Human Language Technologies – The Baltic Perspective
A. Utka et al. (Eds.)

doi:10.3233/FAIA200624

Morfio – A Corpus-Based Perspective on Latvian Morphology

Michal ŠKRABAL[a,1], Pavel VONDŘIČKA[a] and Václav CVRČEK[a]
[a]*Institute of the Czech National Corpus, Charles University, Czech Republic*

Abstract. Our paper introduces Morfio, a corpus-based online tool for the study of derivation and morphological productivity. Originally, Morfio was created for Czech, in this paper, however, we would like to introduce its Latvian implementation. Apart from the tool description, we want to showcase its possibilities for describing Latvian morphology by way of several examples.

Keywords. Morfio, morphological base, formants, alternations

1. Introduction

This demo is a follow-up on the [1] paper presented at the last Baltic HLT conference in Tartu. New corpora for the Baltic languages and tools for exploiting these corpora were introduced in the paper: namely the Latvian component [2] of the InterCorp parallel corpus [3], [4] and Araneum Lettonicum [5], as well as two tools based on these corpora: the translation equivalents database Treq [6] and a word-sketch grammar for Latvian [7]. The current paper presents Morfio, a new tool adjusted for Latvian (and, hopefully, also for other languages, including Lithuanian, in the future).

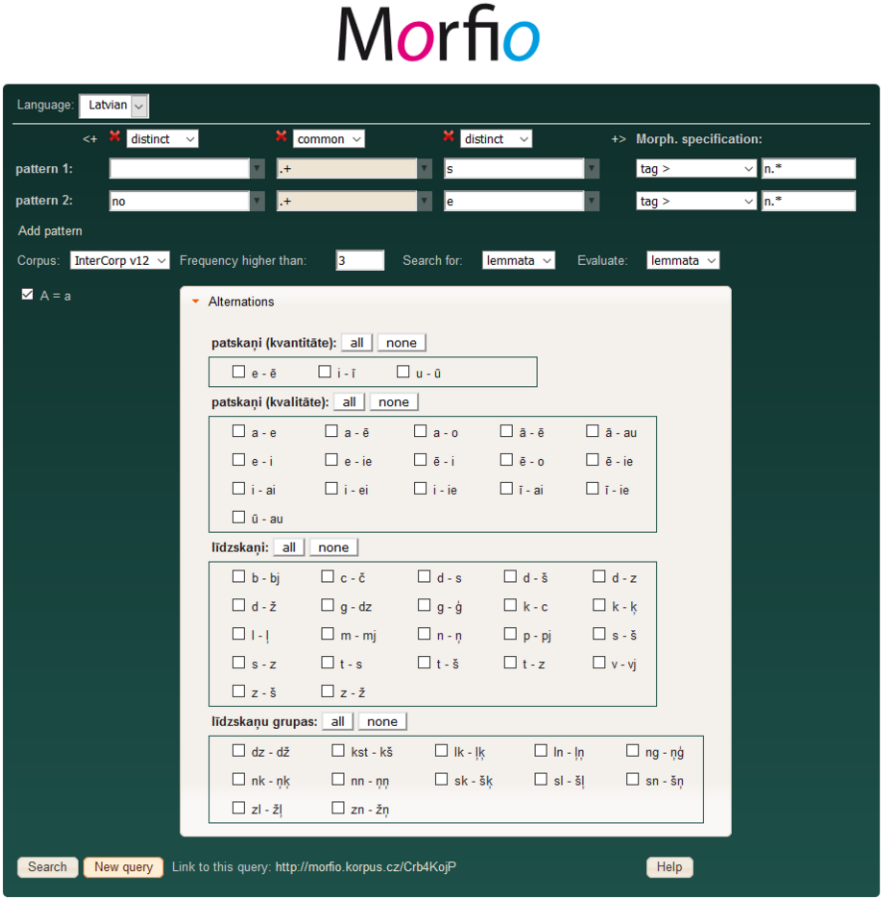
Morfio is a corpus-based online tool for the study of derivation and morphological productivity available within the Czech National Corpus portal www.korpus.cz. It can be used to identify pairs (or triplets/quadruplets) of words which follow the same derivational pattern. This pattern is specified by a user using regular expressions in two ways: 1) "common parts" or the derivational *base* (i.e. parts which are common for both words) and 2) "distinct parts" or the derivational *formants* in which they differ (e.g. ***darb****s* – *no****darb****e*, the bold parts are shared, while the non-bold parts signal the differences). The tool substitutes the common parts with a wild card and searches the corpus for word pairs that a) share the common parts and, at the same time, b) differ in the way specified by distinct parts (using the example introduced above: the *Xs* – *noXe* pattern). This results in a list of word pairs having the same derivational relation (*darbs* – *nodarbe*, *gals* – *nogale*, *jums* – *nojume*, *kalns* – *nokalne*, *kars* – *nokare*, *laids* – *nolaide*, *rīts* – *norīte*, *vakars* – *novakare*, *vietns* – *novietne*, *zars* – *nozare*)[2] with their absolute frequencies in the relevant corpus (also see Figure 2). Furthermore, Morfio estimates the productivity of each word-formation pattern according to an index proposed by [8] (see Figure 3).

