doi.org/10.35321/term28-03

Usage and Empirical Productivity of International Adjectival Suffixes in Slovak Revisited

JANA LEVICKÁ
Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
ORCID id: https://orcid.org/0000-0001-6027-604X

ABSTRACT

This paper represents a sequel to earlier research focusing on the usage and productivity of five Latinate suffixes in Slovak. The analysis focuses on real and potential productivity in a two-stage comparison: 1) tokens and lemmas occurring in a general balanced corpus vs general corpus of specialised and academic texts, 2) general corpus of specialised and academic texts vs specialised (sub)-corpora of medical, legal, economic and religious texts. The first aim of the sequel was to establish if the low-frequency lemmas include new coinages that could contribute to the productivity variation across domains. The second aim was to identify and implement an appropriate statistical measure that would enable the comparison of productivity corpus data across differently sized corpora.

KEYWORDS: productivity, adjective, suffix, general corpus, specialised corpus, hapax legomena, low-frequency lemma, large number of rare events distributions.

ANOTACIJA

Šiame straipsnyje aprašomas tyrimas yra ankstesnio tyrimo, nagrinėjančio penkių lotynų kalbos priesagų vartojimą ir produktyvumą slovakų kalboje, tęsinys. Tyrime siekiama nustatyti realų ir potencialų produktyvumą atliekant palyginimą dviem etapais: 1) žodžių formos (angl. *tokens*) ir antraštinės žodžių formos, arba lemos (angl. *lemmas*), esančios bendrajame subalansuotame tekstyne, lyginamos su esančiomis bendrajame specialiųjų ir akademinių tekstų tekstyne; 2) lyginamos bendrajame specialiųjų ir akademinių tekstų tekstyne ir specialiajame medicininių, teisinių, ekonominių ir religinių tekstų patekstynyje esančios žodžių formos ir lemos. Pagrindinis tęstinio tyrimo tikslas – nustatyti, ar tarp retai vartojamų lemų yra naujažodžių, kurie galėtų prisidėti prie produktyvumo skirtumų skirtingose srityse. Tyrimu taip pat siekiama nustatyti ir pritaikyti statistinę priemonę, kuri leistų palyginti tekstyno medžiagos produktyvumą skirtingos apimties tekstynuose.

ESMINIAI ŽODŽIAI: produktyvumas, būdvardis, priesaga, bendrasis tekstynas, specialusis tekstynas, hapaksas, retai vartojama lema, didelis retų įvykių pasiskirstymų skaičius

1. INTRODUCTION

The motivation for our research arose from an interest in terminological synonymy and, within this broader topic, in synonymous denominal adjectives as constituents of multi-word terms.

The focus on adjectives is not a novelty in terminology research. Detailed analyses aimed at adjectives can be found in the work of a number of terminologists, especially from the 1990s and 2000s[1]. In the history of Slovak terminology, adjective-centred studies, triggered by the need to unify and coin standardised Slovak terminologies, can be traced to the 1950s (see e.g. Ján Horecký 1956).

The importance of adjectives in terminology and languages for specialized purposes can also be illustrated by statistical data provided by corpora. The majority of specialised corpora that emerged from the Slovak National Corpus project (hereinafter only SNC project) features a higher ratio of adjectives compared to the reference corpus (prim-7.0-frk), i.e., the corpus compiled of an even share of fictional, specialised and journalistic texts[2]. As can be seen in Table 1, this difference amounts to almost 2%, except for the religious corpus (blf-2.0) in which the ratio of adjectives is roughly the same as in the reference corpus (7.46%). The second row of Table 1 presents the ratio of gerunds that are also used with differentiating and classifying functions in Slovak multi-word terms. In the framework of the morphological annotation of the SNC project, these are classified as a specific group, which is why they are tagged separately from verbs (Garabík et al. 2004).

The coexistence and competition of adjectives has been identified and analysed in the context of language in general (e.g. Nábělková 1996). However, their entering into multi-word terms represents a rather underresearched topic. A common phenomenon resulting in synonymy and competition of adjectives is the coexistence of Latinate adjectives and their (Slovak) counterparts (*vnútrožilový – intravenózny* "intravenous"). The latter group comprises also a subgroup that features both Slovak and, from the etymological point of view, Latinate or international suffix combined with the same international root (e.g. *bakteri**álny** – baktéri**ový*** "bacterial"). Occasionally, a variation of two international suffixes with the same in-

[1] See, for example, H. Assadi and D. Bourigault 1995; S. Normand and D. Bourigault 2001; B. Daille 2001; F. Maniez 2001; M.-C. L'Homme 2002, 2003; or I. Carrière 2008.

[2] More details concerning the corpora used for presented analysis can be found in Part 3.2.

Table 1. Number of tokens and ratios of adjectives and gerunds in the reference and specialised corpora of the SNC project

CORPUS	REFERENCE CORPUS	SCIENTIFIC SUBCORPUS	MEDICAL SUBCORPUS	LEGAL SUBCORPUS	RELIGIOUS CORPUS	ECONOMIC CORPUS
Number of tokens in the corpus	253,137,609	149,581,785	7,099,555	33,600,183	65,920,357	164,987,015
Number of adjectives in tokens	19,090,396 7.54%	13,415,554 8.97%	660,025 9.3%	4,841,400 9.88%	4,914,860 7.46%	15,540,381 9.42%
Number of gerunds in tokens	3,174,567 1.25%	2,394,914 1.6%	112,164 1.58%	1,267,598 2.59%	761,719 1.16%	2,154,124 1.31%

ternational root can arise (e.g. *hypersonórny – hypersonický, kontradiktorický – kontradiktórny*).

This study, focusing on the usage of specific suffixes in the Slovak language, seeks to identify whether some of them are more likely to facilitate such a coexistence or to occur in terminological adjectives more often, which is why the concept of morphological productivity becomes relevant in this case.

The suffixes selected for the analysis represent a minor group used for coining denominal adjectives within a wide range of Slovak adjectival suffixes[3]. As for their composition, the five suffixes consist of an (adapted) international adjectival suffix combined with semantically equivalent Slovak **-ný/ny**: **-ál**ny, **-ár**ny, **-it**ný, **-ív**ny, **-óz**ny, thus, their adjectival status is somewhat multiplied. It appears obvious that all five suffixes of Latin origin were used to coin adjectives from Latin nouns (**-alis** with the variant **-aris**, **-ītus**, **-ōsus**) or verbs (**-ōrius**). It is worth pointing out, however, that many Slovak adjectives with the analysed suffixes entered

[3] Martin Ološtiak and Lucia Ološtiaková (2015: 230) mention as many as 38 suffixes, though 25 suffixes in their sample – derived from the *Slovník koreňových morfém slovenčiny* comprising 66 500 lexical units – represent only 1% of adjectives.

the Slovak lexicon not only directly from Latin, but more often via French or English in the last decades (Horecký 1999) and underwent the process of adaptation by means of the aforementioned Slovak adjectival suffix. It is significant that this borrowing and subsequent adaptation was enabled and enhanced by extra-linguistic factors. Due to predominantly historical considerations, Slovak has been very open to new words originating in Latin and thus the share of Latinate words is high, even in the general lexicon of Slovak (Horecký 1999: 81). Nowadays, thanks to the impact of mass media, internet, and social media, this combined word-forming process of both borrowing and derivation is exceedingly profitable and thus highly pertinent for a linguistic and terminological study.

Our paper represents a sequel to our previous research on the same research topic, which is currently in print (Levická 2021). For the sake of comparison and clarity for the reader, we considered it necessary to include a summary of the previous reasoning and findings, supplemented with new insights and theoretical background. Thus, some of the tables (1–4), though they appeared in our first paper, are repeated here.

