# Czech Morphological Tagset Revisited

Miloš Jakubíček, Vojtěch Kovář, Pavel Šmerk

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
`{jak,xkovar3,xsmerk}@fi.muni.cz`

**Abstract.** Lot of natural language processing is built on top of some solid morphological annotation. In this paper we present an update of the Czech morphological tagset as given by the analyzer Ajka that has been used for academic as well as commercial purposes for more than dozen years. The revision reacts on rather practical issues that we had to face during development of subsequent tools for NLP, parsers in the first place. We describe the reasoning behind each of the changes and include the full updated tagset reference manual. Finally we provide a comparison and mapping to the Universal tagset as produced by Google.

**Key words:** morphology, tag, tagset, annotation, Czech

## 1 Introduction

Morphology is usually the core of many NLP applications and we are confident that its usefulness heavily depends on the underlying tagset. Despite 20 years of intensive development of NLP applications, there are no conclusions on a widely accepted universal tagset standard across multiple languages, and mostly even within a single language. For Czech, two tagsets have been available since the 90's, provided by two leading NLP groups: one developed in the Institute of Formal and Applied Linguistics at the Charles University in Prague[1] and another one in the NLP Centre at the Masaryk University in Brno[2]. This paper presents a revision of the second tagset together with the underlying morphological database used by the analyzer Majka[3,4]. The main principles of the tagset are outlined in Section 2 and the tagset itself is provided in Appendix A. In Section 3 we describe the changes to the tagset and reasoning and motivation behind them. Finally we provide the current tagset reference together with basic disambiguation rules.

## 2 An Attributive Tagset for Czech

The main properties of the morphological tagset that is described in this paper are as follows:

– *attributive*
  A tag is a sequence of xY pairs, denoting that Y is the value of the attribute x.
– *non-positional*
  The position of the attribute-value pairs in the tag does not matter.

In Figure 1 we provide a sample annotation of the Czech sentence 'Máme zaměstnance , které občas vysíláme na služební cestu.' (*We have employees that we sometimes send on a business trip*). For explanation of the tags meaning, refer to the Appendix A.

```
Máme         (We) have   k5eAaImIp1nP
zaměstnance  employees   k1gMnPc4
,            ,                  kIx,
které        that        k3yRgMnPc4
občas        sometimes     k6eAd1
vysíláme     (we) send   k5eAaImIp1nP
na           on            k7c4
služební     business    k2eAgFnSc4d1
cestu        trip          k1gFnSc4
.            .                 kIx.
```

**Fig. 1.** Example of the annotation using current tagset standard.

## 3 Changing Tagset

We are fully aware of the fact that doing changes to an existing and well-established tagset is an unpopular step that implicates compatibility issues with the old revision, might arouse confusion among current users and definitely should not be carried out without careful consideration of the overall impact. Having that in mind, we briefly outline the most important reasons that led us to take this decision:

– **usability**
  The tagset and Majka have been used extensively in many NLP applications for over dozen years and we can profit from that experience to make the annotation standard more useful in terms of its informativeness and benefit for particular applications (e. g. parsers).
– **consistency**
  Though every effort has been made to eliminate incosistencies in the original tagset as they might be confusing for the users, everyday usage of the tagset showed that one could still make improvements in this respect.
– **simplicity**
  Einstein's famous quote saying that 'everything should be made as simple as possible, but no simpler.' is in the case of any standardization even more

appropriate than otherwise and we took the opportunity to follow it more closely.

– **standardization**

The current description of the tagset is not really up to date since it is quite often the case that different tools use differently evolved versions of the original tagset. This paper aims at creating a new standard that will be subject to further references and common development. We consider the old tagset description to be version 1.0, this new one to be version 2.0 and intend to continue versioning of the tagset in case of future changes.

– **decidability and disambiguability**

For anything in the tagset there must be a clear deterministic procedure saying how a word should be annotated, the goal being a very short manual for disambiguation. This requirement moved us to the decision that the tagset should distinguish between two levels of annotation:

- **restricted (poor) tagset**

  The poor tagset will be restricted to attributes that we expect to be annotated manually (in case of corpus annotation) with very high inter-annotator agreement. In other words it will contain only those attribute where anybody with basic linguistic understanding of the related grammatical notion will be, having the annotation manual available, capable to determine the attribute value.