[1] Corresponding Author: Michal Škrabal, Institute of the Czech National Corpus, Panská 890/7, Prague, Czech Republic; E-mail: michal.skrabal@ff.cuni.cz

[2] The results are accessible within the Morfio tool at http://morfio.korpus.cz/Crb4KojP.

When conducting a derivational research on a corpus which is not semantically annotated, we have to stick to the semasiological approach, i.e. proceeding from the form to function/meaning. This can pose several problems (besides potentially inaccurate morphological annotation and/or lemmatization, also homonymy) whose solutions are outside the scope of this tool and require a thorough manual analysis carried out by linguists (also see Section 4). However, tools such as Morfio can help a researcher by sifting through a large amount of corpus data and identifying potentially relevant candidates for further analysis.

Originally, Morfio was created for Czech [10], [11], yet nothing prevents it from extending its functionality to other languages,[3] including the Baltic ones. For a fully-fledged non-Czech version of the tool, we had to implement configurability for different tagsets [12] and add an inventory of relevant vocal and consonant alternations (according to [13], [14]).



Language: Latvian

	distinct	common	distinct	Morph. specification:	
pattern 1:		.+	s	tag >	n.*
pattern 2:	no	.+	e	tag >	n.*

Corpus: InterCorp v12 Frequency higher than: 3 Search for: lemmata Evaluate: lemmata

☑ A = a

patskaņi (kvantitāte):
□ e - ē □ i - ī □ u - ū

patskaņi (kvalitāte):
□ a - e □ a - ē □ a - o □ ā - ē □ ā - au
□ e - i □ e - ie □ ē - i □ ē - o □ ē - ie
□ i - ai □ i - ei □ i - ie □ ī - ai □ ī - ie
□ ū - au

līdzskaņi:
□ b - bj □ c - č □ d - s □ d - š □ d - z
□ d - ž □ g - dz □ g - ģ □ k - c □ k - ķ
□ l - ļ □ m - mj □ n - ņ □ p - pj □ s - š
□ s - z □ t - s □ t - š □ t - z □ v - vj
□ z - š □ z - ž

līdzskaņu grupas:
□ dz - dž □ kst - kš □ lk - ļķ □ ln - ļņ □ ng - ņģ
□ nk - ņķ □ nn - ņņ □ sk - šķ □ sl - šļ □ sn - šņ
□ zl - žļ □ zn - žņ

Link to this query: http://morfio.korpus.cz/Crb4KojP

Figure 1. Morfio's main menu, with the inventory of relevant morphological alternations for Latvian

[3] In fact, Morfio has already been successfully applied to the Polish part of InterCorp [15]. We chose Latvian next because it is a morphologically rich language, yet a non-Slavic one.

2. Morfio Interface

After entering a valid query in the form, Morfio provides four types of results which are organized in separate tabs: Summary, List, Productivity and Pattern 1 (2, 3, 4).

