Valoda: nozīme un forma / Language: Meaning and Form 15, 2024, 138–151.
https://doi.org/10.22364/vnf.15.09

How lemmatisation and derivational annotation affect productivity measures: The case of deverbal agent nouns in the Joint Corpus of Lithuanian[1]

Kā lematizēšana un derivatīvā anotēšana ietekmē produktivitātes vērtēšanu: darītājvārdi Vienotajā lietuviešu valodas korpusā

Jurgis Pakerys

Institute for the Languages and Cultures of the Baltic
Department of Baltic Studies, Vilnius University
Universiteto St. 5, Vilnius, LT-01131, Lithuania
E-mail: *jurgis.pakerys@flf.vu.lt*

Virginijus Dadurkevičius

Institute of Digital Resources and Interdisciplinary Research
Vytautas Magnus University
V. Putvinskio St. 23–216, Kaunas, LT-44243, Lithuania
E-mail: *virginijus.dadurkevicius@vdu.lt*

Agnė Navickaitė-Klišauskienė

Institute for the Languages and Cultures of the Baltic
Department of Baltic Studies, Vilnius University
Universiteto St. 5, Vilnius LT-01131, Lithuania
E-mail: *agne.navickaite@flf.vu.lt*

We discuss the automatic and manual stages of the lemmatisation and annotation of the Joint Corpus of Lithuanian (1.3 billion words) used to measure derivational productivity. As a case study, we present data of three productive deverbal agent noun suffixes in Lithuanian, *-toj-*, *-ėj-*, *-ik-*, and measure their realized, expanding, and potential productivity. We show that

[1] This article is one of the outcomes of the project "Derivational productivity of Lithuanian suffixed nouns in the Joint Corpus of Lithuanian" funded by the Research Council of Lithuania (LMTLT), agreement No. S-LIP-22-61. The paper is based on our presentation at the 58th International Academic Conference in Honor of Professor Arturs Ozols: "Grammar and Word Formation" at the University of Latvia, Riga (17 March, 2023). We sincerely thank the audience of the conference and the anonymous reviewers of the article for their input and critical remarks. We are also grateful to Cristina Aggazzotti for editing the English of the article.

an additional semi-automatic lemmatisation and a manual derivational annotation significantly increase type and hapax counts. We also note that lemmatisation is affected by an artificially increased number of lemmas due to homographic forms unresolved by the lemmatiser. After the manual disambiguation of hapaxes, the numbers of feminine formations in *-toj-(a)* and *-ėj-(a)* were the most significantly reduced.

Keywords: word formation; derivational productivity; agent nouns; Lithuanian.

1. Introduction

The measures of derivational productivity in corpora have been widely applied since their introduction in the early 1990s, namely realized, expanding, and potential productivity (see overviews in Baayen 2009; Zeldes 2012, 48–95; Gaeta, Ricca 2015, 844–849; Dal, Namer 2016, 73–76). The reliability of these measures depends not only on the size and representativeness of the corpus but also on how the derived lexemes are filtered out. To this end, one usually needs a lemma list of the corpus, and the quality of representation of derivational processes attested in that corpus will depend on the principles of lemmatisation and the derivational annotation (Evert, Lüdeling 2001; Dal et al. 2008; Baayen 2009, 207).

In this paper, we discuss one case study in which a large corpus of 1.3 billion words was used to measure the derivational productivity of three Lithuanian deverbal agent noun suffixes. We aim to demonstrate that step-by-step improvements in lemmatisation and derivational annotation significantly affect the measures of productivity. We also discuss the obstacles that we were largely unable to overcome, such as the disambiguation of homographic forms.

Our paper is structured as follows. In Section 2, we discuss our corpus and lemmatiser, in Section 3, we present the measures of productivity – realized, expanding, and potential –, and in Section 4, we introduce the three productive Lithuanian deverbal agent noun suffixes. Section 5 covers the main results and their discussion: in 5.1 we present the measures obtained from the initial lemmatisation, while in 5.2 we discuss the processes and results of the additional lemmatisation and derivational annotation. Section 6 summarizes the main points of our study.

2. Our corpus and lemmatiser

We chose the Joint Corpus of Lithuanian for our study due to its open access lemma and word-form lists (Dadurkevičius 2020a; 2020b). The corpus has ca. 1.3 billion words and comprises three subcorpora, as shown in Table 1 (Dadurkevičius, Petrauskaitė 2020, 123–124). Despite its convenient open access word lists, the corpus is not accessible for immediate word search in context and one of the subcorpora, namely the collection of Lithuanian web texts, is currently not prepared for online access, but was kindly provided by the developers of the Vilnius University Machine Translation project (https://vertimas.vu.lt/) as a plain text file for the purposes of the present study.

