Ján MAČUTEK
Michaela KOŠČOVÁ
Slovak Academy of Sciences
Emmerich KELIH
University of Vienna
Radek ČECH
University of Ostrava

DOI: 10.14746/bo.2023.1.7

Frequency and morphological behaviour of nouns in Czech and Russian[1]

Keywords: Slavic languages, Czech, Russian, declension, gender, animacy, case

Abstract

Declensional morphology of nouns in Czech and Russian is investigated and compared. It is shown that, in general, word forms which are more similar to their lemmas are preferred, but there are differences between animate and inanimate nouns and also among grammatical genders. The frequency distribution of grammatical cases is also studied, with animacy and gender being again important factors.

1. Introduction

The paper focuses on morphology of nouns in two Slavic languages (Czech and Russian). Specifically, the relation between frequency of word forms and their difference from their lemmas is investigated.

Nouns in Slavic languages (with the exception of Bulgarian and Macedonian) possess quite a rich declensional morphology (Comrie and Corbett, 1993, p. 6), i.e., different grammatical cases are expressed mainly by adding inflectional endings (desinences) to the stem. The endings do not provide only information on grammatical case, but also on grammatical gender[2] and number[3]. Therefore, Slavic languages are typologically ranked among fusional languages (one suffix denotes more than one morphological category). On the other hand, case syncretism, i.e. one inflected word form corresponding to more than one case, can be observed in these languages (Gvozdanović, 2009; Hentchel and Menzel, 2009). They use six or seven cases (nominative, genitive, dative, accusative, locative, instrumental; some of them also vocative; cf. *Mluvnice češtiny* 1986).[4]

Sometimes, also some morpho-phonetic alternations in the stem can be observed, such as elisions, e.g. *pes* (‘dog’, nom.sing.), *psa* (gen.sing.), or alternations, e.g. *Praha* (‘Prague’, nom.sing.), *Praze* (loc.sing.).

We use two parallel texts (a part of a Russian novel and its Czech translation) to gain some insight into the variability of word forms. Only nouns in singular are investigated. Nominative singular is considered the word lemma, then the difference between the lemma and other word forms is evaluated. We show that the frequency behaviour depends on grammatical gender, and that animacy plays an important

[1] Supported by research grants APVV SK-AT-20-0003 (J. Mačutek, M. Koščová, E. Kelih), VEGA 2/0096/21 (J. Mačutek, M. Koščová), and APVV-21-0216 (J. Mačutek).

[2] All Slavic languages distinguish three grammatical genders (masculine, feminine, and neuter). The category of animacy (which is sometimes described as a subgender) is also of considerable importance. A short overview of the interaction among gender, animacy, and case can be found in Comrie and Corbett (1993, pp. 16–17). The grammatical categories of gender and animacy in Slavic languages are introduced in Doleschal (2009) and Klenin (2009), respectively.

[3] All Slavic languages use singular and plural; Slovene, Lower Sorbian and Upper Sorbian preserved also dual (Hentschel and Menzel, 2009). Traces of dual can be found also in other languages, see e.g. examples in Meyer (1973).

[4] In some Slavic languages, vocative practically disappeared. It survives only in a few word forms and mostly in specific contexts, such as e.g. in prayers or fairy tales. See Hentschel and Menzel (2009) for a detailed discussion on cases in Slavic languages.

role. Gender and animacy are factors which influence also frequency distributions of grammatical cases. We note that we work with tokens, i.e. each occurrence of a word is counted (as opposed to types, which would consider only different words).

This study can be seen as a follow-up of the paper by Mačutek and Čech (2013). Here we apply an improved, fully algorithmized method for the evaluation of the differences between word forms and lemmas. We also add the analysis of another Slavic language, namely Russian.

2. Methodology and language material

The size of a difference between a word form and its lemma is quantified using the Levenshtein distance (LD henceforward), which is a measure of difference (or similarity) between two character strings. It was introduced by Levenshtein (1965); see also Deza and Deza (2009, p. 202).[5] The LD between two words is the minimum number of single-character deletions, additions, or substitutions needed to transform one word into the other. It is implemented in the statistical software environment *R* which we used for data analysis. In Table 1 we present several examples from Czech.