2.1. Summary Tab

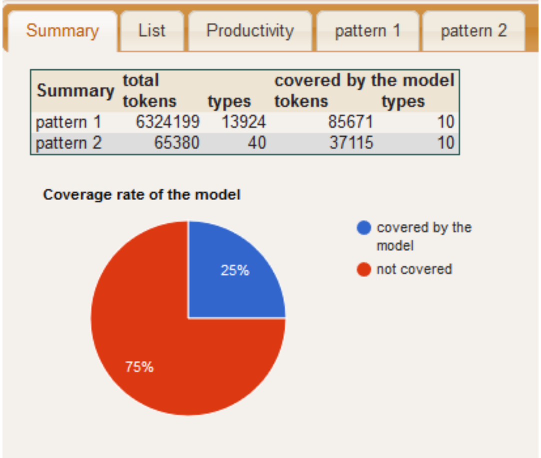
The Summary tab shows three types of information: number of types (with frequency above the limit specified by the user), sum of their occurrences, and an estimation of model completeness. One set of results (column "Total") refers to each of the isolated patterns itself, while the other set (column "Covered by the model") refers to those words following the given pattern which also fall into the analysed word-forming model, i.e. words for which a derivative counterpart was identified by the second pattern. The estimation of model completeness is based on the following assumptions:

a) Each pattern in the model identifies a certain number of items in the corpus (either wordforms or lemmas). We assume that most word-forming or derivational relations are asymmetric: there will be fewer words which are derived than those serving as derivation bases, as not every base produces a derivative. Thus, we can distinguish patterns that are basic in the model (those that include a larger number of items in the corpus) and those which are conditional (with a smaller number of items). In other words, the pattern which identifies a smaller set of word-types is considered a derivative of the pattern that identifies more word-types.

b) The completeness of the model is then calculated as the proportion of the total number of word pairs in the model to the total number of types of the conditional pattern, i.e. how many words in the less represented pattern find a derivative counterpart in the second pattern.

When inspecting our example model that can be characterized by a pair of words *darbs* – *nodarbe*, or by a pair of formants *-s* and *no-e* respectively, we can identify 10 lemma pairs involved in this word-forming process (see their list in Section 1 or in Figure 3). The first pattern (*Xs*) alone provides 13,924 different noun lemmas, while the second pattern (*noXe*) alone provides 40 different noun lemmas. Each of the patterns contains words that do not enter the model (e.g. the noun *vīrs* does not meet our requirements due to the non-existence of the noun **novīre*; similarly, we cannot find the noun **bīds* as a counterpart to the noun *nobīde*, etc.). The pattern with a smaller number of identified words (in our case, *noXe*) represents a greater limitation for the whole model than the pattern identifying more words (in our case, *Xs*). The condition for the existence of the word-forming relation specified by the example model is the existence of the noun ending with *-s*; however, the derivation process is limited mainly by the number of nouns following the pattern *noXe*, i.e. the pattern *noXe* is considered conditioned by the existence of the pattern *Xs*. In such a case, it makes sense to estimate the completeness of the proposed model according to how much the words of the conditional model contribute to it. In this case, the estimate is calculated as the ratio of the types of the pattern *noXe* that enter the model to all types of this pattern, i.e. 10/40, which corresponds to 25 %.

For word-formation analysis, it is obviously optimal if the coverage of the model is close to 100 %. This means that for all words in the conditional pattern, we have identified derivational bases in the other pattern. If such coverage cannot be achieved, the word-formation model explains only a part of the words formally defined by the conditional pattern; to relate the uncovered words with their bases, it is necessary to modify the existing model or create another, a complementary one.



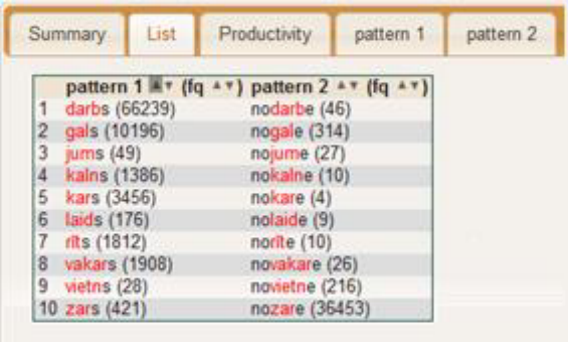
Summary	total		covered by the model	
	tokens	types	tokens	types
pattern 1	6324199	13924	85671	10
pattern 2	65380	40	37115	10

Figure 2. Summary tab

2.2. List Tab

The table in this tab lists all occurrences from all patterns that enter the specified model. The red part of the words indicates a common base (which may differ only if alternations are applied). The numbers in parentheses represent the total frequency of the lemma in the selected corpus. The table can be sorted according to any column using the arrows in the table header, both alphabetically and by frequency. At the same time, each word functions as a link to an example concordance in the selected corpus.

