

What to take as the basic verb form? A pilot study in Czech

Ján Mačutek^{1,2}, Michaela Koščová¹, Michaela Nogolová³, Radek Čech⁴

¹Mathematical Institute, Slovak Academy of Sciences, Bratislava, Slovakia

²Department of Mathematics, Constantine the Philosopher University in Nitra, Slovakia

³Department of Czech Language, University of Ostrava, Ostrava, Czech Republic

⁴Department of Czech Language, Masaryk University, Brno, Czech Republic

jan.macutek@mat.savba.sk, matemiko@savba.sk,
michaela.nogolova@osu.cz, cechradek@phil.muni.cz

Abstract

This paper investigates which Czech verb form can serve as the most suitable for analysing the relationship between the frequency of individual verb forms and the number of formal changes required to derive them. The number of changes is measured by the Levenshtein distance. The results show that the present indicative third person singular occupies the central position among all verb forms. Furthermore, verbs exhibit an inverse relationship between the number of formal changes with respect to the basic form and form frequency. This relation can be interpreted as a consequence of the principle of least effort.

Keywords: principle of least effort, verb forms, conjugation, Levenshtein distance, Czech.

1 Introduction

From the perspective of the principle of least effort (Zipf, 1949), it is plausible to assume that inflectional paradigms are not frequency-neutral systems. Language users tend to favour forms that minimise production and processing costs. In inflectional morphology, these costs may be reflected in the amount of formal changes required to derive a given word form from its basic form. Thus, word forms that are closer to the basic form may be expected to occur more frequently than forms involving a greater number of modifications.

Results presented in a series of recent papers on noun declension in Slavic languages (Kelih et al., 2023; Mačutek et al., 2023, 2025) are consistent with this assumption. More specifically, noun word forms that are more similar to the lemma occur more often. The Levenshtein distance (see Section 2 for its definition) is used as a measure of similarity between word forms and their lemmas.

This pilot study conducted on Czech language material extends the analysis to verbs. However,

an important decision must be made at the very beginning. While the choice of the basic form is relatively obvious for nouns (it is nominative singular), the situation is more complicated with verbs. In the verbal domain, several forms may in principle be considered candidates for the basic form.

One natural candidate is the infinitive. In many grammatical traditions, the infinitive functions as the citation form of the verb, i.e. the form under which verbs are listed in dictionaries and grammars. At the same time, it is morphologically less specified than finite forms. This relative neutrality makes it a plausible candidate for the basic form of the verbal paradigm. From a methodological point of view, the infinitive is also a convenient reference form, since in corpus linguistic annotation it often corresponds to the lemma. On the other hand, as there are languages without the infinitive (e.g., Bulgarian or Modern Greek), it cannot serve as the universal basic verb form.

Another candidate is the present indicative first person singular. The basic reference form of a verb need not always be an infinitive. In the Latin tradition, for example, the first person singular present indicative active is the first of the principal parts under which verbs are conventionally cited.

Yet another option is the present indicative third person singular. As noted by Bybee (2010) in her discussion of markedness, Jakobson (1990) treats the third person of Russian verb as the unmarked member of the person opposition. Specifically, the category of person is structured so that first and second person are treated as participants in the speech event, whereas third person is a non-participant. Accordingly, the feature “personal” is marked, while its absence is unmarked. From this perspective, the present indicative third person

singular may be regarded as a structurally neutral candidate for the basic verb form.

The aim of the present study is therefore to compare these three candidates – the infinitive, the present indicative first person singular, and the present indicative third person singular – and to determine which of them is the most suitable basic form for the analysis of Czech verbs. More specifically, we investigate which of these forms occupies the most central position in the verbal paradigm, in the sense that its mean Levenshtein distance from all other verb forms is minimal.

