Aggregates and Variants in two Czech morphological approaches

Hlaváčová Jaroslava

Charles University, Faculty of Mathematics and Physics, Prague
Institute of Formal and Applied Linguistics
hlavacova@ufal.mff.cuni.cz

Abstract: There exist several morphological dictionaries for Czech. They differ only in solutions of complicated morphological features. Various attempts have been made to unify their approaches, but only some of them were implemented. The paper deals with several such features and compares their solutions taken in two different projects, namely preparation of the new edition of PDT (Prague Dependency Treebank [1]) and NovaMorf [8]. The features presented in this paper are: aggregates (the wordforms without a clear part of speech, e.g. *užs, oč, naň*), and variants – inflectional (more wordforms for a particular combination of lemma and morphological tag) as well as global ones (mainly orthographic variants expressed in all wordforms of a paradigm).

1 Two innovative projects of Czech morphology

There are several "classical" parts of speech covering almost the whole vocabulary of any language. Apart from them, there are words, that are difficult to be assigned an appropriate part of speech (POS). During the times, they were carried from one POS to another. Also, the number of POS was changing.

Some words are difficult to place into a right position among POS, because they might belong to more of them. It results in an artificial homonymy – the same word is included into more POS classes and due to formal requirements they should be considered different words. There were many attempts to solve that situation, but no generally accepted solution has been found.

Two recent projects dealing with the Czech morphology tried to solve at least some issues connected to POS classes. They are the new edition of PDT (Prague Dependency Treebank [1]) and NovaMorf [8]. Among others, they introduced new classes of POS, namely Foreign Word, (Afixal) Segment and Aggregate. The former two POS were implemented in the same way in the both projects, while the latter one, the Aggregate, has different implementations. The second common solution of the both projects, but with different implementations, concerns variants of words.

2 Aggregates

According to [4], an aggregate is a wordform that is created by combining two or more wordforms (components of the aggregate) into one and cannot be simply assigned any part of speech.

We present several examples together with their explanations[1]:

1. *viděls* = *viděl jsi* (*you saw*)
2. *studentas* = *studenta jsi* (e.g. *Toho studentas neviděl? .. You have not seen that student?*)
3. *užs* = *už jsi* (*you already were*)
4. *doň* = *do něj* (*into it*)
5. *nač* = *na co* (*on what*)
6. *načs* = *na co jsi* (*on what you were*)

Apart from the first example, where the both components are verbs, there are at least two different parts of speech contained in all the aggregates. That fact makes it difficult to assign aggregates one of traditional POS. This is the reason why a new part of speech was introduced to the system of the Czech morphology – Aggregate.

The aggregates can be divided into three groups (aggregate types):

1. Verbal aggregates – those aggregates where the second component stands for the wordform *jsi* (*you are*). The first component can be almost any POS (see the examples 1 to 3 above). In NovaMorf, conditional conjunctions *aby, kdyby* together with all their forms *abych, kdybyste, ...* are considered also verbal aggregates, which is not the case of the PDT project.

2. Pronominal aggregates – those aggregates where the second component is the pronoun *co* (*what*) or *něj* (lemma *on* = *he*). The first component is a preposition (see the examples 4 and 5 above).

3. Combined aggregates – those aggregates where the first component is a preposition, the second one is a pronoun *co* or *něj* (as in pronominal aggregates), and the third component is the auxiliar verb *jsi* (as in verbal aggregates). It is a combination of the previous two types – verbal and pronominal (see the example 6 above).

[1] The English translations might seem meaningless without a context.

The previous list contains all sorts of Czech aggregates. Having divided the types of aggregates, let us have a look how to morphologically annotate them.

Every wordform should be morphologically described as a unique couple consisting of a lemma and a morphological tag. However, there are different approaches how to do it in case of aggregates. In the following subsections we will introduce two of them, belonging to the two beforementioned projects.