2. THEORETICAL UNDERPINNING

The theoretical concept of morphological productivity has been developed by quite a number of researchers and scholars. An overview of previous theoretical thinking and studies in the 20th century has been summarized and commented on in a number of studies and articles, e.g. by Jesús Fernández-Dominguez (2013) or Victoria Hulse (2011). Here, we will focus only on a number of theoretical considerations, which both provided us with inspiration and influenced our choice of the approach and methodology.

The oldest definition quoted here was written in 1948 by American linguist Dwight L. Bolinger (quoted by Säily 2018: 198), who concluded that productivity was the "**statistically determinable readiness**[4] with which an element enters into new combinations".

In 1961, Dutch linguist Henk Schulting (quoted by Stefan Evert and Anke Lüdeling 2001: 165) defined productivity as "the possibility for language users to coin **unintentionally** and, in principle, **unlimited numbers** of new formations, by using the morphological procedure that lies behind the form-meaning correspondence of some known words".

[4] The phrases or words marked in bold letters represent the key features for our research.

Last, but not least, we would like to refer to the word-formation research of Czech linguist Miloš Dokulil, who presented a complex theory of word-formation productivity in his work *Teorie tvoření slov* in 1962. First of all, Dokulil defined the productivity of a language device in general as an "ability of this language device (be it a base, affix or word-formation pattern) to actively participate in coining new words" (1962: 78). He specified that this ability can be *absolute*, i.e. its usability in coining new words in any time and within any word-formation pattern, or *relative* which is semantically, or rather style dependent as well as word-formation pattern dependent. In Dokulil's theory, the productivity is a "concept of synchronic nature, while the implementation of this ability – coining new words according to specific rules – is, to the contrary, a concept of diachronic nature" (1962: 80). Dokulil also differentiated between *systemic* and *empirical productivity* (also termed "parole" or "real productivity" by Framntišek Štícha 2012)), which gives a "general idea of the **overall exploitation** of a specific word-formation pattern or language device in the system of a language in a given time" (Dokulil 1962: 80). Dokulil believed that even "**approximative data** concerning the quantitative use of a given word-formation process or element are of paramount importance for the overall picture of a given language in general and for the characteristics of its lexicon in particular" (Item: 77).

From among contemporary Czech linguists building on Dokulil's theory of productivity and verifying his assumptions on corpus data (see e.g. Štícha 2002, 2007, 2009, 2012 or Magda Ševčíková 2014), Štícha suggests the analysis of empirical/parole productivity, not only in big corpora, but also by means of a series of corpora of different sizes and composition (2012: 104).

It could be claimed that Dokulil's assumptions concerning the relative productivity were echoed by Harald Baayen and others whose methods and approaches we will refer to later in this analysis. In 1999, Baayen and his colleagues Ingo Plag and Christiane Dalton-Puffer wrote that "claims about the productivity of a given affix are generally made without differentiating productivity according to the type of discourse, although it is commonly assumed that certain kinds of derivational suffixes are more pertinent in certain kinds of texts than in others" (1999: 209).

With large corpora available, it is possible to measure, identify and analyse the productivity of a word-formation pattern or element not only in general language at a given time, but also in specialised domains and registers.

3. PREVIOUS RESEARCH

In line with the reasoning and assumptions of linguists quoted in the previous part, we formulated three key research questions at the onset of our research aimed at identifying the morphological productivity of adjectival suffixes:

1) Is the morphological productivity of individual adjectival suffixes from the selected group different in general and specialised texts? And are these differences statistically significant?
2) Can the morphological productivity of individual adjectival suffixes from the selected group vary depending on the specific domain? Are some of those five suffixes more "apt" to contribute to the encoding of field or domain specific concepts?
3) What methods are relevant for this specific analysis of corpus data?

First of all, due to the specificity of this kind of word-formation pattern *Latin base + Latin suffix + etymologically native suffix*, we decided to set aside the qualitative analysis involving the limits and rules governing the combination of individual Latin bases with one of the selected adjectival suffixes in the context of the Slovak language.

Although the quantitative approach cannot provide the whole picture of the productivity status of a suffix or any other morphological form, its benefit for this type of research lies in presenting the perspective, the context of usage of words coined according to a specific pattern, along with specific comparisons with semantically and functionally equivalent forms. It is a generally accepted view that a quantitative approach in the analyses of morphological productivity represents a significant, complementary method to qualitative research (see e.g. Evert and Lüdeling 2001). As Plag states: "quantitative and qualitative notions of productivity [...] are closely related. Thus, the idea of potentiality, which is central to qualitative definitions of productivity, can be expressed in the statistical terms of probability" (quoted by Naccarato 2016: 135). Similarly, Baayen and his co-authors advocate that productivity "seems to be a scalar concept [...] with some affixes one is more likely to encounter newly formed words than with other, a fact that makes productivity a probabilistic notion which is susceptible to statistical analysis" (1999: 10).

3.1. Statistical methods

In stage 1 of our analysis, we adopted the well-trodden methodological path of Harald Baayen and his colleagues and followers (1991, 1992, 1993, 1994, 1996, 1997). We identified the so-called *realized productivity*, i.e. the frequency of usage of adjectives coined with selected suffixes as well as the number of different adjectives coined with these suffixes in different corpora. These corpus statistics outline the productivity of suffixes with respect to the past and present linguistic situation.

The next step was to determine the *potential productivity* of the analysed suffixes, i.e. an estimate of the rate at which new types[5] are expected to appear. Baayen (2009: 7) suggests calculating it as the ratio of hapax legomena[6] with affix X and all tokens with affix X in a corpus. The usage of hapax legomena for the calculation of productivity has been widely criticised, as this group does not directly represent new coinages that are supposed to reflect the productivity of an element or pattern. Plag, Dalton-Puffer and Baayen argue that the potential productivity is a probabilistic measure, and that with the increase in size of a corpus, "the proportion of neologisms among the hapax legomena increases and it has been shown that it is precisely among the hapax legomena that the greatest number of neologisms appear" (1999: 12). In the previous stage of our research, we decided against using the hapax/token method, but adopted the hapax/type method in line with van Marle's reasoning (1992) that token frequency is not as relevant a variable in a measure of productivity as the number of lemmas. However, the most serious drawback imposed by this measure is the impossibility of comparing the statistical data of corpora of different sizes (Naccarato 2016: 133). We will return to this consideration in Part 5.

3.2. Corpora used in the analysis[7]

All three corpora and three subcorpora used in this analysis were released by the Department of the SNC in 2013–2020 and are accessible for all registered users.

[5] Types in corpus linguistics represent unique words, they are synonymous to lemmas (Baker et al. 2006).

[6] Hapax legomena (orig. Greek phrase meaning "once said"), also abbreviated to hapax, is a word that occurs only once in a particular corpus (Baker et al. 2006).

[7] Most of this section comes from the paper describing the previous stage of our research (Levická 2021), however, this repetition is necessary in order to explain the nature of the corpus data.

The first, reference corpus (prim-7.0-frk), amounting to more than 253 million tokens, is composed of an even share of journalistic, specialised and fictional texts (64.12% of them are originally written in Slovak, while 29.51% represent translations), written in 1991–2015. The corpus was used in the compilation of two frequency dictionaries of Slovak (2017, 2018) as well as the reverse dictionary (2018).

The second corpus, prim-9.0-public-prf, is a publicly available subcorpus of the primary corpus of the SNC project (hereinafter referred to as the scientific subcorpus). Compiled from specialised, academic and non-fiction texts, this subcorpus features more than 149 million tokens and documents, as well as the general discourse of science and research, including specialised journalism. Its texts were written between 1955–2019. The percentage of texts written in 1955–1991 amounts to 8.7% (more than 13 million of tokens).

The smallest (sub)corpus of all the searched corpora is the result of filtering the primary corpus of the SNC project [9], version 9.0. It consists of texts that belong to the field of medicine written in 1976–2019 and comprising slightly more than 7 million tokens. The percentage of texts written in 1955–1991 amounts to 1.09%. We will refer to it throughout the text as the medical subcorpus.