- **full (rich) tagset**

  The rich tagset is a superset of the poor one and will superseed it by containing also attributes that do not fulfill our strict requirements imposed on the poor tagset, e. g. attributes that are assigned to a very small set of word forms and can be automatically generated from the lexicon (hence do not require manual disambiguation).

Below we list all changes to the tagset that have been performed, together with a detailed explanation of what was the motivation to conduct such a change. The description is structured according to the part-of-speech kinds, i. e. the k attribute. Different change types are marked by the related bullets as follows:

■ an attribute value has been removed
□ an attribute value has been added
● an attribute has been removed
○ an attribute has been added
★ disambiguation note

## 3.1   k1 – Substantives

★ *substantive-adjective collision*

For any word form that is tagged both as an substantive and adjective there must be serious corpus evidence that the word indeed falls into both of these categories, otherwise only substantive or adjective must be chosen. Under adjective usage we understand that the word describes a property of

some other word (e.g. *červený*); the substantial usage means the word can be used as such and does not directly express the property of something else (e.g. *vrátný, pohřešovaný*).

■□ *family gender*
The gR attribute value has been removed and all tags containing gR are transformed to gM with an additional subclassification attribute value xF. This is mainly to simplify further processing since these words behave syntactically as animate masculines.

■ *family number*
The nR attribute value has been removed as duplicate to the gR (see above).

## 3.2   k2 – Adjectives

■ *dual number*
The nD attribute value has been removed as dual should be handled just as a variant of plural. The same applies for pronouns and numerals.

★ *adjective-verb collision*
There are problems on the syntactic layer caused by the duality between a short form of an adjective and a verb in past participle (e.g. *pečen*). We introduced a new disambiguation rule saying that if the relative verb exists, it is always the verb. Also, the morphological database needs manual checking of all these dualities.

## 3.3   k3 – Pronouns

● *person*
The p attribute is to be assigned only for words forms of *já, ty, on, my, vy, oni*. In the other cases it has currently no usage and is rather confusing.

□ *gender*
The g attribute should be specified always except for derivatives of *se, si, kdo, co, někdo, něco, nikdo, nic, já, ty, my, vy*.

## 3.4   k4 – Numerals

■ *xG, xH*
The xG and xH attribute values have been removed as there were no adjectives with such tag in the morphological database.

★ *noun-numeral collision*
Syntactic agreement should be used to disambiguate between a noun and a numeral: if there are usages where agreement applies (e.g. *s tisíci psy*), it is a numeral. Otherwise (if the word is always followed by a genitive phrase) it is a noun.

### 3.5   k5 – Verbs

■ *biaspectual verbs*
The `aB` attribute value has been removed. Instead, both `aI` and `aP` tags will be used in the morphological database for relevant verbs. When tagging particular corpus occurences, the aspect should always be disambiguated.

### 3.6   k6 – Adverbs

■ *xM, xS*
The `xM` and `xS` attribute values have been removed as there are no data with them in the morphological database. The respective information is to be added into the rich tagset under the `t` attribute.

### 3.7   k9, k0 – Particles and Interjections

★ *revise ambiguity*
In the morphological database, there are lot of ambiguous words where one of the options is `k9, k0`. Sometimes this ambiguity is relevant (e.g. *spíš*) but in many other cases it just causes disambiguation problems and brings no added value. Namely, the conjunctions should not be handled as particles simultaneously as even humans are not able to agree on related disambiguation rules. All these ambiguities need to be manually gone through and disambiguated.

### 3.8   kY – Conditionals

●□ *class removed*
The whole class will be split between conjunctions (*aby, kdyby* and their derivatives) and particles (*by* and its derivatives) as this division better corresponds to their syntactic behaviour. `zY` attribute will be added to these words to mark the conditional. Person and number will not be determined for simplicity.

