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. *hradu* ~ *hradě* in the locative case of the noun *hrad* (*castle*), or across the inflectional wordforms and derivatives (global variants), cf. *fantazijní* ~ *fantasijní* in the adjective derived from the noun *fantazie* ~ *fantasie* (*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*), *museum* ~ *muzeum* (*museum*), *peepshow* ~ *peepšou* ~ *pípšou* (*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* ~ *obchodě*. GLOBAL VARIANTS[1] address the variability in all the inflectional forms, often also in derived words, e.g., the prothetic *v-*[2] 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.[3] 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)

ITAT'22: Information technologies – Applications and Theory, September 23–27, 2022, Zuberec, Slovakia

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

[1] Global variants are also called FULL-PARADIGM VARIANTS in the 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.

[2] This variation originates from the common Czech and is not codified. It is sub-standard, but it occurs in the written language.

[3] There are more ways how to interconnect the global variants (see [4], [5], [6]).

style of the variant. There is no strict rule for such selection, as no of the possible criteria is easy to formulate or check. Usually, the more common (frequent) or standard (over non-standard) variant was taken as the basic one. The final selection of the basic variant depended on the lexicographer's opinion.

Until recently, no special attention was given to the completeness of global variants within the whole of the dictionary. That is why we reviewed available resources addressing the issue of orthographic variability (Section 2), and extracted typical patterns for global variants from them. We applied these patterns to the set of lemmas from MorfFlex to find as many global variant candidates as possible. We also exploited the DeriNet lexicon [14], which models word-formation relations in Czech, to search for global variants within derivationally related words (Section 3). 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.

1.1. Terminology

To sum up definitions of the basic terms used in this paper, we provide this section. We frame it to facilitate reading in case readers would like to easily remind some definitions while reading more advanced parts of the text.

— Inflectional paradigm
is a set of all wordforms derived by means of inflection from a citation wordform (so-called lemma); e.g., the inflectional paradigm of the lemma *obchod* consists of wordforms *obchod*, *obchodu*, *obchodě*, …

— Inflectional variants
are a pair (or generally *n*-tuple) of wordforms belonging to the same inflectional paradigm of the same lemma and having the same values of all morphological categories, but different spellings; e.g., singular locative case of the lemma *obchod* is *obchodu* ~ *obchodě*.

— Global variants
are a pair (or generally *n*-tuple) of lemmas whose difference in spellings propagate to all wordforms of their inflectional paradigms and to most of their derivationally related words; e.g., *obchod* ~ *vobchod* → *obchodní* ~ *vobchodní* (apart from the first letter, inflectional paradigms of these words are identical).

— Basic variant
is the representative variant for an *n*-tuple of global variants.

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 dd, e.g., *lavor* ~ *lavór* (*pail*)), common Czech/non-standard (label gc, e.g., *oprášit* ~ *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 *chytit* ~ *chytnout* (*to catch*) as variants although its status is arguable.

Slovník spisovného jazyka českého (SSJČ) [3], the explanatory dictionary of Czech digitised into a semi-structured format,[4] covers wide vocabulary and includes annotations of orthographic variants without any additional classification. However, it uses various options in which these variants are underlined in glosses of the dictionary, e.g., by using

- *v. = viz* (*see*) in *"obepsati v. opsati"* (*copy*),
- *comma* in *"mysliti, mysleti"* (*to think*), or
- *řidč. = řidčeji* (*rarely*) in *"zpěvánka, řidč. zpěvanka, zpívánka"* (*[usually a short] song*).

Except for these language resources, there is also a long tradition of linguistic studies on the topic of orthographic variants, cf. prefixes *s-*/*se-* and *z-*/*ze-* and orthography of loanwords in general in [15, pp. 167–170], morphological variability and its reflection to orthography in [16, pp. 268–276], orthography of loanwords with *s*/*z* characters, such as *analýza* ~ *analýsa* (*analysis*) in [17], and terminological issues of the phenomenon in [7]. They provide extensive lists of variants or patterns that are shared across variants. We extracted some patterns from the studies, which helped us to identify typical general properties of global variants.

