Frequency and Declensional Morphology of Czech Nouns

Ján Mačutek[1] and Radek Čech[2],

[1] Department of Applied Mathematics and Statistics, Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynská dolina, 842 48 Bratislava, Slovakia
jmacutek@yahoo.com
[2] Department of General Linguistics, Philosophical Faculty, Palacký University, Křížkovského 10, 771 80 Olomouc, Czech Republic
cechradek@gmail.com

Abstract. The relationship between frequency and declension of nouns in Czech is analyzed. The nominative is taken as the basic form of nouns. We define the measure of morphological change as the number of phonetic changes in the stem plus the number of phonetic changes in the declensional suffix. Two approaches were examined: 1) nominative singular as the basic form of a noun regardless of its grammatical number, 2) nominative singular as the basic form of a noun in the singular and nominative plural as the basic form of a noun in the plural. In both cases, the relation "the lesser the change, the higher the frequency" is observed.

Keywords: declension, morphonetic change frequency, case, noun.

1 Introduction

The relationship between the frequency and a word form's "behavior" has been well known for almost two centuries. As M. Krug noticed in [18], J. Grimm [8] already observed the correlation between frequency and irregularity: "Auxiliaria, d.h. Verba, welche sehr häufig gebraucht werden und statt ihrer lebendigen Bedeutung abstrakte Begriffe annehmen, tragen gowöhnlich solche Unregelmäßigkeiten an sich." [Auxiliaries, i.e. verbs which are used very frequently and which take on abstract notions instead of their vivid meanings, usually display such irregularities] (cited and translated according to [8], p. 8). However, except for quantitative linguistics [15], [16], Zipf's approach [24], [25] and so called usage-based models [1], [2], [3] [4], [10], where the frequency is considered to be one of the central "powers" shaping properties of language, the focus on frequency effects on word form (or phoneme, syllable, sentence, etc.) is rather exceptional among linguists, cf. "A newcomer to the field of linguistics might be surprised to learn that for most of the twentieth century facts about frequency of use of particular words, phrases, or constructions were considered irrelevant to study of linguistic structure. To the uninitiated, it does not seem unreasonable at all to suppose that high-frequency words and expressions might

have one set of properties and low-frequency words and expressions another" ([2], p. 5).

In the present study, the relationship among frequency, morphology and phonetics is observed. Particularly, it is assumed that frequency has an impact on the number of morphonetic changes in a particular word form. We hypothesize that *the greater the magnitude of a morphonetic change, the lower the frequency of word forms with the magnitude* (the number of changes with respect to the nominative is considered, cf. Section 2). We can find, to our knowledge, two theoretical sources for the reasoning of this kind: 1) synergetic linguistics [13], [14] and 2) a usage-based cognitive approach [3]. As for 1), we preliminarily suppose that the relationship could be viewed as a result of a mutual interaction of so-called requirements, namely "Minimisation of producing effort", "Minimisation of encoding", and "Minimisation of memory effort" for a speaker, and "Minimisation of decoding", "Minimisation of memory effort", and "Minimisation of inventories" for a hearer. As for 2), from the cognitive point of view the effects of high token frequency are considered as follows: "because exemplars are strengthened as each new token of use is mapped onto them, high-frequency exemplars will be stronger than low frequency ones, and high-frequency clusters – words, phrases, constructions – will be stronger than lower frequency ones. The effects of this strength (lexical strength [1]) are several: first, stronger exemplars are easier to access, thus accounting for the well-known phenomenon by which high-frequency words are easier to access in lexical decision tasks. Second, *high-frequency, morphologically complex words show increased morphological stability*." ([3], p. 24, italics by the authors of this paper).

2 Language material and methodology

For the testing of a hypothesis concerning morphonetic changes a language with a rich inflexional system should be used, for obvious reasons. Therefore, Czech has been chosen; in particular, we focused on Czech nouns.