Pairs created only due to the application of alternation rules are highlighted by a coloured background (not shown in Figure 3). If more than one word corresponds to one pattern in a given pair (i.e. due to the application of alternations), all these words are listed collectively in one row of the table.



	pattern 1 ▲▼ (fq ▲▼)	pattern 2 ▲▼ (fq ▲▼)
1	darbs (66239)	nodarbe (46)
2	gals (10196)	nogale (314)
3	jums (49)	nojume (27)
4	kalns (1386)	nokalne (10)
5	kars (3456)	nokare (4)
6	laids (176)	nolaide (9)
7	rīts (1812)	norīte (10)
8	vakars (1908)	novakare (26)
9	vietns (28)	novietne (216)
10	zars (421)	nozare (36453)

Figure 3. List tab

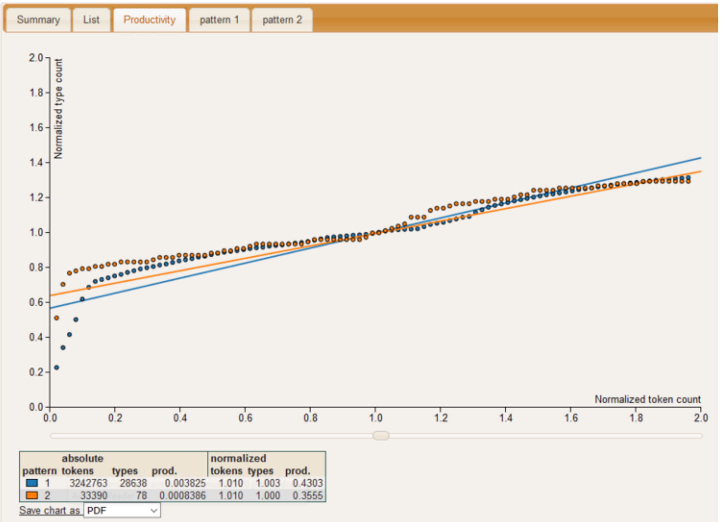
2.3. Productivity Tab

The estimation of the productivity of both patterns and their mutual comparison is based on Baayen's theoretical remarks [8]. Morphological productivity is measured by estimating the increment of new types with the growing number of tokens for each pattern separately. The comparison shows which pattern is more productive, because the number of its types grows faster as new words are being created using its formants and, on the contrary, which pattern is less productive or potentially closed (albeit frequented and large).

Productivity in this approach can be understood as the total probability of all types of a given pattern that are not represented in the corpus. If such a probability is high for a pattern after examining a certain number of occurrences, it means that the pattern is productive; otherwise the pattern seems to be relatively closed. The total probability of unrepresented types for a given pattern can generally be calculated using the Good-Turing estimate [9], as the number of hapaxes related to the total number of tokens. In our case, hapaxes are those types that occur exactly once in a given pattern. If we plot the data of increasing number of types with the growing number of tokens for a given pattern, this total probability will be, as a consequence of the construction of the Good-Turing estimate, the slope of its tangent at the last point.

However, to compare patterns that are of unequal size, type and token data must be normalized. The results shown in the graph (see Figure 4) are thus normalized on both axes, for the median value of tokens and types, respectively. This means that a value of 1 for the normalized number of tokens on the x-axis represents the median for tokens of a given pattern, and, similarly, a value of 1 for a normalized number of types on the y-axis corresponds to the median for the number of different words.

In order to compensate for the influence of the order of texts in the corpus, the data are shuffled several times. The number of random permutations of concordance lines within the corpus is variable and ranges from one shuffle to a maximum of ten repeated randomization cycles.



	absolute			normalized		
pattern	tokens	types	prod.	tokens	types	prod.
🟦 1	3242763	28638	0.003825	1.010	1.003	0.4303
🟧 2	33390	78	0.0008386	1.010	1.000	0.3555