### 3.9   kA – Abbreviations

●□ *class removed*
There is no syntactic or semantic motivation for this part of speech. Rather than that, it causes a lot of problems in automatic syntactic analysis. The words will be divided into the other categories according to their syntactic and semantic properties. `zA` attribute will be added to these words to mark the abbreviation.

### 3.10   kI – Punctuation

○ *class added*
We have added a new `kI` attribute for all types of punctuation. The punctuation was not marked before in any way. An `x` subclassification attribute is to be specified as per the tagset reference in Appendix B.

### 3.11  Common attributes

☐ *frequency characteristics*

We have added a new common ˜ attribute which can be used to store unspecified frequency characteristics (e. g. relative frequency in a particular corpus, normalized to the scale from 0 to 9).

■ *derivative information*

The rD attribute marking the derivation sequence of a word form has been removed. A separate derivative morphological database is currently being prepared, will be available in the future and respective tagset attributes will be added accordingly.

■ *stylistic subclassification*

The wA, wC, wE, wK, wO attributes have been all removed as they are not used anymore in the morphological database.

### 3.12  Gender Problems

We were also concerned with the problem that grammatical gender was not specified for some numerals (e. g. *deset*). At first we were about to add all possible gender values (i. e. M, I, F, N) to the tags of these numerals. However a detailed corpus analysis has shown that in some cases where we expect a syntactic agreement in case, number and gender, there is no real agreement in gender – there are no examples that would distinguish one gender from another by a word form. This concerns adjectives, pronouns and numerals in genitive, dative, local and instrumental case of plural. Every noun phrase we have found in the corpus showed the agreement just in the case and number.

For example, an adjective phrase *s těmi deseti malými* (*with those ten small*) will remain the same no matter if we are talking about masculina (*s těmi deseti malými černoušky, s těmi deseti malými hrady*), feminina (*s těmi deseti malými ženami*) or neutra (*s těmi deseti malými městy*). According to our corpus research, all possible adjective, pronoun and numeral phrases behave in the same way.

Based on this observation, we decided (in contrast to our primary intention) to remove the gender attribute from all the adjectives, pronouns and numerals in the respective cases as it does not reflect the real behaviour.

### 3.13  Canonical ordering

While non-positionality is a handy property when the tagset is used by people, its automated processing resulted into preferring one particular ordering of attributes as a sort of industry standard. This ordering is now part of the tagset reference and it is recommended to be used by applications and their APIs.

## 4  Remaining issues

In this section, we outline some of the problems that have been discussed but so far we were not able to agree on a solution. Many of these problems relate to the extreme complexity of Czech morphology.

### 4.1 Numerals and Pronouns vs. Nouns and Adjectives

From one point of view, almost all numerals and pronouns behave in a very similar way as nouns and/or adjectives, with just marginal differences. However, the differences remain there and so far we were not able to agree on a way this similarity should be expressed on the morphological layer.

### 4.2 Gender of Numerals

It is the nature of the Czech language that many words of the same part of speech behave in a slightly different way. It is then debatable where it is meaningful to mark some properties (see problems with expressing grammatical gender above). Definitely we do not want to mark something that does not describe a real phenomenon. On the other hand, we want the formalism to be simple enough for people to remember it and to be able to manually tag a sentence.

One example for all is the gender of numerals: The numerals *jedna, dva* (*one, two*) do have gender and this is important in syntactic agreements. It is not completely clear if these two are the only ones (and thus the only two that should have gender marked) – this would require another corpus study to reveal the real behaviour and set some sensible rules for the gender assignment.

## 5    Mapping to the Google Universal Tagset

Together with this morphological tagset release, we decided to create a mapping to the universal tagset created by joint effort of Google Research and Carnegie Mellon University [5]. The mapping is given in Table 1.

