Variants and Homographs: Eternal Problem of Dictionary Makers*

Jaroslava Hlaváčová and Markéta Lopatková

Charles University in Prague, ÚFAL MFF,
{lopatkova,hlavacova}@ufal.mff.cuni.cz

Abstract. We discuss two types of asymmetry between wordforms and their (morphological) characteristics, namely (morphological) variants and homographs. We introduce a concept of multiple lemma that allows for unique identification of wordform variants as well as 'morphologically-based' identification of homographic lexemes. The deeper insight into these concepts allows further refining of morphological dictionaries and subsequently better performance of any NLP tasks. We demonstrate our approach on the morphological dictionary of Czech.

1 Introduction and Basic Concepts

In many languages, there are wordforms that may be written in several ways; they have two (or more) alternative realizations. We call these alternatives variants of wordforms. By definition, the values of all morphological categories are identical for the variants. This fact complicates for instance language generation, important part of machine translation. How can a machine decide which variant is appropriate for the concrete task? This is the reason why the variants should be distinguished and evaluated within morphological dictionaries.

Homographs, wordforms that are 'accidentally' identical in the spelling but different in their meaning, can be seen as a dual problem. If their morphological paradigms differ, the necessity to treat them as separate lemmas is obvious. But even if their paradigms are the same, it is appropriate to treat them with special care because their different meanings may affect various levels of language description.

There is not a common understanding of basic concepts among (computational) linguists. Although there have been many attempts to set standards (for instance [1]), they were usually too general, especially for the purposes of inflectionally rich languages. This was the reason why we have decided to specify the basic concepts in a way that allows to cover consistently and meaningfully all special cases that may occur in languages with rich inflection. However, we are convinced that they are useful for other languages too. We support this by examples from English and Czech, as representatives of different types of languages. Our definitions are based on those in [2], [3], [4], [5], and [6].

* This paper is a result of the projects supported by the grant of the Czech Ministry of Education MSM113200006 and the grants of the Czech Academy of Sciences 1ET101120503 and 1ET101120413.

Note: In the following paragraphs we use the concept of meaning. We do not try to present its exact explanation. We consider the lexical meaning an axiomatic concept. Among all the attempts to set its sufficient explanation, we find the following one as the most appropriate for our purposes: "Lexical meaning is a reflection of extralinguistic reality formed by the language system and communicative needs mediated by consciousness.", see [4].

Relations among basic concepts are visualized at the Fig. 1.

Wordform is every string of letters that forms a word of a language, e.g. *flower*, *flowers*, *where*, *writes*, *written*.

Lemma is a basic wordform. Each language may have its own standards, but usually it uses infinitive form for verbs, singular nominative for nouns, ... Lemma is usually used as a headword in dictionaries. Lemmas of the examples from the previous paragraph are the following strings: *flower*, *flower*, *where*, *write*, *write*.

Paradigm is a set of wordforms that can be created by means of inflection from a basic wordform (lemma). E.g. the paradigm belonging to the lemma *write* is the set {*write*, *writes*, *wrote*, *writing*, *written*}. It can be specified either by listing all the wordforms, or by a pattern (a rule according to which all inflectional forms can be derived from the basic form).

Lexical unit is an abstract unit associating the paradigm (represented by the lemma) with a single meaning. In dictionaries, the meaning is usually represented by a gloss, a syntactic pattern and a semantic characteristics (e.g. a set of semantic roles). The lexical unit can be understood as 'a given word in the given sense', see also [3].

Lexeme is a set of lexical units that share the same paradigm. We are aware that especially this term is simplified but it is sufficient for dictionaries containing all necessary information about words but at the same time, easy to use.

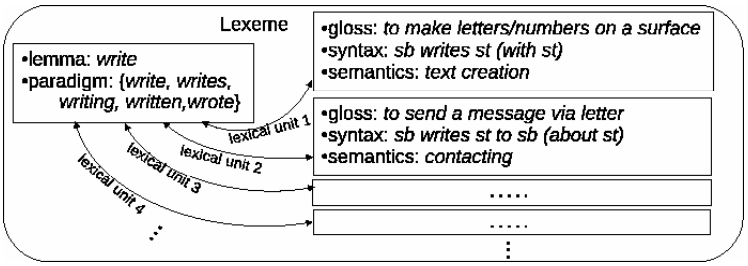

Fig. 1. Relations among basic concepts.

Dictionary is a set of records, called also entries or dictionary entries, that describe individual lexemes. In other words, every dictionary entry contains a complex description of one lexeme, represented by its lemma.

Variants are those wordforms that belong to the same lexeme and values of all their morphological categories are identical. Examples: *colour/color*, *got/gotten* as past participles.