Subcorpus	Online access	Tokens
Lithuanian web texts, Vilnius University, 2014	Not available	779,154,268
Legal documents, courtesy of the Office of the Seimas of the Republic of Lithuania, 2011	https://e-seimas.lrs.lt/portal/documentSearch/lt	443,114,936
Balanced corpus of modern Lithuanian, Vytautas Magnus University, 2008	http://tekstynas.vdu.lt/tekstynas/	112,575,876

Table 1. Subcorpora of the Joint Corpus of Lithuanian

The lemmatiser used to compile the lemma list of the Joint Corpus of Lithuanian functions based on the open-source Hunspell platform, using a set of inflectional rules and a dictionary (Dadurkevičius 2017). The latest version of the lemmatiser recognizes ca. 190,000 word stems (lemmas) but does not perform contextual disambiguation of the homographic forms and returns all possible grammatical analyses of a given form. As a result, the lemma numbers and their frequencies are inflated to some degree. The problem of the homographic forms is further discussed in Section 5.1. The lemmatiser also does not perform derivational analysis and has no word-guessing module.

3. Realized, expanding, and potential derivational productivity

In our study, we prepared corpus data for measuring three types of derivational productivity: realized, expanding, and potential.

Realized productivity reflects the number of derivationally analyzable lexemes with a given affix and informs us of the past productivity of the affix (Baayen 2009, 901–902, 904–905). This measure can be referred to as "lexicographic productivity" because one of the traditional methods of estimating productivity is based on counting derivatives with a given affix listed in a certain lexicographic source, and the lemma list of a given corpus is just one such source.

The main limitation of the realized productivity measure is that it does not distinguish between the old and well-established formations and the recently derived ones. As a result, it cannot be used when one wishes to estimate only the current productivity – the capacity of the affix to derive new formations now. To this end, productivity measures including hapax counts were proposed. Hapaxes are words occurring once in a given corpus and the idea of estimating derivational productivity by including hapax counts is based on the observation that new formations initially have low frequency, and a significant share of them can be found among the hapaxes (Baayen 2009, 902, 905–906).

Hapax counts are used to estimate expanding and potential productivity. Expanding productivity is measured by dividing the total number of derivationally analyzable hapaxes containing a given affix by the total number of all hapaxes in the corpus. This measure estimates the probability of encountering a new type (formation) with a given

affix as we sample the array of potentially new formations that are found in the corpus as hapaxes (Baayen 2009, 902, 905–906). The absolute counts of hapaxes containing a given affix can also be used when one compares competing derivational processes in the same corpus, just as in the case of our present study.

Potential productivity is measured by dividing the total number of hapaxes containing a given affix by the total number of tokens (= total frequency) of all derived lexemes containing that affix (Baayen 2009, 902, 906). It estimates the probability of finding new types (formations) occurring as hapaxes as we sample tokens with a given affix in the corpus. One problem with this measure is that sometimes the total frequencies of the derived lexemes vary so greatly that affixation processes with comparatively low token frequencies become overestimated (Van Marle 1992; Gaeta, Ricca 2006).

4. Productive deverbal agent noun suffixes in Lithuanian

For our study, we chose three suffixes of deverbal agent nouns that exhibit different productivity in modern Lithuanian. According to the major grammars, the most productive suffix is *-toj-*, followed by *-ėj-*, and then *-ik-* (Ulvydas 1965, 317–321; Ambrazas 1994, 104–106). The grammars estimate the productivity of these suffixes based on their use in new formations and, apparently, some counts of lemmas in the dictionaries. The productivity of some of these suffixes is also noted in the latest studies of neologisms (Murmulaitytė 2016; Vaskelienė 2017; Murmulaitytė 2021; Aleksaitė 2022). All three suffixes derive masculine and feminine nouns, and their gender corresponds to specific inflection classes: the citation form in the nominative singular ending *-as* is masculine, while the nominative singular endings *-a* and *-ė* are feminine, as seen in the examples below:

(1)	*vair-uo-ti*	*vairuo-toj-as*
	steering.wheel-VRB-INF	drive-AGN-NOM.SG(M)
		vairuo-toj-a
		drive-AGN-NOM.SG(F)
	‘drive’	‘driver’
(2)	*kep-ti*	*kep-ėj-as*
	bake-INF	bake-AGN-NOM.SG(M)
		kep-ėj-a
		bake-AGN-NOM.SG(F)
	‘bake’	‘baker’
(3)	*plauk-ti*	*plauk-ik-as*
	swim-INF	swim-AGN-NOM.SG(M)
		plauk-ik-ė
		swim-AGN-NOM.SG(F)
	‘swim’	‘swimmer’

The suffixes also show a very strong distribution trend with regard to the morphemic structure of the base verbs: derivations in *-toj-* usually take suffixal verbs as their

input, as in (1), while *-ėj-* and *-ik-* formations are derived from non-suffixal ("primary") verbs, as in (2) and (3) (Ulvydas 1965, 318–321; Ambrazas 1994, 105–106).