**Table 1.** Mapping of the Czech tagset to the Google Universal Tagset

| universal tag | description | attributive tags |
|---|---|---|
| VERB | verbs (all tenses and modes) | k5.* |
| NOUN | nouns (common and proper) | k1.* |
| PRON | pronouns | k3.* |
| ADJ | adjectives | k2.*, k4.*xO, k4.*xR |
| ADV | adverbs | k6.* |
| ADP | adpositions (prepositions and postpositions) | k7.* |
| CONJ | conjunctions | k8.* |
| DET | determiners | (none) |
| NUM | cardinal numbers | k4.*xC |
| PRT | particles or other function words | k9.* |
| X | other: foreign words, typos, abbreviations | k0 |
| . | punctuation | kI |

## 6 Conclusions

We have introduced some practically motivated changes to the attributive tagset for the Czech language. The newly defined tagset should become a new standard that all the tools will be compliant with. With this release, we will also start versioning of the tagset with the hope that we will avoid much confusion in the future. We have provided the mapping from the newly defined standard to the Google universal tagset that is hoped to be an interlingually compatible tagset.

We have also mentioned some remaining open problems and outlined the future research in the dark area of Czech morphology.

## References

1. Hana, J., Zeman, D., Hajič, J., Hanová, H., Hladká, B., Jeřábek, E.: Manual for Morphological Annotation PDT. Technical Report 27, Institute of Formal and Applied Linguistics, MFF UK, Prague, Czech Republic (2005)
2. Pala, K., Rychlý, P., Smrž, P.: DESAM – Annotated Corpus for Czech. In: Proceedings of SOFSEM '97, Springer-Verlag (1997) 523–530
3. Šmerk, P.: Fast Morphological Analysis of Czech. In: Proceedings of the RASLAN Workshop 2009, Brno (2009)
4. Šmerk, P.: Towards Computational Morphological Analysis of Czech. PhD thesis, Faculty of Informatics, Masaryk University, Brno (2010)
5. Petrov, S., Das, D., McDonald, R.: A universal part-of-speech tagset. Arxiv preprint ArXiv:1104.2086 (2011)

# A   Current Tagset (revision from 2006)

### k1 – Substantives

| x | Special paradigm |
|---|---|
| P | půl (half) |

| g | Rod |
|---|---|
| M | Animate masculine |
| I | Inanimate masculine |
| N | Neuter |
| F | Feminine |
| R | Family (surname) |

| n | Number |
|---|---|
| S | Singular |
| P | Plural |
| D | Dual |
| R | Family (surname) |

| c | Case |
|---|---|
| 1–7 | First–Seventh |

| w | Stylistic flag |
|---|---|
| A | Archaism |
| B | Poeticism |
| C | Only in corpora |
| E | Expressive |
| H | Conversational |
| K | Bookish |
| O | Regional |
| R | Rare |
| Z | Obsolete |

| z | Word Form Type |
|---|---|
| S | -s enclitic |

### k2 – Adjectives

| e | Negation |
|---|---|
| A | Affirmation |
| N | Negation |

| g | Gender |
|---|---|
| M | Animate masculine |
| I | Inanimate masculine |
| N | Neuter |
| F | Feminine |

| n | Number |
|---|---|
| S | Singular |
| P | Plural |
| D | Dual |

| c | Case |
|---|---|
| 1–7 | First–Seventh |

| d | Degree |
|---|---|
| 1 | Positive |
| 2 | Comparative |
| 3 | Superlative |

| w | Stylistic Flag |
|---|---|
| A | Archaism |
| B | Poeticism |
| C | Only in corpora |
| E | Expressive |
| H | Conversational |
| K | Bookish |
| O | Regional |
| R | Rare |
| Z | Obsolete |

| z | Word Form Type |
|---|---|
| S | -s enclitic |

### k3 – Pronomina

| x | Type (x) |
|---|---|
| P | Personal |
| O | Possessive |
| D | Demonstrative |
| T | Delimitative |

| y | Type (y) |
|---|---|
| F | Reflexive |
| Q | Interrogative |
| R | Relative |
| N | Negative |
| I | Indeterminate |

| p | Person |
|---|---|
| 1 | First |
| 2 | Second |
| 3 | Third |
| X | First, second or third |