Homographs are wordforms with identical orthographic lettering, i.e. the identical strings of letters (regardless of their phonetic forms), whose meanings are (substantially) different and cannot be connected. E.g. wordform *pen* as a 'writing instrument' and wordform *pen* as an 'enclosure', *bank* as a 'bench', *bank* as a 'riverside' and *bank* as a 'financial institution'.

2 Variants

Variants often violate the so-called Golden Rule of Morphology, see [6]:

lemma + morphological tag = unique wordform

In other words: Given a lemma and a morphological tag, no more than one wordform should exist, belonging to that lemma and having that morphological tag.

This requirement is very important for unambiguous identification of wordforms in morphological and other dictionaries. If satisfied, we can use the pair <lemma, morphological tag> as the unique identifier for each wordform. However, variants often violate this rule because they have the same morphological tag and are assigned the same lemma in morphological dictionaries. That is why it is necessary to make their unique tagging clear.

2.1 Types of variants

We define two types of variants – one affecting the whole paradigm and the second one affecting only wordforms with some combinations of morphological values. The former one is called global, the latter inflectional.

Global variants are those variants that relate to all wordforms of a paradigm, in all cases in the same way. E.g. *colour/color* and all their forms, namely *colours/colors*, and for verbs also *coloured/colored*, *colouring/coloring*. The pairs *zdvihat/zdvíhat* [to lift] or *okno/vokno* [window][1] are examples of global variants in Czech, as all wordforms of respective lemmas demonstrate the same difference, namely *i-í* at the fourth position of the first example and *o-vo* at the beginning of the second example.

Inflectional variants are those variants that relate only to some wordforms of a paradigm. E.g. variants *got/gotten* are inflectional, because they appear only in the past participle of the lemma *get*; other wordforms of the lemma *get* do not have variants. As a Czech example, we can present two forms *jazyce/jazyku* for the locative case singular of the lemma *jazyk* [language] or *turisti/turisté* for the nominative case plural of the lemma *turista* [tourist].

[1] The variants of the first example are both neutral, the latter form of the second example is colloquial.

2.2 Lemma Variants and Multiple Lemma

Inflectional variants may affect any wordform, including the basic one – lemma. That case may lead to their wrong interpretation as global variants but they should be classified as inflectional variants, in accord with the definition introduced above – the variant is not expressed by all the wordforms. Let us take the example of variants *bydlit/bydlet* [to live]. They differ in the infinitive form and in the past tense (*bydlil/bydlel*); the rest of wordforms is the same for both lemmas. Thus the variants *bydlit/bydlet* are classified as the inflectional variants.

Variants of lemmas in general, also called **lemma variants**, can be either global (when they exhibit throughout the whole paradigm), or inflectional.

Lemma variants should be treated with a particular care, as lemmas have a special position among other wordforms – they usually serve as representatives of the whole paradigms and also as labels for lexemes.

To be able to recognize and generate all wordforms of all lemma variants, we have to decide about their representative form. As the selection of a unique representative is an uneasy task (see [6]) we introduce the concept of **multiple lemma** as a set of all lemma variants. A paradigm of the multiple lemma, called **extended paradigm**, is a union of paradigms of individual lemmas constituting the multiple lemma. For example, the lemma *skútr* [scooter] has three different spellings, namely *skútr*, *skůtr* and *skutr*, each having its own paradigm. The multiple lemma {*skútr*, *skůtr*, *skutr*} has an extended paradigm containing all wordforms of all three members of this set.

Implementation of Multiple Lemmas. In the morphological dictionary of Czech [7], the wordforms are not listed separately, they are clustered according to their lemmas. The lemma represents the whole paradigm. However, the multiple lemma cannot represent the extended paradigm straightforwardly because a set cannot serve as unique identifier. Thus, we keep all lemma variants separately but we connect them with pointers (see Fig. 2).

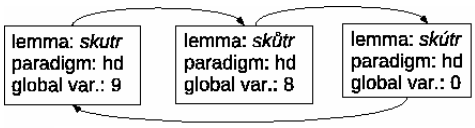

Fig. 2. Schema of implementation of multiple lemma.

For morphological analysis,this arrangement can be used in several ways:

- En bloc: The multiple lemma becomes the representative of all wordforms belonging to all lemma variants during the compilation of the dictionary.
- Stepwise: The multiple lemmas create a separate list. The morphological analysis of input texts is processed in its traditional way, assigning a single

lemma to every wordform. After the analysis, single lemmas are replaced with the multiple ones, when appropriate.
- External: The list of lemma variants is implemented externally into a software tool that processes the output of the morphological analysis.

Each of the approaches has its positive and negative aspects. We will not compare them here, as we concentrate rather on the dictionary building, not on its possible uses.