5. Original and additional lemmatisation

Our study comprised two stages: we first evaluated the agent nouns found in the fully automatically generated lemma list and then proceeded to semi-automatic lemmatisation with the additional procedures of derivational annotation and disambiguation of some homographic forms.

5.1. Original (automatic) lemmatisation

In our first step, we used the lemma lists compiled automatically with the help of the lemmatiser discussed in Section 2. We filtered out the lemmas in both the masculine and feminine citation forms, i.e., those ending in the nominative singular: *-toj-as*, *-toj-a*, *-ėj-as*, *-ėj-a*, *-ik-as*, and *-ik-ė*. Then, we manually reviewed the resulting lists to exclude the following:

(a) lemmas containing character sequences that only formally coincide with the chosen suffixes, and

(b) lemmas that are not suffixal formations (synchronically unanalyzable formations, formations with non-verbal bases, or units that are non-suffixal formations).

For (a), consider the following: *vėj-as* 'wind', *vaik-as* 'child' (character sequences *ėj* and *ik* are parts of the native roots), and *skeptik-as*, *-ė* 'skeptic' (sequence *ik* is etymologically an Ancient Greek suffix).

For (b), consider *švent-ik-as*, *-ė* 'priest, clergy-man/-woman' ← *švent-as*, *-a* 'holy' (deadjectival formation), *foto-mėg-ėj-as*, *-a* 'photo hobbyist' (formed via the addition of the combining form *foto-* 'photo-' to *mėg-ėj-as*, *-a* 'enthusiast, the one who likes smth.' which, in turn, is a deverbal formation based on *mėg-ti* 'like'), and *vain-ik-as* 'wreath' (the base is absent and the suffix can be segmented only etymologically, see Fraenkel 1962, 1182).

We also excluded occasional lemmas based on spelling errors, e.g., *variuotojas* 'the one who plates with copper' instead of *vairuotojas* 'driver'. In this case, the potential lemma was *variuo-toj-as* (← *variuo-ti* 'plate with copper'), but as it is very rare, we were suspicious and decided to check all tokens; we concluded that all of them were just spelling errors.

When the formations were synchronically analysable, we also manually added the bases.

The majority of formations had semantically unproblematic relations with their bases, but in some cases, the links were somewhat obscured due to idiomatisation of the derivatives. To simplify, we considered all lexemes as derived whenever their bases were available, and the semantic links were detectible to a varying degree. For example, *padav-ėj-as* '(restaurant) waiter' is derived from *paduo-ti* (past stem *padav-ė*

is taken as a base) ‘give, serve’, and the formation refers to ‘the one who gives, serves’. The meaning of the lexeme is restricted to ‘person professionally serving food and drinks’, but speakers of Lithuanian have no problem seeing a link between the base and the derivative. The agent noun *(teismo) tar-ėj-as*, *-a* ‘(court) counsellor’ is derived from *tar-ti* ‘say’, and it is likely that at least some speakers see a semantic link, i.e., ‘the one who speaks (and thus gives advice, etc.)’, cf. prefixal verb *pa-tar-ti* ‘advise’. A more complicated case is *pribuv-ėj-a* '(traditional, indigenous) midwife’ ← *pribū-ti* (past stem *pribuv-o* is taken as a base) ‘be present (for a longer time)’. The base verb in this meaning is very rare in the current use and the semantic link may not be evident (‘the one who is present at childbirth’).

We also noted that some of the formations in our lemma lists refer to instruments and not to animate agents, e.g., *prailgint-toj-as* ‘(cord) extender’ ← *prailgin-ti* ‘extend’, *pakrov-ėj-as* ‘(battery) charger’ ← *pakrau-ti* ‘load, charge’, *vilk-ik-as* ‘(cargo) truck’ ← *vilk-ti* ‘tow, pull’, etc. In some cases, the formations can be used in reference to both the instrument and the animate agent, e.g., *pritrauk-ėj-as* ‘(door) closer’ (device) and ‘attractor’ (human agent, e.g., as a raiser of funds, etc.) ← *pritrauk-ti* ‘pull up, attract’. Reviewing all formations and disambiguating between instrument and agent uses would be very time-consuming, so we decided to review only the hapaxes after the stage of additional lemmatisation (see discussion in Section 5.2).

The results of our review of the initial (automatic) lemmatisation and exclusion of non-derived units are presented in Table 2. Here and elsewhere, we refer to lemmas as “types” and to lemmas occurring once as “hapaxes”.

