Modifications of the Czech morphological dictionary for consistent corpus annotation

Jaroslava Hlaváčová, Marie Mikulová, Barbora Štěpánková, and Jan Hajič

Charles University, Prague

Abstract. We describe systematic changes that have been made to the Czech morphological dictionary related to annotating new data within the project of Prague Dependency Treebank (PDT). We bring new solutions to several complicated morphological features that occur in Czech texts. We introduced two new parts of speech, namely foreign word and segment. We adopted new principles for morphological analysis of global and inflectional variants, homonymous lemmas, abbreviations and aggregates. The changes were initiated by the need of consistency between the data and the dictionary and of the dictionary itself.

1 Motivation

Despite recent advances in part of speech (POS) and morphological tagging using Deep Learning, the old truth that more data always gives better results ([1], [9]) still holds. At the same time, consistency in data annotation is a very important factor. For morphological annotation, especially for morphologically rich languages with thousands of possible combinations of morphological values, consistency can only be achieved when a dictionary lists all plausible morphological interpretations of all wordforms [3]. Naturally, such a dictionary must also be consistent with all the annotated data, which is an issue when legacy data are taken into account as annotated with previous – possibly not fully compatible – versions of the dictionary. Therefore, when extending the available set of manually annotated data for POS and morphological tagging, we have to follow the following principles:

(i) use different genre, register, style and/or domain to add diversity to the dataset;

(ii) develop the morphological dictionary in parallel with the annotation process, to ensure consistency among all the annotated data and also between the data and the dictionary.

To meet the requirement (i), we are manually extending the annotated data. We enlarge the morphological annotation of Czech written texts in the Prague Dependency Treebank 3.5 [6] by adding annotation of spoken data (from the Prague Dependency Treebank of Spoken Czech [10]), translation data (Czech part of the Prague Czech-English Dependency Treebank [4]) as well as a small amount of "user-generated" data from the internet translation services (corpus PDT-Faust[1]). This will increase the amount of data available for NLP applications (such as MorphoDiTa [13] or DeriNet [15]) more than twice, genre-diversified (see Tab. 1).

It is important to pursue a manual morphological annotation of large data in parallel with the development of the dictionary (requirement (ii). Therefore, while annotating, we are enriching the dictionary called MorfFlex [5], used in the original annotation,

1 https://ufal.mff.cuni.cz/grants/faust

with words and wordforms stemming from new texts. Moreover, we are making systematic changes in capturing some phenomena in the dictionary. The long-time experience with the usage of the dictionary and the current annotation of real data has shown that several phenomena would be better to capture differently in order to achieve better consistency in the whole dictionary. The changes in the dictionary are being projected back into the data by repeated re-annotation to guarantee full consistency between the dictionary and the data.

Data type	written	spoken	translated	internet	Total
Tokens	1,725,242	742,257	1,162,072	33,772	3,663,343

Table 1. Morphological annotation in the future, consolidated edition of PDT

When formulating the principles of the dictionary and guidelines for annotation, as well as when making changes in the structure of lemmas and tags, it is necessary to find an optimal compromise between linguistic theory (often especially the traditional interpretation) and the needs of practical annotation, for which it is important to have simple and clear rules offering a solution for each token in any real text. We do not want to change the existing structure of MorfFlex, so we are capturing all the changes within the existing dictionary structure. Thus, at this time, we do not include the concept of multiple lemma nor extend the positional tag for marking variants as proposed in [7] and [8].

There are also other approaches to Czech morphology, most notably the NovaMorf project [12] and Universal Dependencies (UD) [11]. However, NovaMorf is still in its specification phase, while in MorfFlex we are bound by the already annotated corpus (PDT), and it is not yet clear if a conversion (both ways) can be lossless. In UD, the morphological features are adapted to the use in multilingual setting, and there is some loss if language-specific features are not used. On the other hand, there is an almost lossless conversion from MorfFlex-based annotation to the UD morphological features system, as described in [14]; future conversion to the UD system should thus be unproblematic.

