Global Variants in the Czech Language

Jaroslava Hlaváčová, Lukáš Kyjánek and Magda Ševčíková

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics Malostranské náměstí 25, 118 00 Prague, Czechia

Abstract

There are words written in several different ways in Czech, e.g., lampion ~ lampión (lampion). This variability may occur in either some inflectional wordforms (inflectional variants), cf. bradu ~ bradě in the locative case of the noun brad (castle), or across the inflectional wordforms and derivatives (global variants), cf. fantazijní ~ fantazijní in the adjective derived from the noun fantazie ~ fantazie (fantasy). It is reasonable to distinguish the global variants as different words but to have formal means that interconnect them in the Natural Language Processing systems and resources. In this paper, we describe the identification of global variants in the Czech vocabulary and summarise new changes in the MorfFlex CZ dictionary and DeriNet lexicon concerning this type of variants. We reviewed several typical patterns within global variants captured in the available resources and combined a set of regular expressions with manual annotations to achieve the highest precision of the identification.

Keywords

global and inflectional variant, morphology, word derivation, Czech

1. Introduction

The written form is one of the possible representations of languages. Czech speakers must learn and use a relevant script, rules, and regularities of the respective writing system because of its substantial standardisation and codification. However, some words can be still written in several slightly different ways in so-called orthographic (spelling) variants, e.g., citron ~ citrón (lemon), muzeum ~ museum (museum), peepshow ~ peepšou ~ píšou (peepshow).

The emergence of orthographic variants in Czech is influenced by various aspects like the spoken representation of Czech, language development, and language contact. Some cases of the variability are only temporary until the use of one of the orthographic variants is established and codified as the preferred one (which can take years or decades). However, codified or not, many of the orthographic variants appear in the texts produced by speakers, which complicates work with language resources and all sorts of NLP applications.

Adhering to the current decisions on annotating this phenomena in the corpus PDT-C (cf. its manual in [10]), we distinguish two types of orthographic variants. Inflectional variants refer to relatively regular variants within a set of wordforms of a given word, e.g., the locative case of some masculine inanimate nouns like obchod (shop): obchodu ~ obchodbě. Global variants1 address the variability in all the inflectional forms, often also in derived words, e.g., the prothetic v- attached to the noun obchod ~ vobchod (shop), and to the derived verb obchodovat ~ vobchodovat (to trade), and the derived adjective obchodní ~ vobchodní (commercial). All those words manifest the same difference in every single wordform of their inflectional paradigms (for instance, genitive cases of the noun (obchodu ~ vobchodu) and adjective (obchodního ~ vobchodního)). On the other hand, somewhere between the two defined types, there are also several cases in which the variability is limited to a few forms, cf. the infinitive and past participles of the verb myslet ~ myslit (to think), while the remaining wordforms are identical.

The inflectional variants are captured in the morphological dictionary MorfFlex CZ (hereafter MorfFlex) by means of the 15th position in the morphological tag describing morphological categories of a given wordform [2]. Until the 2020 edition of MorfFlex, there was no distinction between the description of global and inflectional variants. All the variants were marked at the 15th position of the Prague positional tagset [1]. In the last version, the global variants are annotated by means of links between them. In MorfFlex, one word from the n-tuple of variants is selected as the basic one. All other variants are linked with the basic one by means of additional pieces of information in their lemma. This information contains not only the basic variant, but also the (rough) complete description of the morphological annotation of the corpus PDT-C [10, pp. 36–42]. We will stick to the term global, as it is shorter, but the two terms are equivalent.

1Global variants are also called full-paradigm variants in the

1ITAT’22: Information technologies – Applications and Theory, September 23–27, 2022, Zborov, Slovakia

2hlavacova@ufal.mff.cuni.cz (J. Hlaváčová);
kyjanek@ufal.mff.cuni.cz (L. Kyjánek);
sevcikova@ufal.mff.cuni.cz (M. Ševčíková)

© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)
The final selection of the basic variant depended on the (over non-standard) variant was taken as the basic one. The conclusions of this paper.