2 Language material and methodology

The language material consists of five dependency treebanks from Universal Dependencies 2.13 (Zeman et al., 2023), capturing distinctive domains:

1. FicTree (Jelínek, 2017) represents the literary register, this treebank contains eight texts published between 1991 and 2007 (six works of fiction, alongside one memoir and one children’s book).
2. UD_Czech-CAC is based on the written component of the Czech Academic Corpus 2.0 (Vidová Hladká et al., 2008), it supplies high-density informational texts spanning the scientific, administrative, and journalistic domains.
3. UD_Czech-Poetry is derived from the Corpus of Czech Verse by Plecháč and Kolár (2015). This dataset introduces language governed by strict rhythmic and formal rules.
4. Czech-PDT UD (Bejček et al., 2013) contributes additional journalistic, business, and science-related texts originating from the 1990s.
5. Parallel Universal Dependencies (PUD, Czech part) was originally compiled for the CoNLL 2017 shared task on Multilingual Parsing; this collection features 1,000 sentences from news and Wikipedia sources. The dataset consists of translated texts: 750 sentences were translated directly from English, while the remaining 250 originate from French, German, Italian, and Spanish using English as an intermediary language.

Initially, each sentence from our dataset was represented as a dependency tree following the Universal Dependencies annotation. For the purposes of

this study, we transformed these structures into the Surface Syntactic Universal Dependencies (SUD) format (Gerdes et al., 2018). We adopted this approach because the SUD scheme – by explicitly capturing the structural functions of auxiliaries and prepositions – more accurately aligns with the morphosyntactic structure of the Czech language.

Beyond mapping syntactic dependencies, the annotation framework supplies crucial word-level metadata, including lemmatization, morphological tagging, and part-of-speech (POS) identification. Furthermore, all selected treebanks contain manually annotated language specific extended part-of-speech tags (XPOS). Combined with Universal Features, which assign additional grammatical and lexical properties, this rich annotation allows us to reliably distinguish categories such as grammatical person, gender, tense, and other morphological properties.

Each sentence is represented as a dependency tree with a root. In most cases, this root is a finite verb or an auxiliary (functioning as the predicate), since dependency grammar is inherently verb-centric. As our methodology partly relies on this predicate-driven structure, we exclude sentences lacking an explicit verbal root, such as exclamations, fragments, or elliptical constructions.

The Czech verbal system contains both synthetic and analytic forms (Osolobě, 2017).¹ Because dependency trees represent each individual word as a distinct node, we implemented targeted procedures to properly identify specific analytic forms. For example, past-tense forms in the first and second person require combining the past active participle (also called l-participle) with a present-tense form of the verb *to be*:

psala jsi
‘write-PTCP-SG.F be.PRS.2SG’.

Similarly, the future tense of imperfective verbs is formed analytically by combining the future form of the verb *to be* with the infinitive of the lexical verb:

budeme psát
‘be.FUT.1PL write.INF’.

Further, conditional forms require the past active participle (l-participle) together with a specific conditional form(s) of the auxiliary:

¹All forms of verb *dělat* ‘to do’ are presented in the Appendix A.

napsali bychom

‘write-PTCP-PL be.COND.1PL’.

Another example of an analytic form is the passive voice, which is formed by a passive participle and a form of *to be*:

je napsán

‘be.PRS.3SG write.PASS.PTCP.PRS.3SG.M’.

However, because the passive participle shares behavioral traits with adjectives, the SUD annotation scheme tags these words as adjectives. Consequently, our analysis, which only processes words explicitly tagged as verbs, excludes these passive constructions.²

The annotation labels as verbs also some other parts of speech. For instance, adverbs *lze* (‘it is possible (to)’) and its negative form *nelze* (‘it is not possible (to)’) function syntactically in a verb-like manner and are annotated as verbs. Sentences with such non-verb predicates are excluded from the dataset prior to analysis.

If the infinitive is treated as the basic verb form, all verbs in the sample can be analysed, since the annotation provides the infinitive as the lemma for each verb. By contrast, when the present indicative first person singular or the present indicative third person singular is taken as the basic form, only those verbs for which the corresponding form is attested in the corpus can be included. Some verbs did not occur in the relevant form even once, and therefore the analysis based on these two candidate basic forms are carried out on less rich samples. Table 1 presents the numbers of verbs used in the analysis.

Some conjugated verb forms have two variants considered acceptable, e.g., one can say either *děkuji* or *děkuju* ‘I thank (you)’.³ If such a form is considered the basic form in an analysis, the more frequent variant is taken as the basic one.