In this paper, we describe changes that have been made to MorfFlex related to annotating new data within the project of the consolidated version of PDT.

2 Golden rule of morphology

The MorfFlex dictionary lists more than 100,000,000 lemma-tag-wordform triples. For each wordform, full inflectional information is coded in a positional tag. Wordforms are organized into paradigms according to their formal morphological behavior. The paradigm (set of wordforms) is identified by a unique lemma. Apart from traditional morphological categories, the description also contains some derivational, semantic and stylistic information. The formal specification of the dictionary is in [2].

The so called “golden rule of morphology” (cf. [7], [8]) is applied to the dictionary. The rule says that any pair <lemma, morphological tag> is represented by at most one

wordform.[2] The principle was, however, often violated in the previous version of the dictionary, mainly due to

- homonymy of lemmas;[3]
- different types of wordform variants.

Each of these problematic issues is addressed differently. The former one is solved by adding a numerical index to homonymous lemmas (see Sect. 3), the latter one by distinguishing two types of variants – global and inflectional ones (see Sect. 4). Until recently, both types of variants were marked uniformly at the 15th position of the tag. This did not allow to fully describe the complex variations that can occur for a single wordform.

3 Lemma numbering (indexing)

The problem of homonymy of lemmas is solved by giving numbers to the lemmas with the same spelling. We do not strive to make any distinction between meanings of homonymous words. The only differences we want to capture are those of formal morphological nature. Therefore, we add numbers only to lemmas that differ from the formal point of view. It means that we distinguish lemmas that have either

- different POS, e.g. *růst-1* as noun (‘a growth’) and *růst-2* as verb (‘to grow’), or
- different gender in case of nouns, e.g. *kredenc-1* as masculine and *kredenc-2* as feminine; they have the same meaning (‘a cupboard’), but different paradigms, or
- different aspect in case of verbs, e.g. *stát-1* with perfective aspect (‘to happen’) and *stát-2* with imperfective aspect (‘to stand’).

Thus, we have, e.g., lemma *jeřáb-1* for crane as a bird (animate masculine) and *jeřáb-2* for both a tree and crane as a device for lifting heavy objects (inanimate masculine). We do not distinguish the latter two meanings (tree vs. device), because they do not differ from the inflectional point of view. There might be a difference in derivation. In this case, the word *jeřábník* (a man who works with a crane-device) is derived from *jeřáb* as a device. It is not possible to derive *jeřábník* from *jeřáb* as a tree.

Due to a large number of complicated cases, we have decided not to take into account such derivational, stylistic and semantic differences. Thus we do not distinguish lemmas (if they inflect identically) that have:

- different meaning, e.g. *kohoutek* (‘tap’) and *kohoutek* (‘flower’);
- different derivational model: *matka* (‘nut’) and *matka* (‘mother’ with possessive adjective derivation);
- different style value: *ekonomka* (‘female economist’) and *ekonomka* (‘school of economics’, non-standard).

2 If the pair is meaningful, there is exactly one form, if it is not, there is none of them. There must not exist more than one wordform with the same lemma and tag.

3 The homonymy of wordforms has been resolved sufficiently in the previous versions of the dictionary.

4 Variants

Orthographic and stylistic variants of a word (hereinafter referred to as variants; e.g. archaic variant *these,* standard variant *teze,* and non-standard variant *téze* 'thesis') are the candidates for breaking the golden rule of morphology. We distinguish two types of the variants (see [7]):

- **Inflectional variants** are those variants that relate only to some wordforms of a paradigm defined by a special combination of morphological values, e.g. both *orli* and *orlové* ('eagles') are the wordforms of the noun *orel* ('eagle') and express plural masculine nominative.
- **Global variants** are those variants that relate to all wordforms of a paradigm, and always in the same way, e.g. *vyhýbat* and *vyhejbat* ('to avoid') – the whole paradigms of each verb differ in the distinction *-ý-* vs *-ej-* in the root.

There are two types of information that are used for the description of wordforms: lemma and tag. It is natural to express information about global variants within the lemma, because it is common for all its wordforms, and information about inflectional variants by means of a tag that applies only to specific wordforms.