Second, they contribute to a wider linguistic discussion to the improvement of the Natural Language Processing applications for which the morphological dictionaries have potential in two research directions. First, they lead to the improvement of the Natural Language Processing applications for which the morphological dictionaries serve as the background data, cf. [11], [13], and [12]. After manual filtration of the obtained global variants, our work resulted in interconnecting global variants in MorfFlex. They will appear in its next version. They were also partly uploaded to the DeriNet 2.1 lexicon.

The resulting data and categorisation (Sections 4 and 5) have potential in two research directions. First, they lead to the improvement of the Natural Language Processing applications for which the morphological dictionaries serve as the background data, cf. [11], [13], and [12]. Second, they contribute to a wider linguistic discussion on (not only orthographic) variability in Czech, especially in the context of border cases between inflectional and global variants mentioned above and exemplified before the conclusions of this paper.

2. Available Resources Containing Variants

Since any researcher working on a lexical resource must inevitably process also the orthographic variants, there are some pieces of annotations of them in the existing language resources for Czech. However, capturing variants is not the primary goal in any of the resources, so systematic care is more than needed in this kind of annotation. The available digitised resources provide a good point of departure for the phenomenon of orthographic variants, but we see the following two shortcomings/inconsistencies in them. The number of the captured orthographic variants is often relatively low in the resources. The orthographic variants are not treated across derivation, e.g., úřad ~ ouřad (office) but not úřadovat ~ ouřadovat (to officiate).

There are various language resources and grammar books that include this kind of annotation; however, in the following paragraphs, we describe only those that are available and digitised, and thus machine-readable.

MorfFlex 2.0 [2], the lexicon of the inflectional morphology of Czech, in its currently available version, already includes annotation of global variants. It classifies them into three types of variants: standard (label DS, e.g., lavor ~ lavor (pail)), common Czech/non-standard (label GC, e.g., sprášit ~ voprášit (to dust down)) and distortion/typo (label DS, e.g., Dominigue ~ Dominique). However, as the results of the work herein show, we were able to find many more n-tuples of variants not included in the current version.

VALLEX 3.0 [9] is the valency lexicon of Czech verbs that also interconnects and labels several orthographic variants of verbs. Their representations are systematic, but the definition of being a variant seems different from ours. For instance, VALLEX marks also chytít ~ chytnout (to catch) as variants although its status is arguable.
We started with assembling the existing resources listed in Section 2 and extracting variants from them. While the extractions from MorfFlex and VALLEX were easy, as these resources are designed to be processed automatically, we had to use regular expressions to extract candidates for variants from SSJČ because, in its digitised version, it stores data in a semi-structured file format (see Section 2). Consequently, we checked whether the words extracted from this resource are attested in the vocabulary of MorfFlex to mitigate incorrectly extracted, non-existent, and archaic words. We also wrote down examples and patterns from the relevant linguistic studies like those cited in the previous section.

When comparing the n-tuples of variants extracted from the resources, we have observed that MorfFlex includes more than two thirds of the variants captured in SSJČ. The remaining third of the variants from SSJČ seems questionable, e.g., zvýhodněný ∼ zvýhodnělý (privileged) in which the variability does not affect the individual characters but the affixes (and thus can diverge the word meaning). The extracted variants from VALLEX cover only verbs; most of the variants are not included in the other resources.

### 3. Searching for Global Variants

We developed a semi-automatic procedure that consists of four straightforward subsequent steps to search for Czech global variants. We exploited the available resources mentioned in the previous section, and we extracted frequent patterns that appear in global variants. We applied these patterns to the set of lemmas from MorfFlex 2.0 to obtain n-tuples of global variants, such as extremismus ∼ extrémismus ∼ extremizmus ∼ extrémizmus (extremism).

To get also derivationally related global variants that were not identified on the basis of the extracted patterns, we included DeriNet into the process. Thanks to that we covered also n-tuples like extremistický ∼ extrémistický (extremist) for the already identified variants of the base word extremism mentioned above. The resulting n-tuples were manually annotated to eliminate randomly similar words. In the last step, the data were uploaded as a new type of annotation into DeriNet, and, in parallel, new links were also added to MorfFlex.