In order to have a comparable source originating from other specialised corpora and the reference corpus, the specialised corpus legal-1.1 (hereinafter referred to as the legal subcorpus), built in cooperation with the Slovak Ministry of Justice, was narrowed down to legislative texts created in the period 1991–2011. Its almost 49 million tokens were thus reduced to 33.5 million tokens.

The specialised corpus – blf-2.0 – focusing on the religious domain, was released in 2014. Its texts consist of almost 66 million tokens written between 1989–2014 (hereinafter referred to as the religious corpus). It comprises more than 80% of thematic journals and newspapers. Similarly, specialised corpus ecn-2.0-public (hereinafter referred to as the economic corpus), devoted to the domain of economics, includes as much as 96.24% of specialised texts published in thematic journals and newspapers. Texts of this corpus come from 1992–2014 and comprise almost 165 million tokens.

It must be noted, however, that these corpora are heterogenous and cover a wide variety of texts.

As far as the corpus search is concerned, we did not base the queries on morphological tagging provided for each and every (sub)corpora included in the analysis, on the basis that, with Latinate words, the tagging proved to be inadequate and erroneous. Therefore, we opted for a simple search of the specific ending of a token, e.g. [lemma=".*álny"], combined with the automatic filtering of words with incidentally the same string of characters (see Part 4.1).

3.3. Results of the previous stage

In the previous stage, statistical data from corpora enabled us to create a ranking for the realized productivity of the analysed suffixes. The suffix *-álny* is definitively the most widely used in all (sub)corpora, while *-órny* is at the opposite pole of the frequency axis in 5 (sub)corpora. As can be seen from the normalised frequencies of usage of four out of five suffixes, they appear more frequently in specialised and academic texts (scientific subcorpus) compared to the reference corpus. It is also noteworthy that the normalised frequencies of suffixes in the medical subcorpus either equal or considerably exceed the ipm in the scientific subcorpus, while the ipm of suffixes (instances per million) in the religious corpus is manifestly lower than in the reference corpus. This can be partly explained by the type of texts included in those (sub)corpora.

Table 2. Frequency and normalised frequency (IPM, instances per million) of the analysed suffixes in respective (sub)corpora. The corpora are ordered from the smallest to the largest one

SUFFIX	MEDICAL SUBCORPUS	LEGAL SUBCORPUS	RELIGIOUS CORPUS	SCIENTIFIC SUBCORPUS	ECONOMIC CORPUS	REFERENCE CORPUS
	Tokens	Tokens	Tokens	Tokens	Tokens	Tokens
	IPM	IPM	IPM	IPM	IPM	IPM
-álny	31,408	102,863	128,472	537,682	597 005	547,742
	4423.94	3061.38	1948.9	3594.57	3618.5	2163.81
-árny	4,529	25,980	14,253	94,586	38,178	84,780
	637.93	773.21	216.22	634.14	231.40	334.92
-itný	2,683	2,256	9,176	30,402	52,574	44,655
	377.91	67.14	139.20	203.25	318.66	176.41

SUF-FIX	MEDICAL SUB-CORPUS	LEGAL SUB-CORPUS	RELI-GIOUS CORPUS	SCIENTIF-IC SUB-CORPUS	ECO-NOMIC CORPUS	REFER-ENCE CORPUS
	Tokens	Tokens	Tokens	Tokens	Tokens	Tokens
	IPM	IPM	IPM	IPM	IPM	IPM
-ózny	1,416	482	2,321	8,863	9,440	16,531
	199.45	14.35	35.21	59.25	57.22	65.30
-órny	884	1,959	620	4,537	2,039	4,672
	124.51	58.30	9.41	30.33	12.36	18.46

The ordering of corpora in Table 3 showing the number of different words coined with the analysed suffixes enables us to compare the variability of coinages, first between general and specialised language and, secondly, from the smallest medical corpus up to the largest economic corpus. The table shows Baayen's realised productivity of the analysed suffixes. In all (sub)corpora *-álny* represents the most frequently used suffix appearing in different types, the smallest number of types or lemmas were identified with the suffix *-órny*. Suffixes *-itný* and *-órny* seem to overlap in the majority of (sub)corpora, their more subtle distinction requires a test of significance. However, a real comparison between corpora is not possible as the table features absolute counts of types.

Table 3. Number of lemmas in a given (sub)corpus

SUF-FIX	REFER-ENCE CORPUS	SCIEN-TIFIC SUB-CORPUS	MEDICAL SUB-CORPUS	LEGAL SUB-CORPUS	RELI-GIOUS CORPUS	ECO-NOMIC CORPUS
-álny	1,003	1,178	519	322	611	744
-árny	300	365	179	99	141	200
-itný	90	98	36	41	55	117
-ózny	115	119	72	40	52	82
-órny	25	33	13	10	20	26

The findings concerning the estimate of potential productivity presented in Table 4 are rather varied, as the ordering of suffixes indicates.

Table 4. Hapax/type ratios in a given (sub)corpus. The suffixes in each subtable are listed in the order of decreasing ratio

HAPAX/TYPE RATIO					
reference corpus		scientific subcorpus		medical subcorpus	
-ózny	0.252173913	**-órny**	0.228571429	**-ózny**	0.208333333
-árny	0.201342282	**-árny**	0.210382514	**-itný**	0.138888889
-álny	0.173956262	**-álny**	0.209499576	**-álny**	0.129094412
-itný	0.122222222	**-ózny**	0.201680672	**-árny**	0.1
-órny	0.12	**-itný**	0.195876289	**-órny**	0.076923077
legal subcorpus		religious corpus		economic corpus	
-órny	0.1	**-órny**	0.2	**-itný**	0.237288136
-ózny	0.075	**-álny**	0.188180404	**-ózny**	0.228915663
-álny	0.052795031	**-árny**	0.145833333	**-órny**	0.185185185
-árny	0.05	**-ózny**	0.134615385	**-álny**	0.177954847
-itný	0	**-itný**	0.090909091	**-árny**	0.142156863

The most productive suffix in as many as three corpora is *-órny*. In the remaining three (sub)corpora, the ranking is topped twice by *-ózny* and once by *-itný*. Moreover, the same suffix *-órny* seems to be potentially least productive in general and medical texts. In as many as three (sub)-corpora, it is the suffix *-itný* that has taken the final place in the productivity ranking. Similarly, the last place in the productivity ranking of economic texts is occupied by the suffix *-árny*.

4. RESEARCH QUESTIONS OF PRESENTED ANALYSIS

The previous stage of our research indicated noteworthy differences in morphological productivity of the analysed suffixes, depending both on the type of language and domains. However, one question was answered only partially and gave rise to a new one:

1. Is it worth analysing low-frequency lemmas for the sake of neologisms? In fact, several researchers suggest that the word-frequency distribution of productive affixes is supposed to be distinctly shifted towards low-frequency lemmas comprising new coinages.

2. As we already pointed out in previous sections, results concerning the potential productivity of the analysed suffixes cannot be compared across corpora. Therefore, it is necessary to identify an appropriate statistical method which enables this comparison.