The system of Czech noun declensional morphology contains seven cases (nominative, genitive, dative, accusative, vocative, locative, instrumental). Throughout the paper, we use the following abbreviations for the cases: N – nominative, G – genitive, D – dative, A – accusative, V – vocative, L – locative, I – instrumental. The morphology is expressed by inflectional endings (desinences) added to the stem, cf. the word forms of masculine noun *táta (daddy)*:

Table 1. An example of declensional morphology of Czech nouns: word *táta (daddy)*.

case	singular	plural
N	tát-a	tát-ové
G	tát-y	tát-ů
D	tát-ovi	tát-ům
A	tát-u	tát-y
V	tát-o	tát-ové
L	tát-ovi	tát-ech

I	tát-ou	tát-y

The endings do not merely express information regarding case but also number and gender. Therefore, Czech is typologically ranked among fusional languages (one ending denotes more than one morphological category). Further, morphonetic alternations are typical for Czech, e.g. an elision of "e" in the stem of the word *pes (dog)*

pes (N, singular) ps-a (G, singular)

or an alternation "k" to "c" in word *kluk (boy)*

kluk (N, singular) kluc-i (N, plural).

According to a traditional view (cf. [11] and [23]) the alternations (and other changes) are governed by rules which have no relation to frequency.

Two texts were used for the analysis, namely the book of travel "Obrázky z Holandska" (Pictures from the Netherlands) written by Karel Čapek and the short novel "Krásná Poldi" (Beautiful Poldi) written by Bohumil Hrabal. The language data were taken from *Dictionary of Karel Čapek* [5] and from *Dictionary of Bohumil Hrabal* [6], which are, actually, lemmatized and morphologically tagged authors' corpora. The lemmatization and morphological tagging allow us to process data automatically; for example, all forms of the lemma *kráva (cow)* and their frequencies can be easily obtained as follows.

Table 2. Morphological tagging in [5] and [6].

word form	tag (morphology)	frequency
krávy	NNFP1-----A-----	4
krávy	NNFS2-----A-----	3
krávy	NNFP4-----A-----	2
kráva	NNFS1-----A-----	2
krav	NNFP2-----A-----	2
kravami	NNFP7-----A---1-	1
krávu	NNFS4-----A-----	1
kravách	NNFP6-----A---1-	1

A letter in the first column depends on a part of speech (we chose only nouns denoted by N) The information regarding grammatical number and case, needed for our analysis, is represented by letters (S= singular, P = plural) in the fourth column and by digits (1 = N, 2 = G, etc.) in the fifth columns of the tags.

The nominative form was taken as the basic form of each word. Then, for each lemma the number of phonetic changes/alternations with respect to the nominative in the stem and the number of phonetic changes/alternations with respect to the nominative in the suffix were determined manually. The total number of changes/alternations (i.e., those from the stem plus the ones from the suffix) was used as a magnitude of the morphological change. For example, let us consider all singular cases of the lemma *stůl (table)*.

Table 3. Magnitudes of change in word *stůl (table).*

case	word form	number of changes
N	stůl	0
G	stolu	2
D	stolu	2
A	stůl	0
V	stole	2
L	stolu	2
I	stolem	3

The nominative is regarded as the basic form, therefore it is assigned zero changes (as well as accusative which is represented by the same word form); genitive, dative, and locative display one change in the stem (alternation *ů-o*) and the inflectional ending is represented by one vowel (*u*), whereas the basic form (i.e., nominative) does not have any suffix (or it has a zero-morpheme suffix); so, for all these cases two changes are assigned; instrumental displays one change in the stem (alternation *ů-o*) and the inflectional ending represented by a vowel (*e*) and a consonant (*m*); therefore, three changes are assigned to it.

We remind readers that making a suffix longer (or adding it if there was none) is taken into account (cf. Table 3, the word *stůl* in the nominative and instrumental case). The same applies to an elimination of phonemes (cf. the word *pes* in Table 4).

