Fifth International Workshop on
Resources and Tools for Derivational Morphology
September 4 and 5, 2025, University of Fribourg, Switzerland

Jurgis Pakerys (Vilnius University)
Virginijus Dadurkevičius (Vytautas Magnus University)
Agnė Navickaitė-Klišauskienė (Vilnius University)

Additional lemmatization and measures of derivational productivity: The case of Lithuanian denominal suffixal nouns

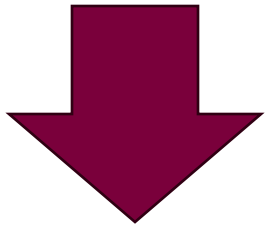
The project Derivational productivity of Lithuanian suffixed nouns in the Joint Corpus of Lithuanian has received funding from the Research Council of Lithuania (LMTLT), agreement No S-LIP-22-61

Outline

1. Measuring derivational productivity in corpora
2. Our aims, data, and methods
3. Results
    1. Quality and status nouns
    2. Personal nouns
    3. Diminutives
4. Discussion and conclusions
5. References

1. Measuring derivational productivity in corpora

Types, hapaxes, total frequencies

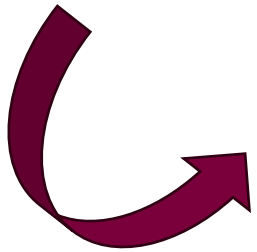


Realized, potential, expanding productivity

Baayen, 1992; Baayen, 1993; overviews in: Baayen, 2009; Gaeta and Ricca, 2015; Dal and Namer, 2016, among others

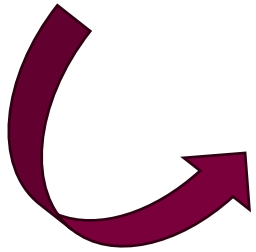
1. Measuring derivational productivity in corpora

Accurate lemmatization tools



Capture both frequent
and rare lexemes

Manual review



Exclude non-analyzable
lemmas and inner
derivational cycles

Evert et al., 2000; Evert and Lüdeling, 2001; Dal et al., 2008;
Gaeta and Ricca, 2006

2. Our aims, data, and methods

- **Expansion of the range of lemmas**
  that are not automatically lemmatized

- **Manual revision and exclusion**
  of non-analyzable lemmas and inner derivational cycles

- **Manual resolution of homographic forms**
  that may distort lemma frequencies

2. Our aims, data, and methods

Corpus

The Joint Corpus of Lithuanian, 1.3 billion tokens

Dadurkevičius, 2020a; Dadurkevičius, 2020b;
Dadurkevičius and Petrauskaitė, 2020

Lemmatization

Hunspell-type lemmatizer operating on the basis of a fixed dictionary and a set of inflectional rules

Dadurkevičius, 2017

2. Our aims, data, and methods

Additional semi-manual lemmatization

1. **Filtering** the forms according to the pattern **SUFFIX +** (all possible) **ENDINGS**
2. Automatic morphological annotation and **grouping** of the forms **into potential lemmas**
3. **Manual revision** and exclusion of derivationally non-transparent items and those with inner derivational cycles
4. **The revisions were mostly done by one annotator**, dubious cases were discussed and resolved by two annotators

3.1. Results: quality and status nouns

*saug-**um**-as* ‘safety’	←	Adj *saug-us* ‘safe’
*kantr-**yb**-ė* ‘patience’	←	Adj *kantr-us* ‘patient’
*krikščion-**yb**-ė* ‘Christianity’	←	N *krikščion-is* ‘Christian’
*nar-**yst**-ė* ‘membership’	←	N *nar-ys* ‘member’
*plokščiapad-**yst**-ė* ‘flat-footedness’	←	Adj *plokščiapad-is* ‘flat-footed’
*individual-**izm**-as* ‘individualism’	←	Adj *individual-us* ‘individual’
*kapital-**izm**-as* ‘capitalism’	←	N *kapital-as* ‘capital’

3.1. Results: quality and status nouns

Suffix	**Automatic** lemmatization		**Additional** lemmatization		Additional lemmatization and manual review			
	V	V_1	V	V_1	V	V_1	Tokens	$P \cdot 10^3$
-um-as	**2,518**	**68**	**22,895**	**9,222**	6,687	1,758	4,027,571	0.4365
-izm-as	**484**	**8**	**4,280**	**1,856**	772	214	7,845,016	0.0273
-yst-ė	**269**	**2**	**2,355**	**1,017**	1,136	365	826,408	0.4417
-yb-ė	**276**	**1**	**3,802**	**1,715**	1,054	332	196,472	1.6898

Table 1: Productivity data of quality and status noun suffixes
(V – types, V_1 – hapaxes, P – potential productivity)

3.1. Results: quality and status nouns

Suffix	Automatic lemmatization		**Additional** lemmatization		Additional lemmatization and **manual review**			
	V	V_1	V	V_1	V	V_1	Tokens	$P \cdot 10^3$
-um-as	2,518	68	**22,895**	**9,222**	**6,687**	**1,758**	4,027,571	0.4365
-izm-as	484	8	**4,280**	**1,856**	**772**	**214**	7,845,016	0.0273
-yst-ė	269	2	**2,355**	**1,017**	**1,136**	**365**	826,408	0.4417
-yb-ė	276	1	**3,802**	**1,715**	**1,054**	**332**	196,472	1.6898

Table 1: Productivity data of quality and status noun suffixes
(V – types, V_1 – hapaxes, P – potential productivity)

3.2. Results: Personal nouns

-as (m.), *-ė* (f.)