2.3 Variants as New Categories

The last thing that is necessary to solve is a compliance with the Golden Rule of Morphology. After adopting the concept of multiple lemma, it would be violated even more. Let us take the multiple lemma {*skútr*, *skůtr*, *skutr*} again. The dative singular has now three possible spellings, namely {*skútru*, *skůtru*, *skutru*} instead of one, that is required by the Golden Rule. In order to distinguish the variants, we introduce a new category, **Global Variant**. The value of this category differentiates a style of individual variants.[2]

As this sort of information is very hard to grasp, we prefer to express the values by means of small integers, to give at least an approximate picture of its scalable nature. Thus, the synchronic standard variants, roughly of the same stylistic characteristic, have the value of 0, 1 or 2. Following integers express the bookish or archaic variants. Integers 9, 8, ... serve for the spoken or colloquial variants, see Fig. 2.[3]

The inflectional variants are marked similarly – there is a new category called **Inflectional Variant** that distinguishes different variants of wordforms of a lemma. This category describes the same property of the wordform – the style, and so have the same set of values.

Each type, inflectional and global, is considered as a special category with its set of values. The necessity of two categories implies from the fact that many variants are both global and inflectional. For example, the wordforms *okny/voknama* (plural instrumental of the multiple lemma {*okno, vokno*} [window]) are global variants (because of the protective *v-* at the beginning shared by all wordforms of the lemma *vokno*) as well as inflectional variants (because of the colloquial instrumental ending *-ama*).

Values of both categories of variants become parts of morphological tags. Thus, the Golden Rule of Morphology always holds true.

3 Homographs

Homographs are wordforms with the same lettering but different meaning. As with variants, two types can be distinguished – inflectional and global ones.

[2] This sort of information should be rather a goal of a linguistic research than to be included in the morphological dictionary but users are accustomed to have it there.

[3] The numbering expresses the relationship among the lemma variants of the multiple lemma, it must not be used for any comparison across different multiple lemmas.

As was stated in Section 2, the pair <lemma, morphological tag> should be unique for every wordform. Homographs, as two (or more) identical wordforms with different meaning, have to differ in this pair. We distinguish two possibilities – with the same lemma (type T) and with different lemmas (type L):

(T) Wordforms with the same lemma and different morphological tags, as e.g. *stopped* being both past tense and past participle. This type, which is very frequent for Czech (e.g. the wordform *hradu* [castle] being both genitive singular and accusative singular), is called syncretism.

(L) Wordforms with different lemmas (as e.g. the wordform *smaž* is imperative of the two verbs – *smažit* [to fry] and *smazat* [to erase]; the wordform *ženu*, which is either the accusative singular of the noun *žena* [woman] or the first person of present tense of the verb *hnát* [to rush]).

Inflectional homographs are those homographs, where at most one homographic wordform is a lemma (all examples in three paragraphs above belong to inflectional homographs as they have no homographic lemma at all).

On the other hand, **global homographs** are those homographs that satisfy the following condition (1) and one of the conditions (2′) or (2″):

(1) They affect at least two lemmas (also called homographic lemmas), i.e. the same string of letters (lemma) represents two (or more) different lexemes.

(2′) The paradigms of the affected lemmas differ (e.g. *hnát* with verbal or nominal paradigm; *žít* with different wordforms *žil*/*žal* for the past tense (and two meanings, 'to live' or 'to mow')).

(2″) The lemmas are derived from different words (e.g. the verb *odrolovat* as 'to roll away' has the prefix *od-* and the stem *-rol-* whereas *odrolovat* as 'to crumble' has the prefix *o-* and the stem *-drol-*).

The reason for distinguishing inflectional and global homographs is obvious. Inflectional homographs do not cause any problem for implementation of dictionaries since particular homographic wordforms belong either to one lexeme (as in case (T)) or to more lexemes with different lemmas (cases (L)). Thus, particular lexemes are represented by unique headwords = lemmas. (However, inflectional homographs represent the central problem for morphological disambiguation).

For lexicographers, global homographs are more problematic. They refer to the cases when the same lemma belongs to two (or more) lexemes. They may violate the Golden Rule of Morphology (see Section 2), as in the case of homographs *žít* or *odrolovat*. Contrary to the case of variants, no morphological category can distinguish them. It is necessary to draw the line between them at the lemma side. The difference between the lemmas is marked with numeral suffixes: *žít*-1 and *žít*-2. Though homographs with different POS do not violate the Golden Rule of Morphology, it is reasonable to deal with them similarly. Thus, we have also *colour*-1 as a verb, *colour*-2 as a noun.

3.1 Global Homography versus Polysemy

Polysemy is usually characterized as the case of a single word having two or more related meanings (based on [2]). Polysemy is treated within a single lexeme.

It is a relation among particular lexical units of a lexeme, contrary to homographs that concern separate lexemes.

