Homonymy and Polysemy in the Czech Morphological Dictionary

Jaroslava Hlaváčová(✉)

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

Abstract. We focus on a problem of homonymy and polysemy in morphological dictionaries on the example of the Czech morphological dictionary MorfFlex CZ [2]. It is not necessary to distinguish meanings in morphological dictionaries unless the distinction has consequencies in word formation or syntax. The contribution proposes several important rules and principles for achieving consistency.

Keywords: Homonymy · Polysemy · Paradigm · Word formation · Czech

1 Introduction — Morphological Dictionary of Czech

The morphological dictionary of Czech used in Prague was designed by Jan Hajič in 1990s. Despite the fast development in the area of NLP, its main features, including the format, are still in use. It contains almost 450,000 lines of coded information on the basis of which more than 100,000,000 wordforms are generated, belonging to almost 900,000 lemmas. Its format is described in the Chap. 4 of the book [1].

We shall briefly introduce only the main features of the dictionary and rules for its entries.

1.1 Dictionary Format

The morphological dictionary consists of lines, that describe a generation of one or more wordforms, together with their morphological tags, belonging to a single lemma. Each line has the following pattern (taken from [1], simplified):

Technical stem (Root)	Model (Paradigm)	Lemma	Tag	AddInfo

Contrary to the original Hajič's work, we have slightly changed his terminology. In the scheme above, the old terms are in parentheses.

P. Sojka et al. (Eds.): TSD 2016, LNAI 9924, pp. 109–116, 2016.
DOI: 10.1007/978-3-319-45510-5_13

We have replaced the original term Root with the term **Technical stem**, because the term root in its strict linguistic sense is something different. Even the term Stem does not reflect accurately the entity in the dictionary. Moreover, there are many definitions that differ. The technical stem is the beginning part (string) of all wordforms that can be generated from the dictionary line. It may contain all the main types of morphemes: prefix, root or its part as well as suffix or its part. Moreover, it may contain also an ending or its part.

The second replacement concerns the term Paradigm. We replaced it with the term **Model**, as the term paradigm is often used in another sense, namely as a set of all wordforms belonging to a certain lemma. We will use it in this sense in the following text. The Model field contains a name of a derivational model. Each derivational model is connected (in a special table) with at least one inflectional model. For instance, the derivational model `mz1` is connected with two inflectional models. The first one, with the same name `mz1` generates the wordforms of a single lemma, in this case a soft masculine animate noun. The second inflectional model is named `uv` and generates all the wordforms of the derived possessive adjective. There is a special model 0, which is used for "exceptions". In that case, the technical stem is always the whole wordform and there must be the unempty field **Tag**, containing all morphological tags belonging to the wordform.

Thus, every dictionary line contains either a non-zero (derivational) model and no tag, or the zero model and a set of morphological tags. Every line may contain also additional information concerning wordforms described by the line. It is often relevant for the whole lemma, even if the line does not describe the whole paradigm.

The **Lemma** field contains a lemma. Especially for human readability, the lemma is just a word in its basic form, there are no precise identifiers. If lemmas were written unambiguously, the word itself could be the unique identifier but it is not so. Different meanings of an ambiguous lemma are distinguished with a number, that becomes part of the lemma itself. Thus, we have the lemma *kohout-1* for an animal (*a cock*), and *kohout-2* for a closure device (*a tap*) for instance of gas or water. If there are more lines for description of a single lemma, the number must be the same for all of them. Moreover, if there is an additional information concerning the lemma (in the dictionary field AddInfo), they all must be the same as well.

The **AddInfo** field contains optional additional information concerning wordforms described by the given line. There might be information concerning derivation, or a semantic explanation (for instance the note "bird" for the lemma *kohout-1*). They are related to the whole paradigm belonging to the given lemma. Another sort of additional information, concerning the style, may be related only to special wordforms. An example is the lemma *téci* (*to flow*) with two forms for the present 3rd person plural: *tekou* and *tečou*. The former form is archaic, which is denoted by a special code on the dictionary line.

1.2 Lemma Numbering

The lines were added to the dictionary mainly manually, by several contributors. They were using data from many various sources, at first from older paper dictionaries of Czech, then from corpora. Despite many different checks, number of errors or inconsistencies entered the dictionary, often due to different opinions of different contributors. One such type of inconsistency is the lemma numbering, described above. There were diligent, punctilious contributors, who tried to put many different senses of individual words to the dictionary, especially from the paper dictionaries [7,8]. However, the morphological dictionary should not be a collection of meanings, especially when some of them are metaphors, or are very close.