First, let us look at realized productivity (type counts). The manual review mostly affected the counts of *-ėj-*, and especially *-ik-*, formations, while the number of types in *-toj-* decreased less. The data are far from perfect (see notes on homographic forms and the improved results in Section 5.2), but it is worth noting that the ranking of suffixes according to realized productivity still corresponds to the one presented in the major grammars. It is tempting to compare the numbers of masculine and feminine formations (the aspect not discussed in the grammars), but caution is required. The problem is that the lemmatiser used in our study does not disambiguate between

Suffix	Before manual review		After manual review	
	Types	Hapaxes	Types	Hapaxes
-toj-as	632	11	627	9
-toj-a	547	11	543	9
-ėj-as	251	2	235	2
-ėj-a	258	3	201	2
-ik-as	293	2	88	0
-ik-ė	159	6	50	4

Table 2. Initial type and hapax counts before and after manual review, which excluded non-derived lexemes

	M	F	M	F	M	F
NOM SG	*-toj-as*	*-toj-a*	*-ėj-as*	*-ėj-a*	*-ik-as*	*-ik-ė*
GEN SG	*-toj-o*	*-toj-os*	*-ėj-o*	*-ėj-os*	*-ik-o*	*-ik-ės*
DAT SG	*-toj-ui*	*-toj-ai*	*-ėj-ui*	*-ėj-ai*	*-ik-ui*	*-ik-ei*
ACC SG	*-toj-ą*	*-toj-ą*	*-ėj-ą*	*-ėj-ą*	*-ik-ą*	*-ik-ę*
INS SG	*-toj-u*	*-toj-a*	*-ėj-u*	*-ėj-a*	*-ik-u*	*-ik-e*
LOC SG	*-toj-uj(e)*	*-toj-oj(e)*	*-ėj-uj(e)*	*-ėj-oj(e)*	*-ik-e*	*-ik-ėj(e)*
VOC SG	*-toj-au*	*-toj-a*	*-ėj-au*	*-ėj-a*	*-ik-e*	*-ik-e*
NOM PL	*-toj-ai*	*-toj-os*	*-ėj-ai*	*-ėj-os*	*-ik-ai*	*-ik-ės*
GEN PL	*-toj-ų*	*-toj-ų*	*-ėj-ų*	*-ėj-ų*	*-ik-ų*	*-iki-ų*
DAT PL	*-toj-am(s)*	*-toj-om(s)*	*-ėj-am(s)*	*-ėj-om(s)*	*-ik-am(s)*	*-ik-ėm(s)*
ACC PL	*-toj-us*	*-toj-as*	*-ėj-us*	*-ėj-as*	*-ik-us*	*-ik-es*
INS PL	*-toj-ais*	*-toj-om(is)*	*-ėj-ais*	*-ėj-om(is)*	*-ik-ais*	*-ik-ėm(is)*
LOC PL	*-toj-uos(e)*	*-toj-os(e)*	*-ėj-uos(e)*	*-ėj-os(e)*	*-ik-uos(e)*	*-ik-ės(e)*
VOC PL	*-toj-ai*	*-toj-os*	*-ėj-ai*	*-ėj-os*	*-ik-ai*	*-ik-ės*

Table 3. Paradigms of *-toj-*, *-ėj-*, and *-ik-* formations with greyed-out homographic cells shared by masculine and feminine derivatives

homographic forms of masculine and feminine formations, and the type counts are inflated to a degree we are currently unable to estimate. For example, let us say a corpus contains only forms of a certain derivative that ends in *-ėj-as*. These word-forms are morphologically interpretable either as the nominative singular of the masculine agent noun or the accusative plural of the feminine agent noun, as shown in Table 3. Our lemmatiser returns a result that includes two interpretations, and to be conservative, adds *two* lemmas to the list. The problem is that we do not know if adding the two lemmas was justified – the forms could have been only nominative singular and only one (masculine) lemma should have been added. We look forward to solutions in the future with the help of lemmatisers that have advanced disambiguation capabilities, but we also understand that a certain margin of error would still be unavoidable.

The problem of homography of our agent nouns is presented in Table 3 where paradigms of the *-toj-*, *-ėj-*, and *-ik-* formations are provided with greyed-out homographic cells: notably, the formations in *-ik-* have fewer homographic cells and their type counts should be less affected than those of *-toj-* and *-ėj-* (formations with the latter suffixes have the same number of homographic cells). The problem of homographic forms also affects the total frequencies needed for estimating potential productivity (see Section 5.2).

As for hapaxes, for a corpus of 1.3 billion words, their counts were too low, and we suspected that the initial lemmatisation missed significant numbers of rare words that were not only hapaxes, but also more frequent lexemes. This was a limitation of the inbuilt dictionary of the lemmatiser, so to resolve this issue, we performed an additional semi-automatic lemmatisation.

