Slovenská akadémia vied
Jazykovedný ústav Ľudovíta Štúra

NLP, Corpus Linguistics, Corpus Based Grammar Research

Fifth International Conference
Smolenice, Slovakia, 25–27 November 2009
Proceedings

Editors
Jana Levická
Radovan Garabík

Tribun

2009

Slovak National Corpus
Ľ. Štúr Institute of Linguistics
Slovak Academy of Sciences
Bratislava, Slovakia 2009
http://korpus.juls.savba.sk/-slovko/

Table of Contents

Inflectional Entropy in Slovak

Adriana Hanulíková[1] and Doug. J. Davidson[2]

[1] Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands
[2] Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Abstract. Statistical measures of word frequency are used in psycholinguistic research to characterize the psychological organization of the mental lexicon, and the processes of retrieving, understanding, and learning words. More recently, researchers have calculated statistics from corpora to gain insights into processing of morphology, based on previous work on Serbian by A. Kostic´ and colleagues. One such statistical measure - the inflectional entropy - has been shown to explain processing costs in word recognition experiments. The inflectional entropy of a word form is the amount of information carried by that inflected form, relative to the statistical distribution of its inflectional paradigm. In this work, we investigate whether it is possible to calculate measures like inflectional entropy for Slovak using the Slovak National Corpus (SNK). This would allow us to compare Slovak with other Slavic languages such as Serbian. The results will be useful for a wide variety of psycholinguistic investigations of comprehension or production of Slovak.

1 Introduction

Many psycholinguistic investigations have shown that the probability of a word has a strong influence on measures of performance (for a recent review see Balota, Yap, & Cortese, 2006). This is true for a wide variety of tasks, such as word recognition, judgement tasks, or picture and word naming. For example, one of the most commonly-used tasks is the lexical decision task. In this task, the time it takes to judge whether a singly-presented word occurs in a language is measured. Response times in this task are faster for more common words relative to less common words (Whaley, 1978). Since a Slovak word like ‘škola’ (book) is used more often than a word like ‘pštros’ (ostrich), lexical decision times should be shorter for ‘škola’.

For the purposes of psycholinguistic studies, the probability (Pr) of a word (w) is often approximated, as in Equation 1, by estimating its unigram frequency count $F(w)$ in a sample of text or speech of size N (Baayen, 2001). These counts are typically derived from non-annotated corpora, which do not provide information about grammatical classes or functions of the individual words.

$$Pr_w = F(w)/N \tag{1}$$

However, more recently researchers have incorporated variables related to morphosyntactic variation in the frequency estimates of words, based on annotated corpora (for review see Milin, Kuperman, Kostic´, & Baayen, in press). This is especially important for Slavic languages, which have richer inflectional morphology than the more-commonly studied West Germanic languages, and thus require more complex probability models. In particular, work on Serbian by A. Kostic´ and colleagues has been instrumental in demonstrating the influence of the inflectional form of a word on lexical decision performance. Since this framework is the point of departure for the present paper on Slovak, we will review some of their findings and conceptual distinctions here.

Kostic´ (1991, 1995) found that the relative frequency of an inflected form within a paradigm, as well as the number of grammatical functions or meanings of a word, was correlated positively with lexical decision times for Serbian nouns. Their measures were based on information theory, quantifying the amount of information that an inflectional suffix provides, relative to its paradigm. More recently, Moscoso del Prado Martín, Kostic´, and Baayen (2004) found that lexical decision times for Dutch nouns were positively correlated with inflectional entropy. Inflectional entropy increases in a paradigm when there are more inflectional variants possible, and/or when the variants have similar probabilities. The key observation of this previous work is that the statistical distribution of word forms within an inflectional paradigm can be factored into two parts: The contribution provided by the stem, and the contribution conveyed by the exponent (i.e., suffix). This is illustrated below in Table 1, which shows a probability model for the Slovak feminine noun 'škola' (school), constructed in a similar way to Milin *et al.* (2009, in press). The columns provide information on the surface frequencies $F(w_e)$ (per million) and surface relative proportions $Pr_\pi(w_e) = F(w_e)/F(w)$, where $F(w)$ is the sum of all $F(w_e)$.

w_e	*F*(w_e)	*Pr(We)*	/ » e	*F(e)*	Prv(e)	I_e
škol-0	211	0.09	3.55	99396	0.11	3.25
škol-*a*	197	0.08	3.65	139469	0.15	2.76
škol-w	248	0.10	3.32	135748	0.14	2.80
škol-i,y	976	0.39	1.34	312564	0.33	1.59
škol-*e*	598	0.24	2.05	146867	0.16	2.68
škol-*ow*	66	0.03	5.23	68712	0.07	3.78
škol-dm	15	0.01	7.36	4890	0.01	7.59
škol-dch	146	0.06	4.09	17630	0.02	5.74
škol-ami	22	0.01	6.81	17576	0.02	5.75

Table 1. Probability distribution for the inflected noun *škola.*

The amount of information conveyed by the inflected words (w_e) and exponents (e) are calculated by applying the base -log2 transformation on the respective relative frequencies of the different exponents, and the relative frequencies of the inflected forms.