2.1 NovaMorf: Multiple tag and multiple lemma for aggregates

In the NovaMorf project, the lemma of an aggregate is the sequence of lemmas of all its segments. We call such a sequence a multiple lemma of aggregates.[2]

According to the previous introductory text, there are at most three different lemmas in the multiple lemma of a Czech aggregate. This may not be the case for other, especially agglutinative languages.

There are only limited number (closed set) of pronominal and combined aggregates in Czech, that is why it is possible to describe those types of aggregates by means of a single morphological tag. The same is possible to achieve for the verbal aggregates, though there are extremely productive. The description of such an annotation is included in [4].

However, there was quite a big opposition against that solution, so the researchers adopted an alternative (equivalent) solution that is possibly more intuitive. They introduced a concept of multiple tag, as a paralel concept to the multiple lemma. The multiple tag is the sequence of the tags describing all segments of an aggregate. The sequence of tags is ordered in the same way as the sequence of lemmas in the multiple lemma of the aggregate.

The morphological description of the examples presented above are given in the table 1.[3]

2.2 PDT: Tagging aggregates within the current framework of Prague tagging system

The project PDT adopted a different solution (see also [6]). The lemma of pronominal and combined aggregates is the lemma of the pronoun. It can be either *on* or *co*, eventually also *copak*. The lemma of verbal aggregates is the lemma of their first component, which can be, as mentioned above, almost any word of almost any POS. The morphological tag of aggregates is enriched by a new category – type of aggregate – wich can have the following values:

- s for verbal aggregates;
- initial letter of the preposition in lower case for the pronominal aggregates;
- initial letter of the preposition in upper case for the combined aggregates.

These values are incorporated to the morphlological tag. The particular placement within the framework of the Prague morphological system is the 14th position of its positional tag.

Such description of aggregates could make use of the existing frame of the morphlological dictionary, without necessity to change the structure of the dictionary. On the other hand, the solution is less intuitive than the previous case of multiple lemmas and tags.

The morphological description of the examples presented above are given in the table 2.

3 Variants of words

Another problem of morphological tagging concerns different types of variants (sometimes called also mutations). There are stylistic variants, orthographic ones, historical etc. No matter what was the origin of the variant, it should be properly tagged as a variant. It is meaningful to join all the variants into a common dictionary record, but it must be done in such a way that does not disturbe the Golden rule of morphology.

The Golden rule of morphology (see [4, 5]) says that every combination of a lemma and a morphological tag should be represented by a single wordform (if the combination of the lemma and tag is meaningful, of course). This statement ensures that, for instance, the two orthographic variants of the lemma *lemon*, namely *citron* and *citrón* should not have the same morphological description – the same lemma and the same tag.

There is a distinction between global variants (manifested in all wordforms of a paradigm) and inflectional ones (applied only for some combinations of morphological values). Until recently, this distinction was not taken into account consistently. For marking a wordform as a variant, the 15th position of the Prague tag was used. Both types of variants were marked in this way.

Inflectional variants are those variants that relate only to some wordforms of a paradigm defined by a particular combination of morphological values for the identical lemma. An example are two forms of the lemma *hrad* (*castle*) in local singular which can be both *hradu* and *hradě*.

Global variants are those variants that relate to all wordforms of a paradigm, and always in the same way. The word *lemon* presented above is the example of the global variants.

[2] There is another use of the concept of multiple lemma for description of variants. It will be introduced later in this paper. The main difference between the two is that the multiple lemma of aggregates is a sequence, while the multiple lemma of variants can be a set.

[3] All the tables in this paper are presented with the morphological tags used in the Prague dictionary Morfflex [3].

