Golden Rule of Morphology and Variants of Wordforms

Jaroslava Hlaváčová

Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague, Czech republic

Abstract: In many languages, some words can be written in several ways. We call them variants. Values of all their morphological categories are identical, which leads to an identical morphological tag. Together with the identical lemma, we have two or more wordforms with the same morphological description. This ambiguity may cause problems in various NLP applications. There are two types of variants – those affecting the whole paradigm (global variants) and those affecting only wordforms sharing some combinations of morphological values (inflectional variants). In the paper, we propose means how to tag all wordforms, including their variants, unambiguously. We call this requirement "Golden rule of morphology". The paper deals mainly with Czech, but the ideas can be applied to other languages as well.

Keywords: morphology, global variants, inflectional variants, multiple lemma, Golden rule of morphology

1 Terminology

As there are quite a lot of different approaches to some basic linguistic terms, let us at the beginning make clear about the terminology.

Wordform[1] is every string of letters that forms a normal word of a language. English examples: *get, gets, sisters, where*, Czech examples *dostal, dostaneš, sestrám, kam*.

Lemma is basic wordform. It is often used in dictionaries as a head word. Lemmas of examples from the preceding paragraph are: *get, get, sister, where*, the Czech ones: *dostat, dostat, sestra, kam*. From the lemma, individual wordforms can be created by means of inflection.

Paradigm is set of wordforms that can be created by means of inflection from their basic wordform (lemma). There can be more than one (variant of a) lemma included in one paradigm (i.e. *color, colour* - see below).

Examples: wordforms belonging to the lemma *get*, namely *get, gets, got, gotten, getting* form one paradigm. Its representative is the lemma *get*. Another paradigm is the set of wordforms *color, colors, colour, colours*, with two lemma variants: *color* and *colour*. Czech example is presented in the Table 1.

Variants are those wordforms that belong to the same paradigm and values of all their morphological categories are identical.

Example: the pair of wordforms *got, gotten* are variants of past participle of the paradigm from the previous paragraph, represented by the lemma *get*.

Morphological category is a morphological property of words, for instance gender, tense, case.

Morphological value is value of a morphological category. For instance there are two values for the morphological category of Number (singular, plural), seven values of the morphological category Case in Czech.

1 It is often written as two words — word form. We have chosen this spelling as it avoids confusion with homographic reading when speaking about word forming.

Every wordform belongs to a paradigm that is represented by a lemma. We can also say that the wordform belongs to the lemma, or is derived from the lemma. Then, zero, one or more wordforms can be derived from a lemma, with a given set of morphological values.[2] In this paper, we will deal with the case of more than one wordforms for a given lemma and set of morphological values.

Maximal set of morphological values is the set that is sufficient for morphological description of a single wordform of a given lemma. What belongs to the maximal set of morphological values, usually depends on part of speech of the given lemma. For example number, gender, case, degree and status of negation are needed to describe a particular wordform of an adjective lemma in Czech.[3]

Morphological tag is a string that puts into code the maximal set of morphological values of a given wordform.[4]

2 Golden Rule of Morphology

Given a lemma and a maximal set of morphological values, no more than one wordform should exist, belonging to that lemma and having those morphological values:

lemma + morphological tag -> single wordform

This requirement is called "Golden rule of Morphology" (see also [7,8]) and is essential especially for generation of wordforms. If there were more than one wordform, a generator (automatic as well as human one) would not know, which variant to choose. Moreover, the variants can differ in a style or other non morphological characteristics, and their replacement could be inappropriate in a given context.

Another reason for accepting the Golden rule is an unambiguous identification of wordforms in morphological (and other) dictionaries. Then, we can use the pair <lemma, morphological tag> as the unique identifier for each wordform.

2.1 Violation of the Golden Rule – example

Let us have a look at the paradigm that is represented by lemma *okénko* in Czech (diminutive of *okno* = *small window*). This lemma has two more variants, namely *okýnko* and *vokýnko*. Every variant have 10 different wordforms, both standard and colloquial. In the Table 1 we can see all of them. Each line of the table represents variants for the same combination of morphological values. Notice especially the last two lines of the table, which are doubled. There we have six different wordforms for one morphological tag. It means that one maximal set of morphological values (one morphological tag) describes six different wordforms.