A morphological dictionary should contain all the wordforms, not meanings, together with their tags and lemmas, and also information of word formation. If two word paradigms share all the morphological properties, including the word formation consequences, there should be only one lemma, even with different meanings. In this sense, it is not necessary to distinguish between the words *kohoutek* as a name of a flower and *kohoutek* as a tap because both have the same derivational as well as inflectional model, namely for masculine inanimate nouns. On the other hand, there is a different lemma *kohoutek* (*a small cock*), that has different derivational (as well as inflectional) model for masculine animate nouns.

In connection with a project *An Integrated Approach to Derivational and Inflectional Morphology of Czech*, we started a deep inspection of the numbers in lemmas. We found out, that they are used often needlessly. On the other hand, we found cases where they were missing. The inconsistent usage of numbers in lemmas leads to wrong evaluations of relations among some wordforms.

One of the reasons of the inconsistencies is a vague distinction between the two crucial linguistic terms — homonymy and polysemy.

2 Homonymy and Polysemy

The definitions seem to be clear [4]: A word with (at least) two entirely distinct meanings yet sharing a lexical form is said to be **homonymous**, while a word with several related senses is said to be **polysemous**.

It is important to note that in this text, we consider only the homonymy on the level of whole paradigms. We do not deal with homonymous wordforms belonging to the same lemma (e.g. *hrad* as nominative as well as accusative of the lemma *hrad* (*castle*)). Very detailed description of homonymy in Czech is in the recently published work [5].

From the above definitions, it is clear, that both, homonymous as well as polysemous words have different meanings. In this sense, homonymous words[1]

[1] The homonymy may be divided into homography for the same spelling and homophony for the same pronouncing. In this work, we use the term homonymy as a synonym for homography, in accordance with the traditional Czech terminology.

are always polysemous. One of the main differences between the two terms is that polysemy always concerns the whole paradigm — a sense of a word is considered to be the same in all its wordforms —, while homonymy might concern only special wordforms. The famous Czech example of homonyms is the word *žeň* as the imperative of the verb *hnát* (*to herd, to chase*), or as the singular nominative of the feminine noun (*a harvest*). The lemmas of those two wordforms are different, there is no homonymy.

However, there are also whole homonymous paradigms. A clear example of such homonymy is the lemma *kolej*, that has two different meanings, namely *a college* and *a track*. These two meanings have different origin and by chance they reached the same spelling after centuries of their development. They have the same inflectional paradigm, but their derivational behaviour differs.

A clear example of polysemy is *průvodce* (*a guide*), that can be a man or a written text. Their origin is the same and it is not surprising that they have the same spelling in their basic form (lemma). Their paradigms differ because of different animateness.

There are many unclear cases that are difficult to distinguish. One of the keys could be a translation to another language. If a word is possible to translate by two different words (like our example of *kolej*), the word is homonymous. If the word is possible to translate with one word, having two meanings even in the other language, like the *guide* from the second example, it is a pure polysemous word. However, this clue is not 100 %.

The distinguishing between homonymy and polysemy is important in the field of NLP, because it has often important consequencies. One of the most important ones concerns derived words. Take for example the homonymous word *kolej*. It is possible to derive an adjective, but each of its meanings will suit another derivation. *Kolej* in the meaning of *college* leads to the adjective *kolejní* (concerning a college, like *kolejní rada = college council*). The meaning of *track* has the adjective *kolejový*, like *kolejový jeřáb = tracked crane*.

However, the precise distinction is impossible, subjective. Lyons [4] proposes two strategies to avoid the problem. We cite from [6]:[2]

1. "Maximise homonymy — associate every meaning of a word with a distinct lemma."
2. "Maximise polysemy — no two lemmas can be entirely distinct when they are syntactically equivalent and when the set of wordforms they are associated with are identical."

Both approaches have their pros and cons. We decided to adopt the strategy 2 and to postpone the resolution of meanings to upper layers of NLP. It should be stated here, that the principle of maximising polysemy is taken not so strictly as it is stated in the above definition which was aimed probably especially for languages with not very rich inflection. We make exceptions, for instance for animate and inanimate nouns — see later.

[2] Available also at ftp://ftp.cogsci.ed.ac.uk/pub/kversp/html/node153.html.

3 Polysemy Within MorfFlex

As stated in the previous section, we decided to maximise polysemy in the morphological dictionary. In other words, if there are two lemmas with the same spelling, they will be considered as one lemma, unless they have different paradigms, or different derived words, or different syntactical behaviour.