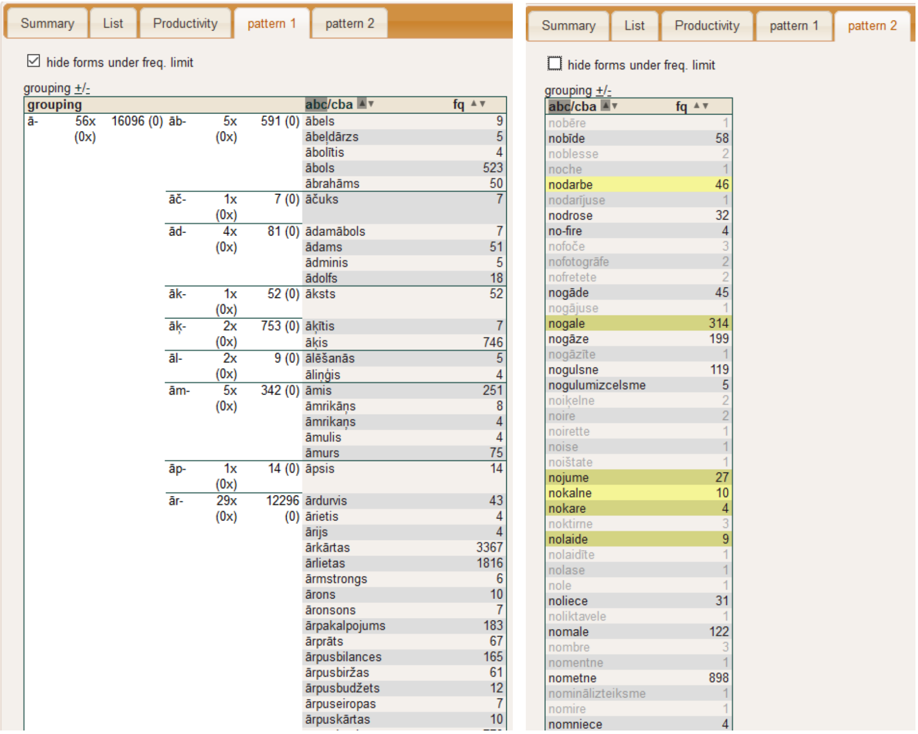
Figure 4. Productivity tab (Pattern 1 has a slightly higher slope and is, therefore, more productive than pattern 2.)

2.4. Patterns Tab

The words (wordforms or lemmas) corresponding to individual patterns are presented in the form of frequency lists in separate tabs. The list can also be supplemented with words that were not taken into account in the model, because their frequency was lower than the threshold set by the user. Data highlighted by a coloured background are involved in the word-forming model (i.e. there is a counterpart with the same base in the second pattern, differing only in formants).

The lists are mainly used to modify the model. If the list contains a word that that is not a part of the model (although it should be), it is signal for the user that it might be appropriate to change the model specification in order to increase its completeness and productivity.

The lists can be sorted in ascending or descending order, not only according to frequency, but also alphabetically, both commonly and retrogradely (i.e. from the end of the word). For better orientation in alphabetically sorted data, it is possible to turn on the grouping switch: the lines are then grouped according to the same start or end sequence of characters (see the left part of Figure 5). The number of grouping levels can be adjusted using the +/- element (each additional letter from the beginning or end, by which the words differ, can form another (sub)level for group division). For each group, data on the number of types and tokens appear, in total as well as those that participate in the word-formation model (shown in parentheses).



☑ hide forms under freq. limit

grouping +/-

grouping						abc/cba	fq
ā-	56x (0x)	16096 (0)	āb-	5x (0x)	591 (0)	ābels	9
						ābeļdārzs	5
						ābolītis	4
						ābols	523
						ābrahāms	50
			āč-	1x (0x)	7 (0)	āčuks	7
			ād-	4x (0x)	81 (0)	ādamābols	7
						ādams	51
						ādminis	5
						ādolfs	18
			āk-	1x (0x)	52 (0)	āksts	52
			āķ-	2x (0x)	753 (0)	āķītis	7
						āķis	746
			āl-	2x (0x)	9 (0)	ālēšanās	5
						āliņģis	4
			ām-	5x (0x)	342 (0)	āmis	251
						āmrikāņs	8
						āmrikaņs	4
						āmulis	4
						āmurs	75
			āp-	1x (0x)	14 (0)	āpsis	14
			ār-	29x (0x)	12296 (0)	ārdurvis	43
						ārietis	4
						ārijs	4
						ārkārtas	3367
						ārlietas	1816
						ārmstrongs	6
						ārons	10
						āronsons	7
						ārpakalpojums	183
						ārprāts	67
						ārpusbilances	165
						ārpusbiržas	61
						ārpusbudžets	12
						ārpuseiropas	7
						ārpuskārtas	10

