really expected to represent very rare words, let alone neologisms. The low frequency of certain types, then, is more indicative of a lower pragmatic usefulness of these words in a given register than it is a reason to claim anything about their status in the mental lexicon (Gaeta & Ricca 2006: 68, Baayen 1994: 453). In fact, this limitation applies to all registers in this study, since the size of the PCESS confines

the size of the samples to analyze. It is therefore difficult to make any strong statements about the different degrees of activation and the role of mental storage concerning the three patterns.

There is another theoretical concern with Gaeta and Ricca's approach. Fixing Nc_f at a given value for all suffixes disregards the fact that the suffixes are simply not equally useful in given corpora or samples. The fixed Nc_f has no statistical meaning for the more frequent suffixes, since it represents a random point at their sampling curve and the values of *P* for the frequent and less frequent suffixes are therefore not anchored in the same corpus. Consequently, for example, the next token added to the *-ja* category does not have the same status as the next token added to the *-mine* category, because the two words would come from different corpora. Hence, while plotting the potential productivity of the suffixes at a fixed number of tokens can give a nice quantitative indication of the dynamics of their respective growth curves, Gaeta and Ricca's variable-corpus approach for comparing the potential productivity of different affixes solves one issue by raising another. As mentioned earlier, when potential productivity *P* is calculated at a fixed number of tokens, the ratios between potential productivity values for different suffixes correspond to those between their expanding productivity values *P**, which are calculated from the total number of tokens. According to Zeldes (2012: 66), instead of comparing *P*, "it is therefore more logical to apply *P** to the full samples and compare values, since its meaning relates to the portion of innovation each process plays as a whole in a certain mass of data".

In the next section, I explore the relationship between the frequency of the derived words and the frequency of their bases to provide some insight into the question of how the individual words affect the productivity of a word-formation pattern and whether base frequency helps to discover the formations which are more likely to be lexicalized. As Hay and Baayen (2002) have shown, derived forms which are more frequent than their bases are more prone to whole word access, regardless of the absolute frequency of the derived form. I will also compare the relative frequencies of bases occurring in both *-mine* and *-us* nouns to examine whether relative base frequency helps to determine the degree of semantic similarity between the derivations.

5.2. Correlations between frequencies

5.2.1. Correlation between base and derivation frequency

Without a doubt, derivations that fall under each pattern analyzed in this study are of different degrees of decompositionality. For example, it is relatively unlikely that the meaning of the word *näitaja* 'estimate, indicator' is processed as a combination of the meaning of the base *näita-* 'show' and the meaning of the suffix *-ja* 'someone doing something', evoking a mental image of someone showing something, and not accessed as a whole word referring to a formal term meaning 'indicator'. The meaning of the formation *unistaja* 'dreamer', however, is more easily parsed into its component meanings.

There is evidence that formations with high-frequency base words are processed more easily than formations with low-frequency base words, irrespective of whether the pattern is productive or not (Baayen 1992). However, Hay (2003) and Hay and Baayen (2002) show that token frequency alone does not suffice to decide whether a form is accessed as a whole or parsed in lexical processing. Instead, it is the relative frequency between the base and the derivation which is a better indicator of the degree of decompositionality. In other words, the correlation between base and derivation might help better distinguish between semantically opaque and transparent forms.

Figure 12 presents the correlation between the base verb stem and the corresponding derivation for each of the three affixes in five registers using the natural logarithm of the absolute frequencies. The reason for this transformation is that humans tend to process frequency in a logarithmic manner and perceive differences between smaller frequencies as more substantial than differences between higher frequencies (Hay & Baayen 2002). All frequencies were increased by 1 in order to avoid negative infinite values when taking a log of 0. The strength and nature of the correlations between base and derivation frequency are expressed via Spearman's nonparametric correlation coefficients (*rs*), which are complemented with GAM curves visualizing the relative frequency effects. Each graph also includes some lexemes representing the bases which are more frequent in verbs, those which are more frequent in derivations, and those which are frequent in both.

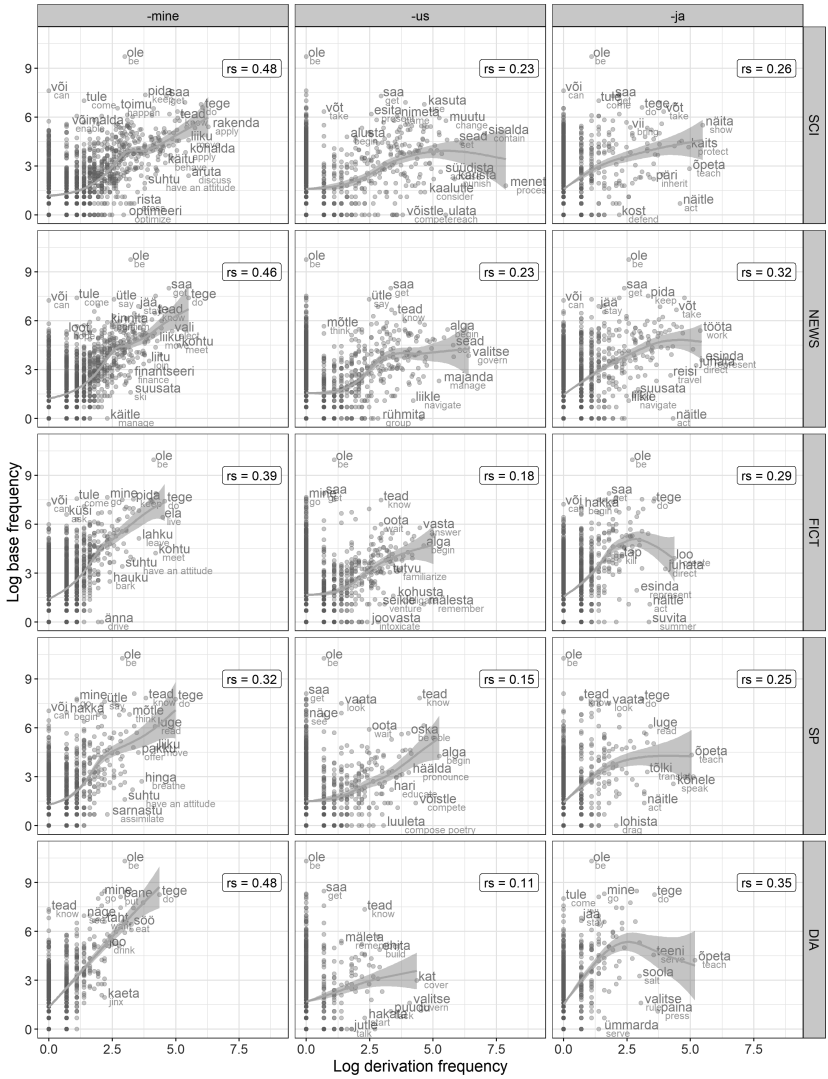

Figure 12. The relation between log base frequency and log derivation frequency for the suffixes *-mine*, *-us* and *-ja* in five registers.