For following explanations, we will need a more detailed terminology:
A **lemma** is a triple (L, N, A), where

- L is a pure lemma (word),
- N is a number, or empty string,
- A is additional info concerning the whole paradigm (represented by lemma), or empty string.

There should hold:
For every two lemmas (L, N, A) and (L, M, B) (with the same pure lemma L) the following should be true:
If $M! = N$, then

- $A! = B$.
- both M and N are nonempty.

If $A! = B$ then $M! = N$.

We have extracted all the dictionary lines that contained a numbered lemma. We ignored abbreviations, in other words those lines that contained a model for an abbreviation. These lemmas cannot be used for studying derivatives or other linguistic topics, as they usually are not “normal words”, but typically only initial letters. For every numbered lemma (without abbreviations), we checked, if there exists another line in the dictionary, containing the same lemma without any number. We added all such lines to the previous ones. Now, we have a set of dictionary lines where every lemma occurs at least once with a number. We want to repair the set in such a way that they conform to the requirements presented above.

According to the requirements, we checked automatically the lines from our file and found violations of the requirements. The next procedure — error corrections — was manual as it was not possible to make them automatically, their variability was considerable.

In the following paragraphs, we will present several main cases of violation the requirements, together with their solutions.

3.1 Uppercase and Lowercase

In the previous version of the dictionary, the lemmas differing in the case of their initial letter were often labeled with a different number, though it was not necessary. The case of the initial letter is a sufficient distinction to consider the two lemmas to be different.

There are three main groups of such lemmas.

Common and Proper Names. The most frequent are proper names having their counterparts as common words.

Example: The lemma *švanda* (*fun*, feminine) had the lemma *švanda-2*, while the proper family name *Švanda* (masculine animate) had the lemma *Švanda-1*. We have preserved the both lemmas, but removed their numbers, because *švanda* with lowercase *š* is different lemma than *Švanda* with uppercase *Š*, there is no need of an additional number.

Car Brands. The second frequent case are car brands. In normal texts they occur in the both variants — with the uppercase as well as lowercase initial letter. The more common car brand, the more often it appears in lowercase, because it became to be perceived as a common name, opposed to a proper name.

An example is the car brand *Lancia*, where only 17.6 % (out of 2,124) are written with lowercase initial letter in Czech texts, while in case of (in the Czech republic more common) *Trabant* the proportion is opposite — almost 3/4 occurrences (out of 9,567) are written with initial letter lowercase. The figures were counted on the corpus SYN [3].

In the majority of occurences, when speaking about a vehicle, the both variants are interchangable. They have also the same derivational and inflectional models. It is tempting to include them under one lemma. However, when speaking about a factory, or a brand, it is necessary to preserve the lemma with uppercase initial letter as well, we cannot consider the cases variants. The solution is the same as in the previous case — we have two lemmas differing in the initial letters (*trabant* and *Trabant*, *lancia* and *Lancia*).

Latin Names. MorfFlex CZ contains a number of Latin names denoting biological species. The reason of their inclusion into the dictionary is unclear, they probably occured frequently in a text that was to become part of a corpus. In the dictionary, they often occur in two variants, with an upper- and lowercase initial letter, this time with different part of speech. That one with the uppercase is usually a noun, the second one an adjective, both with underspecified other morphological categories (or some of them). The part of speech of such words is disputable — though adjectives in Latin, they do not behave as adjectives in Czech, and as parts of a Latin name they both could be considered (in Czech texts) a noun. One of the solutions could even be their omission from the dictionary, as there are only some of the Latin names included in the dictionary, and the selection is arbitrary. However, we decided not to erase anything that could occur in Czech texts, so we left both lemmas (noun, adjective) in the dictionary, but removed their numbers.

Example: There were two lemmas for the partrige (*Perdix perdix* in Latin): *perdix-1* with the tag `AAXXX----1A----` and *Perdix-2* with `NNMXX-----A----`. After the changes, we have the lemmas *perdix* and *Perdix*. The capitalization is sufficient for the lemma distinction. In the future, we should adopt a better solution, especially concerning their part of speech.

3.2 Fluctuating Declension

There are nouns in the Czech language, the gender of which is not strict. The most of them fluctuate between masculine and feminine, there are 173 animate and 93 inanimate nouns with both masculine and feminine inflecional models. There is a set of expressive words, for instance *šmudla* (*a dirty man/woman*). Another example is *privilej* (*a privilege*) which was interperted as *privilej-1* (fem.), *privilej-2* (masc.). After checking, we have one *privilej*, with feminine as well as masculine declension. Another set fluctuates between feminine and neuter (17 lemmas), for instance the bird *káně* (*a buzzard*).