Table 1: Examples of aggregates with their annotation in NovaMorf

Wordform	Stands for	Multiple Lemma	Multiple Tag
viděls	viděl jsi	{vidět, být}	{VpYS----R-AAI–, VB-S---2P-AAI--}
studentas	studenta jsi	{student, být}	{NNMS4-----A-—, VB-S---2P-AAI--}
užs	už jsi	{už, být}	{Db----------—, VB-S---2P-AAI--}
doň	do něj	{do, on}	{RR--2----------, P5ZS2--3------—}
nač	na co	{na, co}	{RR--4----------, PQ--4--------—}
načs	na co jsi	{na, co, být}	{RR--4----------, PQ--4--------—, VB-S---2P-AAI--}

Table 2: Examples of aggregates with their annotation in PDT

Wordform	Stands for	Lemma	Tag
viděls	viděl jsi	vidět	VpYS----R-AAIs-
studentas	studenta jsi	student	NNMS4-----A--s-
užs	už jsi	už	Db----------s-
doň	do něj	on	P5ZS2--3-----d-
nač	na co	co	PQ--4--------n-
načs	na co jsi	co	PQ--4--------N-

For treatment of global variants, the lemmatization is very important. In this respect, the morphological dictionary has been inconsistent. Some global variants shared the same lemma, which violated the Golden rule of morphology, because the same morphological tag combined with that lemma was connected with two different wordforms. Others were lemmatized as distinct lemmas, which made impossible to link the variants. The lemma representing the both (or sometimes more) variants should link them, but at the same time, the Golden rule of morphology must not be violated.

3.1 Variants in NovaMorf

In NovaMorf, the information about variants is added to the morphological tag. Global variant as well as inflectional variant became new (morphological) categories that have a set of predefined values. The values are based preferably on the orthographic features of individual variants; they do not attempt to have any evaluative meaning, such as expressive, vulgar, archaic or others. Contrary to other morphological categories, there can be more values of both variants. That is why a special notation was proposed to express potentially more values of the variants. No matter how the morphological tag is constructed (Prague style [2], Brno style [7], or other), the traditional tag is followed by an additional "subtag" for the variants. The information about variants are marked with a letter G for global variants and F for inflectional ones. Then, a set of codes expressing the appropriate type of the variant(s) follows.

All the global variants are then represented by a common lemma. Here again, the concept of multiple lemma is used. In the case of variants, its members have always the same POS – they are the variants themselves.

Table 3 shows an example of capturing the three global variants of the lemma *thesis* in NovaMorf.

The inflectional variants are tagged similarly; only instead the leading G, the letter F starts the sequense of the inflectional variants. Naturally, the inflectional variants need not to be represented by a multiple lemma.

3.2 Variants in PDT

In the PDT project, the inflectional variants are marked as before, at the 15th position of the tag. The list of possible values was substantially simplified; there are code numbers for marking literal and standard variants (1 – 5), and code numbers for substandard variants (6 – 9). No other finer distinction (archaic, colloquial etc.) is marked.

For annotating global variants, the concept of links, originally created for derivational relations[4], was used. For every set of variants, it is necessary to choose one of them as a basic one. All the other global variants are then linked with that basic variant. A style marker is assigned to every link, but the set of its possible values is limited.

The selection of the basic variant is not crucial, because the set of links to other variants join them all, so that they can be reached all at once. However, it is reasonable to choose always the most common variant (in whatever sense).

The table 4 shows an example of capturing the three global variants of the lemma *thesis* in the PDT project.

4 Summary

We have presented and compared solutions of two problematic features occuring in Czech morphology, namely aggregates and variants. They were adopted by two different projects, PDT and NovaMorf, that are currently being finalized. The both projects used the same ideas, but implementations are different.

[4]The link connects the lemma under consideration to its derivational origin.

Table 3: Example of global variants of the word *thesis* in NovaMorf. Values of the global variants are: 0/h = without h/with h, k/d=short/long variant, s/z=containing s or z.

Variants	Multiple lemma	Global variant subtag
teze	{*teze, téze, these*}	G0kz
téze	{*teze, téze, these*}	G0dz
these	{*teze, téze, these*}	Ghks

Table 4: Example of global variants of the word *thesis* in the PDT project: h means standard variant. In the parentheses, there is the selected basic variant.