Figure 12 shows a positive monotonic correlation between the log of base verb frequency and the log of *-mine* derivation frequency from the corresponding base. Spearman's nonparametric correlation coefficient (*rs*) indicates a moderate to strong correlation, with SCI, DIA and NEWS showing the strongest link. This means that though most *-mine* nouns

and most verbs occur with very low frequency, *-mine* derivations with higher frequency appear to be formed also from higher frequency verbs. This suggests high regularity: the pragmatic necessity for expressing a certain state of affairs also increases the need for the corresponding nominalization. This correlation is less explicit with other suffixes: the association between the base and the derivation frequency is weaker (especially with *-us* nouns in the two spoken registers) and not necessarily monotonic. There are more high frequency bases from which no formations are derived with *-us* or *-ja* (e.g. the copula *olema* 'be; have' or the modal verb *võima* 'can; may'). More importantly, the deviations from the general trend line seem more extreme with *-us* and *-ja* (e.g. *ulatus* 'extent; extending; reaching; handing', *võistlus* 'competition; competing', *valitsus* 'government; ruling; governing' or *näitleja* 'actor', *painaja* 'nightmare'). The former aspect is linked to the morphophonological constraints and semantic vagueness of *-us* derivations and the low pragmatic usefulness of *-ja* derivations (in addition to the agentivity preference). The latter tendency, in turn, suggests that the derivation patterns host several formations for which the degree of decompositionality in human morphological parsing might be lower.

For *-mine*, the lexemes falling below the general trend line mostly fall into two groups: either they are characteristic of the topics discussed in certain registers (e.g. *optimeerimine* 'optimization', *suusatamine* 'skiing', *käitlemine* 'management', *ännamine* 'driving', *sarnastumine* 'assimilation', *kaetamine* 'jinxing'), or they represent more frequent formations which have acquired specialized meanings in addition to the action noun reading (e.g. *liikumine* 'movement; moving', *kohtumine* '(a) meeting; meeting', *suhtumine* 'attitude; having an attitude', *valimine* 'election; electing', *elamine* 'household; living situation; living', *pakkumine* 'offer; offering', *söömine* 'food; eating'). The lexemes above the line represent semantically general bases whose meaning is specified in context (e.g. *tulemine* 'coming', *olemine* 'being', *saamine* 'getting', *tegemine* 'doing; making', *minemine* 'going', *panemine* 'putting').

The same tendency can be seen with *-ja*: the lexemes below the line represent mostly occupations (*näitleja* 'actor', *juhataja* 'manager', *valitseja* 'ruler; governer', *maletaja* 'chess player', *ümmardaja* 'servant', perhaps also *looja* 'creator; god') or less permanent roles (*kostja* 'defendant', *esindaja* 'representative', *suvitaja* 'vacationer', *kõneleja* 'speaker'). The bases above the line, again, are semantically

vague (*pidaja* ‘keeper’, *tulija* ‘the one who comes’, *saaja* ‘the one who gets’, *tegija* ‘the one who does’).

For *-us*, most of the depicted lexemes can occur in some specialized meaning, irrespective of their position relative to the trend line. For some nouns, especially the ones with different suffix allomorphs, the (temporal) action noun reading seems impossible (e.g. *ollus* ‘substance, matter’, *võtlus* (in *ette+võtlus*) ‘enterpreneurship’, *saadus* ‘product; produce’, but also lexicalized *katus* ‘roof’, *ulatus* ‘extent’, *liiklus* ‘traffic’, *teadus* ‘science’, *vaatus* ‘act (of a play)’, *ütlus* ‘(a) saying’). For others, the action noun reading is possible but difficult to evoke due to the strong entrenchment of some semantically related meaning (e.g. *sisaldus* ‘content; containing’, *karistus* ‘punishment; punishing’, *majandus* ‘economy; managing economics, household etc.’, *armastus* ‘love; loving’, *mälestus* ‘memory; commemorating’, *kohustus* ‘obligation; obligating’, *joovastus* ‘exuberance; intoxication; exuberating; intoxicating’, *luuletus* ‘poem; composing poetry’, *puudus* ‘lack; shortage’, *jutlus* ‘sermon; conversing’). Some *-us* nouns, in turn, lend themselves to the action noun interpretation more easily (e.g. *rühmitus* ‘(a) group; grouping’, *võistlus* ‘competition; competing’, *menetlus* ‘procedure; processing’, *kasutus* ‘usage; using’, *valitsus* ‘government; governing; ruling; controlling’, *etendus* ‘performance; performing’, *lõikus* ‘(a) cut; cutting; harvesting’). However, it seems that the latter interpretation would require that the noun somehow be made definite or specific, either by adding a determiner or an agent or patient argument (e.g. *selline arvude rühmitus* ‘such grouping of the numbers’), or that the process be made explicitly temporal and unbounded (e.g. *aastatepikkune kasutus* ‘years of using’).

Most of the formations whose semantics deviate in some respect from the general meaning of the action noun or agent noun schema are not semantically irregular formations. Instead, they instantiate subschemas or polysemic links related to the primary meaning of the action or agent noun schema in a regular way through metonymic associations. For example, as mentioned in Section 2, *-us* formations can systematically express other participants in an event frame evoked by the verbal base: results, instruments, objects, subjects, and even locations. From a constructionalist perspective, it would perhaps make more sense to regard such systematic correspondences simply as extensions

of the prototypical meaning and not as lexicalized, semantically opaque instances of a derivation pattern.

In any case, base frequency is not the best predictor of decompositionality or mental activation of *-us* derivations, though the correlation is stronger for the two regular suffixes *-mine* and *-ja*. For *-mine*, the distribution of bases and derivations roughly corresponds to the so-called NV-scores used in Pilvik (2019), where the relative frequency of the base among all verb stems in the sample was simply subtracted from the relative frequency of the corresponding *-mine* type among all *-mine* tokens. This was done in order to look for the *-mine* types which are more attracted to the nominalized structure than would be expected from the overall frequency of the base verb. Such formations tend to occur in more specialized meanings (events, objects or results of processes), e.g. *kohtumine* 'meeting', *pakkumine* 'offer', *nõudmine* 'demand' (Pilvik 2019: 93).

Determining the threshold line above which the derived forms are likely to be parsed (see Hay and Baayen 2002, who show how this line is related to the affix's productivity) would require more advanced statistical modeling and more data or a different sampling method to balance the influence of individual speakers/writers. Therefore, this remains outside the scope of this article. Since the samples are small, the results here must again be interpreted with caution, because there is obvious topical and idiosyncratic bias in the samples. For example, *teadmine* ('knowledge; knowing'), a common semi-lexicalized *-mine* derivation in other registers, does not occur in the sample of DIA used in this article. This does not mean that the word is not used in dialects or even that it does not exist in the CED (performing a query using the corpus web interface[19] returns 9 hits of *teadmine*), but rather that it is not used frequently enough in that type of discourse to come up in every possible sample of the corpus. Neither the derivation *ännamine* 'driving around with no purpose' nor its base is listed in any dictionary of Estonian, and they occur in only one text by one writer in FICT. In addition, the topics of linguistics are overrepresented in the PCESS, as the more frequent derivations *sarnastumine* 'assimilation' and *kõneleja* 'speaker' imply. Additionally, there is the curious case of *tegema* 'do', which is the base for by far the most *-mine* nouns in almost all registers

19 www.murre.ut.ee/mkweb (Accessed 15.12.2019.)