4.1 Global variants

Global variants were not tackled uniformly in MorfFlex. Some global variants had different paradigms with different lemmas, others were grouped into one paradigm with one common lemma. In the former case there was no connection between the two variant lemmas. The latter case led to the most massive violations of the golden rule because there were different wordforms with the same tags belonging to the same lemma.

Wordform	Lemma
teze	teze
these	these_,a_^(^DD**teze)
téze	téze_,h_^(^GC**teze)

Table 2. Global variants – example

We have decided to select one of the variants as "basic" and interconnect other variants via links to it. We use a notation that was originally designed for marking derivational relations. To distinguish variants from derivations, we introduce new codes for variants. We also simplify and reduce the set of style flags. We are now using only three types of global variants:

- DD – standard variant, including archaic ones,
- GC – non-standard (general Czech) variant, including dialectical, expressive, slang and vulgarisms,
- DS – distortion (a frequent typo, or otherwise distorted spelling).

Every variant, except for the basic one, has to be assigned a single indication of style. See examples in Tab. 2.

There are two main differences when compared to the previous treatment of variants; the global variants are really global – there cannot be a wordform belonging to the same lemma having different (or none) sign of style, and there is at most one indication of style for each paradigm.

4.2 Inflectional variants

For marking inflectional variants, we use the 15th position of the tag, as has been done before. The main difference lies in the fact that now we use this position strictly for inflectional variants. Another change is the simplification of the set of possible values. Numbers 1 to 4 mark standard variants, while numbers 5 to 9 relate to substandard ones. See examples in Tab. 3.

Wordform	Lemma	Positional Morph. Tag
přijdeme	přijít	VB-P---1P-AAP--
přijdem	přijít	VB-P---1P-AAP-6
přídeme	přijít	VB-P---1P-AAP-5
přídem	přijít	VB-P---1P-AAP-7
příjdeme	přijít	VB-P---1P-AAP-8
příjdem	přijít	VB-P---1P-AAP-9

Table 3. Inflectional variants – example

5 New features in the tagset

Czech texts contain not only “normal” words that fit well into traditional categories but also various sorts of strings (e.g. foreign words, abbreviations, etc.) that must be processed as well, and thus they need to be defined more precisely.

5.1 New part of speech: Foreign word

The POS of most foreign words were taken from their original languages. Thus, the wordform *in* was a preposition, *European* was an adjective, etc. However, in Czech texts, these words do not behave as their original POS might suggest. They are usually part of a longer foreign phrase, which may be a citation, a foreign name, etc. It seems inappropriate to assign usual morphological values to foreign wordforms within foreign phrases, since their role in Czech texts differs from their role in foreign texts. Therefore we have adopted a special POS concept of “foreign word” (presented for the first time in [8]).

Foreign word is such word that is not subject to Czech inflectional system and has no meaning of its own in Czech. Lemma of a foreign word is the same as the word itself.

The tag contains special values at the POS and SUBPOS positions, namely F%. There are no other morphological values involved in the tag (see Tab. 4).

Foreign words should not be confused with indeclinable words that are of foreign origin, have already become part of the Czech vocabulary and have their meaning within the Czech language, e.g. the noun *kupé* (‘compartment in a train’) or an adjective *lila* (‘lilac colour’).

Wordform	Lemma	Tag
European	European	F%----------
market	market	F%----------

Table 4. Foreign word – examples

5.2 New part of speech: Segment

Segments are incomplete words. They are parts of words; in order to understand them, they must be joined with another string or word to create a complete word. As they are quite common in Czech texts and they were not previously captured consistently in the dictionary, we have created a new POS with the code S for them. According to their position in the complete word, we distinguish prefixal and suffixal segments.