---

2 The alternative of no wordforms arises when the set of morphological values is not appropriate for the given lemma, for example no verb can be derived with a given value of the morphological category case.

3 Example: maximal set of morphological categories for description of the wordform *nehezkou* (*not pretty* - instrumental), belonging to the lemma *hezký* (*pretty*), is Number (sing), Gender (fem), Case (instr), Degree (1) and Negation (N). Avoiding any of them, more than one wordform would result - for instance without specifying the category Negation, there would be *hezkou* and *nehezkou*.

4 There are several types of morphological tags, but their specific appearance is not important, if they contain all the morphological information needed for a unique wordform description. That is the reason why we present no specific suggestions for tagging the new features.

Morphology (Case & Number)	Wordforms		
nom sg/acc sg	*okénko*	*okýnko*	*vokýnko*
gen sg/nom pl/acc pl	*okénka*	*okýnka*	*vokýnka*
dat sg/loc sg	*okénku*	*okýnku*	*vokýnku*
instr sg	*okénkem*	*okýnkem*	*vokýnkem*
gen pl	*okének*	*okýnek*	*vokýnek*
dat pl	*okénkům*	*okýnkům*	*vokýnkům*
loc pl	*okénkách* *okéncích*	*okýnkách* *okýncích*	*vokýnkách* *vokýncích*
instr pl	*okénky* *okénkama*	*okýnky* *okýnkama*	*vokýnky* *vokýnkama*

Table 1. Paradigm *okénko, okýnko, vokýnko*.

In the table, the first two columns under the title Wordforms include two standard variants, the third (greyish) column is colloquial. Moreover, all columns contain a wordform that is also colloquial, due to its colloquial ending, namely *-ama* (underlined in the Table 1). Thus, the lower rightmost wordform is colloquial twice — once due to its inclusion under a colloquial basic form, secondly due to its colloquial ending.
In our example, the golden rule of morphology does not hold true. Even if we declared each of the three columns a separate paradigm, it would not hold true because of the two lower lines. Moreover, the three basic wordforms really are variants and they should belong to the same paradigm. We need another means how to distinguish all the variants and to ensure the validity of the Golden rule of morphology.

3 Typology of variants

Following the observation from the example, we define two types of morphological variants — one affecting the whole paradigm and the second one affecting only a specific combination of morphological values. The former one is called **global**, the latter one **inflectional**.

Inflectional variants are those variants that relate only to some wordforms of a paradigm defined by a special combination of morphological values.
Global variants are those variants that relate to all wordforms of a paradigm, and always in the same way.
In accordance with the variant types we define two new morphological categories which describe them. Before we formally define their values, let us have a look at their nature.

3.1 Inflectional variants

In Czech, the majority of inflectional variants differ in endings. There are patterns that include the inflectional variants for particular morphological values. In that sense we can say that they are mostly systematic. Examples of inflectional variants are in Table 2. The upper part contains systematic inflectional variants, in the rest there are

more specific variants, that affect only those lemmas mentioned, or, as in the last line, a small set of similar words, in this case some verbs of movement.

Morphology	Czech variants	Czech lemma	English translation
loc pl	*hradu / hradě*	*hrad*	*(in the) castle*
loc pl	*lesu / lese*	*les*	*(in the) forest*
nom pl	*soudcové / soudci*	*soudce*	*(the) judges*
1st person pl	*mažeme / mažem*	*mazat*	*we spread*
1st person pl	*jdeme / deme / jdem / dem*	*jít*	*we are walking*
comparativ	*bílejší / bělejší*	*bílý*	*more white*
imperativ	*běž / poběž*	*běžet*	*run!*(imperativ)

Table 2.Examples of Czech inflectional variants

3.2 Global variants

This type of variants is often not morphological, but orthographic. However, concerning automatic natural language processing, there is no difference between the two. The major point is that the pairs of variants have different spelling, for whatever reason. Thus, we do not distinguish between morphological and orthographic (or even other, e.g. etymological, stylistic) types.
Global variants can also be systematic, but the system always affects all wordforms belonging to a lemma. Examples of the systematic global variants are in the Table 3.
Global variants always differ in one or more letters, no matter whether at the beginning, in the middle, or at the end of the lemma. There can be more letter changes within a single word.