It should be noticed that we followed the morphonetic approach and phonetic (not phonemic or graphemic) changes were taken into account; for example, in the case of the word *dub (oak)*, pronounced [dup], we counted also change of voice (p-b), cf. Table 4:

Table 4. Magnitudes of change in words *dub (oak)* and *pes (dog),* word forms in singular are presented for both nouns. Pronunciation is given in square brackets, elimination of a phoneme is marked by Ø.

case	*dub*		*pes*	
	word form	number of changes	word form	number of changes
N	dub [dup]	0	pes	0
G	dubu [dubu]	2	pØsa	2
D	dubu [dubu]	2	pØsovi	4
A	dub [dup]	0	pØse	2
V	dube [dube]	2	pØsa	2
L	dubu [dubu]	2	pØsovi	4
I	dubem [dubem]	3	pØsem	3

Two approaches were applied: first, the nominative singular was taken as the basic form of a noun regardless of its grammatical number; second, the nominative singular was taken as the basic form of a noun in the singular and the nominative plural as the basic form of a noun in the plural. We thus analyze six datasets (three for each author).

3 Results

In the first step, we modeled the relation between the magnitude of change and the frequency of word forms. We fit all datasets with the function

$$y = a\left(x+1\right)^{b} e^{-cx}, \tag{1}$$

where x is the magnitude, y is the frequency of word forms displaying magnitude x, and a, b and c are parameters. Function (1) is the basic form of the Wimmer-Altmann model [22]. The parameter a is often interpreted as the value which y takes for the smallest x, i.e., in this case a is the frequency of word forms which are the same as the nominative form and they are therefore assigned zero changes. Thus we have

$$y = y\left(0\right)\left(x+1\right)^{b} e^{-cx}. \tag{2}$$

The data obtained, together with parameter values b and c from function (2) and values of the determination coefficient R^2, can be found in Table 5. Throughout the paper we use the following notation:

(S+P)/S nominative singular is the basic form for all nouns regardless of their grammatical number,
S/S nominative singular is the basic form for nouns in the singular,
P/P nominative plural is the basic form for nouns in the plural.

As can be seen, function (2) yields a very good fit in all cases.

Table 5. Magnitude of change and frequency – fitting function (2) to data form [5] and [6].

	Čapek			Hrabal		
	(S+P)/S	S/S	P/P	(S+P)/S	S/S	P/P
magnitude of change	frequency					
0	738	711	426	821	778	295
1	849	456	181	639	395	120
2	566	235	140	382	183	86
3	70	33	44	72	34	25
4	10	3	2	7	1	4
	parameters and determination coefficient					
b	2.953	1.572	0.258	1.836	0.981	-0.353
c	1.857	1.506	0.579	1.484	1.337	0.575
R^2	0.959	0.991	0.976	0.983	0.992	0.985

The data and the respective functions for Čapek's book from Table 5 are presented also in Figures 1, 2 and 3.

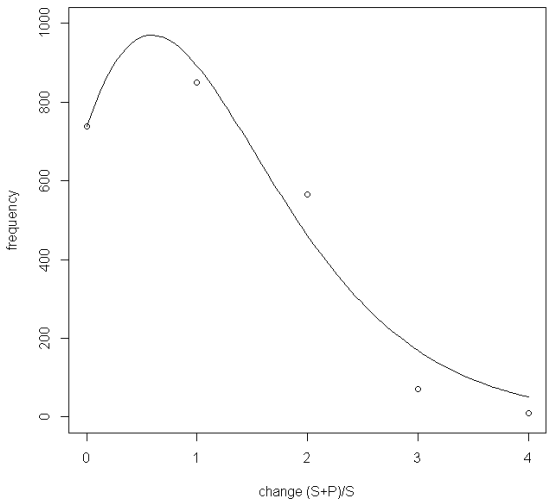

Fig. 1. Fitting function (2) to data from [5], method (S+P)/S.