Variants	Dictionary lemma entry (simplified)
teze	teze
téze	téze_,h_(teze)
these	these_,a_(teze)

The NovaMorf project proposes to change some of the basics of the Prague tagging system. It wants to implement the concept of multiple lemma and use it for description of aggregates (sequence of segments of the aggregate) as well as for variants (set of lemma variants). The second change is creation of a subtag for marking variants (global and inflectional ones).

The PDT project is more traditional and does not introduce special formats. It is also the reason, why the changes described in this paper, have been already implemented to the new version of the morphological dictionary (however not publicly released yet) only within this project.

The treatment of aggregates is for the both projects equivalent. In NovaMorf, a special category (Type of Aggregate) was added for description of aggregates. Its value becomes a new part of the morphological tag. In PDT, the information about the type of the aggregate is incorporated into the existing positional tag. However, all the information about the aggregate, its type and its segments, are present in both solutions. NovaMorf treats them probably more transparently. The multiple lemma lists all the lemmas of the segments, while the mark within the morphlological tag (adopted by PDT) assumes that the users would derive the information about the segments from the mark, which is not so straightforward.

The treatment of variants is not equivalent in the two projects. NovaMorf is again more transparent — the concepts of multiple lemma and multiple tag will enable especially users of corpora not to take care about more possibilities, because the dictionary itself would know them all. Every variant is lemmatized by the same set of variant lemmas as shown in the example presented in the table 3. Individual lemmas (and wordforms) are then distinguished by the variant subtag, that uses a new set of values to mark the variants.

PDT, on the other hand, selects one variant as the basic one. Contrary to previous attempts, there is no predefined rule which variant to select. However, the authors insisted to preserve at least some information about the stylistic features of the individual variants. Lemma of each global variant is the variant itself.

If there was a need to unify the two solutions in the future, the lemma variants could be easily taken out from the dictionary and put together to make the set of a multiple lemma. The only thing that would need to add, would be the type of the variant according to the values prepared for NovaMorf. The opposite conversion, from multiple lemma to lemma variants would be even easier — each member of the set representing the multiple lemma would become an independent lemma. The selection of the central basic lemma for the link can be, as mentioned before, arbitrary. As the two approaches are not equivalent, there would be also needed some handwork, namely to add the information about the style of the variants.

The presented solutions could be possibly used also for other languages, but there were no attempts undertaken to prove it.

Acknoledgements

The research has been supported by the LINDAT/CLARIN and LINDAT/CLARIAH-CZ projects of Ministry of Education, Youth and Sports of the Czech Republic (LM2015071 and LM2018101).

References

[1] Bejček, E., Hajičová, E., Hajič, J. et al.: Prague Dependency Treebank 3.0, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, (2013).

[2] Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). Nakladatelství Karolinum. (2004)

[3] Hajič, J., Hlaváčová, J.: MorfFlex CZ, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague. (2013).

[4] Hlaváčová, J.: Formalizace systému české morfologie s ohledem na automatické zpracování českých textů. Disertační práce. Univerzita Karlova. (2009).

[5] Hlaváčová, J.: Golden Rule of Morphology and Variants of Wordforms. In: Jazykovedný časopis / Journal of Linguistics, Vol. 68, No. 2,Slovak Academic Press, Bratislava, Slovakia, ISSN 0021-5597, pp. 136-144. (2017).

[6] Hlaváčová, J., Mikulová, M., Štěpánková, B., Hajič, J.: Modifications of the Czech morphological dictionary for consistent corpus annotation. In: Proceedings of SLOVKO 2019, to appear.

[7] Osolsobě, K.: Algoritmický popis české formální morfologie a strojový slovník češtiny. Disertační práce. 161 s. (1996).

[8] Osolsobě, K., Hlaváčová, J., Petkevič, V., Svášek, M., Šimandl, J.: Nová automatická morfologická analýza češtiny. Naše řeč 100, No 4, pp 225–234. (2017)

[9] Petkevič, V., Hlaváčová, J., Osolsobě, K., Šimandl, J., Svášek, M.: Microsyntactic Parts of Speech in NovaMorf, a New Morphological Annotation of Czech. In: Proceedings of SLOVKO 2019, to appear.