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 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* $\overset{?}{\sim}$ *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.

3.2. 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 *česač* ~ *česáč* (*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

[4] https://ssjc.ujc.cas.cz/

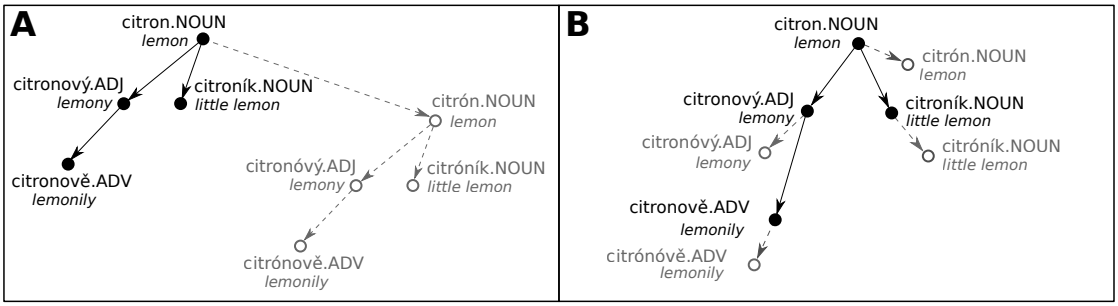

Figure 1: Two possible ways of representing global variants in the rooted trees; (A) making parallel branches, (B) connecting variants to the basic variant (the latter option implemented in DeriNet 2.1).

achieve higher precision, we also exploited the knowledge of morpho-syntactic categories of the candidates for global variants. If these categories, e.g., grammatical gender or animacy, of the candidates differed between the words, these candidates were excluded, cf. the masculine animate noun *car* (*tsar*) ⊀ the masculine inanimate *cár* (*shred*).

In order to obtain more consistent list of variants, we also took derivational morphology from DeriNet into account. For each identified global variant, we observed relevant sub-tree of derivationally related words and tried to identify the same patterns among the derivatives.

The resulting n-tuples of global variants were also manually filtered. The annotator was provided lists of n-tuples of global variant candidates; the task was to go through the lists of variants and exclude those n-tuples which only accidentally met a derivational pattern, but that were not variants, e.g., the pair *fiala* (*wallflower*) ⊀ *fiála* (*pinnacle*) had to be excluded manually, although the same pattern works well in real variants like *neandrtalec* ∼ *neandrtálec* (*Neanderthal*). During the manual work on the global variant candidates we also identified inflectional variants with variant lemma — see the example of the pair of verbs *myslit*, *myslet*.

Table 1 shows how many variant n-tuples were annotated in the MorfFlex 2.0, and how many were added thanks to the new found n-tuples. It is visible, that the main increment is recorded for smaller n, especially for pairs ($n = 2$), triples ($n = 3$) and 4-tuples. The bigger values of n remain the same.

In the following sections, we present a more detailed analysis of the prototypical cases of global variants (Section 4) and, on the other hand, cases that we do not treat as variants (Section 5).

3.4. Global Variants into DeriNet

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.

n	MorfFlex 2.0	after
2	31,919	49,079
3	1,227	2,089
4	121	264
5	16	18
6	187	187
8	4	4
9	1	1
11	1	1
12	1	1

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* → *citronový* → *citronově* parallel to *citrón* →