#### 3.1. Extracting Variants from the Existing Resources

We started with assembling the existing resources listed in Section 2 and extracting variants from them. While

#### 3.2. Formalising Regular Patterns

Having relatively large amount of n-tuples with global variants, we considered whether to use pattern-matching algorithms or to formalise frequent patterns manually. We chose the latter way as it allowed us to have a better overview of the processed data and to create more complicated patterns that would take into account not only character changes but also morpho-syntactic categories of words. For instance, this decision allowed us to avoid interconnection of the masculine animate variant pair česač ∼ česáč (a man harvesting apples) to the masculine inanimate variant pair česáč ∼ česač (an instrument for harvesting fruits of tall trees).

We exploited various types of intersections of the extracted lists of variant n-tuples and their sorting to be able to infer frequent patterns that occur in global variants. We first looked at the global variants extracted from all three resources, then at those that occurred in at least two resources, and only then at those that were in individual resources but not in the others. More than one hundred observed patterns were formalised into the form of regular expressions, e.g. "o.* ↔ "vo.* in obchod ∼ vobchod (shop). The relevant morpho-syntactic categories were also stored with the particular regular expression.

#### 3.3. Applying Patterns to MorfFlex

We applied the formalised patterns to all the lemmas from MorfFlex (we did not search for inflectional variants, so we did not have to take wordforms into account). To
We uploaded the resulting global variants into the newest version of DeriNet 2.1, and we intend to do so also for the next version of the inflectional dictionary MorfFlex.

<table>
<thead>
<tr>
<th>n</th>
<th>MorfFlex 2.0</th>
<th>after</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>31,919</td>
<td>49,079</td>
</tr>
<tr>
<td>3</td>
<td>1,227</td>
<td>2,089</td>
</tr>
<tr>
<td>4</td>
<td>121</td>
<td>264</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>6</td>
<td>187</td>
<td>187</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1. Number of interlinked variants in MorfFlex 2.0 (the second column) and after addition of the new variant annotation (the third column). The first column lists sizes of n-tuples — n = 2 is for pairs.

Both resources differ in the data structures they use for storing their data, but they both share the same set of lemmas.

DeriNet interconnects derivationally related words into so-called DERIVATIONAL FAMILIES. Each family of words is represented in a form of rooted tree (in graph theory terminology), in which words are represented as nodes while derivational relations as edges. In other words, each derived word in DeriNet has at maximum one base word (antecedent), e.g., učitelka (female teacher) ← učitel (teacher) ← učit (to teach).

In the rooted tree data structure, unidentified global variants caused structural inconsistencies. For instance, the adjective citrónový (related to lemon) could be connected to the noun citron, although the noun citrón would be a better antecedent (both global variants of lemon). To tackle this issue, identifying global variants is crucial.

We considered two possible ways of representing global variants in the current rooted tree data structure of DeriNet. In the first approach (see Fig. 1, part A), the global variants would create parallel branches in the tree, e.g., citron → citrónový → citronově parallel to citrón →
The major disadvantage of this approach is that the branches may be disconnected or contain gaps if any variant is missing in the vocabulary.

In the second approach (see Fig. 1, part B), the global variants would be connected to one basic variant to which the derivatives are connected, while the other variants would not have any derivatives connected, e.g., \([\text{citron} \sim \text{citrón}] \rightarrow [\text{citronový} \sim \text{citrónový}] \rightarrow [\text{citrónově} \sim \text{citrónově}]\). The lack of global variants in this approach does not disconnect word(s) from the tree. Therefore, we chose this approach for DeriNet.

The selection of the basic variant followed the similar criteria that were applied in MorfFlex. We tried to do so consistently across the \(n\)-tuples that share the same pattern. The final decision depended on a lexicographer.\(^5\)

As a result, the data from our experiments with global variants has been already uploaded into DeriNet 2.1. If words are variants in this lexicon, one of the words is selected as the basic one and the other ones are connected directly to it by special relation that is labelled as Type=Variant. Fig. 2 illustrates words derived from the variant pair \(\text{úřad} \sim \text{ouřad} (\text{office})\) from DeriNet 2.1; the missing variants of the individual derivatives, such as \(\text{úřadek} \sim \text{ouřadek}\), will be connected in the new release.