□ hide forms under freq. limit

grouping +/-

abc/cba	fq
nobēre	1
nobīde	58
noblesse	2
noche	1
nodarbe	46
nodarījuse	1
nodrose	32
no-fire	4
nofoče	3
nofotogrāfe	2
nofretete	2
nogāde	45
nogājuse	1
nogale	314
nogāze	199
nogāzīte	1
nogulsne	119
nogulumizcelsme	5
noiķelne	2
noire	2
noirette	1
noise	1
noištate	1
nojume	27
nokalne	10
nokare	4
noktirne	3
nolaide	9
nolaidīte	1
nolase	1
nole	1
noliece	31
noliktavele	1
nomale	122
nombre	3
nomentne	1
nometne	898
nominālizteiksme	1
nomire	1
nomniece	4

Figure 5. Pattern 1 and Pattern 2 tabs

3. Demo (Morfio-based Queries)

In the HLT Baltic demo session we aim to present the use of Morfio for the study of Latvian morphology. We use Morfio with the data from Araneum Lettonicum corpus [5] (over 671 M tokens)[4] to extract words involved in the following derivational models:

- prefixes/circumfixes: we are looking for triplets of 1) non-prefixed (*X*), 2) prefixed (*saX*) and 3) both prefixed and reflective verbs (*saXies*) with the same stem (lemmas, minimum frequency 5) – 504 types: *adīt – saadīt – saadīties … žņaugt – sažņaugt – sažņaugties*;[5]
- prefixoids: e.g. non-substantive lemmas *X* × *pašX* – 273 types: *aizdedzināties – pašaizdedzināties … zīmēt – pašzīmēt*;[6]
- suffixes: pairs of lemmas with the same stem, yet different suffix: e.g. nouns *Xums* × *Xība* (255 types: *absurdums – absurdība … žultainums – žultainība*)[7] or adjectives *Xains* × *Xīgs* (24 types: *acains – acīgs … zīdains – zīdīgs*);[8]
- alternations: feminines of the 5th declination class and their (non-)alternation of the stem in genitive plural form – 1565 types: *ābece – ābeču … žubīte – žubīšu*;[9]
- noun diminutives ending with both formants *-iņa* and *-ele* (184 types: *acs – aciņa – ačele … zupe/zupa – zupiņa – zupele*) or *-iņš* and *-elis* respectively (109 types: *auns – auniņš – aunelis … žurnālists – žurnālistiņš – žurnālistelis*);[10] adjective diminutives *X* × *Xiņš* (only 2 types: *kluss – klusiņš*; *mazs – maziņš*).[11]

4. Limits and Advantages of Morfio

It goes without saying that Morfio – as a tool based solely on analysis of the form and ignoring the meaning of the words – cannot produce error-free and ready-to-use results without the need for further manual inspection. The tool focuses on providing maximal *recall* by pre-processing a large amount of data and yielding a list of morphologically related candidates, making analysis faster and more accessible for researchers. Relevance of results (*precision*) is left solely to the judgment of the user: i.e. to the actual query formulation and the subsequent interpretation of the findings.

Yet, these data are hardly accessible by a linguist's introspection, and, especially in some cases, a corpus-driven approach is the only possible way to obtain them. The

[4] The Czech-Latvian components of the parallel corpus InterCorp (IC) [2] are another two searchable datasets, unfortunately, their size is still quite small (v9 – 40,6 M tokens, v12 – 32,7 M tokens). The size of a corpus will understandably affect the size of the results. E.g. the list for the above-mentioned pattern *Xs – noXe* is almost 20 times bigger in Araneum Lettonicum (with the same frequency threshold of 3), yielding 186 word pairs, although the precision itself decreases significantly. See http://morfio.korpus.cz/EWWr9pfi for the results of the query.

[5] http://morfio.korpus.cz/Bv2wcPQT; cf. 69 types for ICv12 (http://morfio.korpus.cz/TfyBMwBs).

[6] http://morfio.korpus.cz/9wQcQKUM; cf. 19 types for ICv12 (https://morfio.korpus.cz/Pa9kBXWE).

[7] http://morfio.korpus.cz/kRH8xniz; cf. 41 types for ICv12 (http://morfio.korpus.cz/OMUCmsbD).