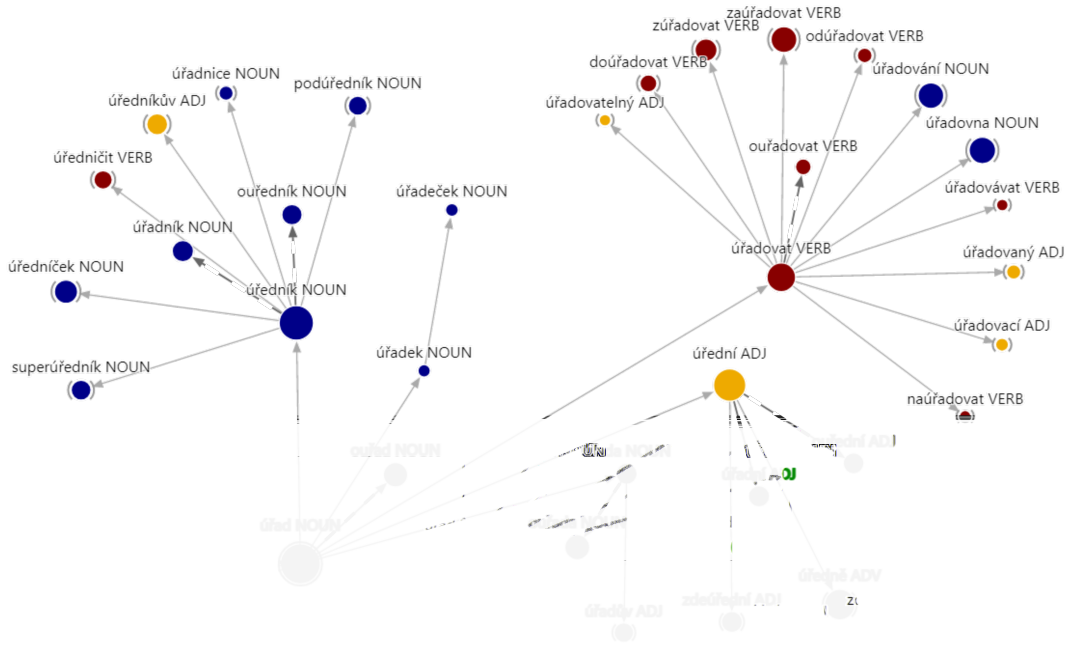

Figure 2: Simplified record of the global variants of the noun *úřad* ~ *ouřad* (*office*) and its derivatives from DeriNet 2.1. Variant relations are represented by dark grey dashed arrows that are shorter than the light grey solid arrows, which represent derivational relations. Size of the nodes corresponds to the token frequency of the lemmas in the corpus SYNv4 [8]. Brackets around nodes indicate that the node's derivatives were hidden for spatial reasons.

citrónový → *citrónově*. 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., [*citron* ~ *citrón*] → [*citronový* ~ *citrónový*] → [*citronově* ~ *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 *úřad* ~ *ouřad* (*office*) from DeriNet 2.1; the missing variants of the individual derivatives, such as *úřadek* ~ *ouřadek*, will be connected in the new release.

[5]Unfortunately, this task was not coordinated between MorfFlex and DeriNet projects but we plan to make a unification.

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 *lítat* ~ *létat* (*to fly*) derives iterative verbs *lítávat* ~ *létávat*, adjectives *lítající* ~ *létající* (*flying*), and/or verbal nouns *lítání* ~ *létání* (*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 *svíčkař* ~ *svíčkář* (*someone who makes candles*), and root variation in *kvikat* ~ *kvíkat* (*to oink/squeak*).

[6]This overview is by no means complete.

4.2. Alveolar vs. Postalveolar/Palatal Consonants

The consonants alternate in the root; the instances are of different origins.
Example: *vlaštovka* ~ *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* ~ *vokno* (*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-, obe-, přede-*). Nevertheless, some words can have both spellings, which makes them variants.
Example: *stříst* ~ *setříst* (*shake off*), *rozsmutnit* ~ *rozesmutnit* (*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*), *vytékat* ~ *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* ~ *Čajkovski*, 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 *kužel* (*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 *stroj*). 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*); *ribstole* (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 *pécéčko* $\overset{?}{\sim}$ *písíčko* (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* (*adjective to the town of Studená*) ≁ *studénský* (*adjective to the town of Studénka*).