There is no clear cut between polysemy and homography as these concepts are based on the vague concept of meaning (see above). Unfortunately, lexicographers hesitate quite often and dictionaries are not consistent in distinguishing homographs from polysemic lexemes. For example, Czech verb *hradit* is treated differently in Czech normative dictionaries: as one polysemic lexeme with two lexical units 'to fence' and 'to reimburse', or as homographic lemma, i.e. lemma representing two different lexemes.

The requirement of the identity of lemmas on one hand and difference in (morphological) paradigms or difference in a word creation on the other hand are rather technical but solid criteria for homographs. Based on these criteria, global homography and polysemy can be distinguished consistently. Polysemy is characterized by the identity of the whole paradigms, while global homography requires identity of lemmas only. Thus we obtain the single lexeme for the verb *hradit* [to fence], [to reimburse] but two lexemes represented by the lemmas *žít*-1, *žít*-2 (see 2′). However, we have the single polysemic lexeme for the verb *odpovídat* regardless of the "distance" of its at least four lexical units 'to answer', 'to react', 'to be responsible', and 'to correspond'.

4 Duality of Variants and Homographs

The basic difference between the two concepts are illustrated on the schemas in Fig. 3. For variants, the shape of the schema resembles the letter A, while for homographs it is the letter Y. The polysemy appears only at the syntactic (if applicable) or semantic levels of the schema (see the right schema). It is not surprising that these schemas resemble those introduced in [8], where they illustrate synonymy and homonymy as relations between separate layers of language description.

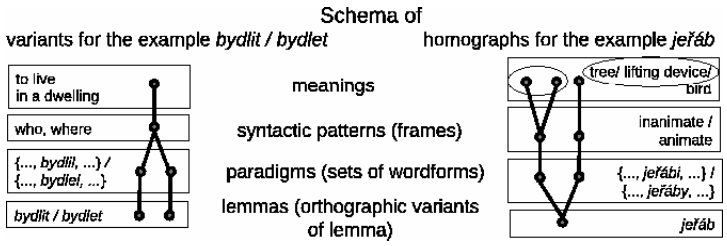

Fig. 3. Schema of variants and homographs. Parts in ellipses concern polysemy.

Let us present another example of variants and homographs to clarify the previous concepts. The word *jeřáb* can be either animate denoting a species

of bird (plural nominative is *jeřábi*, with the inflectional variant *jeřábové*), or inanimate (plural nominative is *jeřáby*) having two meanings, one being a name of a tree, the second one a crane. Thus, there are two homographic lemmas *jeřáb*-1 and *jeřáb*-2. Moreover, the inanimate lemma *jeřáb*-2 is polysemic, with two meanings (but the same paradigm), see Fig. 2.

Summary

We have brought a deeper insight into the problem of variants and homographs, especially those that affect lemmas.

We have also introduced a novel treatment of variants that meets the requirement of a unique wordform for each pair <lemma, morphological tag>, the so called Golden Rule of Morphology. Variants are treated in one (extended) paradigm, specified by (multiple) lemma and morphological tag enriched with information about two types of variants, inflectional and global. A corpus user searching for all occurrences of any of these lemma variants can put into the query any lemma from the multiple lemma.

Similarly, we distinguish inflectional and global homographs. The 'morphologically-based' specification of homographs enables us to distinguish homography from polysemy consistently.

Based on a close examination of the phenomena described in this paper we have proposed an implementation of variants and homographs within a wide coverage morphological dictionary [7] used in various tasks in NLP. The proposed treatment of lemma variants enables both their subsuming under common headword, the so called multiple lemma (e.g. for querying in corpora or for IR tasks) as well as distinguishing particular wordforms (e.g. for language generation).

References

1. ISO/TC 37/SC 4: Language Resources Management - Lexical Markup Framework (LMF). http://www.lexicalmarkupframework.org/ (2007) Rev. 14, date 2007-06-03.
2. Matthews, H.: The Concise Oxford Dictionary of Linguistics. Oxford University Press, Oxford (1997)
3. Cruse, D.A.: Lexical Semantics. Cambridge University Press, Cambridge (1986)
4. Filipec, J.: Lexicology and Lexicography: Development and State of the Research. In Luelsdorff, P.A., ed.: The Prague School of Structural and Functional Linguistics, Amsterdam-Philadelphia, John Benjamins (1994) 163–183
5. Žabokrtský, Z.: Valency Lexicon of Czech Verbs. PhD thesis, Charles University, Prague (2005)
6. Hlaváčová, J.: Pravopisné varianty a morfologická anotace korpusů. In Štícha, F., ed.: Proceedings of 2nd International Conference Grammar and Corpora 2007. (2008) In press.
7. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Charles Univeristy Press, Prague (2004)
8. Panevová, J.: Formy a funkce ve stavbě české věty. Academia, Praha (1980)