Frequency of Words and Their Forms in Contemporary Slovak Language Based on the Slovak National Corpus

Mária Šimková and Miroslav Ľos

Ľ. Štúr Institute of Linguistics,
Slovak Academy of Sciences, Bratislava, Slovakia

Abstract. Highest ranking words, word forms and n-grams based on their absolute frequencies are presented and compared with respect to the size, time of conception and stylistic focus of the source corpora. Corpus coverage by the most frequent words is discussed and distribution of words among morphological categories is analyzed. The dependencies of word use on style and genre of text are uncovered, utilizing the morphological tagging available in the Slovak National Corpus.

1 Introduction

The development of Quantitative (Statistical) Linguistics has seen more than 100 years since publication of the first European frequency dictionary (F. W. Kaeding: Häufigkeitswörterbuch der deutschen Sprache, 1897), and more than forty years since the first computer-aided compilation of language statistics from an electronic corpus was presented (H. Kučera & W. N. Francis: Computational Analysis of Present-Day American English, 1967). Many more languages have since been augmented with a frequency dictionary and/or statistical characteristic of their own, satisfying demand coming not only from disciplines such as NLP, (both monolingual and translation) lexicography, first and second language education, logopeadics, neurology, psychology etc., but also from language hobbyists interested in various quantitative properties of particular language units as well as the language itself. The existence of multiple voluminous corpora and the tools to process them allow for rapid collection of such statistics, often inspiring further research.

2 Word, word form and n-gram frequencies

J. Mistrík ([1] 1969, [2] 1985) manually analyzed word, word form and language construct frequencies in Slovak language, based on a text of 1 million proper words (i. e. without punctuation and any other non-word tokens). Some of his findings have been confirmed by data from the Slovak National Corpus[1], whose latest main corpus – prim-4.0 – made available in early 2009, counts about 550 million tokens. For example, the five lemmas (*a, v, na, sa, byť*) are consistently the top five, or among the top ten, highest-ranking lemmas across all its specializations, even in the spoken corpus.

[1] http://www.korpus.sk

Corpus	FSS 1969	prim1 2004	prim-4.0 2009	s-hovor 2009
Size	1 mil. w	200 mil. t	550 mil. t	434,676 t
1.	a	byť	byť	byť
2.	byť	v	a	to
3.	sa	a	v	a
4.	v	sa	sa	že
5.	na	na	na	sa
6.	on	ten	to	tak
7.	ten	ktorý	ktorý	ja
8.	že	s	s	v
9.	z	z	že	na
10.	ako	že	z	no

Table 1. Ten highest-ranking lemmas in general corpora and in the spoken corpus

Individual style of some authors, or smaller, specific texts, utilizing some characteristic language constructs (cf. Šimková 2008 [3]), could present statistics skewed from those obtained from general corpora. We created three corpora of distinct size and specialization: journalism (one month of the newspaper Smena), nonfiction (selection of texts regarding belief) and fiction (works of A. Habovštiak). The top rankings of lemmas from these corpora were essentially consistent with the general corpora.

Corpus	Smena 1968	belief	Habovštiak
Size	296,917 t	14,697,297 t	1,088,695 t
1.	a	a	sa
2.	v	byť	a
3.	byť	sa	byť
4.	sa	v	on
5.	na	na	na
6.	že	ktorý	v
7.	to	on	že
8.	ktorý	že	aj
9.	s	s	čo
10.	z	to	keď

Table 2. Ten highest-ranking lemmas in specialized corpora

Next to the top five lemmas, the corpora usually contain pronouns *ten* or *to*, *ktorý* (less often used as a conjunction), conjunction *že*, and prepositions *s*, *z*. Fiction works stand apart from other texts, for example in the higher use of the word *on* (he) both in the Habovštiak corpus and the corpus of Mistrík, mainly consisting of fiction.

Also notable is the higher use of the conjunction and particle *aj* (too, also, even) and the diminished conjunctive role of the pronoun *ktorý* (which), that has likely been partially replaced by the pronouns *čo* (that) and even *keď* (as, when), suggesting here a fundamental difference in fiction vs. nonfiction authors' selection of sentence structures.