(in SCI, the most frequent *-mine* noun is *rakendamine* ‘applying; application’, but *tegemine* runs a close second). On the one hand, its frequent use is expected, since it can be classified as a lexically general noun in Estonian whose meaning is specified through a possibly infinite number of complements. In the process of *-mine* nominalization, complements are incorporated into the NP as prenominal genitive modifiers (e.g. *hea nalja tegemine* ‘making a good joke’), in compounds (*naljategemine* ‘joking’) or as adverbial modifiers (e.g. *välja tegemine* ‘treating (someone at one’s own cost)’, *haigeks tegemine* ‘making sick’). However, *tegemine* also occurs in two very frequent constructions. The predicative construction *olema* ‘be’ + *tegemine*$_{\text{PRT}}$ + PRED$_{\text{COM}}$ (8) is very frequent in written language, whereas in spoken language, the so-called *busy*-construction A$_{\text{ADE}}$ + *olema* ‘be’ + *tegemine*$_{\text{PRT}}$ (+ X$_{\text{COM}}$) (9) often occurs.

(8) *Tege-mis-t* *ol-i* *nalja-ga*
do-NMLZ-PRT be-PST.3SG joke-COM
‘It was a joke.’

(9) *Ema-l* *ol-i* *looma-de-ga* *tege-mis-t*
mother-ADE be-PST.3SG animal-PL-COM do-NMLZ-PRT
‘Mother was busy with the animals.’

The syntactic productivity of the aforementioned constructions is limited as the specific semantics of the constructions restrict the lexical set of nominalizations to only *tegemine* (or the semantically equivalent *-u* nominalization *tegu*)[20]. However, different types of action nominal constructions (in the sense of semantically non-compositional syntactic units) with varying degrees of schematicity and generality are a common phenomenon in the Finnic area and also in Estonian (see Neetar 1988, Sahkai 2011, Pilvik 2016, 2017). Therefore, the corpus frequencies and productivity values of both individual types and the derivation patterns as a whole are undoubtedly also affected by the pattern’s ability to participate in different syntactic-semantic structures. Consequently,

20 The latter construction (9) is somewhat more open, since it can sometimes also license other *-mine* types from a semantically coherent group of lexemes (e.g. *jändamine*, *pusimine*, *askeldamine*, *toimetamine*, all of which mean something similar to ‘grinding, working, struggling’) or replace the copula with another semantically general verb (e.g. *tulema* ‘come’, *hakkama* ‘begin’).

in attempts to explore whether a form is accessed as a whole word or parsed in lexical processing, one should ideally also consider the possible environments in which the form may occur. A low-frequency derivation from a high-frequency base might be used only in a specific constructional setting and the construction itself processed as a whole, whereas a high-frequency derivation from an otherwise low-frequency base could be used in highly diverging functions. In the latter case, the exact meaning of the form could be derived in different ways: by parsing the form in a specific context using lexical procedural knowledge; by parsing the construction; or by accessing either or both from the lexical memory.

5.2.2. Correlation between **-mine** *and* **-us** *derivations*

Finally, I explore the potential rivalry of *-mine* and *-us* nouns. It has been suggested that in the case of rivalry (i.e. being used for the same functions), two suffixes tend to choose their bases from complementary domains (e.g. van Marle 1985). Consequently, this means that when a pair of *-mine* and *-us* nouns derived from the same base are equally frequent, it is expected that they fulfill different functions (for example, they have different meanings). Kasik (2015: 187) also notes that lexicalized derivations containing less productive action noun suffixes such as *-us* can be detected in comparison with the corresponding *-mine* derivations, although the latter can also sometimes lexicalize. Based on the assumptions underlying the productivity values described in previous sections, it is also expected that lower frequency derivations represent the pattern's regular use, yielding semantically transparent formations. Therefore, the low-frequency spectrum would be an ideal place to find near-synonymous, rival *-us* and *-mine* derivations.

Figure 13 presents the distribution of bases which occur at least once with both suffixes. The axes have been scaled using base 2 log transformation. This means that the distances between the data points are not equal on the plot, and differences between lower frequencies are highlighted. Not all lexemes are plotted for the sake of visual clarity, but the distribution of data points can be seen in the background. The darker the color of the text, the more similar the frequencies of *-mine* and *-us* nouns (e.g. *lahendamine* 'solving' and *lahendus* 'solution; solving') and the closer the base is to the dashed diagonal *xy*-line.

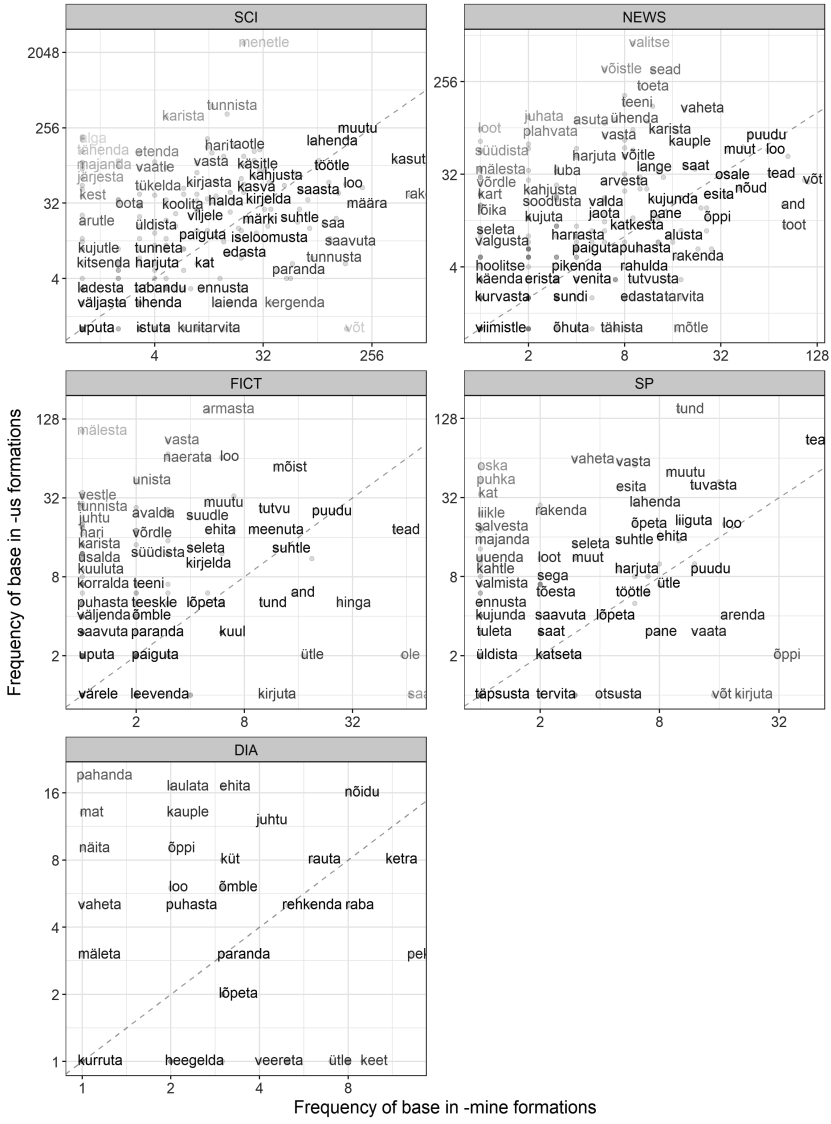

Figure 13. Comparison of the frequencies of *-us* and *-mine* nouns derived from the same base.