5.2. Additional (semi-automatic) lemmatisation

For the additional lemmatisation, we first went through automatized stages. The forms of the corpus were filtered out according to the template SUFFIX + ENDING. For example, for the suffix *-toj-as* (masculine), we filtered out the nominative singular *-toj-as*, genitive singular *-toj-o*, dative singular *-toj-ui*, etc., and for the suffix *-toj-a* (feminine), we picked out the corresponding forms *-toj-a*, *-toj-os*, *-toj-ai*, etc. These forms were later grouped into lemmas, and potential bases were provided when found in the dictionary of our lemmatiser. Grouping into (potential) lemmas was done purely on formal grounds, and we expected some of the formations to be constructed artificially (see example below). We reviewed the lemma lists manually to exclude the words mentioned in Section 5.1 as cases (a) and (b). In addition, we also had to delete items according to (c):

(c) lemmas erroneously listed (constructed) based on certain homographic forms.

For example, our corpus contains the hapax *adaptuotojo*, which is a past passive participle genitive singular definite form ('of the adapted') of the verb *adaptuoti* 'adapt', as seen from the wider context. This form, however, can also be interpreted as a genitive singular of the agent noun *adaptuo-toj-as* 'adapter' – the lemmatiser listed it and we had to delete this entry during review.

We also manually added the derivational bases missing in the dictionary of the lemmatiser and corrected a few cases when automatically provided bases were incorrect. For example, the lemma *atsimojuo-toj-as* 'the one who waves back (i.e., refuses to do smth.)' lacked the base *atsimojuo-ti* 'wave back'; the lemma *pakas-ėj-as* 'the one who buries smth. (in a metaphorical sense)' had the automatically added base *pakas-y-ti* 'scratch' instead of *pakas-ti* 'bury', etc.

The results of our manual review are presented in Table 4. Compared to the initial lemmatisation (Table 2 in Section 5.1), the counts increased significantly: both the type and the hapax counts now appear to be more realistic for a 1.3-billion-word corpus.

First, let us consider the type counts: just as in the case of the initial data, the ranking of the suffixes remains the same and corresponds to the one presented in the grammatical descriptions. If the gender of the formations is considered, one of our earlier problems remains – the inflated number of lemmas due to homographic forms of masculine and feminine nouns. Based on the number of homographic paradigm cells,

Suffix	Before manual review		After manual review	
	Types	Hapaxes	Types	Hapaxes
-toj-as	3,305	822	2,457	530
-toj-a	2,590	637	2,279	516
-ėj-as	2,351	642	687	131
-ėj-a	1,576	384	620	126
-ik-as	911	189	256	67
-ik-ė	857	274	89	26

Table 4. Additional lemmatisation: type and hapax count before and after manual review, which excluded non-derived lexemes

the numbers for formations in *-toj-* and *-ėj-* should be more affected than the ones for formations in *-ik-* because feminine and masculine formations in *-toj-* and *-ėj-* have more homographic forms than formations in *-ik-* do (see Table 3 in Section 5.1).

As a partial solution to this problem, we decided to manually review the contexts of use of all hapaxes because reviewing and disambiguating all homographic forms of non-hapaxes would be a very time-consuming manual task. For example, for the hapax *apčiupinėtojas*, which can be interpreted as either the nominative singular of the masculine agent noun or the accusative plural of the feminine agent noun, our lemmatiser added two lemmas to the list: *apčiupinė-toj-as* (masculine) and *apčiupinė-toj-a* (feminine) 'the one who feels/checks by touching' ← *apčiupinė-ti* 'feel/check by touching'. We looked at the context of the use of this particular form and determined that it was the nominative singular, see (4). As a result, the feminine noun *apčiupinėtoja* was marked as artificially constructed and was excluded from the count of hapaxes.

(4) *O premjeras atvirai džiaugiasi, kad*
apčiupinė-toj-as *mus* *giri-a*
feel.by.touching-AGN-NOM.SG 1PL.ACC praise-PRS.3
"And the prime minister is openly happy that the checker [= the inspecting agency] is praising us" (JCL, web texts, 2014)

Not all cases could be fully resolved due to the lack of wider context, especially in the case of the web text subcorpus, where in many cases, only one sentence was available for review. For instance, consider (5), where the form *švietike* may be the vocative singular of two nouns: the feminine *šviet-ik-ė* or the masculine *šviet-ik-as* 'the one who enlightens' (← *švies-ti* 'shine, enlighten'; past stem *šviet-ė* is taken as a base):

(5) *Ei-k* *tu* *taut-os* ***šviet-ik-e***
go-IMP.2SG 2SG.NOM nation-GEN.SG enlighten-AGN-VOC.SG
'Hey, you, the enlightener of the nation' (JCL, web texts, 2014)

The majority of homographic forms that could not be fully resolved as masculine or feminine appear to be masculine nouns used as generic terms, as in (6) where the genitive plural is the same for both genders. To be on the safe side, however, we marked such cases as unresolved, but one may also count them as potentially masculine.