The four most frequent words in the corpus prim-4.0 after the verb *byť* (to be) are also present in the top ten of both bigrams (*a* one time, 3× *v*, 5× *sa* and 2× *na*) and trigrams (3× *v*, 2× *na* and *sa*). These top-tens also contain the most frequent substantive *rok* (year), ranking 9th among bigrams and 3rd or 4th (ranked by *na druhej strane* if counting ignoring character case) among trigrams. Analysis of the various texts shows that time units in general (e. g. *rok*, *týždeň*, *deň*, *hodina*) have substantial frequencies in all works independent of the work's style and genre.

Notable among the most frequent trigrams are secondary prepositions (ranking 2nd and 4th, respectively), which is likely a result of the high proportion of journalistic texts in the subcorpus prim-4.0-public, the source of the list of bi- and trigrams in table 3.

Rank	Bigram	Frequency	Trigram	Frequency
1.	nie je	392,383	že je to	35,800
2.	sa v	315,831	v súvislosti s	35,021
3.	sa na	298,802	v tomto roku	32,553
4.	je to	295,003	v porovnaní s	27,489
5.	že sa	282,920	to nie je	26,740
6.	som sa	277,091	Na druhej strane	26,255
7.	by sa	258,065	na druhej strane	23,464
8.	a v	245,519	by som sa	21,807
9.	v roku	235,353	Nie je to	20,509
10.	na to	203,130	že by sa	20,411

Table 3. Ten highest-ranking bigrams and trigrams in prim-4.0-public

Whenever corpora differ in size by multiple orders of magnitude, the statistics collected on them naturally diverge in several parameters, such as the accumulated share of top-ranking lemmas (see Table 4).

Ranks	FSS	prim-4.0	Difference
1. – 10.	18.6 %	17.98 %	- 0.62
11. – 20.	6.4 %	5.74 %	- 0.66
1. – 20.	25.0 %	23.72 %	- 1.28
21. – 30.	4.0 %	3.51 %	- 0.49
1. – 30.	29.0 %	27.23 %	- 1.77
1. – 100.	41.5 %	37.81 %	- 3.69

Table 4. Corpus coverage by most-frequent words (excluding punctuation)

While according to J. Mistrík's Frequency Dictionary of Slovak Language (1969), the top one hundred words cover 41.5 % of the text (when discounting random or rare words with zero dispersion, this value rises to 56.13 %), this coverage is almost 4 % lower in prim-4.0. This can be explained by the rising number of random tokens as a corpus grows in size. It is further supported by lemmas from prim-4.0 with frequencies 1 or 2, which represent almost 66.5 % of all lemmas in the corpus, of which an estimated half are numerals, typos, words in foreign scripts or other symbols.

3 Frequencies of morphological categories

In the corpora based on written text, the first two placed are held by substantives and verbs (see Tab. 5). The 3rd place is held by prepositions in contemporary texts, in contrast with pronouns in the corpus of Mistrík, which may signify a higher level of abstraction in the discourse, perhaps as a consequence of the higher share of journalistic and professional works in prim-4.0. The ranks of adjectives, conjunctions, numerals and interjections is the same between these corpora; the ranks of the latter two are also shared with the spoken corpus. The generally low share of interjections in all corpora is nonetheless much higher in the spoken corpus (0.31 %) than in prim-4.0 (0.058 %). As was to be expected, verbs and pronouns take the first two places of the spoken corpus; moreover, the share of conjunctions and particles is quite higher. The statistics on adverbs shows across all corpora, that their potentiality (being derivable from both adjectives and substantives) in the language system is much higher than actual use.

Corpus	FSS 1969	prim-4.0 2009	s-hovor 2009
Size	1 mil. w	550 mil. t	434,676 t
1.	substantives	substantives	verbs
2.	verbs	verbs	pronouns
3.	pronouns	prepositions	substantives
4.	adjectives	adjectives	conjunctions
5.	prepositions	pronouns	particles
6.	conjunctions	conjunctions	prepositions
7.	adverbs	particles	adjectives
8.	particles	adverbs	adverbs
9.	numerals	numerals	numerals
10.	interjections	interjections	interjections

Table 5. Frequency of words by morphological category

The highest share of substantives classified by gender belongs to the masculine. Since the morphological annotation of the SNC distinguishes between animate and inanimate masculines (see `http://korpus.juls.savba.sk/usage/morpho/`), the order of genders thus split is instead as follows: 1st the feminine, 2nd inanimate masculine (i), 3rd animate masculine, and 4th the neuter. This order is consistent between works' styles and genres, with the exception of professional text (prf), having the animate masculine gender last, likely due to the generally abstract and impersonal style of such text (see Fig. 1).