\(^4\)Un fortunately, this task was not coordinated between MorfFlex and DeriNet projects but we plan to make a unification.

\(^6\)This overview is by no means complete.

4. Prototypical Cases of Global Variants

In this section, we will present the most common types of global variants together with typical examples.\(^6\) One of the important properties of global variants is that their derivatives can also become global variants.

Example: The pair of verbs \(\text{lítat} \sim \text{létat} (\text{to fly})\) derives iterative verbs \(\text{lítávat} \sim \text{létávat}\), adjectives \(\text{lítající} \sim \text{létající}\) (flying), and/or verbal nouns \(\text{lítání} \sim \text{létání} (\text{the flying})\). Derivatives in each of the pairs are also global variants.

4.1. Long and Short Vowels

In this type of variants, words vary in the length of a vowel, either in the affix, or in the root.

Example: Suffix variation in \(\text{svíčkař} \sim \text{svíčkář}\) (someone who makes candles), and root variation in \(\text{kvíkat} \sim \text{kvíkat}\) (to oink/squeak).
4.2. Alveolar vs. Postalveolar/Palatal Consonants

The consonants alternate in the root; the instances are of different origins.
Example: vlašťovka ~ vlašťovka (a swallow), student ~ študent (student), mrazený ~ mražený (frozen).

4.3. Soft and Hard Adjectives

There are two types of adjectives — soft and hard, but some of them can vary between the two types. This was quite common in the past, as is visible from the additional information attached usually to one of the variants — it is often archaic or outdated. The basic variant can be soft as well as hard, depending on the lexicographer’s decision. At the beginning of our work, this type of variants was not recorded.
Example: Adjectival variation in the pairs námezdný ~ námezdní (hired), přívodný ~ přívodní (feed, inflow ... e.g. pipe).

4.4. Prothetic v-

Many Czech words starting with the vowel o exist also in the variation with the prothetic v- at their very beginning. Though the latter variant is considered non-standard and is used mainly in spoken Czech, it is very common and penetrating into the written Czech, too.
Example: okno ~ okono (window). This type of variants can appear not only at the beginning of words, but also after a prefix that precedes the o; e.g., zotvírat ~ zvotvírat (to open step by step).

4.5. Vocalized and Non-vocalised Prefixes

The prefixes v-, s-, vz-, roz-, od-, pod-, nad-, ob-, před- can be expanded by e (ve-, se-, vze-, roze-, ode-, pode-, nade-, obeh), pěde-). Nevertheless, some words can have both spellings, which makes them variants.
Example: střást ~ setřást (shake off), rozsmutnit ~ rozsmutnit (make sad), objet ~ obejet (go around).

4.6. Stylistic Variants (ú ~ ou, ý ~ ej, th ~ t, s ~ z)

This type of variants usually puts into opposition standard and non-standard Czech, let it be archaic, colloquial or other sort of style. The most frequent is the variation between s and z, especially within the suffixes -ismus and -izmus.
Example: mechanismus ~ mechanizmus (mechanism), vytekat ~ vytejkat (flow/leak out), úzký ~ ouzký (narrow), ortopedie ~ orthopedie (orthopedics).

4.7. Variants of Foreign Names

Most frequent foreign geographic names have usually a Czech translation.
Example: The Czech variant of Paris is Paříž, Moscow is Moskva, Berlin is Berlín.

Though both words can appear in Czech texts, they are not considered global variants. Moreover, the original of the foreign name is usually not inflected.

Person names are typically not translated, but their spelling is often unusual. In addition, errors or typos frequently occur in their spelling. In such cases, they can be considered variants. Sometimes, one of the variants is a spelling adapted to the pronunciation, as the long variant in the following example.
Example: Abdulah ~ Abdullah ~ Abduláh.