4.1. Manual clean-up of data

Both previous and current stages of our analysis are based on manually cleaned data. The importance of this tedious and time-consuming process has been pointed out by several researchers. Evert and Lüdeling claimed (and proved) that this is a prerequisite for quality statistical evaluation, not *only* when employing more sophisticated measures (2001: 168). The manual cleaning is more than important for three reasons: the first resulting from the nature of corpus data. First of all, the corpus data usually comprise a fairly large share of typos, especially within lemmas of low frequency. Secondly, the corpora feature orthographical errors which occur frequently with words of Latinate origin in particular. Thirdly, it was necessary to merge those lemmas with suffix X differing only in usage/non-usage of a hyphen, as well as lemmas with capitalised and non-capitalised first letters (not being proper names) that are distinguished by automatic lemmatization used within the SNC project. The result of this stage of our manual clean-up can be seen in Table 5, where the average share of hapaxes for the analysed suffixes ranges from 24 to 30% of the overall count of lemmas with a given suffix. However, less than 5% of lemmas (possible neologisms) remained in the case of *-itný* in legal texts, and, on the contrary, 40% or more were left with the suffix *-órny* in religious and general scientific texts and with the suffix *-ózny* in economic texts and, surprisingly, in the texts of the reference corpus. These shares could be indicative of (un)productivity of a given corpus. As Baayen and co-authors claim, the number of hapaxes for productive elements can reach as much as half of the observed vocabulary size in a "sufficiently large corpus" (1999: 11).

The second reason for the manual clean-up was to filter the extracted corpus data in order to obtain lists featuring neologisms with the analysed suffixes. We therefore excluded lemmas that 1) incidentally include the same string of characters as the analysed suffixes; 2) can be found in general Slovak dictionaries and the Dictionary of Foreign Words[8], i.e.

[8] We used the integrated search tool of the dictionary portal https://slovnik.juls.savba.sk/.

Table 5. Ratio of hapaxes (potential neologisms) in a given (sub)corpora after the 1st stage of manual cleaning

	RATIO OF HAPAXES WITH A SPECIFIC SUFFIX AFTER THE 1ST STAGE OF MANUAL CLEANING					
suffix	medical subcorpus	legal sub-corpus	religious corpus	scientific subcorpus	economic corpus	reference corpus
-álny	30.8%	24.2%	29.5%	32.7%	29.7%	30%
-árny	29.6%	25.3%	33.3%	35.1%	31%	36%
-itný	36.1%	4.9%	25.5%	29.6%	31.6%	26.7%
-ózny	34.7%	32.5%	38.5%	30.3%	45.1%	42.6%
-órny	30.8%	30%	40%	42.4%	38.5%	28%

therefore they are not neologisms; 3) were found in the two most extensive SNC corpora[9] – those lemmas must have occurred in at least one of them three or more times, provided that those occurrences came from three different sources and from three different years. This last filtering principle was established due to the fact that our targeted words with one of the analysed suffixes tend to be terms of specialised domains and we lack up-to-date specialised dictionaries in Slovak. It must be admitted that the resulting lists of lemmas represent only potential neologisms. The results can be seen in the columns titled *Ratio of potential neologisms in hapaxes* in tables 6 and 7. Compared to Table 5 the share of potential neologisms was considerably reduced.

4.2. Neologisms in low-frequency lemmas

In order to answer the first research question, we carried out the same filtering procedure as that employed for the list of hapaxes. In tables 6 and 7, we show the percentages (counted from the overall number of lemmas with a given suffix, cf. Table 3) of potential neologisms in the low-frequency lemmas occurring 4–2 times in (sub)corpora and next to it the percentages of potential neologisms within hapax group.

While the share of potential neologisms features small increases in specialised and academic texts for every analysed suffix, due to the inclusion

[9] Those corpora are the largest ones in the SNC project for the time being: general corpus prim-9.0-juls-all and legal corpus legal-1.1.

of potential neologisms from the group of low-frequency lemmas, potential neologisms in the reference corpus experienced only a very modest increase for four suffixes, whereas the percentage of the fifth one – *-órny* – remained stable. This situation could be expected, as new coinages with the analysed suffixes are more likely to appear in specialised texts.

Table 6. Ratio of potential neologisms in the group of low-frequency lemmas (4–1 occurrences) and in the group of hapaxes from the overall count of types after cleaning up in general corpus and general scientific subcorpus

SUFFIX	REFERENCE CORPUS		SCIENTIFIC SUBCORPUS	
	Ratio of potential neologisms in low-frequency lemmas (4–1)	Ratio of potential neologisms in hapaxes	Ratio of potential neologisms in low-frequency lemmas (4–1)	Ratio of potential neologisms in hapaxes
-álny	20.5%	17.3%	27.2%	21%
-árny	24.5%	20%	27.1%	21.1%
-itný	13.3%	12.2%	26.5%	19.4%
-ózny	27.8%	25.2%	24.4%	20.2%
-órny	12%	12%	27.3%	24.2%

As far as the domain specific potential neologisms are concerned, the highest percentage, similar to the situation in the scientific subcorpus, can be seen in the economic corpus where the suffix *-itný* amounts to as many as 35% of potential neologisms. The remaining suffixes in this corpus reaches approximately 25% of potential neologisms within low-frequency lemmas. On the other hand, the lowest percentage of potential neologisms was identified in legal texts. It is also worth mentioning those cases where the percentage was doubled by the inclusion of the low-frequency lemma group: for *-órny* in medical texts, *-árny* in legal texts and *-itný* in religious texts. At the same time, it is no surprise that the suffix *-ózny* tops the ranking in medical texts as it is a prototypically medical suffix. However, almost the same percentage of potential new lemmas with *-ózny* were to be found in economical texts. Zero increase in potential neologisms occurred for the suffix *-itný* in medical and legal texts and for the suffix *-órny,* also in legal texts.

Table 7. Ratio of potential neologisms in the group of low-frequency lemmas (4–1 occurrences) and in the group of hapaxes from the overall count of types after cleaning up in specialised texts of the four domains

SUFFIX	MEDICAL SUBCORPUS		LEGAL SUBCORPUS		RELIGIOUS CORPUS		ECONOMIC CORPUS	
	Ratio of potential neologisms in low-frequency lemmas (4–1)	Ratio of potential neologisms in hapaxes	Ratio of potential neologisms in low-frequency lemmas (4–1)	Ratio of potential neologisms in hapaxes	Ratio of potential neologisms in low-frequency lemmas (4–1)	Ratio of potential neologisms in hapaxes	Ratio of potential neologisms in low-frequency lemmas (4–1)	Ratio of potential neologisms in hapaxes
-álny	16.8%	12.9%	7.8%	5.3%	26.5%	19.7%	26.5%	18%
-árny	12.8%	10.1%	12.1%	5.1%	19.9%	15%	22.5%	14.5%
-itný	13.9%	13.9%	0%	0%	18.2%	9.1%	35.1%	23.1%
-ózny	26.4%	20.8%	10%	7.5%	21.2%	13.5%	25.6%	23.2%
-órny	15.4%	7.7%	10%	10%	30%	20%	23.1%	19.2%

For the sake of better understandability, we also add one more table showing the percentages of potential neologisms in the group of low-frequency lemmas.