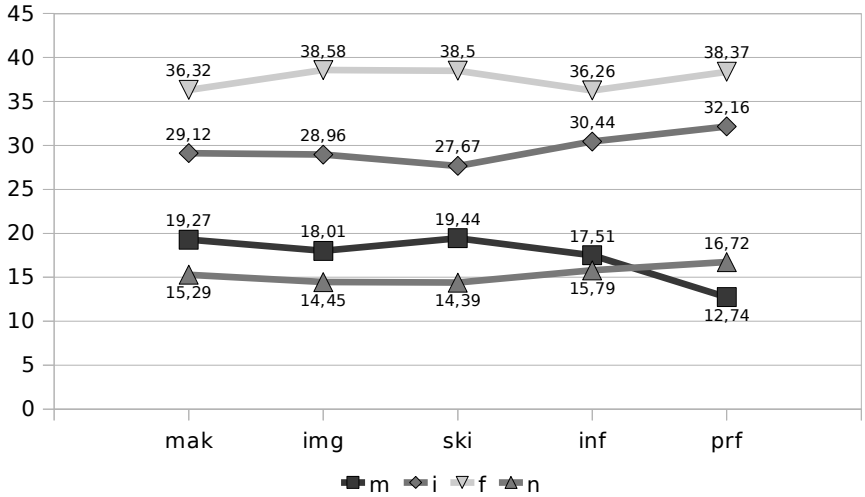

Fig. 1. Distribution of substantives by gender (%)

The manually morphologically annotated corpus (mak), used as reference in present analysis due to its negligible error rate, is closest in its gender distribution to the subcorpora of journalistic and fiction/artistic text (tagged inf and img, respectively). This reflects the bigger share these two types of texts have in mak (of the 1 207 939 tokens, 44.3 % are journalistic, 36.7 % fiction, and 19.0 % professional). The subcorpus ski contains all original text from img created by Slovak authors. Comparing it to its parent corpus, we see the same share for both feminine and neuter substantives, but a shift of some 1.3 percentage points in favor of animate masculines in the Slovak fiction corpus.

The distribution of substantives by number and case (Fig. 2) shows a clear preference towards singular forms, holding top four ranks in all main four corpora analyzed. The natural first is Nsg, the common 4th is Lsg. The second and third places belong to Gsg and Asg, with not so small differences ranging from 2.5 % points in mak to more than 7 % points in prf. In fiction, the accusative wins due to the higher usage of verb-object phrases. The journalistic and professional texts contain more substantives in genitive, which matches the higher use of genitive attributes in noun phrases.

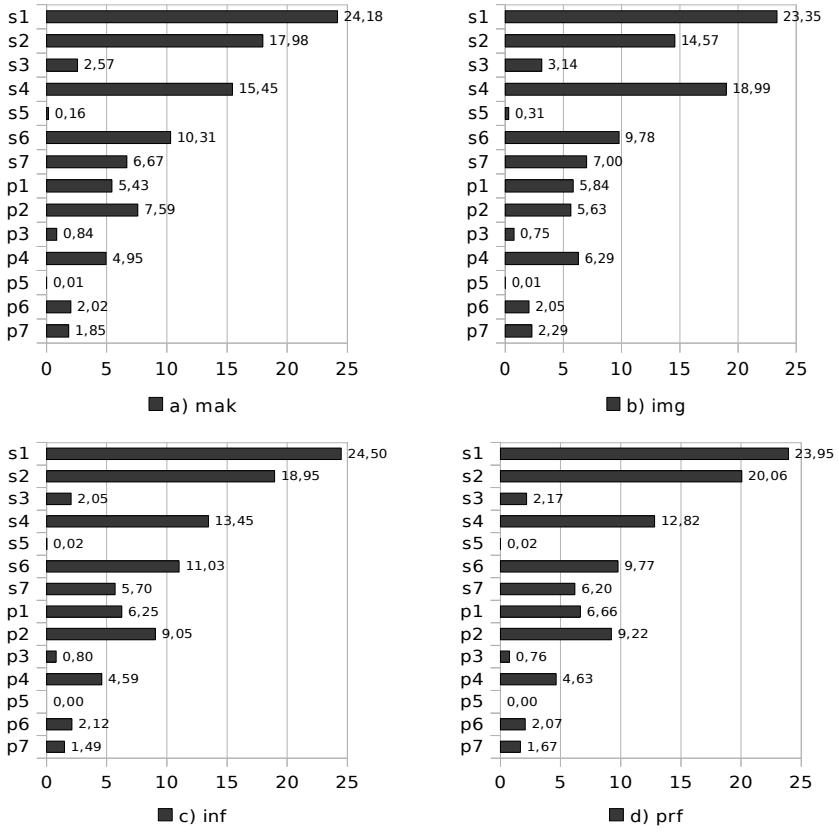

Fig. 2. Distribution of substantives by number and case (%)