Table 1. Levenshtein distances between some word forms and lemmas in Czech

Word lemma	Word form	LD	Operations
škola ('school', nom.sing.)	*škole* (dat.sing.)	1	one substitution
pes ('dog', nom.sing.)	*psa* (gen.sing.)	2	one deletion, one addition
táta ('dad', nom.sing.)	*tátovi* (dat.sing.)	3	one substitution, two additions
pes ('dog', nom.sing.)	*psovi* (dat.sing)	4	one deletion, three additions

[5] Deza and Deza (2009) spell it "Levenstein distance".

In this paper, as announced above, we analyze only nouns in singular.[6] Plural of several frequently used nouns presents problems, as is shown by the following Czech examples. It can be irregular (*člověk* 'man/human' – *lidé* 'men/humans'), which results in very high values of LDs. There are also some words with two word forms for plural (e.g. *muž* 'man/husband' – *muži* or *mužové* 'men/husbands'). One must then choose one of them as the basic form for plural, although both are considered correct. There are similar problems also in Russian.

The first ten chapters from the Russian novel *Kak zakaljalas' stal'* ('How the Steel Was Tempered') by N. Ostrovsky (both the Russian original and its Czech translation) served as the source of data. The texts were annotated using TreeTagger (Benko, 2014) and by UDPipe (Straka, 2018), respectively. The choice of the language material was motivated by the fact that there is a parallel corpus of the novel translations into almost all standard Slavic languages (with the exception of Lower Sorbian), see Kelih (2009). In future it will thus be possible to compare declensional morphology in these languages.

3. Results

In Tables 2 and 3, M, F, and N stand for gender (masculine, feminine, and neuter); A and I denote animate and inanimate nouns, respectively. In Czech, animacy is annotated only for masculine gender, as only there it has an impact on declensional morphology (see Short, 1993, p. 465). In Russian, animacy plays a role in all three genders (see Timberlake, 1993, p. 837). However, only three animate neuter nouns occur in the Russian text, therefore we merged all Russian neuter nouns into one category. Tables 2 and 3 (and Figures 1 and 2) present frequencies of differences (expressed in terms of the LD) between word forms and lemmas in Czech and Russian, respectively.

[6] Mačutek and Čech (2013) used also nouns in plural.

Table 2. Absolute and relative frequencies of differences between word forms and lemmas in Czech

LD	M(A)		M(I)		F		N		All words	
0	640	0.58	1156	0.42	1259	0.33	843	0.62	3898	0.44
1	265	0.24	936	0.34	1938	0.51	424	0.31	3563	0.41
2	144	0.13	596	0.22	544	0.15	89	0.07	1243	0.14
3	58	0.05	51	0.02	41	0.01	5	.01	155	0.01
4	4	.01	10	.01	7	.01	1	.01	22	.01
5			1	.01	1	.01			2	.01

Table 3. Absolute and relative frequencies of differences between word forms and lemmas in Russian

LD	M(A)		M(I)		F(A)		F(I)		N		All words	
0	1504	0.68	1446	0.48	488	0.70	816	0.26	640	0.56	4894	0.48
1	500	0.23	1064	0.35	151	0.22	1961	0.63	0.43	0.43	4167	0.41
2	190	0.09	492	0.16	33	0.05	334	0.11	4	.01	973	0.10
3	26	0.01	30	0.01	29	0.04			14	0.01	99	0.01
4									1	.01	1	.01

There are remarkable similarities and differences both within and between the two languages. First, comparing Czech and Russian, relative frequencies of the differences between word forms and lemmas for all words are similar in both languages. The same is true for masculine animate as well as for masculine inanimate nouns (and, to a lesser extent, also for neuter nouns). As the annotation does not provide information on animacy for feminine nouns in Czech, we are not able to compare the two languages in detail. However, if we merge feminine nouns in Russian (i.e. considering them as one category regardless of their animacy), we again observe relative frequencies quite similar as in Czech.

In Russian, animate nouns strongly prefer their basic forms with *LD*=0 (roughly 70% of both masculine and feminine nouns). Masculine inanimate and neuter nouns also prefer the basic form, but less strongly, and feminine inanimate is the only category where the basic form is not the most frequent.