This is not applied to Slavic names with the ending -ij or -i which are sometimes translated with the ending -ý. As the variation appears only in the nominative singular (lemma) and vocative singular, we consider this type of variants as inflectional.
Example: All the three variants of the name Čajkovský (Tchaikovsky), namely Čajkovský ~ Čajkovskij ~ Čakov’skij, are inflectional variants of the singular nominative and vocative cases. Other cases do not manifest this type of variation. They are not global variants.

Similarly, names of ancient Greeks with the lemma ending -es or -és are not global variants, as this variation appears only in nominative singular. They are inflectional variants.
Example: Empedokles, Empedoklés.

5. Non-variants

The soft–hard type can seemingly be applied to soft and hard declension of nouns with feminine or masculine gender. In reality, in such cases, we should rather speak about a combined paradigm and merge the two variants into one inflectional paradigm. This has been already done for masculine declension of soft–hard pairs, both animate and inanimate.
Example: The lemma kúzél (cone) can be inflected either as a hard noun (following the traditional masculine inanimate declension class hrad) as well as a soft noun (following the traditional masculine inanimate declension class strost). It is reasonable to join wordforms of the two inflected sets and to represent the whole set of wordforms by a single lemma. As there is only one lemma, these words cannot be global variants either.

The feminine gender is different, as there the lemmas differ. However, the difference is always within the ending, so according to the definition of global variants, they are not global variants. Though they are often viewed as global variants, there is rather one inflectional paradigm with inflectional variants affecting all the wordforms.
The new pattern for this type of variation should be added and all the wordforms merged into a single inflectional paradigm with inflectional variants even for the lemma.

**Example:** The lemmas kapuce ~ kapuca (hood) have different inflectional paradigms, but the individual tags (combinations of number and grammatical case) differ only in endings.

Similar cases are variants with different genders.

**Example:** brambora (fem.), brambor (masc. inan.) (both potato); ribstol (fem.), ribstol (masc. inan.) (both wall bars).

Again, the variation manifests itself only in endings, so they cannot be considered global variants. The solution proposed for the nouns with the same gender (merging the inflectional paradigms) cannot be applied here, because of the so-called "Principle of morphological differentiation" introduced in [10]. One of its requirements is that the gender of a noun should stay the same within the whole inflectional paradigm. These examples reveal that the Principle is questionable; it would probably be advisable to reconsider it.

During the work on global variants in Czech resources, we came across several peculiarities.

**Example:** The pair pěcíčka ≅ půsíčko (a sort of abbreviation of personal computer / PC). Is it a pair of global variants, or not? For the time being, the two lemmas are not interlinked.

Sometimes, we found sets of seeming variants, that had a typical variant pattern, but they were not variants because of different meaning.

**Example:** valečka (biol. sort of grass) ≅ válečka (someone [fem.] who rolls something), and/or studenský (adj. to the town of Studená) ≅ studěnský (adj. to the town of Studénka).

Neither we interconnected the onomatopoeic or expressive words.

**Example:** děbounout ≅ děboun (expr. to push).

6. Conclusion

The paper presented the specialised project of looking for global variants in available resources of Czech lexical data. The main aim was to make an "inventory" of Czech global variants and to annotate them. Special attention was paid to the distinction between the global and inflectional ones. This distinction has been already captured in the new edition of MorfFlex 2.0, in which many pairs of global variants still remained unlinked. In particular, it was necessary to reflect the existence of global variant n-tuples in DeriNet. The newly identified global variants now are captured in the recent edition of DeriNet 2.1. Comprising of the new annotation of global variants into MorfFlex is planned for a future edition.

This project included lots of manual work, as the topic of variants is very variable and there are no rules for the really strict distinction of what are and what are not variants. Thus, the manual work discovered some border cases where it had to be decided from scratch. In general, we adopted very strict rules, e.g. we do not consider variants those words that contain formally different affixes (see the example zvůhodný ≅ zvýhodněný (privileged)). All those cases are to be researched in greater detail in the future.

References

of Mathematics and Physics, Charles University, Czechia. URL: http://hdl.handle.net/11234/1-2307.