Description	Czech variants (lemmas)	English translation
protetic *v-*	*okno / vokno*	*window*
-ismus/-izmus	*feminismus / feminizmus*	*feminism*
generally *-s-/-z-*	*kurs / kurz*	*course*
-t-/-th-	*Atény / Athény*	*Athens*

Table 3: Examples of Czech global variants

We have already mentioned the possibility of not distinguishing the global variants. We can assign an individual lemma to both (all of) variants. Thus, we could have two lemmas, e.g. *okno, vokno* (*window*), with their own separate paradigms. However, there are words, especially foreign names with more spellings. For instance *Afgh-anistan* has 8 different spellings occurring in Czech texts[5]. It is reasonable to subsume them all under a single paradigm.
For linguists – corpus users – it is very convenient. If a corpus manager is designed and configured appropriately, they need not remember all the possible variants, but put only one of them into a query, and the resulting concordances will contain all of the

5 *Afghánistán, Afgánistán, Afganistán, Afghanistán, Afghanistan, Afganistan, Afghánistan, Afgánistan*

possible variants. At the same time, it must be naturally possible to exclude some of them, if the user wants so, but this can be done by means of the query language, e.g. regular expressions.
The way how to join the variants together, while allowing their separate tagging to distinguish them, is described in the next section.
For computational linguists, it is also useful to have the variants labelled, because it is often necessary to automatically select one from more possibilities. If the variants were not described individually, there would be no clue for a selection of one of them. The tools even would not "know" that there are more possibilities.

4 How to tag variants

4.1 Present State

Examples from English
In English, there almost do not exist inflectional variants. There are a few exceptions with two wordforms for past participle, for instance the verb *to get*, with two possible wordforms *got* and *gotten*, or past tense and past participle (*hanged* and *hung* for the lemma *to hang*).
There are quite a lot of global variants, especially due to wide area where the English is spoken. Each region can have its specific variants. Well-known differences are between British and American spellings. Probably the most common type of global variant for English is the type *ou-o,* as in pairs *colour/color*, *labour/labor*.
English tagsets, as far as we know, do not take care about variants. For instance, the both global variants *color* and *colour* mentioned in the previous paragraph, have the identical tag in British National Corpus ([2]), namely NN1 for singular nouns and VVI for infinitive form of verbs. The same can be stated about inflectional variants for past participle *got* and *gotten* (both has morphological tag VVN), or *hung* and *hanged* (both VVN as past participles, or VVD as past tense).

Czech Case
In Czech, there are two main morphological tagsets (Prague tagset see [3], Brno tagset see [4]), both taking care about variants, but none of them being entirely satisfactory and consistent. The major point is that neither of them distinguishes between the two types, inflectional and global. They both use a single slot in the morphological tag for their description.
As we have seen, there can be more variants for one combination of morphological values within a given paradigm, some of them differing inflectionally, some globally, and others in both ways. One category of variants is thus not enough. There have to be two of them.
Another problem is values of the two new morphological categories. Both present Czech morphological systems make distinction among styles of the variants. There are variants equipollent, archaic or bookish, colloquial, dialectical, to name just the most common ones. In other words, the values of the variants have an evaluative nature.
There is number of studies about styles of variants for Czech. However, there is often little agreement - e.g. whether a variant is (still) colloquial or whether it has (already) penetrated into the standard vocabulary. Or, vice versa, whether a variant is (still) standard, or whether it is (already) archaic. The result is an inconsistent description.
Thus, the list of variant values should be stated objectively and without evaluating ambitions.

4.2 Values of the morphological categories describing variants

As stated above, our new proposition is to avoid any evaluation. The values of the morphological categories Global and Inflectional variant should be strictly formal. The main reason for their introduction was only their distinction. Then, we add them to the morphological tag in order to ensure the validity of the Golden rule of morphology.