*men-**inink**-as, -ė* ‘artist’	←	N *men-as* ‘art’
*blaiv-**inink**-as, -ė* ‘abstainer’	←	Adj *blaiv-us* ‘sober’
*jaun-**uol**-is*, *-ė* ‘young person’	←	Adj *jaun-as* ‘young’
*turt-**uol**-is*, *-ė* ‘wealthy person’	←	N *turt-as* ‘wealth’
*gitar-**ist**-as*, *-ė* ‘guitar player’	←	N *gitar-a* ‘guitar’
*real-**ist**-as, -ė* ‘realist’	←	Adj *real-us* ‘real’

3.2. Results: Personal nouns

Suffix	**Automatic** lemmatization		**Additional** lemmatization		Additional lemmatization and manual review			
	V	V_1	V	V_1	V	V_1	Tokens	$P \cdot 10^3$
-inink-as	**760**	**18**	**7,412**	**3,138**	2,657	693	4,461,517	0.1553
-inink-ė	**421**	**31**	**1,599**	**613**	700	138	481,476	0.2866
-ist-as	**268**	**1**	**6,665**	**2,980**	996	341	638,651	0.5339
-ist-ė	**151**	**3**	**1,527**	**683**	254	69	75,879	0.9093
-uol-is	**97**	**0**	**1,244**	**545**	271	67	281,323	0.2382
-uol-ė	**85**	**1**	**804**	**342**	155	34	115,766	0.2937

Table 2: Productivity data of personal noun suffixes
(V – types, V_1 – hapaxes, P – potential productivity)

3.2. Results: Personal nouns

Suffix	Automatic lemmatization		**Additional** lemmatization		Additional lemmatization and **manual review**			
	V	V_1	V	V_1	V	V_1	Tokens	$P \cdot 10^3$
-inink-as	760	18	**7,412**	**3,138**	**2,657**	**693**	4,461,517	0.1553
-inink-ė	421	31	**1,599**	**613**	**700**	**138**	481,476	0.2866
-ist-as	268	1	**6,665**	**2,980**	**996**	**341**	638,651	0.5339
-ist-ė	151	3	**1,527**	**683**	**254**	**69**	75,879	0.9093
-uol-is	97	0	**1,244**	**545**	**271**	**67**	281,323	0.2382
-uol-ė	85	1	**804**	**342**	**155**	**34**	115,766	0.2937

Table 2: Productivity data of personal noun suffixes
(V – types, V_1 – hapaxes, P – potential productivity)

3.2. Results: Personal nouns

- Frequencies of some agent nouns (m./f.) may be significantly distorted due to homographic forms (Pakerys et al., 2024)

- Personal nouns (m./f.) also have some homographic forms

- Manual disambiguation of hapaxes with regard to gender:

-inink-as, -ė	1 f. → m., 4 ambiguous
-ist-as, -ė	3 f. → m., 2 ambiguous
-uol-is, -ė	1 f. → m., 2 m. → f., 7 ambiguous

3.3. Results: Diminutives

*vaik-**el**-is* ‘small child’ (m.) ← *vaik-as* ‘child’ (m.)
*rank-**el**-ė* ‘small hand’ (f.) ← *rank-a* ‘hand’ (f.)

*ežer-**ėl**-is* ‘small lake’ (m.) ← *ežer-as* ‘lake’ (m.)
*parduotuv-**ėl**-ė* ‘small shop’ (f.) ← *parduotuv-ė* ‘shop’ (f.)

*rat-**uk**-as* ‘small wheel’ (m.) ← *rat-as* ‘wheel’ (m.)
*kavin-**uk**-ė* ‘small cafe’ (f.) ← *kavin-ė* ‘cafe’ (f.)