The shared bases are typically more frequent in the less productively formed *-us* derivations. This robustly illustrates the complementary relationship between rival suffixes: when a base is very frequent in one morphological construction, it is less likely to be used in another, where

it might potentially compete for the same slot in a sentence. The fact that *-us* nouns are more productively derived from bases which have undergone verbal derivation (e.g. ending in e.g. *-ta*/*-da*, *-le*, or *-tse*) consequently suggests that *-mine* nouns from those stems are less likely to be found.

Figure 13 shows that the common bases do indeed mostly end with sequences matching the above-mentioned verbal affixes. In fact, when digging into all possible *-us* bases in the dataset (not only those which also occur in *-mine* derivations), the bases which end with either of the four abovementioned sequences (*-ta*, *-da*, *-le*, or *-(t)se*) constitute a clear majority, with *-ta* being the base ending for over 50% of the *-us* types in each register (Figure 14). *-ta* is also the most popular base ending among *-mine* nouns, but it only occupies around 20% of all items in each register. Other base structures are distributed more evenly. This is the pattern also seen for verbs in general in the maximal subcorpus samples. Figure 14 presents the proportion of base ending sequences among *-mine* types, *-us* types and verb types in five registers. While the structural distribution of *-mine* type bases closely mirrors that of verbal bases, the distribution of *-us* type bases is visibly more extreme, since there are more structural restrictions, which means that available base structures occupy a larger proportion of type bases. Of course, not all bases ending with those frequent sequences can be analysed as formally decomposable bases, at least not synchronically. For example, while *-ta*/*-da* is a highly productive verbal suffix and occurs in various suffix combinations (*-sta*, *-nda*, *-rda*, *-lda*), *vaata-* 'look' is an old base, developed through the shortening of the base **valvata-*[21]. In turn, *tuvasta-* 'recognize, determine', is an artificial base whose spread is attributed to an accident: the originally suggested *turvasta-* as an Estonian counterpart for the German word *feststellen* was written down incorrectly and became used as such in legal language (Erelt, Erelt & Ross 2007). However, from the perspective of deverbal noun derivation, internal compositionality of the base seems to be of little relevance. Instead, it is the base syllable structure and the character sequence at the end of the base that seem to matter for *-us* derivation.

21 https://www.eki.ee/dict/ety/index.cgi?Q=valvama (Accessed 03.04.2021.)

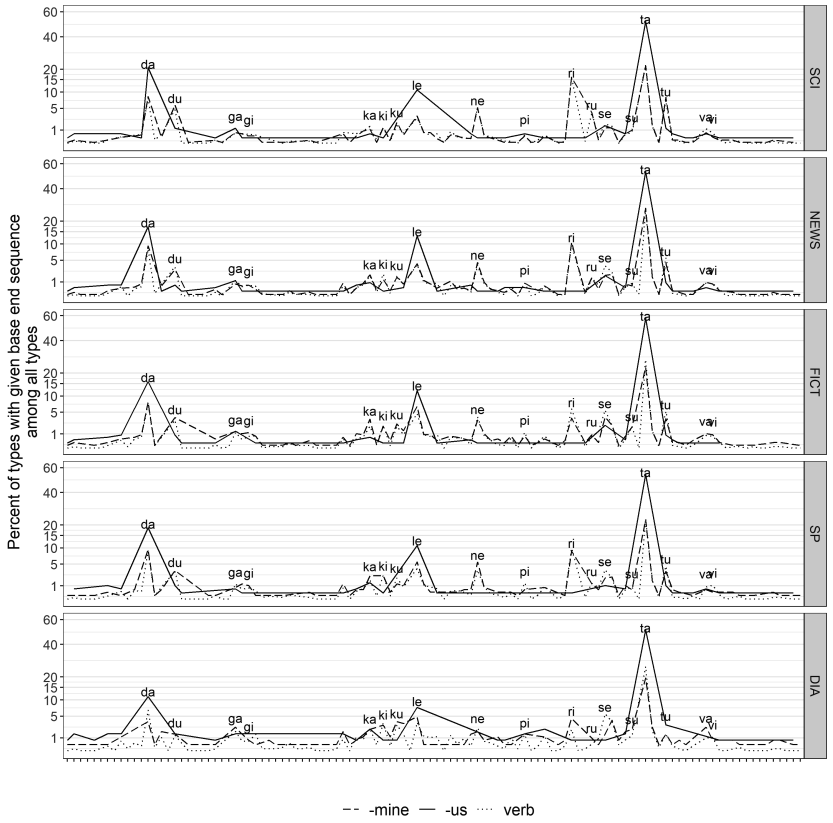

Figure 14. The two last characters of bases in *-mine* derivations, *-us* derivations and verbs.

When analysing the potential meanings of the *-us* and *-mine* derivations formed from the bases in Figure 13, it becomes apparent that without context, a great deal of them could, in principle, be used in the same function. This is particularly obvious for the more frequent base endings. While bare *-us* nouns do seem to present a single, bounded instance of a process and *-mine* nouns an unbounded process, the action noun reading is still present with both suffixes, and the aspectual (un)boundedness of both derivations can be changed with the help of other linguistic elements (e.g. temporal adverbials for a temporal reading of *-us* nouns and pluralization for the bounded instance reading of *-mine* nouns). With regard to the prototypical meanings, however, most of the words can be considered semantically different. The *-us* derivations

from transitive bases very easily lend themselves to interpretations as objects or results of the corresponding actions (e.g. *oskus* 'skill' vs. *oskamine* 'being able to do smth.', *rakendus* 'application' vs. *rakendamine* 'applying', *karistus* 'punishment' vs. *karistamine* 'punishing', *ennustus* 'prediction' vs. *ennustamine* 'predicting'). Still, there are also formations with a higher degree of idiosyncratic lexicalization among *-us* nouns. The items where parallel readings are not possible are mainly cases where either the different suffix allomorphs are used for *-us* derivation or where the base stem is in weak gradation in *-us* derivations and strong gradation in *-mine* derivations. Such pairs share the least common semantic characteristics, since the *-us* nouns are always used in a semantically specialized meaning that differs from that of the action noun (e.g. *a**nn**us* 'dose' and *a**nd**mine* 'giving', *peksa**ndus*** 'grain from threshing' and *peksmine* 'beating; fighting', *kee**d**us* '(boiled) food' and *kee**t**mine* 'boiling', *tu**nn**us* 'characteristic, marker, trait' and *tu**nd**mine* 'feeling; knowing', *hari**dus*** '(an) education' and *harimine* 'educating; cultivating (land)', *sega**dus*** 'mess' and *segamine* 'mixing; stirring', *saa**dus*** 'product; produce' and *saamine* 'getting'). There are also synchronically completely lexicalized instances containing the *-us* suffix (compare *katus* 'roof' and *katmine* 'covering', *matus* 'funeral' and *matmine* 'burying').