(6) *Jau* *buv-o* *daug* ***"griežtin-toj-ų"***
already be-PST.3 many make.stricter-AGN-GEN.PL
'We already had many persons who make things [the rules, etc.] stricter' (JCL, web texts, 2014)

Our disambiguation results are presented in Table 5. We see that the counts of suffix *-ik-* were less affected, which is explained by the lower number of homographic forms. The counts of feminine agent noun hapaxes were most significantly reduced for the suffixes *-toj-* and *-ėj-*. First, this can be explained by a higher number of homographic forms. Then, it appears that the proportion of numbers of masculine and feminine hapaxes may depend on the overall expanding productivity: the greater the total

Suffix	Before disambiguation (M/F)	After disambiguation (M/F)	Unresolved (M/F)	After instruments are excluded
-toj-as	530	464	129	448
-toj-a	516	69		69
-ėj-as	131	113	30	100
-ėj-a	126	25		25
-ik-as	67	63	4	62
-ik-ė	26	23		23

Table 5. Hapax counts of agent nouns before and after manual disambiguation of homographic forms (masculine/feminine) and exclusion of instrument nouns

number of hapaxes, the bigger the difference between the masculine and feminine formations, such as relation 2.74 for *-ik-* (63 masculine hapaxes, 23 feminine hapaxes, 86 in total), 4.52 for *-ėj-* (113 masculine hapaxes, 25 feminine hapaxes, 138 in total), and 6.72 for *-toj-* (464 masculine hapaxes, 69 feminine hapaxes, 533 in total).

The manual disambiguation of hapaxes demonstrates that the real type counts for agent nouns should also be lower, especially for *-toj-*, *-ėj-*, and their feminine formations. We can now subtract the eliminated hapaxes from the current totals of types. To arrive at more precise figures, a full disambiguation of forms should be done because some lemmas were included in the lemma list based on the homographic forms that were not hapaxes.

As noted in Section 5.1, some formations in *-toj-as* and *-ėj-as* (i.e., only the masculine variants of the suffixes) are used to derive not only animate agents, but also instrument nouns. Separating agents from instruments by reviewing the tokens of all potential instrument formations with *-toj-as* and *-ėj-as* would be a very time-consuming task, and we decided to focus on the hapaxes once again. After another round of review to remove instruments, the hapax counts of *-toj-as* and *-ėj-as* decreased by 16 and 13 lemmas, respectively. This reduced the counts of the agent noun hapaxes of *-toj-as* to 448 and of *-ėj-as* to 100 (see Table 5).

Now let us consider the ranking of the suffixes according to the revised numbers of hapaxes and see how they reflect expanding productivity. If gender is considered, the ranking is, as follows: *-toj-as* (M) > *-ėj-as* (F) > *-toj-a* (F) > *-ik-as* (M) > *-ėj-a* (F) > *-ik-ė* (F). The feminine formations are apparently less productive due to the use of the masculine in generic contexts, but further research is needed. If differences according to gender are ignored, the overall ranking *-toj-* > *-ėj-* > *-ik-* corresponds to both the ranking presented in the major grammars and the ranking according to realized productivity. Outstanding productivity of *-toj-* formations is also noted in the studies of recent neologisms in Lithuanian (Murmulaitytė 2016, 6–7, 12–13, 16–18; 2021, 151–153, 155–157; Aleksaitė 2022, 57–58, 62). Some formations in both *-toj-* and *-ėj-* are also found among new coinages by Lithuanian fiction authors (Vaskelienė 2017, 5).

We should recall that the suffixes under consideration have a morphological distribution with respect to bases: *-toj-* is added to suffixal bases, while *-ėj-* and *-ik-* are

Suffix	Hapaxes	Total frequency	Potential productivity
-toj-as	448	6,520,044	0.0687
-toj-a	69	4,710,297	0.0146
-ėj-as	100	3,899,405	0.0256
-ėj-a	25	2,973,383	0.0084
-ik-as	62	205,327	0.3020
-ik-ė	23	21,306	1.0795

Table 6. Hapaxes, total frequencies, and potential productivity

added to non-suffixal bases, so a competition under equal conditions is seen only for *-ėj-* and *-ik-*, where *-ėj-* wins. The outstanding productivity (both expanding and, apparently, realized) of *-toj-* is due to the fact that the suffixed verbs are a productive type and outnumber the non-suffixed ones (Ulvydas 1971, 247), so the input array of potential bases of *-toj-* is simply larger than that of *-ėj-* and *-ik-*. It is interesting to note in this context that the expanding productivity of masculine agent nouns in *-ėj-as* is still higher than that of the feminine formations in *-toj-a*, apparently due to the above-mentioned tendency to derive more masculine agent nouns used as both generic terms and in reference to male agents.

Finally, we estimated potential productivity using the hapax counts after the manual disambiguation of masculine and feminine formations and the exclusion of the instrument nouns. The results are presented in Table 6. Again, we should keep in mind that actual total frequencies should be lower to some degree due to unresolved homographic forms, especially for *-toj-* and *-ėj-*. Despite this shortcoming, let us look at the results.