Czech and Russian can be said to behave similarly with respect to the frequency distribution of the differences between word forms and lemmas. We now shift our attention to the frequency behaviour of grammatical cases. The distributions are presented in Tables 4 and 5.

Table 4. Absolute and relative frequencies of grammatical cases in Czech

Case	M(A)		M(I)		F		N		All words	
nom.	647	0.58	596	0.22	1041	0.2	268	0.20	2552	0.30
gen.	99	0.09	668	0.24	783	0.21	354	0.26	1304	0.15
dat.	86	0.08	116	0.04	114	0.03	41	0.03	457	0.05
acc.	140	0.13	638	0.23	958	0.25	352	0.26	2088	0.25
voc.	19	0.02	4	.01	0	0	0	0	23	.01
loc.	25	0.02	369	0.13	472	0.12	208	0.15	1074	0.13
ins.	95	0.09	359	0.13	422	0.11	139	0.10	1015	0.12

Table 5. Absolute and relative frequencies of grammatical cases in Russian

Case	M(A)		M(I)		F(A)		F(I)		N		All words	
nom.	1496	0.68	655	0.22	473	0.67	647	0.21	247	0.21	3520	0.34
gen.	347	0.16	666	0.22	60	0.09	631	0.20	233	0.20	1931	0.19
dat.	112	0.05	137	0.05	46	0.07	145	0.05	69	0.06	509	0.05
acc.	164	0.07	872	0.29	71	0.10	934	0.30	344	0.30	2385	0.23
voc.	3	.01	0	0	2	.01	0	0	0	0	5	.01
loc.	13	0.01	316	0.10	4	0.01	331	0.11	110	0.10	774	0.05
ins.	85	0.04	386	0.13	45	0.06	432	0.14	144	0.13	1084	0.11

The relative frequencies for all words are again not too different (Czech uses locative more frequently, while in Russian there are higher proportions of nominative and genitive). In Russian, there are striking similarities in frequencies of nominative for animate nouns on the one hand, and inanimate and neuter on the other. Nominative clearly dominates among animate noun cases, whereas the frequency distribution is more uniform for inanimate (with accusative being the most frequent). While masculine animate and feminine animate nouns differ in frequencies of other cases (see especially in genitive), mascu-

line inanimate, feminine inanimate, and neuter[7] form one homogenous group. The relative frequencies of Czech masculine inanimate and neuter nouns are very similar as well. Thus, it seems that animacy is a decisive factor which shapes the distribution of cases.

4. Conclusion and discussion

Our results reveal an intrinsic regulation of the morphology of word forms, with a strong tendency towards forms more similar to their basic forms. The difference between a word form and its lemma correlates negatively with frequency in both Czech and Russian – word forms which differ more from the lemma occur less often.[8]

This finding can be interpreted as another manifestation of the least effort principle (Zipf, 1949) – a word form is the easier to reproduce the more similar to its lemma it is (e.g. Levelt et al., 1999, present a model according to which lemmas are retrieved from memory when one speaks or writes, and the lemmas are then morphologically encoded). Interestingly enough, the subgender of animacy seems to be crucial, whereas gender is of secondary importance.

The same is true also for the frequency distribution of cases. For animate nouns, nominative is by far the most frequent case, which means that they tend to be subjects. On the other hand, inanimate nouns occur most frequently in accusative, and thus they are often objects.

These observations lead to further tentative formulations of hypotheses which will be addressed in future research. First, inanimate nouns occur in nominative (which corresponds to the lemma) less frequently than animate. Therefore, in order to keep the production of their word forms "cheap", their declension paradigms should be less complicated (e.g. nominative and accusative are expressed by the

[7] We remind that all but three neuter nouns in the text under analysis are inanimate.

[8] This is true for a text taken as a whole, but not for all particular word lemmas.

same word form). Second, less frequent nouns are more difficult to retrieve from memory. It follows that their declension paradigms should be simple – thus, that the cognitive effort needed to proces them is kept relatively low. But it is well known that less frequent words are longer. We therefore expect shorter words to have a more complicated declension paradigms than longer ones. Finally, if animate nouns can have a more complicated morphology of word forms, and the same is true for shorter words, we allow ourselves to formulate a hypothesis that animate nouns are on average shorter than inanimate.