| g | Gender |
|---|---|
| M | Animate masculine |
| I | Inanimate masculine |
| N | Neuter |
| F | Feminine |

| n | Number |
|---|---|
| S | Singular |
| P | Plural |
| D | Dual |

| c | Case |
|---|---|
| 1–7 | First–Seventh |

| w | Stylistic Flag |
|---|---|
| A | Archaism |
| B | Poeticism |
| C | Only in corpora |
| E | Expressive |
| H | Conversational |
| K | Bookish |
| O | Regional |
| R | Rare |
| Z | Obsolete |

| z | Word Form Type |
|---|---|
| S | -s enclitic |

## k4 – Numerals

| x | Type (x) |
|---|---|
| C | Cardinal |
| O | Ordinal |
| R | Reproductive |
| G | Grammar |
| H | Grammar |

| y | Type (y) |
|---|---|
| N | Negative |
| I | Indeterminate |

| g | Gender |
|---|---|
| M | Animate masculine |
| I | Inanimate masculine |
| N | Neuter |
| F | Feminine |

| n | Number |
|---|---|
| S | Singular |
| P | Plural |
| D | Dual |

| c | Case |
|---|---|
| 1–7 | First–Seventh |

| w | Stylistic Flag |
|---|---|
| A | Archaism |
| B | Poeticism |
| C | Only in corpora |
| E | Expressive |
| H | Conversational |
| K | Bookish |
| O | Regional |
| R | Rare |
| Z | Obsolete |

| t | Grammar Terminal |
|---|---|
| A–F | Terminal A–F |
| I–R | Terminal I–R |
| S | Q@ |
| T | O@ |
| U | L@ |
| V | jedno |
| W | sto |
| X | dvě |
| Y | stě |
| Z | tři/čtyři |

| z | Word Form Type |
|---|---|
| S | -s enclitic |

## k5 – Verbs

| e | Negation |
|---|---|
| A | Affirmation |
| N | Negation |

| a | Aspect |
|---|---|
| P | Perfect |
| I | Imperfect |
| B | Biaspectual |

| m | Type (Mode) |
|---|---|
| F | Infinitive |
| I | Present indicative |
| R | Imperative |
| A | Active part. (past) |
| N | Passive part. |
| S | Adv. part. (present) |
| D | Adv. part. (past) |
| B | Future indicative |

| p | Person |
|---|---|
| 1 | First |
| 2 | Second |
| 3 | Third |

| g | Gender |
|---|---|
| M | Animate masculine |
| I | Inanimate masculine |
| N | Neuter |
| F | Feminine |

| n | Number |
|---|---|
| S | Singular |
| P | Plural |

| w | Stylistic Flag |
|---|---|
| A | Archaism |
| B | Poeticism |
| C | Only in corpora |
| E | Expressive |
| H | Conversational |
| K | Bookish |
| O | Regional |
| R | Rare |
| Z | Obsolete |

| z | Word Form Type |
|---|---|
| S | -s enclitic |

## k6 – Adverbs

| e | Negation |
|---|---|
| A | Affirmation |
| N | Negation |

| x | Pron. Adv. Type (x) |
|---|---|
| D | Demonstrative |
| T | Delimitative |
| M | Modal |
| S | Status |

| y | Pron. Adv. Type (y) |
|---|---|
| Q | Interrogative |
| R | Relative |
| N | Negative |
| I | Indeterminate |

| d | Degree |
|---|---|
| 1 | Positive |
| 2 | Comparative |
| 3 | Superlative |

| w | Stylistic Flag |
|---|---|
| A | Archaism |
| B | Poeticism |
| C | Only in corpora |
| E | Expressive |
| H | Conversational |
| K | Bookish |
| O | Regional |
| R | Rare |
| Z | Obsolete |

| z | Word Form Type |
|---|---|
| S | -s enclitic |

**k7 – Preposition**

| c | Case |
|---|------|
| 1 | First |
| 2 | Second |
| 3 | Third |
| 4 | Fourth |
| 5 | Fifth |
| 6 | Sixth |
| 7 | Seventh |