Wordform	Lemma	Tag	Example
česko	česko	S2--------A----	*česko-ruská kniha* ‘Czech-Russian book’
tří	tří	S2--------A----	*tří až pětiletý* ‘three to five year old’
nepoliticko	politicko	S2--------N----	*nepoliticko-politické* ‘nonpolitical-political’

Table 5. Prefixal segment – examples

Wordform	Lemma	Tag	Example
kou	ka	SNFS7-----A----	*s manželem/kou* ‘with husband/wife’
tice	tice	SNFS1-----A----	*n-tice* ‘n-tuple’
a	a	SpQW----R-AA---	*řekl(a)* ‘he or she said’

Table 6. Suffixal segment – examples

Prefixal segments are strings that appear at the beginning of words. They are followed with a space or another separator, most often with a hyphen.

Lemma of prefixal segment is the string itself, unless it is in negative form. In that case, the positive form (without the prefix *ne-*) is considered to be the lemma. The tag of all prefixal segments has the code 2 at the 2nd position. Moreover, we specify for them also the 11th position concerning negation (see Tab. 5).

Suffixal segments are strings that may appear at the end of a wordform. They are usually attached directly to the word they combine with. The separator is most often

a hyphen, parenthesis or a slash (/).

The suffixal segments express an affiliation to a specific POS. Thus, all the inflectional categories that describe the whole wordform, except for the first one (= the code for POS, which is S), are filled in the tag (with the exception of the aspect for verbs). The lemma is the closest "basic wordform" (see Tab. 6).

Wordform	Decomposed	Lemma	Tag
zač	*za co*	co	PQ--4--------z-
začs	*za co jsi*	co	PQ--4--------Z-
doň	*do něj*	on	P5ZS2--3-----d-
dobřes	*dobře jsi*	dobře	Dg-------1A--s-
promluvils	*promluvil jsi*	promluvit	VpYS----R-AAPs-
kdyžs	*když jsi*	když	J,----------s-

Table 7. Aggregate – examples

5.3 Aggregates

An aggregate is a wordform that is created by joining two or more wordforms (components of the aggregate) into one and cannot be simply assigned any POS. Aggregates are common especially in agglutinative languages, but there are two aggregate types in Czech, too:

- pronominal aggregates consisting of a preposition and the pronoun *on* ('he') or *co, copak* ('what');
- verbal aggregates consisting of a wordform of almost any POS with the string *s* added to the end. It stands for the wordform *jsi* ('you are').

The lemma of pronominal aggregates is the lemma of the pronoun. The lemma of a verbal aggregate is the lemma of its first component. The fact that a wordform is an aggregate is coded at the 14th position of the tag. The code of pronominal aggregates corresponds to the initial letter of the preposition that forms their first component, verbal aggregates are coded with the letter s (see Tab. 7). Verbal and pronominal aggregates can combine; such aggregates are marked with the initial letter of the preposition, but in an uppercase letter (see the example *začs* in Tab. 7).

In the original MorfFlex, the pronominal aggregates were signaled by means of the second position in the tag. The lemma of pronominal aggregates was always the aggregate itself, the lemma of verbal aggregates with a verb at the beginning was the infinitive of the leading verb. Verbal aggregates composed of other POS (e.g. *kdyžs* 'when you are') were not treated as aggregates at all.

5.4 Abbreviations

An abbreviation that abbreviates a single word (e.g. *str - strana* ‘p - page’) is captured as a special wordform of the paradigm of that word. Only those categories that are valid for each use of the abbreviation are coded in the tag. The fact that it is an abbreviation is expressed at the 15th position by the letters *b* or *a* (see examples of the lemma *strana* in Tab. 8).