In the annotation used here, affirmative and negative verb forms are traditionally lemmatized under the same basic form, namely the affirmative lemma. For the purposes of the present study, however, this would be problematic, since in Czech negation is usually expressed by the prefix *ne-*, which would

increase the Levenshtein distance of all negative forms by 2. We therefore decided to treat affirmative and negative forms separately as two distinct lemmas (i.e., all basic forms in our study). Negative counterparts of the relevant basic forms were generated by adding the prefix *ne-*. The only irregular case is the verb *být* ‘to be’: instead of the expected form *neje*, Czech uses the form *není*, which was therefore used in the analysis. This is the only irregularity in the negation of verbs in Czech.

Finally, all words were phonologically transcribed following the rules presented in Pelegriová (2023: 111–112).

The Levenshtein distance (Levenshtein, 1966) is the distance between two character strings (in our case, between two verb forms) measured by the minimum number of single-character deletions, additions, or substitutions necessary to change one string to another.⁴ We exemplify it in Table 2 for several forms of verb *dělat* ‘to do’, with its present indicative third person singular *dělá* considered as the basic form.

3 Results

The distributions of the Levenshtein distance values for all three considered basic verb forms – the infinitive, the present indicative first person singular, and the present indicative third person singular – are presented in Tab. 3 and in Figs. 1–3, respectively.

We note that there are some verb forms for which the annotation used (c.f., Section 2) is not able to determine some grammatical categories; e.g., verb *získat* ‘to get’ has the same form *získala* for both the past indicative third person singular in the feminine gender⁵ and for the past indicative third person plural in the neuter gender.⁶ Such verb forms disregarded in Table 1 because the category of grammatical number cannot be determined from the verb form alone. On the other hand, these ambiguities do not present any problems when the Levenshtein distance is evaluated;⁷ therefore they are included in computations resulting in values presented in Table 3.

⁴It is known also as the edit distance. Deza and Deza (2016: 216) spell it Levenstein distance.

⁵*Moje dcera získala potřebný souhlas.* ‘My daughter got the necessary approval.’

⁶*Všechna místa získala dobré hodnocení.* ‘All the places got good reviews.’

⁷The Levenshtein distance depends on the verb form only. If a verb has the same form, e.g., for singular and plural, the distance is the same regardless of grammatical number.

²Considering passive voice forms as adjectives is the only influence of the annotation schema. Otherwise, we take into account verb forms as they appear in the corpora/treebanks.

³See the online handbook of Czech language provided by the Czech Language Institute of the Czech Academy of Sciences, <https://prirucka.ujc.cas.cz/?slovo=d%C4%9Bkuju#bref3>.

	present		past		future		total	
	f.	r.	f.	r.	f.	r.	f.	r.
first person singular	4,342	0.05	2,805	0.06	122	0.03	7,269	0.05
second person singular	919	0.01	87	<0.01	37	0.01	1,043	0.01
third person singular	59,885	0.64	32,727	0.67	2,979	0.65	95,591	0.65
first person plural	5,731	0.06	1,581	0.03	284	0.06	7,596	0.05
second person plural	1,898	0.02	251	0.01	47	0.01	2,196	0.01
third person plural	20,376	0.22	11,155	0.23	1,130	0.25	32,661	0.22
participle singular	171	<0.01	4	<0.01			175	<0.01
participle plural	22	<0.01					22	<0.01
total	93,344	1.00	48,610	1.00	4,599	1.00	146,553	1.00
infinitive							26,198	

Table 1: Frequencies (f.) and relative frequencies (r.) of verb forms.

basic form	another form	LD	operations on the basic form
<i>dělá</i>	<i>dělám</i> (present indicative first person singular)	1	-m added
	<i>dělali jsme</i> (past indicative third person plural)	7	-á substituted with a, -li jsme added
	<i>budeš dělat</i> (future indicative second person singular)	7	budeš added, -á substituted with a, -t added
	<i>dělej</i> (imperative second person singular)	2	-á substituted with e, -j added

Table 2: Examples of Levenshtein distances (LD) for some forms of the verb *dělat* ‘to do’, its present indicative third person singular *dělá* is the basic form.