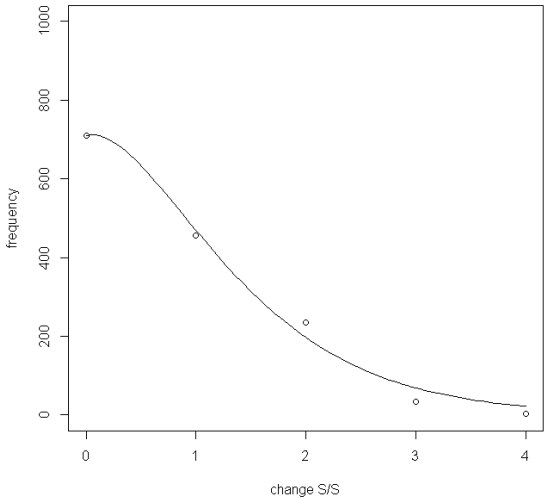

Fig. 2. Fitting function (2) to data from [5], method S/S.

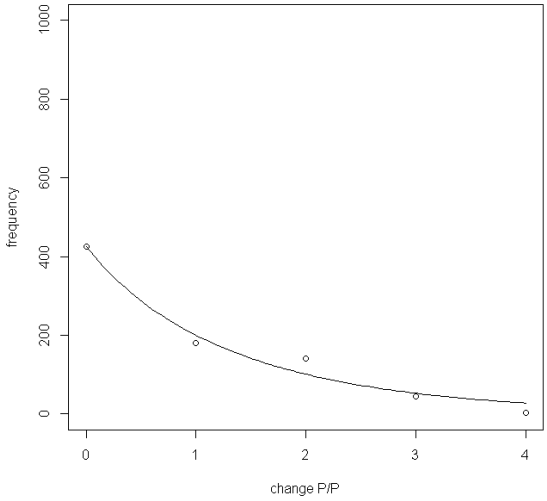

Fig. 3. Fitting function (2) to data from [5], method P/P.

The data and respective curves for Hrabal's short story are quite similar (cf. Table 5).

We also paid attention to a similar hypothesis introduced by Fenk-Oczlon in [7] (cf. also a collection of open problems in quantitative linguistics [21]): "the more frequent a case in a particular language, the more it tends toward zero coding". Frequencies of cases in the texts by Čapek and Hrabal can be easily determined from the morphological tags (cf. Table 2). The magnitude of coding was defined as the mean number of changes with respect to the basic form (cf. Section 2). We applied two approaches from Section 2 again, i.e., first, the nominative singular was taken as the basic form of a noun regardless of its grammatical number; second, the nominative singular was taken as the basic form of a noun in the singular and the nominative plural as the basic form of a noun in the plural. The mean number of changes per case was then computed as the total number of changes per case (i.e., the sum of all changes per case) divided by the frequency of the case. Results are presented in Tables 6 and 7.

Table 6. Case frequency and magnitude of coding in the text by Čapek [5] (*f* – frequency, *mc* – magnitude of coding).

	S+P/S		S/S		P/P	
c	*f*	*m*	*f*	*m*	*f*	*m*
ase		*c*		*c*		*c*
N	7	0	4	0	2	0
	07	.45	62	.00	45	.00

G	5	1	3	1	2	1
	70	.21	52	.01	18	.31
D	6	1	4	1	2	2
	2	.63	0	.23	2	.23
A	3	0	2	0	1	0
	82	.68	37	.34	45	.14
V	3	1	2	1	1	0
		.67		.50		.00
L	2	1	1	1	5	2
	39	.61	83	.43	6	.00
I	2	1	1	1	1	1
	70	.76	62	.77	06	.25

Table 7. Case frequency and magnitude of coding in the text by Hrabal [6] (*f* – frequency, *mc* – magnitude of coding).

