The Biennial of Czech Linguistics, Data-based research in word formation
Charles University, Prague, September 20, 2024

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

How the measures of derivational productivity depend on lemmatization quality: the case of Lithuanian deverbal 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-22-6

Outline

1. Derivational productivity measures
2. Our corpus and lemmatizer
3. Lemmatization and productivity measures
   - Fully automatic: agent nouns
   - Semi-automatic: agent nouns
   - Other categories
4. Conclusions
5. References

1. Derivational productivity measures

- **Realized** productivity = lemmas with affix *X*
- **Expanding** productivity = hapaxes with affix *X* or:

$$\frac{\text{hapaxes with affix X}}{\text{total hapaxes}}$$

- **Potential** productivity = $\frac{\text{hapaxes with affix } X}{\text{total frequency of lemmas with affix } X}$

Baayen (1992; 1993); overviews: Baayen (2009); Gaeta & Ricca (2015: 844–849), Dal & Namer (2016: 73–76), etc.

1. Derivational productivity measures

- Our claim: in the case of Lithuanian, **these measures are heavily dependent on the quality of lemmatization** (including automatic and manual stages)

- Similar observations:
  - German: Evert & Lüdeling (2001)
  - Italian: Gaeta & Ricca (2006)
  - French: Dal et al. (2008)
  - ...
  - In general: Baayen (2009: 207)

2. Our corpus and lemmatizer

- The Joint Corpus of Lithuanian, **1.3 billion tokens**
  Dadurkevičius (2020a; 2020b), Dadurkevičius & Petrauskaitė (2020)

Subcorpus	Tokens
Lithuanian Internet content, 2014	779 M
Legal documents, Parliament of Lithuania, 2011	443 M
Balanced corpus of Vytautas Magnus University, 2008	113 M

- Lemmatized with Hunspell-based **lemmatizer**: **fixed dictionary + inflectional rules** (Dadurkevičius 2017)

3. Lemmatization & productivity measures

3.1. Fully automatic lemmatization

3.2. Additional semi-automatic lemmatization

- **Lithuanian deverbal suffixed nouns**: actions, agents, instruments, results/objects, places
- **Focus on the agent nouns** (3.1–3.2), brief notes on other categories (3.3)
- Only **realized and expanding productivity**

3.1 Fully automatic lemmatization

- Initial automatic lemmatization: **low type and hapax counts**
- For agent nouns, we still reviewed the data to determine how many non-transparent formations and accidentally included items have to be excluded
- For other categories, we omitted this stage and proceeded to additional semi-automatic lemmatization

Suffix	Before manual review		After manual review	
	V	V_1	V	V_1
-toj-as (m)	632	11	627	9
-toj-a (f)	547	11	543	9
-ėj-as (m)	251	2	235	2
-ėj-a (f)	258	3	201	2
-ik-as (m)	293	2	88	0
-ik-ė (f)	159	6	50	4

3.1 Fully automatic lemmatization

Agent nouns

V – types

V_1 – hapaxes

m – masculine

f – feminine

3.2 Additional semi-automatic lemmatization

- Tokens were automatically filtered according to pattern SUFFIX + ENDING and grouped into (potential) lemmas
- Lemma lists were manually reviewed, artificially constructed and derivationally non-transparent items were excluded
- Additional lemmatization significantly increased type and hapax counts

Suffix	Before manual review		After manual review	
	V	V_1	V	V_1
-toj-as (m)	3,305	822	2,456	532
-toj-a (f)	2,590	637	2,299	525
-ėj-as (m)	2,351	642	686	130
-ėj-a (f)	1,576	384	640	128
-ik-as (m)	911	189	254	65
-ik-ė (f)	857	274	92	27

3.2 Additional semi-automatic lemmatization

Agent nouns

V – types

V_1 – hapaxes

m – masculine

f – feminine

3.2 Additional semi-automatic lemmatization

- **Type and hapax counts were** found to be somewhat **inflated due to homographic forms** of masc./fem. nouns
- Morphologically ambiguous tokens yielded two lemmas due to the lack of morphological disambiguation stage during the lemmatization

- **We manually disambiguated the hapaxes**
- This step significantly reduced hapax counts of fem. formations in *-toj-a* and *-ėj-a* due to high number of their shared homographic forms

3.2 Additional semi-automatic lemmatization

Suffix	Before manual review		After manual review		
	V	V_1	V	V_1 (before m/f disam-bigua-tion)	V_1 (after m/f disam-bigua-tion)
-toj-as (m)	3,305	822	2,456	532	464
-toj-a (f)	2,590	**637**	2,299	**525**	**65**
-ėj-as (m)	2,351	642	686	130	112
-ėj-a (f)	1,576	**384**	640	**128**	**24**
-ik-as (m)	911	189	254	65	61
-ik-ė (f)	857	**274**	92	**27**	**20**

Agent nouns

V – types

V_1 – hapaxes

m – masculine

f – feminine

3.3 Other categories

Suffix	Initial automatic lemmatization		**Additional semi-automatic lemmatization**			
	No manual review		Before manual review		After manual review	
	V	V_1	V	V_1	V	V_1
-im-as(-is)	9,634	610	17,134	3,591	15,274	2,855
-ym-as(-is)	1,770	234	2,101	418	1,943	351
-acij-a	795	15	1,145	113	966	69
-es-ys	66	0	350	105	154	27

Action nominals

V – types

V_1 – hapaxes

3.3 Other categories

Suffix	Initial automatic lemmatization		**Additional semi-automatic lemmatization**			
	No manual review		Before manual review		After manual review (**cases of agent/instrument/ place nouns still not disambiguated**)	
	V	V_1	V	V_1	V	V_1
-tuv-as	192	3	592	127	488	87
-ikl-is	134	4	332	67	293	54
-atori-us	239	6	468	69	208	16
-tuv-ė	37	0	248	51	123	20

Instrument nouns

V – types

V_1 – hapaxes

Agent/instrument categories in 3.2 were also not disambiguated

3.3 Other categories

Suffix	Initial automatic lemmatization		**Additional semi-automatic lemmatization**			
	No manual review		Before manual review		After manual review	
	V	V_1	V	V_1	V	V_1
-in-ys	193	1	**837**	**143**	**230**	**16**
-al-as	260	5	618	126	124	31
-at-as	466	16	2,222	514	72	5

Object/result nouns

V – types

V_1 – hapaxes

3.3 Other categories

Place nouns

V – types

V_1 – hapaxes

Suffix	Initial automatic lemmatization		**Additional semi-automatic lemmatization**			
	No manual review		Before manual review		After manual review (and **disambiguation of place/instrument /celebration**)	
	V	V_1	V	V_1	V	V_1
-ykl-a	89	0	218	52	189	41
-tuv-ė	37	0	**248**	**51**	**33**	**7**

4. Conclusions

- The quality of lemmatization significantly influences the estimates of derivational productivity

- Lemmatization parameters are very important and should be listed in all corpus descriptions:

  - Built-in dictionaries, word-guessing modules, capabilities of solving morphological ambiguity, machine learning techniques, etc.

Thank you!

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,* vol. 328. Amsterdam: IOS Press, 123–126 (https://doi.org/10.3233/FAIA200613).

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, Stefan & 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).