LD	infinitive		pres. ind. 1st pers. sing.		pres. ind. 3rd pers. sing.	
	f.	r.	f.	r.	f.	r.
0	29,170	0.15	3,914	0.03	60,351	0.33
1	16,827	0.09	30,210	0.24	10,465	0.06
2	62,600	0.32	38,692	0.31	45,855	0.25
3	54,069	0.28	17,587	0.14	35,343	0.19
4	20,628	0.11	22,817	0.18	16,005	0.09
5	5,203	0.03	2,981	0.02	5,003	0.03
6	3,941	0.02	4,020	0.03	5,168	0.03
7	1,530	0.01	1,616	0.01	2,296	0.01
8	530	<0.01	1,053	0.01	1,573	0.01
9	237	<0.01	329	<0.01	674	<0.01
10	58	<0.01	108	<0.01	188	<0.01
11			16	<0.01	29	<0.01
12			15	<0.01	15	<0.01
total	194,793	1.00	123,358	1.00	182,965	1.00

Table 3: Frequencies (f.) and relative frequencies (r.) of Levenshtein distances, three approaches.

basic form	mean LD
infinitive	2.33
present indicative 1st person sing.	2.55
present indicative 3rd person sing.	2.00

Table 4: Mean values of Levenshtein distance for three different approaches.

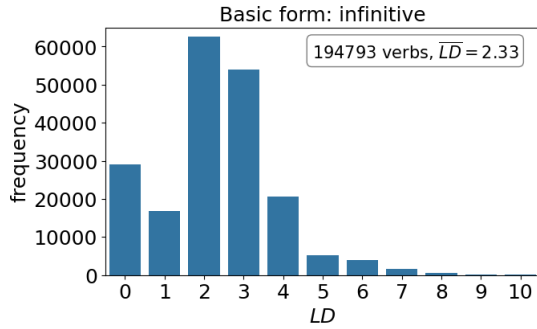


Figure 1: Frequency distribution of LD considering infinitive as the basic verb form. The frequencies are listed in Tab. 3, left.

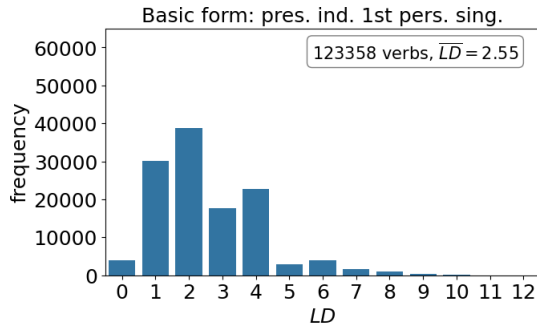


Figure 2: Frequency distribution of LD considering present indicative first person singular as the basic verb form. The frequencies are listed in Tab. 3, middle.

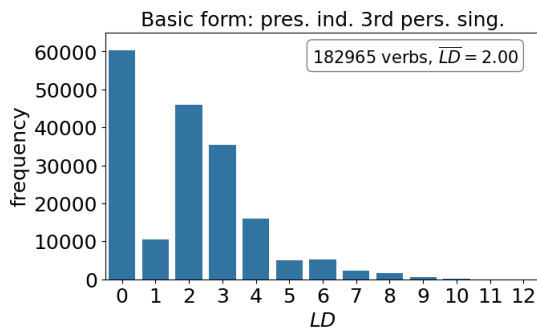


Figure 3: Frequency distribution of LD considering present indicative third person singular as the basic verb form. The frequencies are listed in Tab. 3, right.

The approach which considers the present indicative third person singular as the basic form yields the lowest mean Levenshtein distance across the verb form tokens (i.e., frequencies of particular verb forms are taken into account) in corpus (see Table 4). Therefore, it is used for the analysis that follows. This form is also the only one of the three candidates which exhibits an inverse relationship between the number of formal changes with respect to the basic form and the form frequency (i.e., the higher value of the Levenshtein distance from the basic form, the lower the frequency). However, the decrease of the frequencies is not as regular as for nouns (see Kelih et al., 2023; Mačutek et al., 2023, 2025). The frequency of 1 is clearly underrepresented, see the last two columns in Table 3 and Figure 3.