For example, the amount of information conveyed by the exponent 'u' (2.80) is calculated from the probability of the exponent $Pr_\pi(e)$

$$I_e \quad = \quad -\backslash og_2Pr\hat{}(e) \qquad (2)$$

where e = *u* (0.1439), estimated from the frequency of the exponent *F(e)* (135748) relative to the sum of the frequencies of the exponents in the paradigm (942852)

$$Pr\text{\textasciicircum}ye) = \qquad (3)$$

There are also other statistical measures which represent properties of the entire paradigm. The *entropy* of an inflectional paradigm, *H*, is calculated as

$$H = -E_e Pr\text{\textasciicircum}\{w_e)\backslash og_s \; Pr_v(w_e) \qquad (4)$$

For the values shown in Table 1 for'škola', this is calculated as: i7('škola')= —[0.0851x *log20.0851*.. .0.0089 x Zo<?20.0089], which amounts to 2.46. Informally, this index captures the degree to which the paradigm is unevenly distributed over the different forms.

In sum, these metrics characterize the contribution of stems and exponents to the probability that a word form will occur. These measures are made practically possible with the availability of relatively large morphosyntactically-annotated corpora such as the Slovak National Corpus (SNK).

Here we want to investigate whether it is possible to calculate inflectional entropy using the SNK, and if so, characterize how the results differ from previously reported results from Serbian. These comparisons would support future empirical research on word processing in Slovak, and help characterize differences between these two closely related languages.

Number	Case	Serbian	Slovak
Singular	Nominative	planin-*a*	planin-*a*
	Genitive	planin-*e*	planing
	Dative	planin-*i*	planin-*e*
	Accusative	planin-*u*	planin-*u*
	Instrumental	*planin-om*	planin-*o*w
	Locative	planin-*i*	planin-*e*
Plural	Nominative	planin-*e*	planin-^
	Genitive	planin-*a*	planín-0
	Dative	pl*anin-ama*	planin-*ám*
	Accusative	planin-*e*	planing
	Instrumental	*planin-ama*	planin-*ami*
	Locative	*planin-ama*	planin-*ách*

Table 2. Slovak and Serbian regular feminine inflectional exponents, illustrated with the noun 'planina' (meaning mountain in Serbian and plain in Slovak).

Despite the differences between surface exponents used in Serbian and Slovak (see Table 2 above for an example), there are many similarities between the morphosyntactic systems of Slovak and Serbian. Both languages have relatively complex inflectional systems, in which nouns are marked for number (singular and plural) and grammatical case (nominative, genitive, accusative, dative, instrumental, locative; the vocative is archaic in Slovak and its status is disputed in Serb). In addition, the inflectional endings depend on the gender of the noun (feminine, masculine, neuter) and the inflectional class.

Given such similarities, we would expect that statistical distribution of the Serbian and Slovak terms would be similar. If we take the example of a base-level term used in Milin *et al.*, such as 'žena' (woman), we should observe a similar statistical distribution as their Slovak counterpart 'žena', because they would be expected to have a similar distribution of grammatical functions and meanings. If this is the case for most of the terms in Slovak, then many of the psycholinguistic results obtained from the study of Serbian should also generalize to Slovak.

On the other hand, there might be some reasons to expect differences between these (and also other Slavic) languages. First, some of the basic-level terms in the two languages have different meanings, gender or inflectional class. For example, the primary meaning of 'planina' (mountain in Serbian) does not correspond to the same meaning as its Slovak counterpart 'planina' (plain in Slovak). Second, the statistical estimates for Serbian are based on a *sample* of text, as is the case with all statistical parameter estimates. It may be the case that the parameter estimates for a given measure like inflectional entropy will be conditioned on the data source. This would suggest that the Slovak and Serbian parameter estimates could be different, either due to real differences in the usage of the two languages, or to differences in the samples used to estimate the parameters.