With regard to the relative frequency between *-mine* and *-us* nouns, it seems that regular semantic correspondences (even with aspectual differences) are more likely to be found in the low frequency range for both *-mine* and *-us* derivations. For example, *täpsustus* and *täpsustamine* 'specifying, specification', *arendus* and *arendamine* 'development', *väljastus* and *väljastamine* 'issuing', *tihendus* and *tihendamine* 'sealing; thickening', *hoolitsus* and *hoolitsemine* 'taking care', *eristus* and *eristamine* 'differentiating', *istutus* and *istutamine* 'planting', *edastus* and *edastamine* 'transmission, transmitting', *katsetus* and *katsetamine* 'testing', *parandus* and *parandamine* 'fixing'. Near-synonymy, or at least a higher degree of semantic relatedness, should be more apparent in the low-frequency range, since this is where the use of both constructions is more likely to operationalize a productive pattern. However, on the one hand, not all low-frequency derivation pairs are necessarily semantically similar. On the other hand, regular correspondences can also be found in the higher frequency range, especially when the base frequencies are more or less equal in both derivations (e.g. *kasutus* and

kasutamine ‘using; usage’, *töötlus* and *töötlemine* ‘processing’, *osalus* and *osalemine* ‘participating; participation’, *suhtlus* and *suhtlemine* ‘communicating; communication’, *rautus* and *rautamine* ‘shoeing (of a horse)’). Moreover, as mentioned in the previous paragraph, it is difficult to see why systematically deriving objects and results with the action noun schema should not be considered as instantiating a productive construction through established polysemic links. Therefore, if there was a novel base *pohista-* meaning ‘make funny shapes out of foam’, then we could use the derivation *pohistus* ‘funny shape made out of foam’ without it having to be interpreted as an action noun first. This, in turn, means that the low-frequency type range need not contain only action noun readings and consequently, that semantic comparison with *-mine* nouns having the action noun reading is not a completely reliable way to detect lexicalized *-us* nouns.

To conclude, it is not clear that it is specifically similar frequency which helps to detect *-mine* and *-us* nouns with diverging semantics. In fact, one could make the following two propositions about pairs of derivations formed from the same base: one, that they are typically used for different functions, and two, that despite this, they may evoke similar readings depending on the syntactic context. Those derivations with similar semantics are more likely to be found in the lower frequency spectrum as low frequency promotes higher regularity of derivation patterns. However, this current analysis is based on a subjective assessment of simple lemma frequencies with no context provided. Therefore, the semantic analysis is and can only be rather superficial. Including information about case forms and the syntactic context of the derivations, combining corpus linguistics with experimental techniques, or applying the methods of distributional semantics (see Shen & Baayen 2021) could provide a more adequate account of the relationship between rival affixes in Estonian.

6. Summary and discussion of the results

In this article, I have examined how well the frequency-based quantitative measures of morphological productivity capture the aspects of productivity of the three most frequent deverbal suffixes in Estonian across different registers. The underlying assumption of this kind of

analysis is that corpus frequencies are an adequate enough reflection of how complex words are used and processed in natural linguistic communication settings within the language community.

Using data from five different registers confirmed that there is indeed a distinction between the written and spoken registers with regard to both the raw number of different derivations used as well as the extent to which the morphologial constructions can be extended to include more formations not yet attested in the samples. However, this distinction is also apparent from the degree of (in)formality characteristic of the communication types represented by the registers: written fiction aligns with the more formal written registers (media texts and scientific texts) in the higher extent to which it uses the derivations, while it is more similar to spoken registers in its higher potential to include new formations. The different productivity measures applied in this study, therefore, seem to highlight different aspects of productivity. The realized productivity of a pattern, expressed by the type count of that derivation pattern, reflects first and foremost the pragmatic usefulness of a morphological category in different registers. This is the measure linked to the different natures of spoken and written communication that are manifested *inter alia* by the distribution of the part-of-speech classes in the corresponding corpora. In turn, potential productivity, which assesses the proportion of types occurring only once among all tokens representing a derivation pattern, highlights the structural and semantic generality of a pattern, i.e. the likelihood of the morphological construction to be instantiated by new items. This is where the informal registers appear to license a more productive use of the derivation patterns, while the formal written registers seem to exhaust the list of available bases rather quickly. Since this also seems to apply for a derivation pattern with a theoretically infinite number of available bases, such as *-mine*, it is again pragmatic factors which enforce constraints on the list of potential bases in formal registers. Finally, expanding productivity, which assesses the proportion of types formed with a given suffix among all hapaxes in the corpus, is connected to both pragmatic and paradigmatic relations, namely the likelihood of the derivation pattern to be chosen for expressing any new concept, given the pragmatic need and all alternative ways of saying the same thing. Again, fiction is the register where all derivation patterns are the most profitable for the creation of new concepts. The analysis in this article presupposes

uniformity within registers, i.e. that the language within each register is produced in similar topical and situational context and conforms to similar participant relations, purpose, and production circumstances. This is naturally a great simplification. Even a strict division between the linguistic characteristics of speech and writing would be impossible as both generalize over several situational (and processing) constraints and a variety of communicative tasks (see Biber, 1988: 45). However, I do believe that there is more uniformity within than across registers, which is why registers are seen as macro-categories here and taken as predefined by the corpora.

With regard to ranking the 3 affixes by their potential productivity, the three approaches generally gave similar results. With a few minor exceptions, *-mine* was ranked as the most productive suffix in all registers, followed by *-ja* and *-us*. This in itself is not surprising, since *-mine* has been considered a borderline case between inflection and derivation. However, in addition to fiction, the *-us* pattern was used somewhat more productively in the spoken registers, especially in dialects. This might result from the use of more diverse vocabulary among the hapaxes, but the results also suggest that in dialects, the pattern is indeed likely to instantiate types that are not used in other registers. *-ja*, on the other hand, was used less productively in scientific texts. This results from the fact that scientific texts mostly use *-ja* nouns in some fixed meanings (professions, instruments, certain specific participant roles) and less for actively identifying an actor or experiencer by their action or state.