The suffix *-ik-* clearly stands out and this must be the result of the vast difference in total frequencies. As noted in Section 3, the potential productivity of formations with comparatively low total frequencies may be overestimated. This is especially evident in the case of masculine *-ik-as* and feminine *-ik-ė* formations where the difference in total frequencies is large, both between *-ik-as* and *-ik-ė* and in comparison, to formations with *-toj-* and *-ėj-*. If only the suffixes *-toj-* and *-ėj-* are considered, the ranking according to potential productivity is the same as the one according to expanding productivity: masculine *-toj-as* and *-ėj-as* followed by feminine *-toj-a* and *-ėj-a*.

Conclusions

To achieve reliable measures of derivational productivity in corpora, the lemmatisation principles need to be examined in detail. If the lemmatiser relies only on the inbuilt dictionary, a sizable portion of the formations in a large corpus may not be recognized and thus not lemmatised – this depends on the sheer size of the lemmatiser dictionary. In such cases, additional semi-automatic lemmatisation needs to be performed and forms of the (potential) derivatives must be filtered out according to the pattern

AFFIX + INFLECTION and grouped into lemmas. Then, the lemma lists need to be manually reviewed to exclude non-derived items. In our case, semi-automatic lemmatisation and review of the lemma lists produced significantly larger numbers of types and hapaxes.

Further, one needs to consider how homographic forms may influence the results of the lemmatisation. It is preferable to use lemmatisers with good disambiguation capabilities, but if no disambiguation is available, a feasible task is to perform a manual disambiguation of the hapaxes. In our case, the difference between non-disambiguated and disambiguated hapax counts was significant, especially for the feminine agent nouns in *-toj-(a)* and *-ėj-(a)*. A manual review of hapaxes can be used for annotating further aspects when the revision and annotation of more frequent formations is not possible. We used an additional annotation step to separate agent formations with the suffixes *-toj-as* and *-ėj-as* from instrument nouns with the same suffixes.

As a result of manual review, disambiguation, and additional annotation, our most reliable data were the hapax counts that reflect expanding productivity, which ranked our suffixes as follows: *-toj-as* (M) > *-ėj-as* (M) > *-toj-a* (F) > *-ik-as* (M) > *-ėj-a* (F) > *-ik-ė* (F). The same ranking is seen according to potential productivity measures for the suffixes *-toj-* and *-ėj-*, but the potential productivity of *-ik-as* and *-ik-ė* is overestimated due to comparatively low total frequencies of the formations. The ranking according to realized productivity differs: *-toj-as* (M) > *-toj-a* (F) > *-ėj-as* (M) > *-ėj-a* (F) > *-ik-as* (M) > *-ik-ė* (F). If gender is ignored, the ranking corresponds to the one presented in the major grammars.

Abbreviations

1	1st person
3	3rd person
ACC	accusative
AGN	agent noun
DAT	dative
F	feminine
GEN	genitive
INF	infinitive
INS	instrumental
JCL	The Joint Corpus of Lithuanian (see Sources below)
LOC	locative
M	masculine
NOM	nominative
PL	plural
PRS	present
PST	past
SG	singular
VOC	vocative
VRB	verbalizer

Sources

Dadurkevičius, Virginijus. 2020a. *Wordlist of lemmas from the Joint Corpus of Lithuanian*. CLARIN-LT digital library in the Republic of Lithuania. Available at: https://clarin.vdu.lt/xmlui/handle/20.500.11821/41

Dadurkevičius, Virginijus. 2020b. *Assessment data of the Dictionary of Modern Lithuanian versus Joint Corpora*. CLARIN-LT digital library in the Republic of Lithuania. Available at: https://clarin.vdu.lt/xmlui/handle/20.500.11821/36

References

Aleksaitė, Agnė. 2022. *Lietuvių kalbos naujažodžių daryba (2011–2019 m. Naujažodžių duomenyno pagrindu).* Daktaro disertacija. Vilnius: Lietuvių kalbos institutas. Available at: https://talpykla.elaba.lt/elaba-fedora/objects/elaba:132642831/datastreams/MAIN/content

Ambrazas, Vytautas (ed.). 1994. *Dabartinės lietuvių kalbos gramatika*. Vilnius: Mokslo ir enciklopedijų leidykla.