The proposed values of the variant categories are based on differences between pairs of variants. Thus, we define the opposite values long and short (according to long and short vowels, e.g. *é/e*), hard and soft (according to hard and soft consonants, e.g. *s/š*), etc. The Table 4 lists the most common types of global variants together with examples and values of the morphological category Global variant.

Inflectional variants have similar set of values.

Values of variants that are not common (see the example of *Afghanistan*) are tagged with numbers. If needed, the set of values might be enlarged.

5 Lemmatization and variants

We have solved the problem of variants, but created another one: Which of the possible global variants should be the representative of the whole paradigm? In other words: What is lemma of a paradigm with global variants? Which properties are essential for its selection?

We present several requirements that seem naturally and reasonably at first glance.

It should be neutral, it must not be archaic nor colloquial. The problem is that the styles are not (and cannot) be defined exactly, they are changing and there is no agreement, as we have already mentioned. Moreover, there are often two, or even more equipollent variants. There are also variants that do not have a neutral counterpart. Ultimately, it was the reason why we do not use these concepts for their tagging.

It should be the most frequent (or better, most common). According to our language feeling, the basic variant should be that one, which is more common, but this characteristic is also very difficult to detect. Of course, we can use the frequency, or better, the average reduced frequency (see [5]) calculated from a referential corpus (Czech National Corpus for Czech — see [1], for instance), but it often happens, that the corpus does not contain some, or even none of the variants, or the difference between their (average reduced) frequencies is negligible. There does not exist entirely representative corpus in which we could trust with respect to frequency or commonness.

We could find more requirements for such a representative, but none of them is entirely neutral. There is another solution — multiple lemma.

Type	Example	Values of the morphological category Global variant
o-vo	*okno --- vokno*	0 --- v
ý-ej	*mýdlo --- mejdlo*	0 --- j
z-s	*klauzule --- klausule*	z --- s
t-th	*tema --- thema*	0 --- h
é-í	*kolébka --- kolíbka*	e --- i
é-ý	*okénko --- okýnko*	e --- y
á-e	*originální --- originelní*	a --- e
á-a	*Abrahám --- Abraham*	long --- short
é-e	*acetylén --- acetylen*	
ó-o	*salón --- salon*	
ý-y	*apetýt --- apetyt*	
í-i	*alexandrín --- alexandrin*	
ů-u	*přezůvky --- přezuvky*	
ú-u	*Plútarchos --- Plutarchos*	
s-š	*student --- študent*	hard --- soft
t-ť	*vlaštovka --- vlašťovka*	
n-ň	*šnůra --- šňůra*	
d-ď	*dolík --- ďolík*	
e-ě	*Bardejov --- Bardějov*	
z-ž	*zbrzďování --- zbržďování*	

Table 4. List of most common types of global variants.

5.1 Multiple Lemma

Every paradigm can have not only one representative, but as many as there are global variants of its lemma. In other words, the lemma of a paradigm with global variants is a set of lemmas (see also [6,7]). Then, if a corpus user asks for a lemma in his/her query, he/she needs not to care about a "basic" global variant, but can use whichever lemma, that come under the desired paradigm. They even need not to know all possible lemmas that could belong under the paradigm.

Multiple lemma		*okénko*	*okýnko*	*vokýnko*
Code of Global variant		e0	y0	yv
Morphology		Wordforms		
Case & Number	Inflectional Variant			
nom sg/acc sg		*okénko*	*okýnko*	*vokýnko*
gen sg/nom pl/acc pl		*okénka*	*okýnka*	*vokýnka*
dat sg/loc sg		*okénku*	*okýnku*	*vokýnku*
instr sg		*okénkem*	*okýnkem*	*vokýnkem*
gen pl		*okének*	*okýnek*	*vokýnek*
dat pl		*okénkům*	*okýnkům*	*vokýnkům*
loc pl	a	*okénkách*	*okýnkách*	*vokýnkách*
loc pl	i	*okéncích*	*okýncích*	*vokýncích*
instr pl	y	*okénky*	*okýnky*	*vokýnky*
instr pl	m	*okénkama*	*okýnkama*	*vokýnkama*

Table 5. The example with the multiple lemma {*okénko, okýnko, vokýnko*}. Every wordform has distinguished Global variant (columns) and Inflectional one, where necessary (4 bottom lines). The global variant "ev" (*vokénko*) is not included, though theoretically possible.