This apparent irregularity prompts us to explore the reasons why there are so few verb forms with the Levenshtein distance of 1 from the present indicative third person singular. Browsing through Czech verb classes (Komárek et al., 1986: 427–495), it becomes obvious that this value is very typical for the present indicative singular in the first and second person throughout the classes.⁸ However, according to Jakobson (1990), the first and second person are ‘personal’, as they are participants in the speech event, while the third person, as a non-participant, is ‘impersonal’. The feature of ‘personal’ is marked, as opposed to the unmarked ‘impersonal’, which is reflected in the tendency to use zero morpheme in third person (see also Bybee, 2010). Thus, verb forms used for the third person are shorter than the other forms. Zipf’s law of brevity (Zipf, 1949) predicts a positive correlation between length and frequency, i.e., shorter language units occur more often. Indeed, the frequency of the third person clearly dominates over the first two, and verb forms in the first or second person in the present tense constitute less than 3.6% of all verb forms for which the categories of person and tense can be determined (see Tables 1 and 5).

There are also instances of the Levenshtein distance of 1 for some other combination of grammatical categories, such as

- *hrají* ‘play’ in the present indicative third person plural from the basic form *hraje*,
- *vedl* ‘lead’ in the past indicative third person

⁸See also forms for present indicative of the verb *dělat* ‘to do’ in Table 9. However, there are some exceptions, like, e.g., the verb *být* ‘to be’.

masculine from the basic form *vede*,

- *myjte* ‘wash’ in the imperative second person plural from the basic form *myje*.

However, these instances are specific to particular verb classes rather than general. Moreover, plural occurs less often than singular, past tense less often than present, and imperative much less often than indicative (see Tables 6, 7, and 8, respectively). With the frequencies of these verb forms being relatively low, the Levenshtein distance of 1 cannot achieve a frequency high enough for the distribution from Figure 3 to behave more regularly.

4 Conclusion

This study has two main aims. First, it examines whether Czech verbs follow the expected inverse relation between the number of formal changes and form frequency. Second, it attempts to determine which verb form can best serve as the basic form for this type of analysis.

As for the identification of the basic verb form, three candidates were compared: the infinitive, the present indicative first person singular, and the present indicative third person singular. The results show that the present indicative third person singular yields the lowest mean Levenshtein distance from the remaining verb forms. This finding is in line with the view that third person represents the unmarked form with respect to the person opposition (Jakobson, 1990). Further, the results corroborate the expected inverse relation between the number of formal changes with respect to the basic form and form frequency, as previously observed for Slavic noun inflection (Kelih et al., 2023; Mačutek et al., 2023, 2025). Compared with nouns, however, verbs have a richer paradigm (in particular, Czech verbs have both synthetic and analytic forms) and behave less regularly with respect to this relation. Finally, our findings are also compatible with the general observation by Haspelmath and Sims (2010: 265–276) that more frequent grammatical forms tend to be shorter.

It remains an open question whether the same results would be achieved for spoken corpora. They can contain more dialogues where one can suppose a higher frequency of the second person forms.

We are aware of the fact that this paper mostly remains at the descriptive level. On the other hand, such a pilot study suggests a methodological approach and brings first basic insights into the topic.

Further investigations (ideally, on material from several typologically diverse languages) are necessary to embed the results into a broader linguistic theoretical framework.

Acknowledgments

This research was supported by projects APVV-25-0505 and VEGA 2/0120/24 (J. Mačutek and M. Koščová), by Moravian-Silesian Region, grant programme “Podpora vědy a výzkumu v Moravskoslezském kraji 2023”, project “Podpora talentovaných studentů doktorského studia na Ostravské univerzitě VII”, project No. 00516/2024/RRC (M. Nogolová) and by European Regional Development Fund project “A lifetime with language: the nature and ontogeny of linguistic communication (LangInLife)” (reg. no.: CZ.02.01.01/00/23_025/0008726) (R. Čech).

<i>LD</i>	first person		second person		third person	
	f.	r.	f.	r.	f.	r.
0					60,349	0.44
1	3,487	0.23	638	0.20	6,038	0.04
2	5,410	0.36	1,636	0.51	22,669	0.16
3	213	0.01	213	0.07	29,250	0.21
4	307	0.02	223	0.07	11,392	0.08
5	355	0.02	72	0.02	4,378	0.03
6	1,621	0.11	147	0.05	3,397	0.02
7	1,729	0.12	147	0.05	420	<0.01
8	908	0.06	97	0.03	568	<0.01
9	624	0.04	45	0.01	5	<0.01
10	169	0.01	19	0.01		
11	27	<0.01	2	<0.01		
12	15	<0.01				
total	14,865	1.00	3,239	1.00	138,466	1.00

Table 5: Frequencies (f.) and relative frequencies (r.) of *LD* distinguished by **person**.