	(S+P)/S		S/S		P/P	
c	*f*	*m*	*f*	*m*	*f*	*m*
ase		*c*		*c*		*c*
N	7	0	4	0	2	0
	07	.45	62	.00	45	.00
G	5	1	3	1	2	1
	70	.21	52	.01	18	.31
D	6	1	4	1	2	2
	2	.63	0	.23	2	.23
A	3	0	2	0	1	0
	82	.68	37	.34	45	.14
V	3	1	2	1	1	0
		.67		.50		.00
L	2	1	1	1	5	2
	39	.61	83	.43	6	.00
I	2	1	1	1	1	1
	70	.76	62	.77	06	.25

The tendency from Fenk-Oczlon's hypothesis can be seen (it can be tested, e.g., by the Kendall correlation coefficient, the p-values are higher than 0.05 in all cases). Nevertheless, the data are not "smooth" and therefore they are difficult to model. Function (2), e.g., does not fit them well (which is true especially for the text by Čapek). Figure 4 presents the data from Čapek's text, with nouns in the singular being considered, i.e., S/S. One obtains a very low value of the determination coefficient ($R^2 = 0.486$).

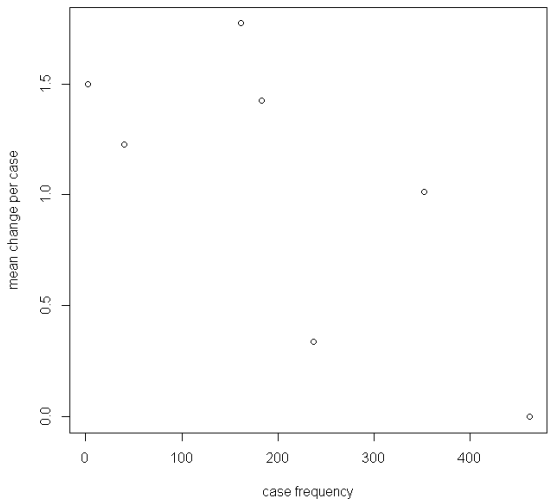

Fig. 4. Case frequency and magnitude of coding, Čapek S/S (cf. Table 6).

This is the most irregular behavior among the datasets from Tables 6 and 7, but only two of the datasets achieve a determination coefficient higher than 0.8 (for Čapek's text one obtains $R^2 < 0.7$ for all three approaches). However, applying a different methodology (i.e., a different definition of the magnitude of change/case coding, e.g., the length of the suffix only) could lead to different results.

3 Conclusion and ideas for further research

The paper shows that declensional morphology of Czech nouns is clearly related to frequency – if the nominative is taken as the basic form, it holds that "the lesser the magnitude of change, the higher the frequency" (or, in other words, the more similar the word forms are to their nominative case, the more frequently they occur).

This line of investigation should be broadened in two directions. First, similar studies must be undertaken also for other (and not only Slavic) languages with a rich inflexional system. Second, other parts of speech, especially adjectives, should be studied. Results obtained should be in future interpreted within the synergetic linguistics framework [13], [14]. We suppose that especially connections and interrelations with other properties of morphology and syntax will be established.

The data presented in this paper, especially Tables 6 and 7, can serve as material for building a model for case frequencies (a study on case diversification [20] contains rank-frequency distributions of cases from many texts in German, Russian, Slovak and Slovene).

Relations among the case frequency, the frequency of inflexional suffixes and the length of inflexional suffixes (which can be understood as a measure of

inflexion/change) are exploited in psycholinguistics (we mention works [17] for Polish, [12] and [19] for Serbian, and [9] for Slovak). A theoretically based model would therefore also be useful in this research area.

Acknowledgment. The authors were supported by research grants VEGA 2/0038/12 (J. Mačutek), ESF OPVK 2.2 - Innovation of the General linguistics and theory of communication in cooperation with the natural sciences (CZ.1.07/2.2.00/28.0076) and ESF OPVK 2.3 - Linguistic and lexicostatistic analysis in cooperation of linguistics, mathematics, biology and psychology (CZ.1.07/2.3.00/20.0161) (R. Čech).