**k8 – Conjunction**

| x | Type |
|---|------|
| C | Coordinate |
| S | Subordinate |

| z | Word Form Type |
|---|----------------|
| S | Word form with -s enclitic |

**k9 – Particle**

| z | Word Form Type |
|---|----------------|
| S | Word form with -s enclitic |

**k0 – Interjection**

**kA – Abbreviation**

**kY – by, aby, kdyby**

| m | Relation to the Verb Mode |
|---|---------------------------|
| C | conditional |

| p | Person |
|---|--------|
| 1 | First |
| 2 | Second |
| 3 | Third |

| n | Number |
|---|--------|
| S | Singular |
| P | Plural |

| w | Stylistic Flag |
|---|----------------|
| A | Archaism |
| B | Poeticism |
| C | Only in corpora |
| E | Expressive |
| H | Conversational |
| K | Bookish |
| O | Regional |
| R | Rare |
| Z | Obsolete |

**Notes**

| Tag | Note |
|---|---|
| wH | 795 |
| rD,rD | INF : ADJ-cí |
| rD,rD | INF : ADJ-ší |
| rD,rD,rD,rD | INF : ADJ-ý : SUBST-í : ADJ-n//-t |
| rD,rD,rD | INF : SUBST-í : ADJ-cí |
| rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-cí : SUBST-í : ADJ-ý : ADJ-n//-t |
| rD,rD,rD | INF : SUBST-í : ADJ-ší |
| rD,rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ší : ADJ-ší : SUBST-í : ADJ-ý : ADJ-n//-t |
| rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-cí |
| rD,rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-cí : SUBST-í : ADJ-ý : ADJ-n//-t |
| rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-ší |
| rD,rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-ší : SUBST-í : ADJ-ý : ADJ-n//-t |
| rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-n//-t : ADJ-cí |
| rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-n//-t : ADJ-cí : ADJ-cí |
| rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-n//-t : ADJ-ší |
| rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-n//-t : ADJ-ší : ADJ-ší |
| rD,rD,rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-n//-t : SUBST-í : ADJ-ý : ADJ-n//-t : ADJ-cí |
| rD,rD,rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-n//-t : SUBST-í : ADJ-ý : ADJ-n//-t : ADJ-ší |
| rD,rD,rD,rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : ADJ-n//-t : SUBST-í : ADJ-ý : ADJ-n//-t : ADJ-ší : ADJ-ší |
| rD,rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : SUBST-í : ADJ-ý : ADJ-n//-t : ADJ-cí |
| rD,rD,rD,rD,rD,rD,rD | INF : SUBST-í : ADJ-ý : SUBST-í : ADJ-ý : ADJ-n//-t : ADJ-ší |
| _,hF | SUBST : FEMPOSS |
| _,hM | SUBST : MASKPOSS |
| _,_,hM,hF,_,hR | M : F : Mpřivl : Fpřivl : rodina : Rpřivl |
| wZ | Obsolete |
| wB | Poeticism |
| tQ | Expresses extent |
| tA | Expresses respect |
| tL | Expresses place |
| tT | Expresses time |
| tC | Expresses reason |
| tM | Expresses manner |
| tD | Modal adverb |
| tS | Status adverb |
| wR | Rare |
| hT | Represents thing |
| hP | Represents person |

| xC | Cardinal numeral |
|---|---|
| xO | Ordinal numeral |
| xR | Reproductive numeral |
| yQ | Interrogative |
| yR | Relative |
| xD | Demonstrative |
| yN | Negative |
| xT | Delimitative |
| yI | Indeterminate |
| xP | Personal pronomina |
| yF | Reflexive pronomina |
| xO | Possessive pronomina |
| xC | Coordinate conjunction |
| xS | Subordinate conjunction |
| c1 | Preposition with first case |
| c2 | Preposition with second case |
| c3 | Preposition with third case |
| c4 | Preposition with fourth case |
| c6 | Preposition with sixth case |
| c7 | Preposition with seventh case |
| aP | Perfect |
| aI | Imperfect |
| aB | Biaspectual |
| wH | Conversational |
| wN | Dialectal |