Neither we interconnected the onomatopoeic or expressive words.

Example: *ďoubnout* ≁ *ďubnout* (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 *zvýhodněný* ≁ *zvýhodnělý* (*privileged*)). All those cases are to be researched in greater detail in the future.

Acknowledgement

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, and the Grant No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935), and LINDAT/CLARIAH-CZ project of the Ministry of Education (LM2015071, LM2018101).

References

[1] Hajič, J. 2004. Disambiguation of Rich Inflection (Computational Morphology of Czech). Nakladatelství Karolinum, Charles University, Czechia.

[2] Hajič, J.; Hlaváčová, J.; Mikulová, M.; Straka, M.; Štěpánková, B. 2020. MorfFlex CZ 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Czechia. URL: http://hdl.handle.net/11234/1-3185.

[3] Havránek, B. (ed.) 1960–1971. Slovník spisovného jazyka českého. Academia, Prague, Czechia.

[4] Hlaváčová, J. 2009. Formalizace systému české morfologie s ohledem na automatické zpracování českých textů. Ph.D. thesis, FF UK, 146 pp.

[5] Hlaváčová, J. 2011 Problém variantních tvarů slov při automatickém zpracování jazyka. In: Information Technologies – Applications and Theory, pp. 75-78.

[6] Hlaváčová J. 2019. Aggregates and Variants in Two Czech Morphological Approaches. In: Proceedings of the 19th Conference ITAT 2019: Slovenskočeský NLP workshop (SloNLP 2019), pp. 120-124.

[7] Hrbáček, J. 1974. Lexikální ekvivalenty, dublety a varianty Naše řeč 57(1), pp. 28–33.

[8] Křen, Michal et al. 2016. Corpus SYN, version 4. Prague, Institute of the Czech National Corpus, Faculty of Arts, Charles University; http://www.korpus.cz.

[9] Lopatková, M.; Kettnerová, V.; Bejček, E.; Vernerová, A.; Žaborktský, Z. 2016. VALLEX 3.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty

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

[10] Mikulová, M.; Hajič, J.; Hana, J.; Hanová, H.; Hlaváčová, J.; Jeřábek, E.; Štěpánková, B.; Vidová Hladká, B.; Zeman, D. 2020. Manual for Morphological Annotation. Revision for Prague Dependency Treebank – Consolidated 2020 release. Technical Report TR-2020-64. Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Czechia. ISSN: 1214-5521. URL: https://ufal.mff.cuni.cz/techrep/tr64.pdf.

[11] Richter, M.; Straňák, P.; Rosen, A. 2012. Korektor – A System for Contextual Spell-checking and Diacritics Completion. In: Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012), pp. 1–12. Coling 2012 Organizing Committee, Mumbai, India.

[12] Straka, M.; Straková, J. 2017. Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 88–99. Association for Computational Linguistics, Vancouver, Canada.

[13] Straková, J.; Straka, M.; Hajič, J. 2014. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 13–18. Association for Computational Linguistics, Baltimore, Maryland.

[14] Vidra, J.; Žabokrtský, Z.; Kyjánek, L.; Ševčíková, M.; Dohnalová, Š.; Svoboda, E.; Bodnár, J. 2021. DeriNet 2.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Czechia. URL: http://hdl.handle.net/11234/1-3765.

[15] Mluvnice češtiny 1: Fonetika, Fonologie, Morfonologie a morfemika, Tvoření slov. 1986. Academia, nakladatelství Československé Akademie věd, Prague, Czechia.

[16] Mluvnice češtiny 2: Tvarosloví. 1986. Academia, nakladatelství Československé Akademie věd, Prague, Czechia.

[17] Pravopis a výslovnost přejatých slov se s – z. Internetová jazyková příručka [online] (2008–2022). Ústav pro jazyk český AV ČR, Prague, Czechia. Cit. 28. 5. 2022. URL: https://prirucka.ujc.cas.cz