PARTS OF SPEECH IN NOVAMORF, A NEW MORPHOLOGICAL ANNOTATION OF CZECH

VLADIMÍR PETKEVIČ[1] – JAROSLAVA HLAVÁČOVÁ[2] – KLÁRA OSOLSOBĚ[3] – MARTIN SVÁŠEK – JOSEF ŠIMANDL

[1] Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University, Czech Republic

[2] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic

[3] Institute of the Czech Language, Faculty of Arts, Masaryk University, Czech Republic

Abstract: A detailed morphological description of word forms in any language is a necessary condition for a successful automatic processing of linguistic data. The paper focuses on a new description of morphological categories, mainly on the subcategorization of parts of speech in Czech within the NovaMorf project. NovaMorf focuses on the description of morphological properties of Czech word forms in a more compact and consistent way and with a higher explicative power than approaches used so far. It also aims at the unification of diverse approaches to morphological annotation of Czech. NovaMorf approach will be reflected in a new morphological dictionary to be exploited for a new automatic morphological analysis (and disambiguation) of corpora of contemporary Czech.

Keywords: NovaMorf, morphological annotation, parts of speech, morphological categories, subcategorization

1 INTRODUCTION

We present a repertoire of morphological categories and, mainly, the parts of speech (POS) and their subcategorization distinguished in NovaMorf, the project of an innovated description of Czech morphology as a linguistic base for a new morphological analysis and subsequent disambiguation of Czech texts. For over 25 years, morphological dictionary and analysis of Czech were based on (almost) unchanged annotation systems. After years of experience with the use of language corpora, it turned out that these systems, designed many years ago, have become somewhat obsolete and need to be amended, concerning the system itself, the tagset, and morphological dictionary used by morphological analysis.

The NovaMorf starting point is [3]. Other suggestions for solving partial problems are based on it ([5], [6], [7], [10], [11]). NovaMorf is also based on the Prague morphological annotation system ([1], [2]) and Ajka/Majka system developed in Brno ([4], [8], [9], [12]) on recent grammars of Czech and corpus data. NovaMorf critically evaluates these systems and creates a new one so that the resulting description is linguistically adequate, economical and consistent (e.g., unlike current systems, it consequently applies the so-called "golden rule of morphology"[1]). It also takes into account the Universal Dependencies approach, trying to be very close to it.[2] This new annotation system, as a national standard, is to be used for morphological annotation of corpora of contemporary Czech and in various NLP applications (e.g., parsing, spell and grammar checking) dealing with Czech.

We only briefly introduce:
(i) the repertoire of suggested morphological categories (Chap. 2),
(ii) the repertoire of parts of speech and their subcategorization (Chap. 3).

2   MORPHOLOGICAL CATEGORIES

We design the following categories with each category being assigned a set of values. Each morphological interpretation of a word form is assigned just one value for a given category (except for global and inflectional mutations).

1.   Part of speech – POS (cf. 3.1)
2.   POS subcategorization – SUB (cf. 3.2.1)
3.   Deixis – DEI (cf. 3.2.2)
4.   Aspect – ASP
5.   Abbreviation – ABR
6.   Gender – GEN
7.   Number – NUM
8.   Case – CAS
9.   Person – PER
10. Degree of Comparison – DEG
11. Negation – NEG
12. Verb form – VRB
13. Adjectival Short Form – NOM
14. Aggregate/Compound Type – AGR
15. Global mutation – GMU
16. Inflectional mutation – FMU
     Let us now characterize individual parts of speech.

---

[1] The rule states that two different word forms cannot be assigned an identical annotation (= [lemma, tag] pair).

[2] The detailed description of the NovaMorf approach also contains a detailed comparison with the UD approach. For scope reasons, no details can be presented here.

3   PARTS OF SPEECH

3.1   POS category

The part of speech category (POS) is the basic one since each morphological interpretation of a word form is assigned a POS value.