If a single global variant is desired, it has to be selected from the set by adding another condition to the query. There are two possibilities:

– to specify spelling of the lemma or wordform, or

– to specify the type of the global variant.

Software tools used for searching the corpus (corpus managers) are able to deal with multiple values of attributes.

Let us show our new suggestions on the example that we have used as the introduction into the problem of variants. The lemma representing the whole paradigm presented in the Table 1 is the set {*okénko*, *okýnko, vokýnko*}. If we wanted to search for all occurrences of this multiple lemma in a corpus, we need not to write our query using regular expression like

[lemma="v?ok[éý]nko"][6]

We can just state any of the three lemmas and will get what we wanted.

The Table 5 presents the example from the Table 1 with all the wordforms distinguished by means of Global and Inflectional variants.

We do not intentionally specify the shape of the morphological tags, because there are more ways how to code the information about the variants. There are several types of morphological tags, even for Czech, and each system can subsume the new morphological categories differently. One of possible suggestions can be found in [7].

6 Theoretically, this regular expression search also for possibly non-existing, absurd lemma *vokénko*.

6 Final remarks

We have proposed how to deal with variants of wordforms and lemmas. The main reason, why to distinguish among them, is the effort to support the Golden rule of morphology, which ensures an unambiguous description of each wordform of a language. Without proper tagging variants it could not hold, which would cause problems in various fields of natural language processing — generating text, machine translation, indexing wordforms, to name just a few.
There are two types of variants, inflective and global. They should be treated as two different morphological categories, as they may be combined in many ways. The existing systems do not distinguish between them, which causes a violation of the Golden rule of morphology.
The existence of global variants leads to a multiple lemma — set of all global lemma variants. This concept is more general and objective than choosing one representative from the set of variant lemmas, as there is no entirely objective criterion for that.

However, there is no general way how to deal with the variants, each application has to choose its own way. There can be an application where the preference is given to the variant with the substring *-t-* over the variant with the (oldish one) *-th-* for modern translations. With a strict description of all wordforms, especially their variants, such a preference is easy to implement.
Although we decided not to tag the variants with any semantic or stylistic labels, such as emotional, colloquial, archaic etc., it might be useful to do so. The main reason why we do not include any type of evaluation into the morphological description is that there is no exact rule how to define individual values of such labels. Even experts are not able to agree on objective criteria. Moreover, the meaning of those labels changes in time. However, if it was needed, the formal non-evaluative tagging of variants enable to make decisions tailored to various special and detailed requirements. Without a strict unambiguous description of each wordform, there would be not possible to make extensions mentioned above.

Acknoledgements

The work on this paper was supported by the grant number 16-18177S of the Grant Agency of the Czech republic (GAČR) and LM2015044 of the Ministry of Education, Youth and Sports.

References

[1] Czech National Corpus: http://ucnk.ff.cuni.cz/
[2] British National Corpus. http://www.natcorp.ox.ac.uk/
[3] Hajič, J. (2004). Disambiguation of Rich Inflection. (Computational Morphology of Czech). Praha, Karolinum.
[4] Brno morphological analyzer ajka. http://nlp.fi.muni.cz/projekty/ajka/index.htm
[5] Savický, P.; Hlaváčová, J. (2002): Measures of Word Commonness. Journal of Quantitative Linguistics. , Swets & Zeitlinger, Vol. 9, No. 3, pp.215-231.
[6] Hlaváčová, J. (2011): Problém variantních tvarů slov při automatickém zpracování jazyka. In: Information Technologies – Applications and Theory, Univerzita Pavla Jozefa Šafárika v Košiciach, Slovakia, ISBN 978-80-89557-01-1, pp. 75-78.

[7] Hlaváčová J. (2009): Formalizace systému české morfologie s ohledem na automatické zpracování českých textů. Ph.D. thesis, FF UK

[8] Hlaváčová J. (2008): Pravopisné varianty a morfologická anotace korpusů. In: Grammar & Corpora / Gramatika a korpus 2007, Academia, Praha, ISBN 978-80-200-1634-8, pp. 161-168.