<i>LD</i>	singular		plural	
	f.	r.	f.	r.
0	57,250	0.55	3,101	0.07
1	5,270	0.05	4,911	0.12
2	15,333	0.15	13,043	0.31
3	12,773	0.12	12,470	0.29
4	5,747	0.06	3,617	0.09
5	2,336	0.02	1,484	0.03
6	3,113	0.03	1,380	0.03
7	1,138	0.01	1,109	0.03
8	814	0.01	758	0.02
9	247	<0.01	427	0.01
10	36	<0.01	152	<0.01
11	12	<0.01	17	<0.01
12	9	<0.01	6	<0.01
total	104,078	1.00	42,475	1.00

Table 6: Frequencies (f.) and relative frequencies (r.) of *LD* distinguished by **number**.

<i>LD</i>	present		past		future	
	f.	r.	f.	r.	f.	r.
0	60,351	0.64				
1	8,974	0.10	1,207	0.02		
2	12,217	0.13	17,430	0.30	126	0.03
3	7,050	0.07	21,397	0.37	1,317	0.29
4	1,028	0.01	10,274	0.18	644	0.14
5	1,982	0.02	2,776	0.05	54	0.01
6	2,168	0.02	1,709	0.03	1,288	0.28
7	225	<0.01	1,754	0.03	317	0.07
8	141	<0.01	689	0.01	743	0.16
9	186	<0.01	460	0.01	28	0.01
10	27	<0.01	81	<0.01	80	0.02
11	3	<0.01	24	<0.01	2	<0.01
12			15	<0.01		
total	94,352	1.00	57,816	1.00	4,599	1.00

Table 7: Frequencies (f.) and relative frequencies (r.) of *LD* distinguished by **tense**.

<i>LD</i>	indicative		conditional		imperative	
	f.	r.	f.	r.	f.	r.
0	60,349	0.62				
1	8,702	0.09			255	0.17
2	11,736	0.12			553	0.37
3	7,969	0.08	22	<0.01	288	0.19
4	768	0.01	530	0.10	359	0.24
5	195	<0.01	1,930	0.37	42	0.03
6	2,858	0.03	2,168	0.41		
7	2,070	0.02	225	0.04		
8	1,431	0.01	141	0.03		
9	488	0.01	186	0.04		
10	161	<0.01	27	0.01		
11	26	<0.01	3	<0.01		
12	15	<0.01				
total	96,768	1.00	5,232	1.00	1,497	1.00

Table 8: Frequencies (f.) and relative frequencies (r.) of *LD* distinguished by **mood**.