We hypothesized that the factors governing the paradigm distribution of nouns in Slovak and Serbian would be similar. We predicted that the measures of inflectional entropy and paradigm entropy of Slovak and Serb would therefore also be similar.

2 Method

For a global comparison with Serbian results, we created two figures as in Milin *et al.* (2009:55). We made a query from the SNK for all feminine and masculine nouns in all respective cases and numbers. We then extracted statistical information for all feminine exponents. Masculine nouns were not further analyzed. Milin and colleagues focused on dominant regular inflectional subclasses in their paper; we consider all feminine exponents. Note that (y, i**)** exponents were not computed separately, since in modern Slovak they both express the same phoneme */i/*. The function of **(**y**)** is to indicate that the preceding sound is not palatalized.

For the comparison of inflectional entropy between the two languages, we selected words from the word list provided in Milin *et al.* (2009) for which there was (almost) complete form overlap with their Slovak counterparts, and used these for the query in the manually morphologically annotated subcorpus *r-mak-3.0* from the SNK. For the analysis, we used only those words that were present two or more times in the SNK sample, and we did not include diminutives. The frequencies and relative frequencies of inflected variants and inflectional exponents were computed in the same way as in Milin *et al.* (2009) and as described earlier in the Introduction.

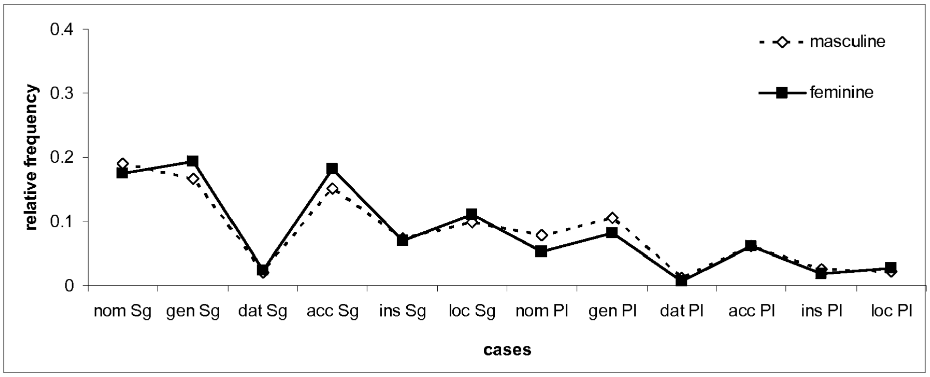

Fig. 1. The relative frequencies of feminine and masculine nouns for Slovak according to case and number.

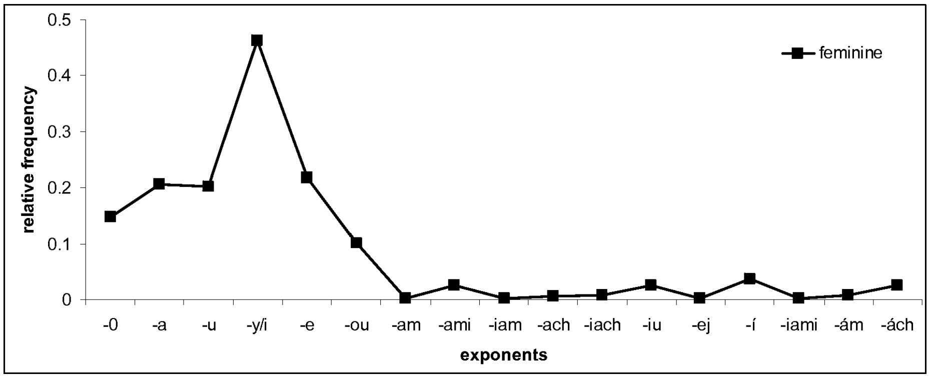

Fig. 2. Relative frequency of feminine nouns in Slovak according to inflectional suffix.

3 Results and discussion

Figure 1 shows for each case-number combination the distribution of relative frequencies within each inflectional class (here, the masculine and feminine nouns). Except for the values of relative frequencies, the picture is almost identical to the Serbian results. This is a good example of how different corpora can still be representative with respect to morphological aspects of language use, irrespective of whether it is of a smaller or larger size. Figure 2 plots the relative frequency of individual exponents within the feminine inflectional classes. These are also considerably similar to Serbian.