Until now, we have considered the different genders as a distinguishing mark for different lemmas but this conclusion appears to be wrong. The fluctuating gender has no impact on meaning, there is no homonymy, no polysemy. It is only a matter of inflection. Therefore we removed the numbers from such lemmas. There could arise an objection, that in certain contexts it is not possible to decide the gender. The answer is Yes, that is true. If it is not possible to decide, there are two options — either to decide arbitrarily, or not to try to decide at all. It may depend on an application, what is better, but from the theoretical point of view, the both strategies are good.

3.3 Animate vs Inanimate Nouns

A shift in meaning may happen between an inanimateness and animateness. In our dictionary, there are almost 500 lemmas of this sort. If there is a common meaning for both, we consider them as one lemma. In this case, there could be an objection against merging the two lemmas into one, concerning derivation of possessive adjectives. Naturally, they can be derived only from animate nouns. On the other hand, their common origin and our principle of maximising polysemy speak in the favor of their identity.

Example: *baryton* as a bariton voice or a man singing in bariton voice, *recyklátor* (*a recycler*) as a man or a device, tool. Another example is the lemma *průvodce* (*guide*) from the Introduction section of this paper.

3.4 Predicatives

Predicatives[3] are described as neuter nouns or adverbs. Thus, we have *jasno-1* as noun and *jasno-2* as adverb (both translated into English as *clear*, for instance as in the sentence *Today will be clear.*). It would be natural to have only one lemma for such words, because the meaning is always the same. However, the fluctuation among parts of speech appears to be more severe than fluctuation among genders. Therefore, the situation of predicatives remains unchanged — we have two lemmas differing in number. Again, a better solution should be adopted in future.

[3] Czech predicatives form a class of words usually ending in -o. They are problematic in terms of POS classification.

3.5 Figurative Meanings

Virtually all words may be used in their normal or usual context as well as in an unusual one. In the latter case, it may happen, that another — figurative meaning of the original word arises from its usage in an "unusual" context. As there are no strict definitions of "usualness", it is often hard to decide, whether a certain context is "usual", or "newly usual" or completely "unusual". This implies that there is no generally accepted procedure to decide how many meanings a word has. It can be demonstrated by comparisons of some dictionary entries in several explanatory dictionaries. For instance, the dictionary [8] gives an explanation for 8 meanings of the Czech adjective *černý* (*black*), while another dictionary [7] presents only 6 meanings.

If there are no syntactical or derivational differences, it is not reasonable to make any distinction and have a single lemma for all such meanings in the morphological dictionary. The morphological dictionary contained two lemmas for this adjective: *černý* (color) and *černý-2* (*illegal*). We merged them into one lemma *černý*.

4 Conclusion

There are more inconsistencies in MorfFlex CZ, but those concerning homonymy and polysemy seemed to be the most important, as their influence on complex NLP applications may appear crucial. We want to quantify the impact of the actual dictionary cleaning on several results in tasks that use morphological dicionary as one of their bases. And we will continue in fixing other, less visible, errors and inconsistencies that are still present in the dictionary.

We hope that the principles adopted for Czech solution of homonymy and polysemy in morphological dictionaries could be inspirative for other languages.

Acknowledgment. Work on this paper was supported by the grant number 16-18177S of the Grant Agency of the Czech republic (GAČR).

References

1. Hajič, J.: Disambiguation of rich inflection (Computational Morphology of Czech). Nakladatelství Karolinum (2004)
2. Hajič, J., Hlaváčová, J.: MorfFlex CZ, LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague (2013)
3. Křen, M., Čermák, F., Hlaváčová, J., Hnátková, M., Jelínek, T., Kocek, J., Kopřivová, M., Novotná, R., Petkevič, V., Procházka, P., Schmiedtová, V., Skoumalová, H., Šulc, M.: Corpus SYN, version 3. Institute of the Czech National Corpus FF UK, Prague (2014). http://www.korpus.cz
4. Lyons, J.: Semantics. Cambridge University Press, Cambridge (1977)
5. Petkevič, V.: Morfologická homonymie v současné češtině. Studie z korpusové lingvistiky 22. Nakladatelství Lidové noviny (2016)
6. Verspoor, C.M.: Contextually-dependent lexical semantics. Ph.D. thesis, University of Edinburgh (1997)
7. Slovník spisovného jazyka českého. Nakl. Československé akademie věd (1960)
8. Příruční slovník jazyka českého. Státní nakl (1937)