Of plural forms, the genitive enjoys the highest usage (i. e. 5th place overall), most visibly again in journalistic and professional texts, followed in these by Npl by a margin of 3 % points. In fiction, Isg holds the 5th place, then followed by Npl and Apl. Gpl ranks 8th in fiction. The rarest case is dative, both in singular (9th overall, 10th in pub) and in plural (12th).

Places 9th–11th, containing Dsg, Lpl and Ipl., have only small differences among each other (0.5 % points in pub and prf to about 1 % point in img). A category of its own belongs to the vocative, regarded as just a relic in contemporary Slovak language system due to its near absolute homonymy with nominative. The few exceptions (e. g. *otče, majstre, bratu, priateľu*) represent a negligible part of the corpus, mostly only present in fiction (0.31 % in sg and 0.01 % in pl).

The distribution of prepositions by case (Fig. 3) is naturally distinct from that of the substantives, as the most frequent nominative completely lacks prepositions, and both genitive and accusative are often used without any preposition. The first place is thus taken by locative prepositions, mainly used by journalistic and professional works. The frequencies of genitive and accusative prepositions, chiefly used in fiction, are rather close to each other. Fiction works also use instrumental and dative prepositions more heavily, to express predicate complements.

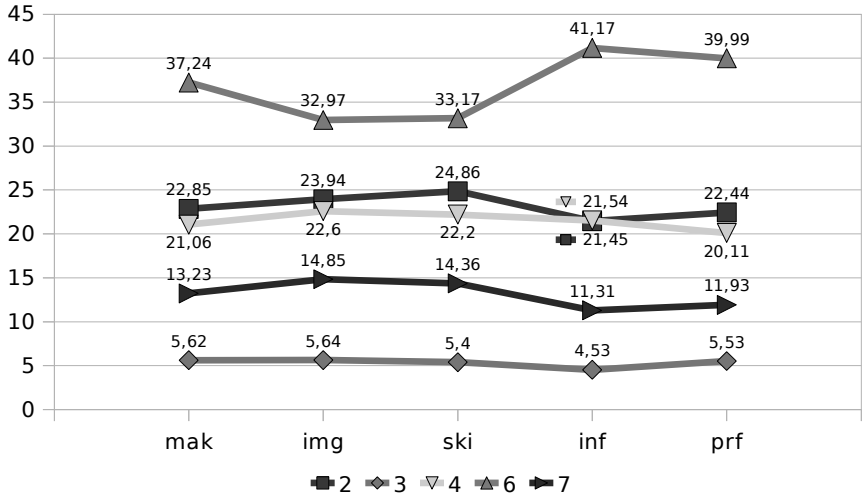

Fig. 3. Distribution of prepositions by case (%)

From the basic verb forms recognized by the morphological annotation in the SNC (i. e. B – future of *byť*, H – gerund, I – infinitive, K – indicative, L – l-participle, M – imperative), the l-participle (i. e. past tense and conditionals), the present indicative and the infinitive are the most frequent (Fig. 4). The other three forms are very specific, with only little applicability. The indicative is the most frequent in journalistic and professional text (represented by up to one half of all verbs in the subcorpus), while the l-participle is used the most in fiction (again nearly 50 % of all verbs).