3.3. Results: Diminutives

Suffix	**Automatic** lemmatization		**Additional** lemmatization		Additional lemmatization and manual review			
	V	V_1	V	V_1	V	V_1	Tokens	$P \cdot 10^3$
-(i)uk-as	**1,182**	**7**	**15,646**	**6,530**	3,598	950	687,339	1.3821
-(i)uk-ė	**147**	**0**	**3,807**	**1,521**	686	209	24,372	8.5754
-ėl-is	**483**	**3**	**6,467**	**2,500**	1,900	566	376,211	1.5045
-ėl-ė	**302**	**2**	**4,813**	**1,785**	1,026	257	203,848	1.2607
-el-is	**481**	**2**	**14,364**	**5,885**	755	74	1,354,266	0.0546
-el-ė	**319**	**5**	**6,862**	**2,673**	645	65	1,098,362	0.0592

Table 3: Productivity data of diminutive noun suffixes
(V – types, V_1 – hapaxes, P – potential productivity)

3.3. Results: Diminutives

Suffix	Automatic lemmatization		**Additional** lemmatization		Additional lemmatization and **manual review**			
	V	V_1	V	V_1	V	V_1	Tokens	$P \cdot 10^3$
-(i)uk-as	1,182	7	**15,646**	**6,530**	**3,598**	**950**	687,339	1.3821
-(i)uk-ė	147	0	**3,807**	**1,521**	**686**	**209**	24,372	8.5754
-ėl-is	483	3	**6,467**	**2,500**	**1,900**	**566**	376,211	1.5045
-ėl-ė	302	2	**4,813**	**1,785**	**1,026**	**257**	203,848	1.2607
-el-is	481	2	**14,364**	**5,885**	**755**	**74**	1,354,266	0.0546
-el-ė	319	5	**6,862**	**2,673**	**645**	**65**	1,098,362	0.0592

Table 3: Productivity data of diminutive noun suffixes
(V – types, V_1 – hapaxes, P – potential productivity)

3.3. Results: Diminutives

- Manual disambiguation of hapaxes with regard to gender:

-el-is, -ė	1 m. → f., 1 f. → m.
-ėl-is, -ė	1 f. → m., 10 ambiguous
-(i)uk-as, -ė	1 m. → f., 7 f. → m.

4. Discussion and conclusions

- **Additional lemmatization** based on simple search strings significantly **increases the counts of types and hapaxes**
- However, a substantial amount of random data is introduced, making **manual review essential**

- Lemma **frequencies** may be **affected by** non-disambiguated **homographic forms**
- Disambiguation of the homographic forms of hapaxes, however, did not lead to significant changes in the present study, but cf. Pakerys et al., 2024

Merci vielmal!

jurgis.pakerys@flf.vu.lt

5. References

Baayen, R. Harald (1992). Quantitative Aspects of Morphological Productivity. In: Booij, Geert E. & van Marle, Jaap (eds.), *Yearbook of Morphology 1991*. Dordrecht: Kluwer Academic Publishers,109–149 (https://doi.org/10.1007/978-94-011-2516-1_8).

Baayen, R. Harald (1993). On Frequency, Transparency, and Productivity. In: Booij, Geert E. & van Marle, Jaap (eds.), *Yearbook of Morphology 1992*. Dordrecht: Kluwer Academic Publishers,181–208 (https://doi.org/10.1007/978-94-017-3710-4_7).

Baayen, R. Harald (2009). Corpus linguistics in morphology: Morphological productivity. In: Lüdeling, Anke & Kytö, Merja (eds.), *Corpus Linguistics: An International Handbook,* vol. 2. Berlin, New York: Mouton de Gruyter, 899–919 (https://doi.org/10.1515/9783110213881.2.899).

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

Dadurkevičius, Virginijus (2020a). Wordlist of Lemmas from the Joint Corpus of Lithuanian. *CLARIN-LT digital library in the Republic of Lithuania* (http://hdl.handle.net/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* (https://clarin.vdu.lt/xmlui/handle/20.500.11821/36).

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

Dal, Georgette, Fradin, Bernard, Grabar, Natalia, Namer, Fiammetta, Lignon, Stéphanie, Plancq, Clément, Zweigenbaum, Pierre & Yvon, François (2008). Quelques préalables au calcul de la productivité des règles constructionnelles et premiers résultats. In: Durand, Jacques, Habert, Benoît & Laks, Bernard (eds.), *Actes du premier Congrès mondial de linguistique française, Paris, 9–12 juillet 2008*. Paris: Institut de Linguistique Française, 1587–1599 (https://doi.org/10.1051/cmlf08184).

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

Evert, Stephanie, Heid, Ulrich & Lüdeling, Anke (2000). On Measuring Morphological Productivity. In: Schukat-Talamazzini, Ernst Günter & Zühlke, Werner (eds.), *KONVENS-2000 Sprachkommunikation*, Ilmenau: VDE-Verlag, 57–61.

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

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

Pakerys, Jurgis, Dadurkevičius, Virginijus & Navickaitė-Klišauskienė, Agnė (2024). How lemmatisation and derivational annotation affect productivity measures: The case of deverbal agent nouns in the Joint Corpus of Lithuanian. *Valoda: nozīme un forma / Language: Meaning and Form*, 15, 138–151 (https://doi.org/10.22364/vnf.15.09).