Values of the POS category:

- Noun (code N)
- Adjective (A)
- Pronoun (P)
- Numeral (C)
- Verb (V)
- Adverb (D)
- Preposition (R)
- Conjunction (J)
- Interjection (I)
- Particle (T)
- Foreign word (F)
- Affix segment (S)
- Symbol (Z)
- Unknown word (X)
- Aggregate (G[3])

In addition to traditional POS (noun ... particle, cf. 3.2.1ff.), we introduce the following "parts of speech" in a broader sense.

Foreign word – code F

Foreign word is a string that does not have its own meaning in Czech; typically a word of a foreign language occurring in a Czech sentence, usually within a quotation or saying (*we shall overcome*; *per se*) or as part of a proper name (*New York City*). Typical examples: *the, you, du, to*. This does not concern loan words which are part of the vocabulary of Czech (*image, khaki*).

Its lemma = the form itself. Foreign words have no subcategories.

Affix segment – code S

Affix segment is a string that is typically marked with a separator: a hyphen, space, slash. We distinguish:

- numeral prefix segment (C)
- postfix segment (p)
- other prefix segments (e)

---

[3] The code of this POS is not used since aggregates are described by their components only, see below.

(i) **Numeral prefix segment**: an initial numeral segment of a word standing alone and being added to the full word further on in the text:

●   *tří* až *čtyřprocentní* 'three to four percent' (prefix segment: *tří*)

(ii) **Postfix segment**: a final segment of a word that becomes a full word by adding a part of a previous string in front of it. Its lemma = the form itself.

●   *řekl/a* 'he said / she (said)' (postfix segment: *a*, lemma(*a*) = *a*)

Unlike a prefix segment, it can be assigned values of some morphological categories being derived from the entire word form they abbreviate (thus *a* is a postfix segment that is assigned past participle values; this may sound paradoxical, but a relation to the full verb *řekl* is established).

(iii) **Other prefix segments**: an initial non-numeral segment of a word standing alone, but forming a full word with a substring of another string further on in the text:

Examples:

●   *česko*- a *rusko-německý* 'Czech- and Russian-German' (prefix segments: *česko*, *rusko*)

Symbol – code Z
Linguistically, symbol is not one of the traditional parts of speech, but it is very useful to assign – along the lines of existing annotation systems – each symbol the same code (Z), and to understand the set of symbols as a special word class. The lemma of a symbol is typically a symbol itself, but inverted commas, apostrophes and various kinds of brackets are normalized. Symbols are divided into:
(i) punctuation marks (z) (e.g. comma: ",")
(ii) other signs (J) (e.g. *%* or math symbols)

Unknown word – code X
An unknown word is a string whose POS cannot be recognized, typically a typo or a foreign language word not being classified as a foreign word (see above); it is not contained in the morphological dictionary. Its lemma = its form.
This value is already present in existing annotation systems.

Aggregate
An aggregate is a special word class reserved for describing a word form, consisting of a combination of 2 or 3 word forms – aggregate components that may

belong to different classical parts of speech. Therefore, none of them can be assigned to the aggregate, which is described via its components only, each component being assigned its own tag including POS. The lemma of an aggregate is a *multiple lemma* – an ordered set of component lemmas.[4] For instance:

- *připravilas* 'you prepared' consists of two verbal components: *připravila* 'prepared' and the 2nd pers. sg. enclitic *-s* (= *jsi* 'you_are'), with each component being assigned its own tag; the lemma is a multiple one: lemma(*připravilas*) = {*připravit*, *být*} '{prepare, be}'
- *abyste* 'so that you' consists of the conjunction *aby* 'so_that' and the present tense 2nd pers. pl. form of the verb *být* 'be': each component is assigned its respective tag, the multiple lemma(*abyste*) = {*aby*, *být*} '{so_that, be}'

The aggregate is not considered a compound word made up via a usual word formation process; a compound word is assigned a classic POS (adjective *černobílý* 'black and white') and has other morphological properties as well.

3.2 POS categories

The majority of parts of speech is further subcategorized, some of them (pronouns, numerals, adverbs, some nouns and adjectives of the "numeral type") being subcategorized into two categories:

- subcategory SUB (3.2.1)
- subcategory, called Deixis (DEI, 3.2.2).