Baayen, Harald Rolf. 2009. Corpus linguistics in morphology: Morphological productivity. *Corpus Linguistics: An International Handbook*. 2. Lüdeling, Anke, Kytö, Merja (eds.). Berlin, New York: De Gruyter Mouton, 899–919. https://doi.org/10.1515/9783110213881.2

Dadurkevičius, Virginijus. 2017. Lietuvių kalbos morfologija atvirojo kodo "Hunspell" platformoje. *Bendrinė kalba.* 90, 1–17. Available at: https://journals.lki.lt/bendrinekalba/article/view/156

Dadurkevičius, Virginijus, Petrauskaitė, Rūta. 2020. Corpus-based methods for assessment of traditional dictionaries. *Human Language Technologies–The Baltic Perspective*. *Frontiers in Artificial Intelligence and Applications.* 328. Utka, Andrius, Vaičenonienė, Jurgita, Kovalevskaitė, Jolanta, Kalinauskaitė, Danguolė (eds.). Amsterdam: IOS Press, 123–126. https://doi.org/10.3233/FAIA200613

Dal, Georgette et al. 2008. Quelques préalables au calcul de la productivité des règles constructionnelles et premiers résultats. *Actes du premier Congrès mondial de linguistique française, Paris, 9–12 juillet 2008*. Durand, Jacques, Habert, Benoît, Laks, Bernard (eds.). Paris: Institut de Linguistique Française, 1587–1599. https://doi.org/10.1051/cmlf08184

Dal, Georgette, Namer, Fiammetta. 2016. Productivity. *The Cambridge Handbook of Morphology*. Hippisley, Andrew, Stump, Gregory (eds.). Cambridge: Cambridge University Press, 70–90. https://doi.org/10.1017/9781139814720.004

Evert, Stefan, Lüdeling, Anke. 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? *Proceedings of the Corpus Linguistics 2001 Conference*. Rayson, Paul, Wilson, Andrew, McEnery, Tony, Hardie, Andrew, Khoja, Shereen (eds.). Lancaster: Lancaster University, 167–175.

Fraenkel, Ernst. 1962. *Litauisches etymologisches Wörterbuch*. Heidelberg: Carl Winter.

Gaeta, Livio, Ricca, Davide. 2006. Productivity in Italian word formation: a variable-corpus approach. *Linguistics.* 44(1), 57–89. https://doi.org/10.1515/LING.2006.003

Gaeta, Livio, Ricca, Davide. 2015. Productivity. *Word-Formation: An International Handbook of the Languages of Europe*. 2. Müller, Peter O., Ohnheiser, Ingeborg, Olsen, Susan, Rainer, Franz (eds.). Berlin/Boston: De Gruyter Mouton, 842–858. https://doi.org/10.1515/9783110246278-003

Murmulaitytė, Daiva. 2016. Naujieji asmenų pavadinimai darybos ir semantiniu aspektu. *Lietuvių kalba.* 10, 1–22. https://doi.org/10.15388/LK.2016.22591

Murmulaitytė, Daiva. 2021. *Naujažodžių darybos ir morfemikos tyrimų perspektyvos (Lietuvių kalbos naujažodžių duomenyno atvejis)*. Vilnius: Lietuvių kalbos institutas. https://doi.org/10.35321/e-pub.16.naujadaros-tyrimu-perspektyvos

Ulvydas, Kazys (ed.). 1965. *Lietuvių kalbos gramatika*. 1. Vilnius: Mintis.

Ulvydas, Kazys (ed.). 1971. *Lietuvių kalbos gramatika*. 2. Vilnius: Mintis.

Van Marle, Jaap. 1992. The relationship between morphological productivity and frequency: a comment on Baayen's performance-oriented conception of morphological productivity. *Yearbook of Morphology 1991*. Booij, Geert, Van Marle, Jaap (eds.). Dordrecht: Kluwer, 151–163.

Vaskelienė, Jolanta. 2017. Lietuvių rašytojų naujadarų darybos ir semantikos ypatumai. *Bendrinė kalba*. 90, 1–30. Available at: http://www.bendrinekalba.lt/Straipsniai/90/Vaskeliene_BK_90_straipsnis.pdf

Zeldes, Amir. 2012. *Productivity in Argument Selection: From Morphology to Syntax*. Berlin/Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110303919

Kopsavilkums

Šajā rakstā aplūkoti Vienotā lietuviešu valodas tekstu korpusa (1,3 miljardi vārdu) automātiskās un manuālās lemmatizēšanas un marķēšanas posmi, pēc kuriem tiek vērtēta derivatīvā produktivitāte. Kā piemērs prezentēti dati par trim produktīviem lietuviešu valodas deverbālu lietvārdu piedēkļiem – *-toj-*, *-ėj-*, *-ik-* – un mērīta to realizētā, paplašināšanas un potenciālā produktivitāte. Autori cenšas parādīt, ka papildu pusautomātiskā lemmatizēšana un manuālā derivatīvā marķēšana ievērojami palielina gan lemmu, gan hapaksu skaitu. Tāpat atzīmēts, ka lemmatizēšanas procesu ietekmē mākslīgi palielināts lemmu skaits, kas rodas tādēļ, ka lemmatizators neatpazīst homogrāfiskas formas. Pēc manuālās hapaksu pārbaudes visbūtiskāk ir samazinājies sieviešu dzimtes lemmu ar *-toj-a* un *-ėj-a* skaits.

Atslēgvārdi: vārddarināšana; derivatīvā produktivitāte; darītājvārdi; lietuviešu valoda.