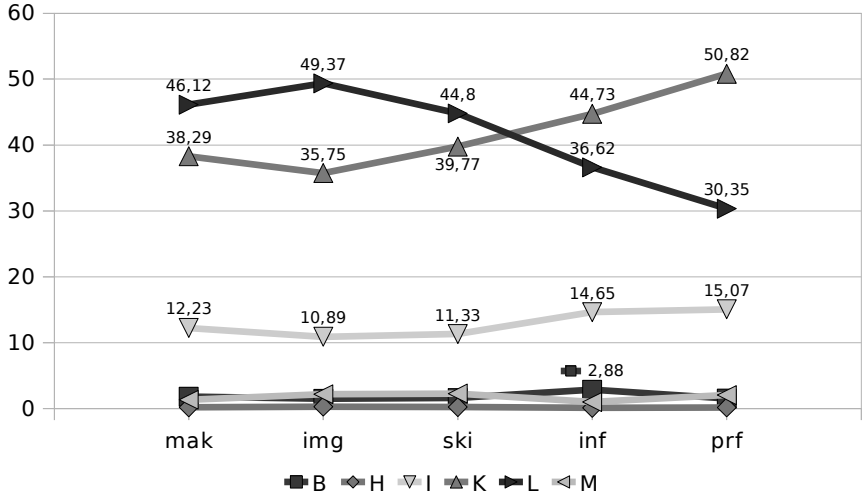

Fig. 4. Distribution of verbs by form (%)

The division of verbs by person-agreement is dominated by the third person singular (Fig. 5), having more than 50 % share across all corpora except Slovak fiction, where they are just below. In fiction, the second place is held by 1st person sg, which drops sharply in other styles towards almost zero in professional text. Journalistic and professional text has instead established 3rd person pl. as the second. Considerable is also the share of the indeterminate verb forms (wrt. number and person), marked under '--' in the chart, where the person is concealed, e. g. due to omitted subject pronoun (e. g. *Prišiel*. – ja, ty, on?; *Prišli*. – my, vy, oni?).

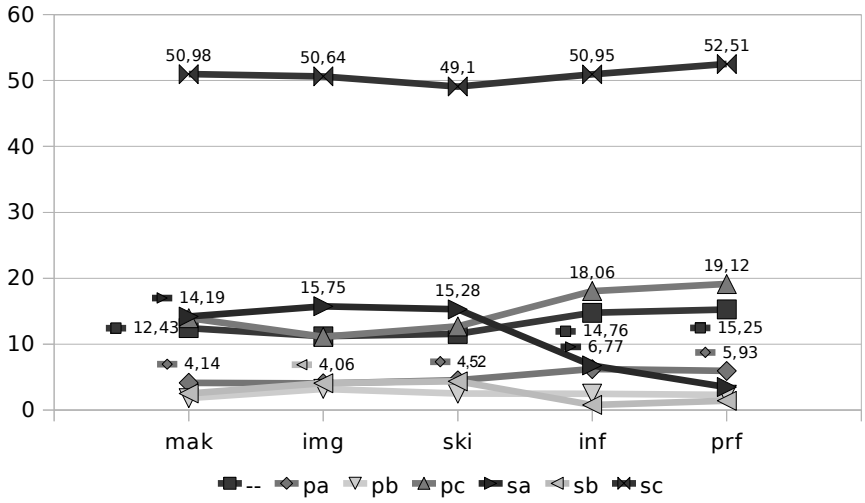

Fig. 5. Distribution of verbs by number and person (%)

4 Conclusion

After six years of building the primary database of Slovak National Corpus (made of texts originating in years 1955 – 2009), its quantitative (550 million tokens), as well as qualitative (i. e. style and genre structure, lemmatization and morphological tagging) properties allow for quantitative analysis with significant results. Much data can also be compared to observations of Slovak language made available 40 years ago, or analogous latest results from other languages.

Linguistic interpretation of frequency-based indicators of select lexical categories and word forms present in contemporary Slovak language presents one of the first contributions to this discipline based on the SNC. We have researched absolute frequencies of presented phenomena based on all written text in the corpus, as well as their distribution among the three main styles: fiction/artistic, journalistic and professional. Even the partial results published in present paper signal a marked tendency of the journalistic and professional styles to merge. We also contrasted some of the results with data from the early Slovak Spoken Corpus, which is too small at the moment to have reliable conclusions drawn from.

Detailed information on frequency-based parameters of contemporary Slovak language will be published in the Frequency dictionary currently in preparation at the Slovak National Corpus.

References

[1] Mistrík, Jozef: Frekvencia slov v slovenčine. Bratislava: Vydavateľstvo SAV 1969. 728 p.

[2] Mistrík, Jozef: Frekvencia tvarov a konštrukcií v slovenčine. Bratislava: Veda 1985. 320 p.

[3] Šimková, Mária: Jazykové prostriedky vo vybraných dielach Martina Rázusa (Analýza na báze textov Slovenského národného korpusu). In: Martin Rázus – politik, spisovateľ a cirkevný činiteľ. Ed. M. Pekník. Bratislava: Veda 2008, p. 238 – 249.