Table 8. Percentage difference of potential neologisms between the group of low-frequency lemmas (4–1 occurrences) and the group of hapaxes from the overall count of types after cleaning up

	RATIO OF POTENTIAL NEOLOGISMS WITHIN LOW-FREQUENCY LEMMAS (4–2 OCC.)					
suffix	reference corpus	scientific subcorpus	medical subcorpus	legal subcorpus	religious corpus	economic corpus
-álny	3.2%	6.2%	3.9%	2.5%	6.8%	8.5%
-árny	4.5%	6%	2.7%	7%	4.9%	8%
-itný	1.1%	7.1%	0%	0%	9.1%	12%
-ózny	2.6%	4.2%	5.6%	2.5%	7.7%	2.4%
-órny	0%	3.1%	7.7%	0%	10%	3.9%

Finally, the last table shows the ranking of suffixes in respective (sub)-corpora after repeated usage of adapted formula from the previous stage – ratio of low-frequency lemmas with a given suffix/number of types with

Table 9. Difference in ratio of potential neologisms between the group of low-frequency lemmas (4–1 occurrences) and the number of types. The ranking of suffixes that are highlighted changed compared to Table 4

LOW-FREQUENCY LEMMAS/TYPE RATIO					
reference corpus		scientific subcorpus		medical subcorpus	
-ózny	0.2689075	**-órny**	0.2727272	**-ózny**	0.2638888
-árny	0.2466666	**-álny**	0.2716468	**-álny**	0.1682785
-álny	0.2053838	**-árny**	0.2712328	**-órny**	0.1538461
-itný	0.1333333	**-itný**	0.2653061	**-itný**	0.1388888
-órny	0.12	**-ózny**	0.2436974	**-árny**	0.1284916
legal subcorpus		religious corpus		economic corpus	
-árny	0.1212121	**-órny**	0.3	**-itný**	0.3504273
-órny	0.1	**-álny**	0.2651391	**-álny**	0.2647849
-ózny	0.1	**-ózny**	0.2115384	**-ózny**	0.2560975
-álny	0.0776397	**-árny**	0.1985815	**-órny**	0.2307692
-itný	0	**-itný**	0.1818181	**-árny**	0.225

a given suffix. Compared to Table 4, the ranking changed in all (sub)-corpora except for the reference corpus. Thus, it seems that the inclusion of low-frequency lemmas proved to be worthwhile.

The suffix *-órny* is most productive in two corpora (compared to three in Table 4), while *-ózny* remained at the top of two rankings and *-itný* at the top of the ranking in the economic corpus. In the legal corpus, the most productive suffix seems to be *-árny* while this very suffix is simultaneously least productive in medical texts and economic texts. The last place in the productivity rankings is occupied by the suffix *-órny* both in the reference corpus and medical texts, by the suffix *-ózny* in general scientific texts and the suffix *-itný* remained least productive in legal and religious corpora.

5. COMPARING PRODUCTIVITY ACROSS (SUB)CORPORA

As we pointed out in Part 3.1, statistical measures using raw frequencies and counts of types prevented us from comparing previous results originating from different (sub)corpora. If we agree with the claim or rather the hypothesis that the productivity of any word-formation element or

pattern can be concluded from new coinages (and, at the same time, we proved in Part 4.2 that these new coinages could occur not only within hapax groups, but also among low-frequency types), it is possible to base further analysis on word-frequency distributions and frequency spectra. By "frequency spectrum" we mean a list reporting how many types in a frequency list can be observed to occur (Baroni 2009: 807).

From the perspective of word-frequency distribution, Evert and Lüdeling (2001: 166), in line with Schultink's definition, claim that "a productive pattern is, in theory, characterised by an infinite vocabulary (cf. the notion of un-limitedness in Schultink's definition), whereas a totally unproductive pattern is expected to have a finite, and often quite small, vocabulary". They point out that "low frequency types (including hapax legomena, but also types occurring two, three, etc. times) account for a major part of the vocabulary" of productive patterns and conclude that this kind of comparison cannot be based on standard statistical models (2001: 167).

For this purpose of statistical modelling, Baayen introduced in 2001 the so-called LNRE[10] distributions. The benefit of these specialised models lies in the possibility of extrapolating the type-token statistics from a specific sample to larger values of tokens and thus estimate potential neologisms outside this sample. They are based on the predicted vocabulary sizes obtained from a count of low-frequency types in the corpus.

One of the outputs of this modelling are vocabulary growth curves or rather than type-token growth curves, which count the number of types as a function of the number of tokens. A typical shape of this curve for an unproductive process starts with a rise, but then it "flattens out and converges to a constant value, the full vocabulary size" (2001: 168). On the contrary, a typical productive pattern features an infinite vocabulary, i.e. the curve seems to grow indefinitely.

By employing the LNRE distributions, it is possible to estimate the number of types in larger quantities of texts, and even the number of types

[10] The LNRE theory (theory of large number of rare events, as an independent area of statistics) has its origin in 1988 when Georgian statistician Estate Khmaladze published *The statistical analysis of a large number of rare events*. As his student Giorgi Kvizhinadze wrote: "The common feature of examples we will present below is that along with several frequent events there is also a very large number of very rare events, say, with frequency 0,1,2. The total amount of these rare events compared to the number of observations typically is not large but the number of these events among all the different observed events is always very significant. These rare events are usually very important. For instance, the number of words used in the book only once can be considered not of vital importance for this book, but it is very clear that these words are absolutely important because they constitute half of the author's vocabulary" (2010).

of a word-formation pattern in the entire "population", i.e. in a language as such. At the same time, it is possible to estimate or predict the number of types in samples of the same size derived from larger corpora.

As mentioned in Part 4.1, this statistical procedure also requires the quality counts of types and tokens, otherwise, the calculation of the estimated vocabulary size or types would be erroneous – instead of flattening out and converging to a constant value, the curve could continue to grow.

5.1. The LNRE distributions

We decided to apply the LNRE models to identify[11] the productivity over differently-sized corpora, partly due to the fact that there is also an open-source tool available online – ZipfR developed by Evert and Marco Baroni (2007). This open source tool includes several models, from which we used finite Zipf-Mandelbrot model (fZM) based on the Zipf-Mandelbrot law[12]. As the authors state, its usage goes beyond linguistics. ZipfR accepts several input formats, including simple frequency lists and plain samples, in a one-token-per-line format. We could use frequency lists from our previous research that had undergone the manual cleaning of the 1st stage (cf. Part 4.1), which represented another advantage of this method.

Our frequency lists underwent 4 basic operations (Evert, Baroni 2007):

(i) parameter estimation, where the parameters of the LNRE model M are determined from a training sample of size by comparing the expected frequency spectrum with the observed frequency spectrum;
(ii) goodness-of-fit evaluation based on the covariance matrix of the number of types and the number of types occurring exactly *m* times;
(iii) interpolation and extrapolation of vocabulary growth, using the expected values;
(iv) prediction of the expected frequency spectrum for arbitrary sample size.

As Evert and Baroni pointed out (2007), this modelling does not focus on the frequencies of individual word types, "but rather on the distribution of such frequencies (in a sample) and probabilities (in the population)".

[11] We would like to thank for the invaluable help from our colleague Ján Mášik who assured the implementation of the fZM model for our corpus data.

[12] For a comprehensive description of the Zipf-Mandelbrot law, the reader is referred to, e.g. Baroni 2009.

5.2. Application of the LNRE distributions and evaluation

The ZipfR tool provided us with word frequency spectra and type-token curves for each and every suffix in every (sub)corpus. For the lack of space, we only present the graphs showing type-token curves that were adjusted to have the same scale on the x-axis. The y-axis differs, as it depends, and results, from the actual count of types for a specific suffix in a given (sub)corpus. We propose to compare the pairs of graphs and finally evaluate the productivity of suffixes across our corpus data.

The x-axis represents the number of tokens, while the y-axis shows the (expected) number of different types produced by the word-formation pattern featuring one of the analysed suffixes.

Both corpora, the reference and scientific ones, show the same ranking of suffixes topped by the suffix *-álny* and ending with the suffix *-órny*. In the reference corpus, the curve of three suffixes *-álny*, *-árny*, *-ózny* grows and thus indicates their productivity, while the curves of *-itný* and, in particular, *-órny* flatten out. In the scientific subcorpus, two suffixes can be considered productive: *-álny* and *-árny*. The curves of the remaining three (*-ózny*, *-itný*, *-órny*) grow very slowly, but show the tendency to flatten out at the end of the x-axis included in the graph.

The graphs for the medical subcorpus and legal corpus show only one productive suffix: *-álny*[13]. In the medical subcorpus, both *-árny* and *-ózny* flatten very quickly (possibly, they reach their maximum of vocabulary size), and while *-itný* and *-órny* grow slowly, but steadily, they cross the 100-type point and possibly flatten out under the 200-type point.