[8] http://morfio.korpus.cz/KLSrOvJH; cf. 0 types for ICv12 (http://morfio.korpus.cz/JK0CGAIx).

[9] http://morfio.korpus.cz/mt11TzQg; cf. 253 types for ICv12 (http://morfio.korpus.cz/RBtmCGjb).

[10] http://morfio.korpus.cz/CvHsAxq9; cf. 6 types for ICv12 (http://morfio.korpus.cz/LaKgeXWg) and http://morfio.korpus.cz/dseYxsxC; cf. 3 types for ICv12 (http://morfio.korpus.cz/oQzn3MVS)

[11] Cf. 1 type (*mazs – maziņš*) for ICv12 (http://morfio.korpus.cz/aEvOVVdW).

frequency of word pairs gives an overall idea about the productivity of the respective phenomena in the contemporary Latvian lexicon and may differ significantly from existing descriptions of Latvian.

Acknowledgements

This work has been supported by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734). During its creation we used the tools developed within the Czech National Corpus project (LM2015044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

[1] Škrabal M, Benko V. Czech & Slovak Corpus Resources Go (not only) Latvian. In: Muischnek K, Müürisep K, editors. Frontiers in Artificial Intelligence and Applications. Proceedings of the 8th International Conference Baltic HLT 2018. Amsterdam: IOS Press; 2018. p. 158-165.

[2] Lazar M, Škrabal M, Vavřín M. Korpus InterCorp – lotyština, verze 12 z 12. 12. 2019. Praha: Ústav Českého národního korpusu FF UK; 2019. Available at: http://www.korpus.cz.

[3] Čermák F, Rosen A. The case of InterCorp, a multilingual parallel corpus. International Journal of Corpus Linguistics 2012;13(3):411-427.

[4] Rosen A. InterCorp – a look behind the façade of a parallel corpus. In: Gruszczyńska E, Leńko-Szymańska A, editors. Polskojęzyczne korpusy równoległe. Polish-language Parallel Corpora. Warszawa: Instytut Lingwistyki Stosowanej; 2016. p. 21-40.

[5] Benko V. Aranea: Yet Another Family of (Comparable) Web Corpora. In: Sojka P, Horák A, Kopeček I, Pala K, editors. Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland; 2014. p. 257-264.

[6] Škrabal M, Vavřín M. The Translation Equivalents Database (Treq) as a Lexicographer's Aid. In: Kosek I et al., editors. Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference. Leiden: Lexical Computing CZ s. r. o.; 2017. p. 124-137.

[7] Benko V. Compatible Sketch Grammars for Comparable Corpora. In: Abel A, Vettori C, Ralli N, editors. Proceedings of the XVI EURALEX International Congress: The User in Focus. 15-19 July 2014. Bolzano/Bozen: Eurac Research; 2014. p. 417-430.

[8] Baayen H. Quantitative aspects of morphological productivity. In: Booij GE, van Marle J, editors. Yearbook of Morphology 1991. Dordrecht: Kluwer Academic Publishers; 1992. p. 109-149.

[9] Good IJ. The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika. 1953;40:237-264.

[10] Cvrček V, Vondřička P. Morfio. Praha: Ústav Českého národního korpusu FF UK; 2013. Available at: http://morfio.korpus.cz.

[11] Cvrček V, Vondřička P. Nástroj pro slovotvornou analýzu jazykového korpusu. In: Gramatika a korpus 2012. Hradec Králové: Gaudeamus; 2012.

[12] Paikens P, Rituma L, Pretkalnina L. Morphological analysis with limited resources: Latvian example. In: Oepen S, Hagen K, Johannessen JB, editors. Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA) 2013. Linköping: Linköping University Electronic Press; 2013. p. 267-277.

[13] Auziņa I et al. Latviešu valodas gramatika. Rīga: LU Latviešu valodas institūts; 2015.

[14] Laua A. Latviešu literārās valodas fonētika. Rīga: Zvaigzne ABC; 1997.

[15] Zasina AJ. Konkurence koncovek -a a -u v genitivu singuláru neživotných maskulin v polštině. In: Stluka M, Škrabal M, editors. Liſka a czban – Sborník příspěvků k 70. narozeninám prof. Karla Kučery. Praha: NLN; 2017. p. 90-98.