References

- Eduard Bejček, Eva Hajičová, Jan Hajič, Pavlína Jínová, Václava Kettnerová, Veronika Kolářová, Marie Mikulová, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Magda Ševčíková, Jan Štěpánek, and Šárka Zikánová. 2013. [Prague Dependency Treebank 3.0](#).
- Joan L. Bybee. 2010. [Markedness: Iconicity, economy, and frequency](#). In Jae Jung Song, editor, *The Oxford Handbook of Linguistic Typology*, pages 131–147. Oxford University Press, Oxford.
- Michel M. Deza and Elena Deza. 2016. [Encyclopedia of Distances](#), 4th edition. Springer, Berlin / Heidelberg.
- Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. [SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD](#). In Marie-Catherine de Marnette, Teresa Lynn, and Sebastian Schuster, editors, *Proceedings of the Second Workshop on Universal Dependencies Workshop (UDW 2018)*, pages 66–74. ACL, Brussels.
- Martin Haspelmath and Andrea D. Sims. 2010. [Understanding Morphology](#), 2nd edition. Routledge, London / New York.
- Roman Jakobson. 1990. Shifters and verbal categories. In Linda R. Waugh and Monique Monville-Burston, editors, *On Language: Roman Jakobson*, pages 386–39. Harvard University Press, Cambridge (MA).
- Tomáš Jelínek. 2017. [FicTree: A manually annotated treebank of Czech fiction](#). In *ITAT 2017 Proceedings: Information Technologies – Applications and Theory*, volume 1885 of *CEUR Workshop Proceedings*, pages 181–185.
- Emmerich Kelih, Ján Mačutek, Michaela Koščová, and Vladimír Benko. 2023. [Nouns more similar to the nominative form are more frequent: A case study in Slovak](#). *Glottology*, 14(1):69–80.
- Miroslav Komárek et al. 1986. *Mluvnice češtiny 2. Tvarosloví [A Grammar of Czech 2: Morphology]*. Academia, Praha.
- Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710.
- Ján Mačutek, Emmerich Kelih, Michaela Koščová, and Mirjam Sepesy Maučec. 2025. [Noun declension in Slavic languages: Animacy has a stronger influence than gender](#). In Adam Pawłowski, Sheila Embleton, Ján Mačutek, and Aris Xanthos, editors, *Mathematical Modelling in Linguistics and Text Analysis: Theory and Applications*, pages 27–42. Benjamins.
- Ján Mačutek, Michaela Koščová, Emmerich Kelih, and Radek Čech. 2023. [Frequency and morphological behaviour of nouns in Czech and Russian](#). *Bohemistika*, 23(1):109–117.
- Klára Osolsobě. 2017. [Sloveso \[Verb\]](#).
- Kateřina Pelegrinová. 2023. [Menzerath-Altmann law in Czech](#). Ph.D. thesis, University of Ostrava.
- Petr Plecháč and Robert Kolár. 2015. [The corpus of Czech verse](#). *Studia Metrica et Poetica*, 2(1):107–118.
- Barbora Vidová Hladká, Jan Hajič, Jirka Hana, Jaroslava Hlaváčová, Jiří Mírovský, and Jan Raab. 2008. [The Czech Academic Corpus 2.0 guide](#). *The Prague Bulletin of Mathematical Linguistics*, 89:41–96.
- Daniel Zeman et al. 2023. [Universal Dependencies 2.13](#).
- George K. Zipf. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge (MA).

Appendix A Variety of verb forms in Czech language

The verb forms of infinitive *dělat* ‘to do’ are shown in Tables 9–28. Less frequent verb forms (considered archaic) are denoted by *.

	sg.	pl.
1.	<i>dělám</i>	<i>děláme</i>
2.	<i>děláš</i>	<i>děláte</i>
3.	<i>dělá</i>	<i>dělají</i>

Table 9: Present tense, indicative.

	sg.	pl.
1.	<i>dělal bych</i>	<i>dělali bychom</i>
2.	<i>dělal bys</i>	<i>dělali byste</i>
3.	<i>dělal by</i>	<i>dělali by</i>

Table 10: Present tense, conditional, masculine.

	sg.	pl.
1.	<i>dělala bych</i>	<i>dělaly bychom</i>
2.	<i>dělala bys</i>	<i>dělaly byste</i>
3.	<i>dělala by</i>	<i>dělaly by</i>

Table 11: Present tense, conditional, feminine.

	sg.	pl.
1.	<i>dělalo bych</i>	<i>dělala bychom</i>
2.	<i>dělalo bys</i>	<i>dělala byste</i>
3.	<i>dělalo by</i>	<i>dělala by</i>

Table 12: Present tense, conditional, neuter.

	sg.	pl.
masculine	<i>dělaje</i>	<i>dělajíce</i>
feminine, neuter	<i>dělajíc</i>	

Table 13: Present tense, transgressive.*

	sg.	pl.
1.	<i>dělal jsem</i>	<i>dělali jsme</i>
2.	<i>dělal jsi</i>	<i>dělali jste</i>
3.	<i>dělal</i>	<i>dělali</i>

Table 14: Past tense, indicative, masculine.

	sg.	pl.
1.	<i>dělala jsem</i>	<i>dělaly jsme</i>
2.	<i>dělala jsi</i>	<i>dělaly jste</i>
3.	<i>dělala</i>	<i>dělaly</i>

Table 15: Past tense, indicative, feminine.

	sg.	pl.
1.	<i>dělalo jsem</i>	<i>dělala jsme</i>
2.	<i>dělalo jsi</i>	<i>dělala jste</i>
3.	<i>dělalo</i>	<i>dělala</i>

Table 16: Past tense, indicative, neuter.