The suffix *-álny* in legal texts is heading to the 350-type point, while in medical texts it is 800-type point (cf. the scale on the y-axis). However, in legal corpus the ranking slightly differs – the suffix *-itný* comes in the third place getting past the suffix *-ózny,* which appears almost completely unproductive with the curve running parallel with x-axis. The final suffix *-órny* features a very slow rise and a subsequent tendency to flattening.

While the graph of the religious corpus shows three productive suffixes *-álny, -árny* and *-ózny,* it is noteworthy that all 5 analysed suffixes seem to be productive in economic texts. The ranking of suffixes in both corpora remains the same.

[13] These two graphs miss the interpolation curve for the suffix *-álny*. The reason for this lies in the adjustment and unification of the scale of the x-axis for all (sub)corpora which did not permit to create this type of curve.

Fig. 1. Vocabulary growth curves of analysed suffixes in the reference corpus

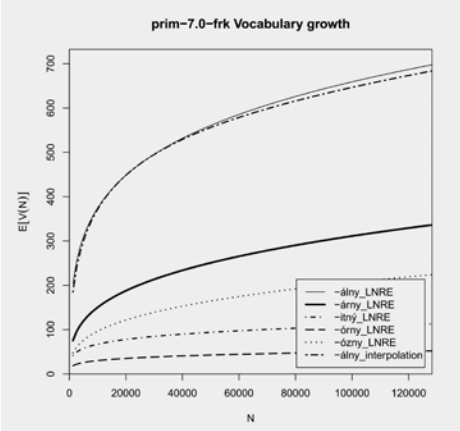


Fig. 2. Vocabulary growth curves of analysed suffixes in the scientific subcorpus

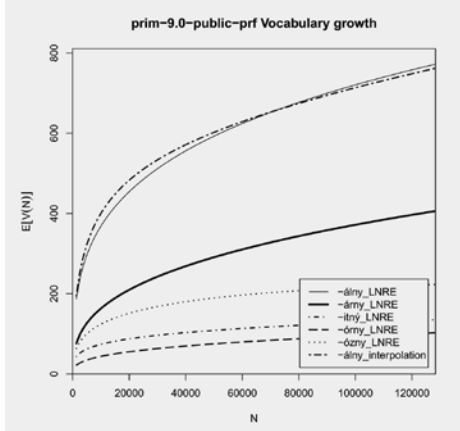


Fig. 3. Vocabulary growth curves of analysed suffixes in the medical corpus

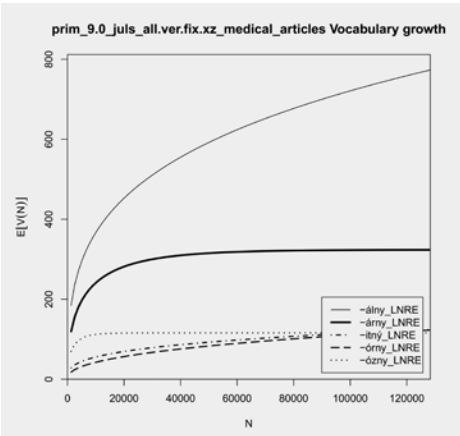
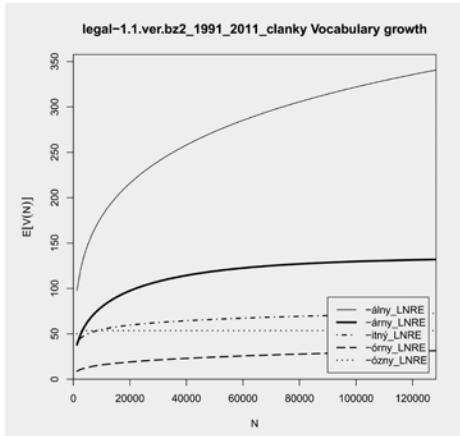


Fig. 4. Vocabulary growth curves of analysed suffixes in the legal corpus



In the religious corpus, both curves of *-itný* and *-órny* flatten out well under the 100-type point.

If we sum up the abovementioned findings concerning the productivity of the analysed suffixes across our corpus data, we can claim that the most productive suffix is ***-álny***, topping the ranking in all the (sub)corpora. Its productivity seems to be highest in the subcorpus of scientific and academic texts and in the medical subcorpus, followed by the reference corpus (the curve is heading to the point of 700 types), the religious

Fig. 5. Vocabulary growth curves of analysed suffixes in the religious corpus

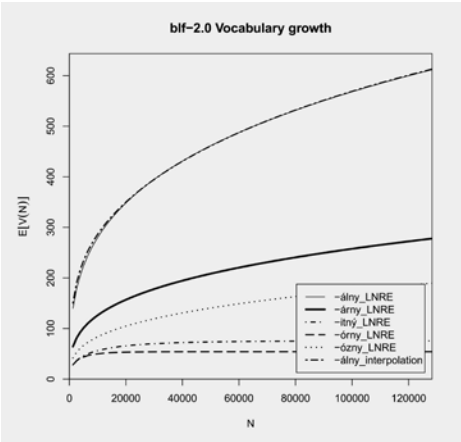
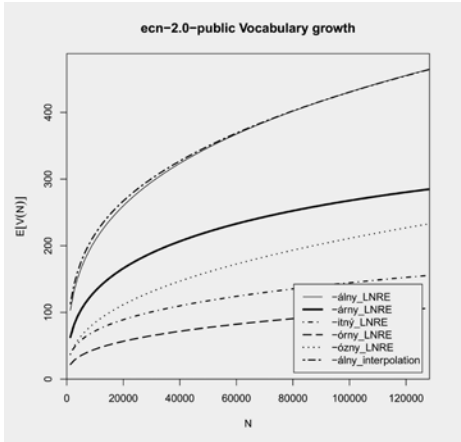


Fig. 6. Vocabulary growth curves of analysed suffixes in the economic corpus



corpus (the curve crosses the 600-type point), the economic corpus (the curve is heading to the 500-type point) and finally the legal corpus (at its highest, the curve is close to 350-type point).

The suffix *-**árny*** comes in second and features productivity in 4 out of 6 (sub)corpora. The shape of the curves suggests that it is unproductive in legal and medical texts. The highest productivity of this suffix can be seen in the subcorpus of scientific and academic texts (the curve crosses the 400-type point), then in the reference corpus (the curve crosses the 300-type point), while in religious and economic texts the curve runs close to the 300-type point.

Similarly, the suffix *-**ózny*** appears to be productive in 4 out of 6 corpora. Interestingly enough, it seems to be unproductive in the same (sub)-corpora as the suffix *-árny* – in legal and medical corpora. It is most productive in economic texts where the curve runs well over the 200-type point, then in the reference corpus where the curve crosses the same point, while in the remaining two corpora the curve either touches the same point, but seems to flatten out there (prim-9.0-public-prf) or only heads for this point in religious corpus.

Also, the suffix *-**itný*** seems to be productive in 4 out of 6 corpora. The flattening curve indicating unproductivity can be seen in graphs for legal and religious corpora. It is most productive in economic texts where the curve runs well over 100-type points, while in three remaining (sub)-corpora it crosses this point and runs lower.

The final suffix in the ranking ***-órny*** – is productive in 4 out of 6 corpora, also. Its unproductivity was identified in the reference and religious corpora. Its highest productivity is in medical and economic texts where the curve crosses the 100-type point, in academic and scientific texts the curve runs close to this point and in legal texts it hardly crosses the 25-type point.

In conclusion, it may be said that the analysed suffixes are most productive in scientific and academic texts or in the texts of a special domain (economy or medicine), which comes as no surprise considering their etymology. Their productivity, or rather degree of productivity, differs considerably across (sub)corpora. The most interesting observation, perhaps, is the unproductivity of three suffixes in the legal domain and the lowest productivity of the remaining two suffixes in the same domain. It could indicate that legal texts do not favour lexical creativity in coining new adjectives of this kind.