References

1. Bybee, J.: Morphology: A Study of the Relation Between Meaning and Form. Benjamins, Amsterdam (1985)
2. Bybee, J.: Frequency of Use and the Organization of Language. Oxford University Press, Oxford (2007)
3. Bybee, J.: Language, Usage and Cognition. Cambridge University Press, Cambridge (2010)
4. Bybee, J., Hopper, P. (eds.): Frequency and the Emergence of Linguistic Structure. Benjamins, Amsterdam/Philadelphia (2001)
5. Čermák, F. (ed.): Slovník Karla Čapka. Nakladetelství Lidové noviny, Praha (2007)
6. Čermák, F., Cvrček, V. (eds.): Slovník Bohumila Hrabala. Nakladetelství Lidové noviny, Praha (2009)
7. Fenk-Oczlon, G.: Familiarity, information flow, and linguistic form. In: Bybee, J., Hopper, P. (eds.) Frequency and the Emergence of Linguistic Structure, pp. 431--448. Benjamins, Amsterdam/Philadelphia (2001)
8. Grimm, J.: Deutsche Grammatik. Dieterich'sche Buchhandlung, Göttingen (1822)
9. Hanulíková, A., Davidson, D.J.: Inflexional entropy in Slovak. In: Levická, J., Garabík, R. (eds.) NLP, Corpus Linguistics, Corpus Based Grammar Research, pp. 145—151. Tribun, Brno (2009)
10. Hopper, P.: Emergent grammar. In: Aske, J., Beery, N., Michaelis, L., Filip, H. (eds.) Proceedings of the 13th Annual Meeting of the Berkeley Linguistic Society, pp. 139--157. Berkeley Linguistic Society, Berkeley (1987)
11. Karlík, P., Nekula, M., Rusínová, Z: Příruční mluvnice češtiny. Nakladetelství Lidové noviny, Praha (1995)
12. Kostić, A., Mirković, J.: Processing of inflected nouns and levels of cognitive sensitivity. Psihologija 35, 287—297 (2002)
13. Köhler, R.: Zur linguistischen Synergetik: Struktur und Dynamik der Lexik. Brockmeyer, Bochum (1986)
14. Köhler, R.: Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) Quantitative Linguistics. An International Handbook, pp. 760--774. de Gruyter, Berlin/New York (2005)
15. Köhler, R., Altmann, G.: Aims and methods of quantitative linguistics. In: Altmann, G., Levickij, V., Perebyinis, V. (eds.) Problems of Quantitative Linguistics, pp. 13--43. Ruta, Chernivtsi (2005)
16. Köhler, R., Altmann, G.: Quantitative linguistics. In Hogan, P.C. (ed.) The Cambridge Encyclopedia of the Language Sciences, pp. 695--697. Cambridge University Press, New York (2011)
17. Krajewski, G., Lieven, E.V.M., Theakston, A.L.: Productivity of a Polish child's inflexional noun morphology: a naturalistic study. Morphology 22, 9—34 (2012)

18. Krug. M.: Frequency as a determinant in grammatical variation and change. In: Rohdenburg, G., Mondorf, B. (eds.) Determinants of Grammatical Variation in English, pp. 7--67. de Gruyter, Berlin (2003)
19. Milin, P., Filipović Đurđević, D., Moscoso del Prado Martín, F.: The simultaneous effects of inflexional paradigms and classes of lexical recognition: Evidence from Serbian. Journal of Memory and Language 60, 50—64 (2009)
20. Popescu, I.-I., Kelih, E., Best, K.-H., Altmann, G.: Diversification of the case. Glottometrics 18, 32—39 (2009)
21. Strauss, U., Fan, F., Altmann, G.: Problems in Quantitative Linguistics 1. RAM-Verlag, Lüdenscheid (2008)
22. Wimmer, G., Altmann, G.: Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.) Quantitative Linguistics. An International Handbook, pp. 791--807. de Gruyter, Berlin/New York (2005)
23. Ziková, M.: Alternace vokálů s nulou v současné češtině – laterální autosegmentální analýza. Dissertation thesis. Filozofická fakulta Masarykovy university, Brno (2008)
24. Zipf, G.K.: The Psycho-Biology of Language. MIT Press, Cambridge (MA) (1935)
25. Zipf, G.K.: Human Behavior and the Principle of the Least Effort: An Introduction to Human Ecology. Addison-Wesley, Cambridge (MA) (1949)