Wordform	Lemma	Tag	Example
s	strana	NNFXX-----A---a	*na s. 12* ‘at page 12’
str	strana	NNFXX-----A---b	*na str. 12* ‘at page 12’
l	letopočet	NNIS2-----A---b	*n. l.* ‘of AD’
V	V-88_;B	NNXXX-----A----	*V. Havel*
ČR	ČR_:B_^(Česká_republika)	NNXXX-----A----	*ČR* ‘Czech Republic’

Table 8. Abbreviation – examples

Lemmas of other abbreviations, especially those that are composed of uppercase letters only (e.g. *USA*), is the abbreviation itself with a special flag B. They are assigned the tag of a maximally subspecified noun (with the value X for any gender, case, number at the positions 3-5). The same holds for one-letter abbreviations that stand for a single word but it is not clear for which of the many alternatives. This is, e.g., the case of initials of proper names (e.g. *V. Havel*, *V. Mrštík*). The abbreviations of this type have usually added the number 88 to their lemma as a human-redable indication of their status. There are some exceptions – very common abbreviations with only one meaning. Lemma of such abbreviations does not have the indexing number 88, as they cannot be mistaken for anything else. They have a semantic explanation as a note attached to the lemma (see Tab. 8).

6 Conclusion

We have described a project of manual morphological annotation on new text types within the new version of PDT. The need for consistency between the treebank(s) and within the dictionary has triggered deep and extensive changes in the Czech morphological dictionary MorfFlex. The release of the new version of MorfFlex together with the new dataset is planned for the end of 2019. Thanks to the newly achieved higher consistency, we believe that the resulting larger, high-quality dataset and dictionary will contribute to better usability of the treebanks for linguistic inquiries, for new annotation projects using Czech, and also an increased accuracy of the NLP tools that learn from them.

Acknowledgements: The research has been supported by the Czech Science Foundation under the project GA17-12624S. The research has also 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] Banko, M., Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *Proceedings of the 39th annual meeting on ACL*. Association for Computational Linguistics, 26-33.

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

[3] Hajič, J. (2000). Morphological Tagging: Data vs. Dictionaries. In *Proceedings of the 6th Applied Natural Language Processing and the 1st NAACL Conference*, Seattle, 94-101.

[4] Hajič J., Hajičová E., Panevová J., Sgall P., Bojar O., Cinková S., Fučíková E., Mikulová M., Pajas P., Popelka J., Semecký J., Šindlerová J., Štěpánek J., Toman J., Urešová Z., Žabokrtský Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In P*roceedings of the 8th International Conference on LREC 2012*, European Language Resources Association, İstanbul, 3153-3160.

[5] Hajič, J., Hlaváčová, J. (2013). *MorfFlex CZ*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11858/00-097C-0000-0015-A780-9.

[6] Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J.,Poláková, L., Rysová, M., Sgall, P., Spoustová, D. J., Straňák, P., Synková, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š., Žabokrtský, Z. 2018, *Prague Dependency Treebank 3.5*. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-2621

[7] Hlaváčová, J. (2017). Golden Rule of Morphology and Variants of Wordforms. *Jazykovedný časopis / Journal of Linguistics*, 68(2), 136-144.

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

[9] Church, K., Mercer, R. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics*, 19(1), 1-24.

[10] Mikulová M., Mírovský J., Nedoluzhko A., Pajas P., Štěpánek J., Hajič J. (2017). PDTSC 2.0 – Spoken Corpus with Rich Multi-layer Structural Annotation. In *Lecture Notes in Computer Science*, No. 20th International Conference TSD 2017, Prague, Springer International Publishing, Cham / Heidelberg / New York / Dordrecht / London, 129-137.

[11] Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C.,

McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on LREC 2016*, Paris, 1659-1666.

[12] Petkevič, V., Hlaváčová, J., Osolsobě, K., Šimandl, J., Svášek, M. (2019). Microsyntactic Parts of Speech in NovaMorf, a New Morphological Annotation of Czech. In: *Proceedings of SLOVKO 2019* (this volume).

[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 ACL: System Demonstrations*, Association for Computational Linguistics, Baltimore, 13-18.

[14] Zeman, D. (2018). *The World of Tokens, Tags and Trees*. Studies in Computational and Theoretical Linguistics, Charles University, Prague.

[15] Žabokrtský Z., Ševčíková M., Straka M., Vidra J., Limburská A. (2016). Merging Data Resources for Inflectional and Derivational Morphology in Czech. In *Proceedings of the 10th International Conference on LREC 2016*, European Language Resources Association, Paris, 1307-1314.