3.2.1 SUB category

The SUB category is relevant to all parts of speech, except for prepositions, interjections, foreign words, aggregates, and unknown words. For each relevant POS, we specify what SUB category values are distinguished and how they are encoded. The corresponding one-letter code is usually the same as the second code in the Prague system ([1], [2]), which is also called Subcategory. In NovaMorf, however, this category is interpreted quite differently. Subcategory in the current Prague system is a mixture of values describing various characteristics of word forms. Some values relate to individual forms, other ones to whole paradigms. On the contrary, the SUB category in NovaMorf is **strictly global**, i.e. it is always relevant to the whole paradigm of a given word form. The same applies to the second subcategory: Deixis (3.2.2).

The SUB category values are different for different parts of speech since they describe different properties. However, some properties (some SUB and DEI

---

[4] Multiple lemmas remain assigned to specific word forms (aggregates, passive participles and forms described by global or inflectional mutations) even after disambiguation in POS tagging: they are not to be further disambiguated. A user can search via each component of a multiple lemma in a corpus.

category values) are shared by more parts of speech, hence they are assigned the same value (code). Unshared features are encoded with different, unambiguous codes. Below is an overview of the shared values of the SUB category.

- **Other** (code 0 – zero). It concerns nouns, adjectives, verbs, adverbs and indicates that a word form does not have any of the other properties distinguished in the category, thus it is not necessary to express this property by different codes depending on POS.
- **Possessive** (code U) is shared by adjectives and pronouns.
- **Deverbal** (code V) is shared by nouns, adjectives and adverbs. It specifies that a given word is derived from a verb (code V is proposed being also the code for Verb as POS). The value is assigned to a specific group of words, rather than to all words derived from verbs.
- **Numeral** (code C) is relevant to nouns, adjectives, adverbs, and numeral prefix segments. It specifies that a given word contains an element common to numerals, i.e. expressing a number, or possibly the word is used instead of a numeral (e.g. *tisícovka* návštěvníků 'a thousand visitors'). Thus, the code C is proposed being also the code for Numeral as POS. In some grammars, such words are referred to as belonging to specific subcategories of numerals.

3.2.1.1 Noun subcategorization
For nouns, the following values of the SUB category are distinguished only:

- Deverbal (V): *pokrytí* 'covering'...
- Numeral (C): words expressing an association with numerals, e.g. *pětka* 'five'…
- Other (0 – zero): *město* 'town'…

The V value is assigned to deverbal nouns ending with the *-ní / -tí* suffix. These nouns have specific (morpho)syntactic properties and can behave differently than other nouns in a sentence: unlike the other nouns they can have a reflexive particle associated with them (***štítění se*** *práce* 'loathing work') or they can be modified by adverbs regularly derived from adjectives (***zpívání*** *falešně* 'singing out of the tune').

We do not distinguish between deverbal nouns and lexicalized deverbal nouns that do not express a verbal action: ***vázání*** *ječmene* 'tying barley' vs. *lyžařské* ***vázání*** 'ski bindings'. In the morphological dictionary, they will constitute a single entry: *vázání*, with SUB=V.

The nouns irregularly derived from verbs are not considered deverbal: e.g. *utrpení* 'suffering' (other is *utrpění* 'suffering'), *uvědomění* 'awareness'… Nor the compound words such as *krupobití* 'hailstorm' are considered deverbal, although their second component is an action noun (= *bití* 'beating'), since corresponding compound verbs usually do not exist (*krupobít 'to hailstorm').

3.2.1.2 Adjective subcategorization

For adjectives, we distinguish the following SUB category values:

- Possessive (U): *matčin* 'mother's'...
- Derived from present transgressives (G): *sedící* 'sitting'…
- Derived from past transgressives (M): *přeživší* 'having survived'...
- Other deverbal (V): passive participle forms (*namazán* 'lubricated'[5]; adjectives ending with *-ný* and *-tý* (*namazaný* 'lubricated') and adjectives ending in *-telný* (*rozpoznatelný* 'recognizable')...
- Numeral (C): lexemes expressing association with numerals, e.g. *dvojkový* 'binary'...
- Other (0 – zero): *starý* 'old'...