	sg.	pl.
1.	<i>budu dělat</i>	<i>budeme dělat</i>
2.	<i>budeš dělat</i>	<i>budete dělat</i>
3.	<i>bude dělat</i>	<i>budou dělat</i>

Table 17: Future tense, indicative.

	sg.	pl.
1.	<i>byl bych dělal</i>	<i>byli bychom dělali</i>
2.	<i>byl bys dělal</i>	<i>byli byste dělali</i>
3.	<i>byl by dělal</i>	<i>byli by dělali</i>

Table 18: Past tense, conditional, masculine, variant 1.*

	sg.	pl.
1.	<i>byl bych býval dělal</i>	<i>byli bychom bývali dělali</i>
2.	<i>byl bys býval dělal</i>	<i>byli byste bývali dělali</i>
3.	<i>byl by býval dělal</i>	<i>byli by bývali dělali</i>

Table 19: Past tense, conditional, masculine, variant 2.*

	sg.	pl.
1.	<i>býval bych dělal</i>	<i>bývali bychom dělali</i>
2.	<i>býval bys dělal</i>	<i>bývali byste dělali</i>
3.	<i>býval by dělal</i>	<i>bývali by dělali</i>

Table 20: Past tense, conditional, masculine, variant 3.*

	sg.	pl.
1.	<i>byla bych dělala</i>	<i>byly bychom dělaly</i>
2.	<i>byla bys dělala</i>	<i>byly byste dělaly</i>
3.	<i>byla by dělala</i>	<i>byly by dělaly</i>

Table 21: Past tense, conditional, feminine, variant 1.*

	sg.	pl.
1.	<i>byla bych bývala dělala</i>	<i>byly bychom bývaly dělaly</i>
2.	<i>byla bys bývala dělala</i>	<i>byly byste bývaly dělaly</i>
3.	<i>byla by bývala dělala</i>	<i>byly by bývaly dělaly</i>

Table 22: Past tense, conditional, feminine, variant 2.*

	sg.	pl.
1.	<i>bývala bych dělala</i>	<i>bývaly bychom dělaly</i>
2.	<i>bývala bys dělala</i>	<i>bývaly byste dělaly</i>
3.	<i>bývala by dělala</i>	<i>bývaly by dělaly</i>

Table 23: Past tense, conditional, feminine, variant 3.*

	sg.	pl.
1.	<i>bylo bych dělalo</i>	<i>byla bychom dělala</i>
2.	<i>bylo bys dělalo</i>	<i>byla byste dělala</i>
3.	<i>bylo by dělalo</i>	<i>byla by dělala</i>

Table 24: Past tense, conditional, neuter, variant 1.*

	sg.	pl.
1.	<i>bylo bych bývalo dělalo</i>	<i>byla bychom bývala dělala</i>
2.	<i>bylo bys bývalo dělalo</i>	<i>byla byste bývala dělala</i>
3.	<i>bylo by bývalo dělalo</i>	<i>byla by bývala dělala</i>

Table 25: Past tense, conditional, neuter, variant 2.*

	sg.	pl.
1.	<i>bývalo bych dělalo</i>	<i>bývaly bychom dělaly</i>
2.	<i>bývalo bys dělalo</i>	<i>bývala byste dělala</i>
3.	<i>bývalo by dělalo</i>	<i>bývala by dělala</i>

Table 26: Past tense, conditional, neuter, variant 3.*

	sg.	pl.
1.		<i>dělejme</i>
2.	<i>dělej</i>	<i>dělejte</i>

Table 27: Imperative.

	masculine	feminine
Present tense, indicative	<i>děláte</i>	<i>děláte</i>
Present tense, imperative	<i>dělejte</i>	<i>dělejte</i>
Present tense, conditional	<i>dělal byste</i>	<i>dělala byste</i>
Past tense, indicative	<i>dělal jste</i>	<i>dělala jste</i>
Past tense, conditional, variant 1*	<i>byl byste dělal</i>	<i>byla byste dělala</i>
Past tense, conditional, variant 2*	<i>byl byste býval dělal</i>	<i>byla byste bývala dělala</i>
Past tense, conditional, variant 3*	<i>býval byste dělal</i>	<i>bývala byste dělala</i>

Table 28: Honorific variations of specific forms in 2. person singular.