Compared to Part 4, the difference in the ranking of suffixes can be partially explained by the phenomenon observed by Václav Cvrček (2012) concerning hapax-type ratio. According to Cvrček's experiments, this ratio tends to decrease from its maximal value 1 to its local minimum (ibid: 5). However, after this point, the ratio starts to increase again. Cvrček claims it to be "some sort of general quantitative principle of large collections of texts" (ibid: 14). The shape of the hapax-type function seems to be roughly the same even for typologically different languages, however, the size of a textual sample depends on the type of the language. In order to reach the minimal point of hapax-type ratio, Cvrček states that an English corpus should comprise at least 3 million tokens, while a Czech one (and we believe a Slovak one as well due to typological relatedness of the two languages) should comprise as many as 58 million tokens. As two subcorpora from our analysis are smaller than Cvrček's limit, their hapax-type ratio is situated in the decreasing part of the hapax-type function.

6. CONCLUSION

In the sequel to our research focusing on the productivity of selected suffixes in different corpora and domains, we managed to answer both questions presented in Part 4.

1. We were able to prove that low-frequency lemmas extracted from a corpus also include potential neologisms, and, though their distribution

may vary across corpus data, they can alter the productivity ranking based on hapax/type formula.

2. We identified and applied statistical modelling for the evaluation of productivity across corpora – LNRE distributions. By feeding the open-source tool with our corpus data we received a visual modelling of productivity that enabled us to state that the productivity of the analysed suffixes differs not only when general and scientific texts are compared, but also between different specialised domains. We believe that the modelling clearly shows the real potential of each and every analysed suffix to produce new types in a respective domain.

However, the results and ranking from Part 4 and 5 of this paper are rather mutually incomparable because the first one is based directly on raw data – the count of types and hapaxes or low-frequency types in a specific (sub)corpus, while the second one on the expected frequency spectra for every (sub)corpus. In short, while the first one represents actual data, the second one seeks to estimate the productivity of an element in the specific language as such. Therefore, for larger corpora of medical and legal texts, we should identify more hapaxes, which might result in a different suffix ranking. Our assumptions could eventually be verified in future research by analysing more extensive corpora comprising a representative sample of relevant text types for specific domains because text genres can also play a role in morphological productivity of an element.

Given that even the biggest corpus would not comprise all the words, we believe that it is more reasonable to evaluate the morphological productivity of any element by means of more sophisticated methods than the raw count of types or tokens.

In conclusion, it could be said that we are able to identify the past productivity in a word-formation pattern or element. We can also provide different estimates of its future exploitation, i.e., Baayen's potential productivity. Furthermore, even if we complete the statistical measures with a qualitative analysis, we should take into consideration one more fact: as Fernández-Domínguez rightly argues, probabilistic predictions of morphological productivity should be perceived in the context of extralinguistic factors and, in particular, those involving the naming needs of a speech community. "If no naming need exists, no productive word-formation can take place" (Fernández-Domínguez 2013: 438).

ACKNOWLEDGEMENTS

The paper has been written within the Slovak National Corpus project supported by the Slovak Academy of Sciences, Ministry of Education, Science, Research and Sport of the Slovak Republic, Ministry of Culture of the Slovak Republic and the Ľudovít Štúr Institute of Linguistics, Slovak Academy of Sciences.

LIST OF SOURCES

Slovenský národný korpus. Korpus prim-7.0-frk. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2018. Accessible at: https://korpus.sk.

Slovenský národný korpus. Korpus prim-9.0-public-prf. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2020. Accessible at: https://korpus.sk.

Slovenský národný korpus. Korpus prim-9.0-juls-all. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2020. Accessible at: https://korpus.sk.

Slovenský národný korpus. Korpus legal-1.1. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2013. Accessible at: https://korpus.sk.

Slovenský národný korpus. Korpus blf-2.0. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2014. Accessible at: https://korpus.sk.

Slovenský národný korpus. Korpus ecn-2.0-public. Bratislava: Jazykovedný ústav Ľ. Štúra SAV 2016. Accessible at: https://korpus.sk.

Slovníkový portál Jazykovedného ústavu Ľ. Štúra SAV https://slovnik.juls.savba.sk/.

The ZipfR package. Available at: http://zipfr.r-forge.r-project.org/.

REFERENCES

Assadi Houssem, Bourigault Didier 1995: Classification d'adjectifs extraits d'un corpus pour l'aide à la modélisation de connaissances. – *Troisièmes journées internationales d'analyse des données textuelles*, 313–320.

Baayen R. Harald, Lieber Rochelle 1991: Productivity and English derivation: a corpus-based study. – *Linguistics* 29, 801–843.

Baayen R. Harald 1992: Quantitative aspects of morphological productivity. – *Yearbook of Morphology 1991*, eds. G. E. Booij, J. van Marle, Dordrecht: Kluwer Academic Publishers, 109–149.

Baayen R. Harald 1993: On frequency, transparency, and productivity. – *Yearbook of Morphology 1992*, eds. G. E. Booij, J. van Marle, Dordrecht: Kluwer Academic Publishers, 181–208.

Baayen R. Harald 1994: Productivity in production. – *Language and Cognitive Processes* 9, 447–469.

Baayen R. Harald, Renouf Antoinette 1996: Chronicling The Times: Productive Lexical Innovations in an English Newspaper. – *Language* 72, 69–96.

Baayen R. Harald, Lieber Rochelle 1997: Word frequency distributions and lexical semantics. – *Computers and the Humanities* 30, 281–291.

Baayen R. Harald 2009: Corpus linguistics in morphology: morphological productivity. – *Corpus Linguistics. An international handbook*, eds. A. Lüdeling, M. Kyto, Berlin, Mouton De Gruyter, 900–919.

Baker Paul, Hardie Andrew, McEnery Tony 2006: *A Glossary of Corpus Linguistics*, Edinburgh: Edinburgh University Press Ltd. Available at: https://pdfs.semanticscholar.org/856a/9640311005ff8a97ab98976c9209aa12120a.pdf.

Baroni Marco 2009: Distributions in text. – *Corpus Linguistics. An international handbook*, eds. A. Lüdeling, M. Kyto, Berlin, Mouton De Gruyter, 803–822. Available at: https://home.sslmit.unibo.it/~baroni/publications/hsk_39_dist_rev2.pdf.

Carrière Isabelle 2008: Méditerm: encodage des adjectifs médicaux. – *Corpus et dicionnaires de langues de spécialités*, eds. F. Manize, P. Dury, N. Arlin, C. Rougemont, Grenoble: Presses Universitaires de Grenoble, 175–196.

Cvrček Václav 2012: How Large is the Core of Language. – *Corpus Linguistics* 2011 [online], Birmingham. Available at: https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2011/Paper-145.pdf.

Daille Béatrice 2001: Qualitative terminology extraction: Identifying relational adjectives. – *Recent advances in computational terminology*, eds. D. Bourigault, C. Jacquemin, M.-C. L'Homme, Amsterdam: John Benjamins Publishing Co, 149–166.