# B   New Tagset

**Common attributes**

| k | Part-of-speech |
|---|---|
| 1 | Substantives |
| 2 | Adjectives |
| 3 | Pronomina |
| 4 | Numerals |
| 5 | Verbs |
| 6 | Adverbs |
| 7 | Prepositiona |
| 8 | Conjunctions |
| 9 | Particles |
| 0 | Interjections |
| I | Punctuation |

| g | Gender (k1–k4) |
|---|---|
| M | Animate masculine |
| I | Inanimate masculine |
| F | Feminine |
| N | Neuter |

| c | Case (k1–k4, k7) |
|---|---|
| 1–7 | First–Seventh |

| n | Number (k1–k4) |
|---|---|
| S | Singular |
| P | Plural |

| e | Negation (k2, k5, k6) |
|---|---|
| A | Affirmation |
| N | Negation |

| d | Degree (k2, k6) |
|---|---|
| 1 | Positive |
| 2 | Comparative |
| 3 | Superlative |

| p | Person (k3, k5) |
|---|---|
| 1–3 | First–Third |

| w | Stylistic subclassification (k0–k9) |
|---|---|
| B | Poeticism |
| H | Conversational |
| N | Dialectal |
| R | Rare |
| Z | Obsolete |

| z | Common subclassification (k1–k9) |
|---|---|
| S | Contains *-s* enclitic |
| Y | Word form of *aby, kdyby, by* |
| A | Abbreviation |

| ⋆ ~ | Statistical characteristics |
|---|---|
| 0–9 | E. g. frequency |

**k1 subclassification**

| ⋆ x | Type |
|---|---|
| P | Word form of *půl* |
| F | Family surname |

**k3 subclassification**

| x | Type |
|---|---|
| P | Personal |
| O | Possessive |
| D | Demonstrative |
| T | Delimitative |

| y | Type |
|---|---|
| F | Reflexive |
| Q | Interrogative |
| R | Relative |
| N | Negative |
| I | Indeterminate |

**k4 subclassification**

| x | Type |
|---|---|
| C | Cardinal |
| O | Ordinal |
| R | Reproductive |

| y | Type |
|---|---|
| N | Negative |
| I | Indeterminate |

**k5 subclassification**

| m | Type (mode) |
|---|---|
| F | Infinitive |
| I | Present Indicative |
| R | Imperative |
| A | Active part. (past) |
| N | Passive part. |
| S | Adv. part. (present) |
| D | Adv. part. (past) |
| B | Futreu indicative |

| a | Aspect |
|---|---|
| P | Perfect |
| I | Imperfect |

**k6 subclassification**

| x | Type |
|---|---|
| D | Demonstrative |
| T | Delimitative |

| y | Type |
|---|---|
| Q | Interrogative |
| R | Relative |
| N | Negation |
| I | Indeterminate |

| ⋆ t | type |
|---|---|
| S | Status |
| D | Modal |
| T | Expresses time |
| A | Expresses respect |
| C | Expresses reason |
| L | Expresses place |
| M | Expresses manner |
| Q | Expresses extent |

**k8 subclassification**

| x | Type |
|---|---|
| C | Coordinate |
| S | Subordinate |

**kI subclassification**

| x | punctuation list |
|---|---|
| . | . ! ? |
| , | , : ; |
| " | " ; ‚ „ " |
| ( | ( { [ < |
| ) | ) } ] > |
| ~ | ~ $ % ^ & - _ += \ | / # etc. |

**Attribute-to-PoS assignment**

| k | attributes |
|---|---|
| 1 | gnczw~ |
| 2 | egncdzw~ |
| 3 | gncpxyzw~ |
| 4 | gncxyzw~ |
| 5 | eampgnzw~ |
| 6 | edxytzw~ |
| 7 | cw~ |
| 8 | xzw~ |
| 9 | zw~ |
| 0 | w~ |
| I | x~ |

**Canonical ordering**

kegncpamdxytzw~