A nominal (short) adjectival form is not a SUB category value since SUB is a **global category**, specifying the **whole** paradigm of some lemma. The short form is assigned the same lemma as the long one, e.g. lemma(*stár*) = *starý* 'old'.

3.2.1.3 Pronoun subcategorization

For pronouns, two subcategories are distinguished: SUB and Deixis (DEI). The usual subcategorization of pronouns does not reflect the dual classification view: e.g., *něčí* 'someone's' is both possessive and indefinite. The second property is the value of the DEI category. The double subcategorization of pronouns was first used in the Ajka system in Brno (cf. [4], [8], [9]).

Since DEI is a category that is common to several parts, we discuss it as a special category (cf. 3.2.2).

The following SUB values are distinguished for pronouns:

- Personal (o): *já* 'I', *oni* 'they', *se* (reflexive)…
- Nominal (N): *kdo* 'who', *nikdo* 'nobody'...
- Possessive (U): *můj* 'my', *čí* 'whose'...
- Other, mainly delimitative (v): *každý* 'every', *týž* 'same'…

3.2.1.4 Numeral subcategorization

Also for numerals, we distinguish two subcategories: SUB and DEI. The following SUB values are distinguished:

---

[5] Compared to previous systems, we consider the forms of the passive participle to be short adjectives derived from verbs, rather than verbal forms. A passive participle is assigned a (non-disambiguated) multiple lemma whose components are (i) the long form of the corresponding adjective and (ii) the infinitive of the underlying verb. In corpora, a user can search for passive participles via both the adjectival or verbal component; the verbal component can be used, i. a., for solving diathesis problems in deep syntax (the connection between active forms of a verb and its passive participle forms is not lost, since they are represented by the verbal component).

- Cardinal (z): *pět* 'five' ... *kolik* 'how many'...
- Ordinal (r): *pátý* 'fifth', *poprvé* 'first'…
- Multiple (n): *dvakrát* 'twice'…
- Fractional (h): *půl* 'half', *třetina* 'third', *čtvrt* 'quarter'...
- Relative to the whole (u), including the following numerals:
    - aggregate: *dvé* 'two', *patero* 'five'...
    - ensemble: *dvoje* 'two', *patery* 'five'...
    - group: *dvojice* 'pair', *pětice* 'group of five'…
    - generic: *dvojí* 'two kinds', *paterý* 'five kinds'…
- Number written in Arabic or Roman digits (=): *586*...

3.2.1.5 Verb subcategorization
For verbs, only the following SUB values are distinguished:

- Auxiliary (b): only *být* 'be', *bývat* 'used to be'
- Other (0 – zero): *navštívit* 'visit', *koupat* 'bathe'...

The verbs *být* and *bývat* are considered auxiliary in case they participate in past tense constructions, pluperfect, present and past conditional mood constructions and future of imperfective verbs; and also conditional forms *by* 'would'... The forms of the verb *být* a *bývat* used
(a) to coform periphrastic passive constructions: ***Jsem/Bývám*** *často překvapen.* 'I am / I am (usually) often surprised.',
(b) as autosemantic words to express existence: *Bůh* ***je***. 'God is.', and
(c) as a copula: *Dělník* ***je*** *doma / v lese / k dispozici / překvapený.* 'The worker is at home / in the forest / available / surprised.'
are not considered auxiliary for annotation purposes, i.e. they are assigned the SUB=0 value.

3.2.1.6 Adverb subcategorization
For adverbs, we distinguish the following values for SUB:

- Pronominal (P), i.e.
    - local: *kudy* 'where', *tudy* 'this way', *odkud* 'from where', *nikam* 'nowhere'...
    - temporal: *kdy* 'when', *kdykoli* 'whenever', *nikdy* 'never', *pokaždé* 'each time'...
    - modal: *jak* 'how', *všelijak* 'in various ways, anyhow', *nijak* 'in no way'...[6]
- Compound (s): *dopředu* 'forward', *namodro* 'blue'…
- Numerical (C): *napůl* 'half', *vedví* 'asunder'...