Dokulil Miloš 1962: *Teorie tvoření slov*, Praha: Nakladatelství ČSAV.
Evert Stefan, Lüdeling Anke 1991: Constraining psycholinguistic models of morphological processing and representation: the role of productivity. – *Yearbook of morphology 1991*, eds. G. Booij, J. van Marle, Dordrecht: Kluwer Academic Publishers, 165–183.
Evert Stefan, Baroni Marco 2007: *The zipfR library: Words and other rare events in R. Presentation at useR!* 2006: The Second R User Conference, Vienna, Austria. Available at: https://zipfr.r-forge.r-project.org.
Evert Stefan 2004: A simple LNRE model for random character sequences. – *Proceedings of JADT 2004*, 411–422.
Fernández-Domínguez Jesús 2013: Morphological Productivity Measurement: Exploring Qualitative versus Quantitative Approaches. – *English Studies* 94, 4, 422–447.
Garabík Radovan, Gianitsová Lucia, Horák Alexander, Šimková Mária 2004: *Tokenizácia, lematizácia a morfologická anotácia Slovenského národného korpusu*: internal document for manual morphological annotation. Unpublished. Available at: https://korpus.sk/attachments/publications/2004-garabik-gianitsova-horak-simkova-tokenizacia.pdf.
Horecký Ján 1956: *Základy slovenskej terminológie*, Bratislava: VEDA.
Horecký Ján 1999: Internacionalizácia a europeizácia slovenčiny. – *Internacionalizácia v súčasných slovanských jazykoch: za a proti*, ed. J. Bosák, Bratislava: Veda, 80–82.
Hulse Victoria 2011: *Productivity in morphological negation: a corpus based approach*, The University of Manchester. Available at: https://www.research.manchester.ac.uk/portal/en/theses/productivity-in-morphological-negation-a-corpusbased-approach(266d2241-a266-4b99-8fab-e19571381d8f).html.
Kvizhinadze Giorgi 2010: *Large number of rare events: Diversity analysis in multiple choice questionnaires and related topics*, Victoria University of Wellington doctor thesis. Available at https://core.ac.uk/download/pdf/41336663.pdf
Levická Jana 2021: Usage and empirical productivity of international adjectival suffixes in Slovak based on general and specialised corpora. – *Jazykovedný časopis* 72, 2 (in print).
L'Homme Marie-Claude 2002: Fonctions lexicales pour représenter les relations entre termes. – *Traitement automatique des langues* 43, 1, 19–41.
L'Homme Marie-Claude 2003: Adjectifs dérivés sémantiques dans la structuration des terminologies. – *Journées d'étude Terminologie, Ontologie et représentation des connaissance*, Lyon.
Maniez François 2002: Distinguer les termes des collocations: études sur corpus du patron < Adjectif – Nom > en anglais médical. – *Actes du colloque TALN de Nancy* 1, 24–27 juin 2002, 345–350.
van Marle Jaap 1992: The relationship between Morphological Productivity and Frequency: A Comment on Baayen's Performance-Orientated Conception of Morphological Productivity. – *Yearbook of Morphology 1991*, eds. G. Booij, J. van Marle, Dordrecht: Kluwer Academic Publishers, 151–163.
Nábělková Mira 1996: Variantnosť ako prvok dynamiky v adjektívnej paradigme. – *Slovenská reč* 61, 257–266.
Naccarato Chiara 2016: A corpus-based quantitative approach to the study of morphological productivity in diachrony: The case of samo-compounds in Russian. – *A Blend of MaLT Hanna Christ*, eds. D. Klenovšak, L. Sönning, V. Werner, Bamberg: University of Bamberg Press, 133–153.
Normand Sylvie, Bourigault Didier 2001: Analysing adjectives used in a histopathology corpus with NLP tools. – *Terminology* 7, 2, 155–164.
Plag Ingo, Dalton-Puffer Christiane, Baayen Harald 1999: Morphological productivity across speech and writing. – *English Language and Linguistics* 3, 2, 209–228.
Säily Tanja 2018: Change or variation? Productivity of the suffixes -ness and -ity. – *Patterns of Change in 18th Century English. A Sociolinguistic Approach*, eds. T. Nevalainen, M. Palander-Collin, T. Säily, Amsterdam: John Benjamins, 197–218.
Ševčíková Magda 2014: Zjišťování slovotvorné produktivity z korpusových dat: přípony odvozující názvy vlastností. – *Naše reč* 97, 228–240.
Štícha František 2012: Jak v epoše elektronických korpusů následovat Miloše Dokulila (Miloši Dokulilovi ke stému výročí narození). – *Jazykovědné aktuality* 49, 95–107.
Štícha František 2002: K Dokulilovu pojmu slovotvorné produktivity (z hlediska korpusové analýzy). – *Čeština doma a ve světě* 4, 302–310.
Štícha František 2007: Korpusové statistiky a slovotvorná produktivita. – *Grammar & Corpora/Gramatika a korpus 2005*, eds. F. Štícha, J. Šimandl, Praha: Academia, 250–257.
Štícha František 2009: Slovotvorná produktivita a gramatičnost: gradační expresivní adjektiva s prefixy pra-, pře- a vele- v současné psané češtině. – *Eslavística Complutense* 9, 145–170.

NAUJAS ŽVILGSNIS Į TARPTAUTINIŲ BŪDVARDŽIŲ PRIESAGŲ VARTOJIMĄ IR EMPIRINĮ PRODUKTYVUMĄ SLOVAKŲ KALBOJE

Santrauka

Šio straipsnio objektas – morfologinis penkių lotynų kalbos priesagų, dažnai vartojamų slovakų kalbos būdvardžiams sudaryti, produktyvumas. Straipsnyje aprašomas tyrimas yra ankstesnio tekstynų duomenimis paremto tyrimo tęsinys. Tyrime buvo nustatyti reikšmingi priesagų vartojimo ir produktyvumo skirtumai skirtingose srityse. Analizei atlikti naudojami šeši tekstynai ir patekstyniai apima ir bendrąjį tekstyną, ir specialiuosius tekstynus (medicinos, teisės, ekonomikos ir ypač religijos srities). Ankstesnės statistinės analizės atspirties taškas buvo neologizmų dalis hapakso lemų grupėje su analizuojamomis priesagomis. Klausimas, ar tarp retai vartojamų lemų yra neologizmų, liko atviras. Beje, keli tyrėjai teigia, kad produktyviąsias priesagas turinčių žodžių dažnumo pasiskirstymas turėtų būti nukreiptas į retai vartojamas lemas, sudarančias naujažodžius. Tokiu atveju statistinis morfologinio tiriamų priesagų produktyvumo įvertinimas galėtų būti kitoks. Be to, ankstesnės analizės rezultatai negalėjo būti palyginti tarp skirtingų tekstynų (patekstynių), nes jie rėmėsi pirminiais duomenimis, labai priklausomais nuo tekstyno (patekstynio) dydžio. Todėl antrasis šio tyrimo tikslas buvo nustatyti tinkamą statistinį metodą, kurį taikant toks palyginimas būtų įmanomas.

Tiek ankstesnis, tiek dabartinis analizės etapai rėmėsi rankiniu būdu ištrinamais tekstyno duomenimis, t. y. lemų vartojimo dažnumo sąrašais, kuriuose nėra nei „tekstyno triukšmo“, nei bendrinės slovakų kalbos žodynuose jau esančių žodžių. Straipsnyje pateiktos analizės rezultatai rodo, kad tarp tekstyne rastų retai vartojamų lemų yra ir neologizmų, ir nors jų pasiskirstymas tekstyno duomenyse gali įvairuoti, jie gali pakeisti produktyvumo eiliškumą pagal hapakso žodžių santykį su visais žodžiais (angl. *hapax / type formula*).

Antroje analizės dalyje buvo identifikuojami ir taikomi statistiniai modeliai, vadinami LNRE (angl. *large number of rare event*s, liet. *didelis retų įvykių skaičius*) pasiskirstymais. Naudojantis atvirojo kodo įrankiu, buvo sukurtas vaizdinis priesagos produktyvumo modelis, kuriame atsiskleidžia analizuotų priesagų skirtumai, pirma, tarp mokslinių ir bendrųjų tekstų ir, antra, tarp specialiųjų sričių tekstų. Taigi, šis modelis įrodo realų kiekvienos analizuotos priesagos potencialą sudaryti naujus žodžius atitinkamose srityse.

Gauta 2021-07-08

Jana Levická
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Panská 26
811 01, Bratislava, Slovakia
E-mail janal@korpus.sk