The results concerning the productivity of the suffixes, while intuitively plausible, must be interpreted with some caution. As Gaeta and Ricca (2006: 68) themselves point out, the variable-corpus approach can undervalue the productivity of very frequent suffixes (like *-mine* and *-us*) when estimated at a low number of suffix tokens. So while Baayen's original approach is criticized for overestimating the productivity of less frequent suffixes, the approaches of both Gaeta and Ricca (2003, 2006) and Plag, Dalton-Puffer & Baayen (1999) can essentially do the same thing, especially when the samples are small. It remains, then, an open question as to what would be an adequate method to compare the productivity of very frequent and very infrequent suffixes. It seems that although Gaeta and Ricca's variable-corpus approach is a useful way of examining the growth curves of different suffixes, it lacks some statistical groundedness. With affixes of significantly different

token counts, the expanding productivity measure anchored in the total samples would be preferable. As shown, however, expanding productivity can be quite unstable in small samples. Therefore, the issue is also linked to what is perhaps the most significant pitfall of the current analysis, namely the extremely small samples used in this study. As noted vigorously in the corpus linguistic literature, this becomes a theoretical problem when using hapaxes for assessing the proportion of rare words among a group of derivations. Nowhere in the literature are hapaxes equated with neologisms, but it has been shown that the greatest number of neologisms do appear among hapax legomena, and that the more a corpus size is increased, the more the number of hapaxes starts to approximate the actual number of neologisms in the corpus (Baayen & Renouf 1996, Plag, Dalton-Puffer & Baayen 1999). In small samples, however, some words can occur only once but still be familiar to the language user without being productively formed neologisms or even simply rare formations. Indeed, most of the hapaxes in the samples used in this study are established formations likely to occur either in a large dictionary or in an average speaker's vocabulary. For example, in the NEWS sample, hapaxes include formations like *küsimine* 'asking', *unustamine* 'forgetting', *ennustamine* 'predicting', all of which are known, regular formations and unlikely to surprise anybody in a discourse (see Baayen & Lieber 1991). This means that such formations might still be stored in the mental lexicon, unlike true neologisms created out of a pragmatic need (Baayen 1994: 453, Gaeta & Ricca 2006: 68). This makes the measures of potential and expanding productivity unreliable for assessing the extensibility of different derivation patterns, i.e. their ability to include new formations. Let it be stressed that the small samples used in this study were not the result of a deliberate choice, but an inevitability. We currently lack larger, accessible, morphologically annotated corpora for contemporary spoken Estonian. It is therefore highly desirable that the analyses be repeated on a considerably larger dataset. Even though the results obtained in this study seem intuitively coherent, the situation could be different with other, less common affixes. That being said, this particular study's use of hapaxes as substitutes for rare formations has provided a useful and accessible heuristic which seems to work fairly well for estimating how well the empirical data correlates with intuitions about productivity. If we relieve ourselves of the demand that hapaxes represent neologisms

in their pure form, created only by a one-time pragmatic demand and never found in a dictionary, and instead treat them as an approximation of the patterns that speakers are more likely to find transparent, regular, and useful in a given register, then even studies conducted on very small samples can provide interesting insights.

Examining the correlations between base frequencies and derivation frequencies showed the highest correlation for the suffix *-mine*, perhaps indicating a more regular correspondence between the meaning of the base and the meaning of the derivation. For all suffixes, high-frequency bases which are very infrequent in derivations are subject to some structural, semantic or pragmatic restrictions, while low-frequency bases with high-frequency derivations are more likely to have specialized meanings. When both are frequent, the base tends to have general semantics, which is specified in context by the inclusion of complements. In the case of *-us*, however, most derivations have a specialized meaning, different or metonymically extended from that of the base. Base frequency in the sense of *verbal stem frequency* cannot, in general, be used as a strong predictor of whether a derivation is more likely to be processed as a whole word or parsed into component meanings. Indeed, recent evidence has suggested that both can occur at the same time. Furthermore, the fact that there are some frequent derivations which never occur as verbs (at least in the samples) suggests that although the word *derivation* implies that verbs are somehow more primary than deverbal nouns, if only from the perspective of word formation, such a hierarchical relationship between base and derivation is unlikely to exist in actual language use. In the case of Estonian, this has already been pointed out by e.g. Vare (1991) and Erelt et al. (1995: 479).

Finally, the comparison of *-us* and *-mine* derivations formed from the same base did not reveal a clear functional distinction based on simple stem frequencies. The *-mine* and *-us* formations which share similar semantics are more likely to be found in the low-frequency range. In most cases, however, the prototypical meanings of *-mine* and *-us* nouns do diverge, irrespective of the correlation between their frequencies. If an *-us* noun does provide multiple interpretations, including one as an action noun, it is context that helps to resolve the actual meaning: there are several linguistic devices which can reinforce the action noun reading, even if the primary, most prototypical meaning of the noun is something else.

Some further remarks should be made with regard to the methodological aspects of this study. In this article, only the simple stems of derivations were analyzed and all compound structures were stripped to their minimal base forms. It has been suggested, however, that also including the outer cycles of word-formation would considerably change the outcome of the analysis (Gaeta & Ricca 2006). Considering complex formations would significantly increase the hapax count while leaving the token count the same. Therefore, this approach would have a more drastic effect on the potential productivity measure *P* and would considerably raise the productivity of those suffixes which, though they are the least available for expanding their basic category, are still able to participate in numerous compound structures (e.g. *-us*). In this article, the analysis of compound structures was disregarded for two main reasons. First, the structure and annotation principles in the corpora used in this study are not homogenous, and a structure written as a compound in one register (e.g. *ära+minemine* in NEWS) might be written differently in another (e.g. *ära minemine* in DIA). Second, it is not clear whether the quantitative measures would reflect the productivity of the derivation pattern, the productivity of compounding, or the ability to retain the base verb's argument structure. This would make it considerably harder to compare different suffixes adequately in different registers.

In the analysis, I have not disregarded the creative or playful formations as irrelevant for the study of morphological productivity as discussed by many researchers (see an overview in Bauer 2001). First of all, it would be impossible to determine which of the formations in the corpora were coined intentionally and which were not. Comparing the types against a large dictionary would not be particularly helpful, as dictionaries do not include a large number of regular derivations. It is also important to remember that dictionaries mediate the linguistic knowledge of an "ideal speaker" through some linguistically trained individual speakers, i.e. the editors of the dictionaries. Secondly, even if playful extensions of a morphological process could be adequately detected and disregarded as not contributing to the derivation pattern's actual productivity, the proportion of "playful" formations in this small sample of only 426,000 tokens per register, where only derivations from simple bases are analyzed, is low enough for them not to have a significant effect on the productivity of the patterns. Thirdly and most importantly, from a usage-based perspective, I do not think such intentional

formations should be disregarded as not contributing to the productivity of a morphological pattern. On the contrary, being able to extend a pattern, construction, process or a rule to include formations never encountered before but still completely understandable, even if intended as wordplay, is exactly what makes a pattern productive. It can produce new words because there exists a transparent analogy (Barðdal 2008: 172–173) strong enough to help the receiver to decode the formation's meaning.