---

[6] Local, temporal and modal adverbs are not distinguished in the SUB category since it is often very difficult to disambiguate between them.

- Regularly derived from verbal adjectives (V): *zamyšleně* 'thoughtfully'…
- Other (0 – zero): *dobře* 'well'

For adverbs, we also specify the Deixis category (DEI, cf. 3.2.2).

3.2.1.7 Conjunction subcategorization

For conjunctions, the following values for SUB are distinguished:

- Coordinate (^ – circumflex): *a* 'and', *ale* 'but'...
- Subordinate (, – comma): *protože* 'because'...
- Mathematical operations (* – asterisk): *krát* 'times'...

Note. We interpret the word forms *abych* 'so_that_I', *abys* 'so_that_you'...,
kdybych 'if_I', *kdybys* 'if_you'... as aggregates (cf. 3.1).

3.2.1.8 Particle subcategorization

For particles, the following SUB values are distinguished:

- Desiderative (p): *ať* 'let'...
- Responsive (o): *ano* 'yes'...
- Discursive marker (d): *bohužel* 'unfortunately'...

3.2.1.9 Affix segment subcategorization

For affix segments, the following SUB values are distinguished (cf. 3.1):

- numeral prefix segment (C)
- postfix segment (p)
- other prefix segments (e)

3.2.2 DEI (deixis) category

In the DEI category, as the second POS subcategory, we distinguish the following values:

- Definite (U): personal pronouns, definite numerals: *dva* 'two'
- Indefinite (N): *někdo* 'someone', *několik* 'several', *někdy* 'sometime'...
- Negative (Z): *nikdo* 'nobody', *nijak* 'in no way'...
- Interrogative (T): *kdo*(*ž*) 'who', *jaký* 'which', *kolik* 'how many', kde 'where'...
- Relative (V): *jehož* 'whose'...
- Reflexive (S): *se*, *si* '-self'...
- Demonstrative (D): *ten* 'that', *takový* 'such'...; numerals *tolik* 'so many'...; pro-nominal adverbs *tady* 'here'…

The DEI category is primarily relevant to pronouns, numerals, pronominal adverbs, although not all values are used for all these parts of speech. For other parts of speech, this property is undefined (-). Every unambiguous pronoun, numeral, and pronominal adverb receives exactly one of these values.

The values of both SUB and DEI categories are combined, but not arbitrarily. Possible combinations of their values are shown in Table 1 (3.2.2.1), Table 2 (3.2.2.2) and Table 3 (3.2.2.3).

3.2.2.1 Pronouns

Table 1 lists combinations of possible values of SUB and DEI categories for pronouns.

PRONOUNS	definite	indefinite	negative	interrog.	relative	reflex.	demonstr.
Personal	*ty* 'you'	-[7]	-	-	-	*si, se*	-
Nominal	-	*leccos* 'anything'	*nic* 'nothing'	*kdo* 'who'	*jenž* 'who'	-	-
Possessive	*jejich* 'their'	*něčí* 'someone's'	*ničí* 'nobody's'	*čí* 'whose'	*jejíž* 'whose'	*svůj*	-
delimitative, other	*každý* 'each'	*všelijaký* 'sundry'	*žádný* 'no'	*jaký* 'which'	-	-	*ten* 'this'

Tab. 1. Pronouns. Combination of SUB (column headings) and DEI (row headers) values and representative lexemes

Similarly as the Prague system, we do not distinguish ambiguous (interrogative / relative) pronouns: *jaký* 'which', *který* 'which', *kdo* 'who'...

Some lexemes are relative only: *jenž* 'who', *jehož* 'whose'...

3.2.2.2 Numerals

Table 2 lists combinations of possible values of SUB and DEI categories for numerals.