Needless to say, similar research must be conducted on other (not only Slavic) languages with a relatively rich declension morphology before one can make a decision on the validity of these hypotheses.

References

Benko, V., 2014, Aranea: Yet another family of (comparable) web corpora. In: P. Sojka, A. Horák, I. Kopeček & K. Pala (eds.), *Text, Speech and Dialogue*, Cham: Springer, pp. 247–254.

Comrie, B., Corbett, G.G., 1993, Introduction. In: B. Comrie & G.G Corbett (eds.), *The Slavonic Languages,* London, New York: Routledge, pp. 1–9.

Deza, M.M., Deza, E., 2009, *Encyclopedia of Distances*, Berlin, Heidelberg: Springer.

Doleschal, U., 2009, Nominale Kategorien: Genus. In: Kempgen, S., Kosta, P., Berger, T. & Gutschmidt, K. (eds.), *Die slavischen Sprachen. Ein internationales Handbuch zu ihrer Geschichte, ihrer Struktur und ihrer Erforschung. Band 1*, Berlin, New York: de Gruyter, pp. 143–152.

Gvozdanović, J., 2009, Synthetismus and Analytismus im Slavischen. In: Kempgen, S., Kosta, P., Berger, T. & Gutschmidt, K. (eds.), *Die slavischen Sprachen. Ein internationales Handbuch zu ihrer Geschichte, ihrer Struktur und ihrer Erforschung. Band 1*, Berlin, New York: de Gruyter, pp. 129–142.

Hentschel, G., Menzel, T., 2009, Nominale Kategorien: Kasus. In: Kempgen, S., Kosta, P., Berger, T. & Gutschmidt, K. (eds.), *Die slavischen Sprachen. Ein internationales Handbuch zu ihrer Geschichte, ihrer Struktur und ihrer Erforschung. Band 1*, Berlin, New York: de Gruyter, pp. 161–176.

Kelih, E., 2009, Slavisches Parallel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In E. Kelih, V. Levickij & G. Altmann (eds.), *Methods of Text Analysis*, Chernivtsi: ČNU, pp. 106–124.

K l e n i n, E., 2009, Animacy, personhood. In: Kempgen, S., Kosta, P., Berger, T. & Gutschmidt, K. (eds.), *Die slavischen Sprachen. Ein internationals Handbuch zu ihrer Geschichte, ihrer Struktur und ihrer Erforschung. Band 1*, Berlin, New York: de Gruyter, pp. 152–161.

L e v e l t, W.J.M., R o e l o f s, A., M e y e r, A.S., 1999, A theory of lexical access in speech production. *Behavioral and Brain Sciences* 22(1), pp. 1–38.

Levenshtein, V.I., 1965, Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10(8), pp. 707–710.

M a č u t e k, J., Č e c h, R., 2013, Frequency and declensional morphology of Czech nouns. In: I. Obradović, E. Kelih (eds.), *Methods and Applications of Quantitative Linguistics*, Belgrade: Akademska Misao, pp. 59–68.

M e y e r, G.L., 1973, Common tendencies in the syntactic development of "two", "three", and "four" in Slavic. *The Slavic and Eastern European Journal* 17(3), pp. 308–314.

Mluvnice češtiny, 1986, díl 2. *Tvarosloví*, 1986, Praha: Academia.

S h o r t, D., 1993, Czech. In: B. Comrie & G.G Corbett (eds.), *The Slavonic Languages,* London, New York: Routledge, pp. 455–532.

S t r a k a, M., 2018, UDPipe 2.0 prototype at CoNLL 2018 UD Shared Task. In: D. Zeman, J. Hajič, M. Popel, M. Potthast, M. Straka, F. Ginter, J. Nivre & S. Petrov (eds.), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Stroudsburg (PA): ACL, pp. 197–207.

T i m b e r l a k e, A., 1993, Russian. In: B. Comrie & G.G Corbett (eds.), *The Slavonic Languages,* London, New York: Routledge, pp. 827–886.

Z i p f, G.K., 1949, Human Behavior and the Principle of the Least Effort. Cambridge (MA): Addison-Wesley Press.