I have also included forms which could synchronically be considered simplexes with no eventive semantics (e.g. *katus* 'roof') but diachronically have been derived from the respective verb stems (*kat-* 'cover'). Such formations "might induce the activation of the respective suffixes, thus influencing their availability in the mental lexicon" as Gaeta and Ricca (2006: 75) point out, although they exclude such formations themselves. Gaeta and Ricca (2006: 75) also exclude lexicalized items, which from a synchronic perspective cannot be semantically related to the base verb form because they doubt that the use of those words would really activate the suffix. However, Hay and Baayen (2005: 342) take into account evidence from psycholinguistic experiments and argue that "morphological structure is inherently probabilistic, experience always leaves traces in memory irrespective of irregularity, and the meanings of complex words can be affected in subtle ways by similarity." As such, an activation effect can never be ruled out when using corpus data and therefore, lexicalized items are also included in this study. Another, more practical reason for doing so is the difficulty of drawing the line between lexicalized and unlexicalized items, especially without looking at their context. In turn, going through the contexts of the over 66,000 derivations used in this study would be needlessly time-consuming considering the possible benefits for this particular analysis.

To sum up, morphological and syntactic productivity is a fascinating research subject, and no quantitative measure can gauge the full complexity of this diverse and constantly changing system. However, there are aspects of productivity that do reveal themselves through frequencies and correlations, and these call for an empirical investigation in Estonian linguistics as well. There are multiple prospects for further research, especially with respect to recent methodological advances and the constantly growing body of data accessible to a linguist. For example, it is possible to analyse samples of equal size

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 analysing 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, like 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 enable us to factor in more variables in studying the productivity of one or more patterns. Artificial neural networks, such as Word2Vec (Mikolov et al. 2013) could be used for assessing the (de) compositionality of a given derivation pattern, examine whether potentially rival suffixes occur in similar contexts, or detect polysemous derivations, given a sufficiently large amount of high quality data. 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, although register comparison would be more difficult. In this article, morphological productivity was analyzed as a function of suffix and register. However, the productivity of affixes in natural language use is affected by many more linguistic and extralinguistic factors. For example, Keune, van Hout and Baayen (2006) compared different statistical modeling techniques to investigate morphological productivity in spoken Dutch as a function of socio-geographic forces (country, sex, education level, and age). Finally, in addition to the stylistic, communicative, and socio-geographic aspects of derivational patterns, the role of individual preferences in productivity should also be considered. These have been shown to be anything but trivial (see De Smet 2020).

7. Conclusion

In this article, I examined the morphological productivity of three Estonian deverbal suffixes, *-mine*, *-us* and *-ja*, in five registers of Estonian (scientific texts, newspaper texts, fiction, spoken spontaneous language, and spoken regional dialects). Using corpus data and the

quantitative measures *realized productivity*, *potential productivity* and *expanding productivity* developed by Baayen and his collaborators (Baayen 1989, 1992, 1993, 2001, 2003, Baayen & Renouf 1996, Plag, Dalton-Puffer & Baayen 1999), I showed that the derivation patterns demonstrate varying degrees of productivity in different registers. The general distinction between written and spoken registers was apparent only on the level of realized productivity, while other aspects of productivity seemed to be conditioned more by the distinction between formal and informal communicative settings as well as by the pragmatic usefulness of the suffixes in different registers. Fiction is the register which seems to facilitate the most productive use of the derivation patterns in all aspects of productivity: it makes frequent use of *-mine*, *-us*, and *-ja* nouns, is not limited by a high proportion of lexicalized formations, and is likely to include new, previously unencountered formations.

Baayen's original procedure of assessing potential productivity at the total number of suffix tokens was compared against the modification of that approach proposed by Gaeta and Ricca (2003, 2006), where the productivity of all suffixes was calculated at an equal number of suffix tokens. While this approach provided some more intuitive results, it suffered from underestimating the productivity of the more frequent suffixes *-mine* and *-us* in the small sample used in this study. However, the general ranking of the suffixes conforms well to Baayen's expanding productivity measure in most registers: *-mine* as the bordeline case between inflection and derivation shows by far the highest productivity, while *-ja* and *-us* (in that ranking order) are more similar in terms of the likelihood of instantiating new concepts. The relatively low productivity of *-us* is related both to the numerous structural restrictions that govern its formation as well as to its semantic vagueness. The productivity of *-ja*, in turn, is restricted by the semantics of the base and the category's overall lower pragmatic usefulness in regular derivations.

The comparison of base and derivation frequencies suggested that while there are significant correlations between the two, base frequency cannot be used as an accurate and reliable estimation of the derivation's regularity, although the association is stronger for *-mine* derivations. Comparing the frequencies of *-mine* and *-us* derivations, in turn, indicated that while most *-mine* and *-us* nouns occur in different semantic functions, irrespective of their frequency, the two are more likely to show semantic similarity in the low-frequency range.

Although the samples used in this study are very small, I have shown that even with a small sample, a quantitative account of productivity enables us to empirically validate a linguist's intuitions. Additionally, I have demonstrated that productivity is not a strictly global property or a general rule, but is an interplay between structural, semantic, paradigmatic and pragmatic restrictions which are also conditioned by the register and the communicative situations which the register presents.

Acknowledgements

This study was supported by the European Union through the European Regional Development Fund (Centre of Excellence in Estonian Studies). I am also deeply grateful to the anonymous reviewers and to professor Harald Baayen for their invaluable comments.

Abbreviations

ADE – adessive, ALL – allative, COM – comitative, ELA – elative, NMLZ – nominalization, PL – plural, PRT – partitive, PST – past tense, SG – singular

References

Aronoff, Mark. 1976. *Word formation in generative grammar*. Cambridge, MA: The MIT Press.

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

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. https://doi.org/10.1515/9783110213881.2.899.

Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Revista Brasileira de Linguística Aplicada* 11(2). 295–328. http://doi.org/10.1590/S1984-63982011000200003.

Baayen, R. Harald & Rochelle Lieber. 1991. Productivity and English derivation: a corpus-based study. *Linguistics* 29. 801–834. http://doi.org/10.1515/ling.1991.29.5.801.

Baayen, R. Harald & Antoinette Renouf. 1996. Chronicling The Times: productive lexical innovations in an English newspaper. *Language* 72(1). 69–96. http://doi.org/10.2307/416794.

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. http://doi.org/10.1037/a0023851.

Baayen, R. Harald, Yu-Ying Chuang, Elnaz Shafaei-Bajestan & James P. Blevins. 2019. The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de)composition but in linear discriminative learning. *Complexity* 2019. 1–39. https://doi.org/10.1155/2019/4895891.

Baker, Paul, Andrew Hardie & Tony McEnery. 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.

Bauer, Laurie. 2001. *Morphological productivity*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511486210.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511621024.

Biber, Douglas. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519871.