NUMERALS	Definite	Indefinite	interrogative	Demonstrative
Cardinal	*sto* 'hundred'	*několik* 'several'	*kolik* 'how many'	*tolik* 'so many'
Ordinal	*pátý* 'fifth'	*několikátý* 'ord. numb. for *several*'	*kolikátý* 'ord. numb. for *how many*'	*tolikátý* 'ord. numb. for *so many*'
Multiple	*dvakrát* 'twice'	*několikrát* 'several times'	*kolikrát* 'how many times'	*tolikrát* 'so many times'
Fractional	*půl* 'half'	-	-	-
Relative to the whole	*dvojí* 'two kinds'	*několikero* 'several kinds'	*kolikerý* 'how many kinds'	*tolikerý* 'so many kinds'

Tab. 2. Numerals. Combination of SUB (column headings) and DEI (row headers) values and representative lexemes

---

[7] "-" means *undefined*. Similarly in the other tables.

3.2.2.3 Adverbs

Table 3 lists combinations of possible values of SUB and DEI categories for adverbs.

PRONOMINAL ADVERBS	definite	indefinite	negat.	interrog.	relat.	demonstr.
	všude 'everywhere', *vždy* 'always'	*někde* 'somewhere', *poněkud* 'somewhat'	*nikdy* 'never'	*kdy* 'when'	*kdež* 'where'	*tady* 'here'

Tab. 3. Adverbs. Combination of SUB (column headings) and DEI (row headers) values and representative lexemes

4 CONCLUSION

We have presented the repertoire of categories and parts of speech and their subcategorization proposed in the NovaMorf project, which focuses on the innovation of the morphological description of Czech. We have shown two POS subcategorizations, SUB and DEI, exemplifying appropriate combinations of their values for pronouns, numerals and adverbs. Only marginally, we have dealt with a complex issue of lemmatization. We consider the outlined concept of the annotation system more systematic and consistent than existing Prague and Brno annotation systems; moreover, our system does not contain less information than they do. The proposed description of POS subcategorizations, together with a detailed description (not presented here) of all morphological categories, will be reflected in a new morphological dictionary, which will be used, i. a., for annotating new corpora of Czech (and possibly also for reannotating existing ones).

R e f e r e n c e s

[1] Hajič, J. (2000). Přehled morfologických značek. Available at: `https://ucnk.ff.cuni.cz/doc/popis_znacek.pdf`.

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

[3] Hlaváčová, J. (2009). Formalizace systému české morfologie s ohledem na automatické zpracování českých textů. Dissertation thesis. Praha: Univerzita Karlova. Available at: `http://utkl.ff.cuni.cz/phpBB3/viewtopic.php?f=11&t=1`.

[4] Osolsobě, K. (1996). Algoritmický popis české formální morfologie a strojový slovník češtiny. Dissertation thesis. Brno, Filosofická fakulta MU.

[5] Osolsobě, K. (2015). Korpusy jako zdroje dat pro úpravy nástrojů automatické morfologické analýzy (Slovotvorné varianty adjektiv na [(ou)|í]cí z hlediska morfologického značkování). Časopis pro moderní filologii, 97(2), pages 136–145.

[6] Osolsobě, K., and Žižková, H. (2016). Automatická morfologická analýza z hlediska pokrytí a nepokrytí morfologických variant. Available at: `http://ucnk.ff.cuni.cz/kl2016/abstract-detail.php?id=151`.

[7] Osolsobě, K., Hlaváčová, J., Petkevič, V., Svášek, M., and Šimandl, J. (2017). Nová automatická morfologická analýza češtiny. Naše řeč 100(4), pages 225–234.

[8] Rychlý, P., Šmerk, P., Pala, K., and Sedláček, R. (2008). Morphological Analyzer Ajka. Available at: `https://nlp.fi.muni.cz/projects/ajka/`.

[9] Sedláček, R. (2010). Morphematic analyser for Czech. Dissertation thesis. Brno, Fakulta informatiky MU.

[10] Šimandl, J. (2015). Slovotvorný přehled slov s číselným významem I: číslovky určité. KGA 12, pages 54–74.

[11] Šimandl, J. (2016). Slovotvorný přehled slov s číselným významem II: číslovky neurčité. KGA 13, pages 48–60.

[12] Šmerk, P. (2010). K počítačové morfologické analýze češtiny. Dissertation thesis. Brno, Fakulta informatiky MU.