Now we turn to the question, whether the inflectional entropy of individual cases is comparable as well. Table 3 shows the inflectional entropy, *H*, calculated for the words we selected from the Serbian lists. The average entropy for Slovak *(/j,* = 1.70), in this sample, was less than Serbian *(/j,* = 2.11), *t(18)* = 2.011, *p* = 0.059. The correlation between the two samples was relatively low, *r* = 0.2. This result would suggest that the deviation from the paradigm pattern is, on average, greater for Serbian than for Slovak.

Slovak	***H***	Serbian	***H***
kniha	2.63	knjiga	2.17
rieka	2.28	reka	2.22
búrka	1.30	bura	2.23
tráva	1.52	trava	2.23
brigáda	0.65	brigada	1.89
fabrika	0.86	fabrika	2.12
škola	2.46	škola	2.20
náuka	0.88	nauka	1.98
ruža	1.24	ruža	1.90
stanica	1.72	stanica	2.05
ulica	3.04	ulica	2.39
dolina	0.59	dolina	2.43
duša	2.36	duša	2.28
ryba	1.71	riba	1.79
sila	3.27	sila	2.03
potreba	2.74	potreba	2.13
vŕba	0.24	vrba	1.86
hlava	0.80	glava	2.34
hviezda	2.01	zvezda	1.83

Table 3. Comparison of Slovak and Serbian word pairs.

This result suggests that despite the similarities between Serbian and Slovak, their inflectional entropy differs. However, several caveats should be kept in mind. This comparison was based on a relatively limited number of words, and in order to maintain strict comparability, we only examined words with overlapping surface forms. Despite this overlap, preferences for certain terms, or differences in meaning in the respective languages, could lead to differences in the frequencies of some terms. Future work could examine larger samples, and other inflectional classes.

Despite the small sample, the results offer some suggestion that individual measures of entropy are needed for each language, even for languages as typologically similar as Serbian and Slovak. In practical terms, it appears that the use of morphologically-annotated corpora are very helpful for calculating these measures for each language. A useful framework for future comparisons of Slavic languages (or other languages that have similar inflectional classes) might include measures like inflectional entropy in order to guage the similarties and differences between languages.

4 Summary

In this paper we have described how inflectional entropy can be estimated from the Slovak National Corpus. The obtained estimates were compared to results reported previously for Serbian. The results showed that overall, the distribution of feminine and masculine inflected nouns (grouped according to case and number) is almost identical for both languages. The comparison of relative frequencies for feminine nouns, grouped by inflectional suffixes, showed a considerable amount of similarity with Serbian, despite the differences in suffix forms. Given this outcome, we expected inflectional entropy measures for a selected number of Slovak and Serbian (high frequency) nouns to be comparable. However, the results showed that the estimates differ. This implies that morphologically-annotated corpora could be very useful for cross-linguistic comparisons.

5 Acknowledgements

We would like to thank the producers of the Slovak National Corpus for making their work widely available. This work was made possible by the support of the Max Planck Gesellschaft. Both authors contributed equally.

References

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Balota, D. A., Yap, M. J., & Cortese, M. J. (2006). Visual word recognition: The journey from features to meaning (A travel update). In M. Traxler and M. Gernsbacher (Eds.) *Handbook of Psycholinguistics, 2nd Edition*. Pp. 285–375. Amsterdam: Elsevier.

Kostic´, A. (1991). Informational approach to the processing of inflected morphology: Standard data reconsidered. *Psychological Resarch, 53*, 62–70.

Kostic´, A. (1995). Informational load constraints on processing inflected morphology. In L. B. Feldman (Ed.) *Morphological Aspects of Language Processing*. Pp. 317–344. New Jersey: Lawrence Erlbaum Inc. Publishers.

Milin, P., Kuperman, V., Kostic´, A., & Baayen, R. H. (in press). Words and paradigms bit by bit: An information-theoretic approach to the processing of inflection and derivation. In J.P. Blevins & J. Blevins (Eds.), *Analogy in grammar: form and acquisition*. Oxford University Press: Oxford.

Milin, P., Durdevic, D. F., & Moscoso del Prado Martín, F. (2009). The simultaneous effects of inflectional paradigms and classes on recognition: Evidence from Serbian. *Journal of Memory and Language, 60*, 50–64.

Moscosos del Prado Martín, F., Kostic´, A., Baayen, R. H. (2004). Putting the bits together: An informational theoretical perspective on morphological processing. *Cognition, 94*, 1–18.

Slovenský národný korpus, – r-mak-3.0. Bratislava: Jazykovedný ústav. L'. Štúra SAV 2008. http://korpus.juls.savba.sk.

Whaley, C. P. (1978). Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior, 17*, 143–154.