Booij, Geert. 2010. *Construction morphology*. Oxford: Oxford University Press.

Booij, Geert & Rochelle Lieber. 2004. On the pragmatic nature of affixal semantics in English and Dutch. *Linguistics* 42(2). 327–357. https://doi.org/10.1515/ling.2004.011.

Bybee, Joan. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780195301571.001.0001.

De Smet, Hendrik. 2020. What predicts productivity? Theory meets individuals. *Cognitive Linguistics* 31(2). 251–278. https://doi.org/10.1515/cog-2019-0026.

Denistia, Karlina & R. Harald Baayen. 2019. The Indonesian prefixes *PE-* and *PEN-*: A study in productivity and allomorphy. *Morphology* 29. 385–407. https://doi.org/10.1007/s11525-019-09340-7.

Diessel, Holger. 2019. *The grammar network: How linguistic structure is shaped by language use*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108671040.

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.

Erelt, Mati, Tiiu Erelt & Kristiina Ross. 2007. *Eesti keele käsiraamat*. Tallinn: Eesti Keele Sihtasutus.

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.

Fonteyn, Lauren & Stefan Hartmann. 2016. Usage-based perspectives on diachronic morphology: A mixed-methods approach towards English *ing*-nominals. *Linguistics Vanguard* 2(1). 1–12. https://doi.org/10.1515/lingvan-2016-0057.

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. https://doi.org/10.1515/LING.2006.003.

Hardie, Andrew & Tony McEnery. 2006. Statistics. In Keith Brown (ed.), *Encyclopaedia of language and linguistics*, 138–146. Oxford: Elsevier. https://doi.org/10.1016/B0-08-044854-2/04759-3.

Hay, Jennifer. 2003. *Causes and consequences of word structure*. New York: Routledge. https://doi.org/10.4324/9780203495131.

Hay, Jennifer & R. Harald Baayen. 2002. Parsing and productivity. In Geert Booij & Jaap van Marle (eds.), *Yearbook of Morphology 2001*, 203–235. Dordrecht: Kluwer. http://doi.org/10.1007/978-94-017-3726-5_8.

Hay, Jennifer & R. Harald Baayen. 2005. Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences* 9(7). 342–348. https://doi.org/10.1016/j.tics.2005.04.002.

Hothorn, Torsten, Peter Buehlmann, Sandrine Dudoit, Annette Molinaro & Mark van der Laan. 2006. Survival ensembles. *Biostatistics* 7(3). 355–373. https://doi.org/10.1093/biostatistics/kxj011.

Kasik, Reet. 1975. Verbide ja verbaalsubstantiivide tuletusvahekorrad tänapäeva eesti keeles. *Keele modelleerimise probleeme* 5. 4–162. Tartu: Tartu Riiklik Ülikool.

Kasik, Reet. 2004. *Eesti keele sõnatuletus*, 2., parandatud trükk. Tartu: Tartu Ülikooli Kirjastus.

Kasik, Reet. 2006. Nominalisatsioon meediauudiste 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. 2014. Eesti sõnamoodustus ja süntaks. *Keel ja Kirjandus* 2. 100–111.

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.

Kerge, Krista. 2002. Kirjakeele kasutusvaldkondade süntaktiline keerukus. In Reet Kasik (ed.), *Tekstid ja taustad: artikleid tekstianalüüsist* (Tartu Ülikooli eesti keele õppetooli toimetised 23), 29–46. Tartu: Tartu Ülikooli Kirjastus.

Kerge, Krista. 2003. *Keele variatiivsus ja mine-tuletus allkeelte süntaktilise keerukuse tegurina* (Tallinna Pedagoogikaülikool. Humanitaarteaduste dissertatsioonid 10). Tallinn: TPÜ kirjastus.

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.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar. Vol. 1. Theoretical prerequisites*. 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". *CogniTex*tes 19. https://doi.org/10.4000/cognitextes.1616.

Mangiafico, Salvatore. 2020. *rcompanion: Functions to support extension education program evaluation*. R package version 2.3.26. https://CRAN.R-project.org/package=rcompanion.

Marle, Jaap van. 1985. *On the paradigmatic dimension of morphological creativity*. Dordrecht: Foris. https://doi.org/10.1515/9783111558387.

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.

Perek, Florent. 2015. *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives* (Constructional Approaches to Language 17). Amsterdam: John Benjamins. https://doi.org/10.1075/cal.17.

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

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. https://doi.org/10.12697/jeful.2017.8.2.10.

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

Plag, Ingo. 1999. *Morphological productivity: Structural constraints in English derivation*. Berlin, New York: Mouton de Gruyter. https://doi.org/10.1515/9783110802863.

Plag, Ingo, Christiane Dalton-Puffer & R. Harald Baayen. 1999. Morphological productivity across speech and writing. *English Language and Linguistics* 3(2). 209–228. https://doi.org/10.1017/S1360674399000222.

R Core Team. 2020. *R: A language and environment for statistical computing* (R Foundation for Statistical Computing). Austria: Vienna. https://www.R-project.org/.

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 konstruktsioonipõhises perspektiivis* (Tallinna Ülikool. Humanitaarteaduste dissertatsioonid 25). Tallinn: Tallinna Ülikool.

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*. https://doi.org/10.1515/cllt-2020-0059.

Vare, Silvi. 1991. Kazuto Matsumura ettekandest Debreceni kongressil ning verbi ja verbaalnoomeni vahekorrast. *Keel ja Kirjandus* 7. 408–415.

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. http://dx.doi.org/10.1201/9781315370279.

Zeldes, Amir. 2012. *Productivity in argument selection. From morphology to syntax* (Trends in Linguistics. Studies and Monographs 260). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110303919.

Kokkuvõte. Maarja-Liisa Pilvik: Eesti keele deverbaalsufiksite *-mine*, *-us* ja *-ja* produktiivsuse võrdlus viies registris: kvantitatiivne kasutuspõhine lähenemine. Artikkel annab empiirilise, kasutuspõhise ülevaate kolme eesti keele deverbaalsufiksi – *-mine*, *-us* ja *-ja* – produktiivsusest viies eesti keele registris. Produktiivsuse eri aspekte näitlikustavad kvantitatiivsed mõõdikud (Baayen 1989, 1992, 1993), mis on leidnud morfoloogilise produktiivsuse uurimustes laialdast kasutust. Analüüsi tulemused näitavad, et selleks uurimuses kasutatud valimi võrdlemisi väike maht mõjutab mõnevõrra mõõdikute tõlgendamist ning nende usaldusväärsust. Samas haakuvad kvantitatiivse analüüsi tulemused hästi keeleteadlase intuitsiooniga ning kasutatud morfoloogilise produktiivsuse mõõdikud võivad seega olla eri registrite empiiriliseks võrdlemiseks kasulikud isegi siis, kui valimid on korpuslingvistilises mõistes väikesed.

Märksõnad: nominalisatsioon, morfoloogia, produktiivsus, varieerumine registrites, korpuslingvistika, eesti keel