A New Data Format
for Czech Morphological Analysis

Pavel Šmerk

Faculty of Informatics, Masaryk University
Botanická 68a, CZ-60200 Brno, Czech Republic
smerk@mail.muni.cz

Abstract. The paper presents a new data format for computational morphology of Czech. The new format allows for a significant reduction of a redundancy yielded by existing formats. It is also much more linguistically interpretable and acceptable. The paper shows that there is no need to develop any computer-specific description of morphology, but that the traditional linguistic description suffices quite well.

1 Introduction

At the first sight, the morphological analysis and synthesis of Czech seems to be a well-solved task. For more than a decade there are available even two well established and broadly used systems for computational morphology of Czech. One of them is developed in Prague [1,2], the other one in Brno [3,4]. These two systems are completely independent, which means that there are two distinct language data sets which describe Czech morphology, two distinct sets of morphological tags, two data formats, and two analyzers.

Despite of many particular differences, the general principle of the language data description is the same. In both solutions the data consist of so-called paradigms, i. e. sets of word endings and corresponding morphological tags, and of a list of lemmata or word stems. Each word stem is assigned to some paradigm in such a way that concatenations of the stem with the paradigm's endings yield all forms of the word along with appropriate morphological tags. The thing is, the stems and the endings are never modified, but only concatenated during a synthesis or separated from a word form during an analysis.

Such an approach is rather inadequate for a language like Czech which has a rich set of graphemic, phonological and morphological alternations. The problem is that these alternations require to set up distinct paradigms even for words which are inflected quite equally but which differ in some—although completely regular—alternations. For example, surnames *Staněk*, *Hromek*, and *Polák* with genitive singular forms *Staňka*, *Hromka*, and *Poláka* obey exactly the same rules within the inflection, but they have to be described by means of three paradigms which contain endings *něk* and *ňka*, *ek* and *ka*, or *0* and *a* respectively.

As a consequence, the number of the paradigms is very high and the paradigm system is therefore very redundant. This redundancy inevitably leads

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 3–8, 2010. © Tribun EU 2010

either to an increase in inconsistencies or even errors in the data, or to a strong need of powerful tools which inhibit emergence of the inconsistency. For a more detailed discussion see [5].

In the following section we offer a proposal of a new data format which lowers the redundancy of the data. Then, in the Section 3, we show results of utilization of the new data format in a description of masculine animate nouns. Finally we sum up the conclusions and sketch out some necessary future work.

2 The New Data Format

As the current data formats do, the new format also divides the data into two parts: a lexicon and paradigms. What is rather new, is the intention to let the lexicon cover the idiosyncracies, whereas the paradigms, and also some rules in the program which interprets the format, should describe only the regularities in the data.

The very basic principle of the data organization remains unchanged: the lexicon contains the stems with names of paradigms, e. g. `slon:pán`, and the paradigms are set of endings and appropriate tags, e. g.

```
pán
        k1gMnSc1        0
        k1gMnSc2        a
        ...
```

The endings are appended to the stems, but as a result, and this is the essential difference, we obtain only structures like `pán-0`, `pán-a`, ... along with tags `k1gMnSc1`, `k1gMnSc2`, ... To derive the "surface" word forms from these structures, some additional rules have to be applied.

Obviously, the most trivial rules have to remove the `-` (which can be interpreted as a morpheme boundary) and `0` (zero ending). Other rules deal with graphemic alternations like `ňe` → `ně`, e. g. `tuleň-e` → `tuleňe` → `tuleně`. Another rules describe the phonological alternations like `k-i` → `c-i`, e. g. `vlk-i` → `vlc-i` → `vlci`. And yet another rules are used to handle some morphological (but in fact phonological as well) alternations like vowels alternating with zero `.VC-0` → `VC-0` and `.VC-V` → `C-V`, e. g. `ďáb.el-a` → `ďábl-a` → `ďábla`.

The paradigm may allow for more than one ending for a particular tag. In such a situation, regular expressions describe a context (possible stem ends) in which the given ending may be used. Omitting the regular expression denotes the default option (an unmarked ending):

```
        k1gMnPc6        ech, ích/[kgc]|ch
```

Even these few above mentioned simple improvements allows us to replace a big portion of the former paradigm system with fewer more general paradigms, but the redundancy would still remain high. To lower it, the new format offers the following possibilities (among others and only in brief):

- the paradigm can be defined as a modification of another paradigm:

```
soudce:muž
        k1gMnPc1        e
        k1gMnPc5        e
```

- inflection of a stem can be described not only by one paradigm, but also by a list of paradigms in which the latter overwrites—or, if the paradigm's name is prepended by a plus sign, is added to—the former;
- the previous has a sense only if the paradigm is allowed to be "incomplete". One can either define a "paradigm" even for a single ending like

```
-ové
        k1gMnPc1        ové
        k1gMnPc5        ové
```

  or use a regular expression to select only a subset of endings of a given paradigm, e. g. `pán_nP` selects only endings whose tags contain `nP`. As examples of these posibilities, consider the following lexicon entries:

```
dřevokaz:pán,+muž
Marcel:pán,-ové,muž_nSc5
```

- if a word or its stem has some irregular forms, these forms have to be explicitely listed in the lexicon, e. g.

```
přítel:muž
        přátel:muž_nP,-é
        přátel-0       k1gMnPc2
```

  where, again, the more specific overwrites—or is added to—the more general;
- we also need a possibility to describe differences between the written form and pronunciation, especially for words of foreign origin, because the analyser deals with the former, but the inflection is driven by the latter. The format uses the following notation:

```
Smith[t:pán,-ové
        +Smith[s:muž,-ové
```

  where the regular expressions, and the rules which derive the word form from the structure "see" the stem-final `t` or `s`, but while deleting the `-`, the whole "pronunciation" part between `[` and `-` is also deleted.

2.1 From the Lexicon with Paradigms to the Lexicon with Features

Up to now, the new format allows us to reproduce the information contained in traditional grammar books quite closely. We can describe the inflection of words by means of the traditional paradigms, eventually with some exceptions, just like the grammar books do. But may be this is not the way people have organized the language data in their heads. It is unlikely that the speakers deal with any paradigms in such a way that they would have some inventory of stems each one "explicitely" linked to some paradigm. More likely they infer the proper inflectional paradigm from some features or properties of the stem. For instance, the native speakers of Czech know that masculine animate nouns ended up with a hard consonant belong to a "hard" declination. Thus there is no need to have an extra information on the paradigm in the lexicon: such an information would be redundant for these nouns.

To implement this idea we allow for an addition of "implicit" rules like these:

```
[sxz]/qJO       muž,pán_nPc[67],+pán_nPc4
$T\Ka           žena_nS,-ovi,pán_nP,-ové
```

where `$T` is a shorthand for a regular expression which defines hard consonants and `qJO` is a tagset extension which denotes proper names of persons.

On the left side of the rule, there are the conditions which have to be satisfied if the rules are to be applied. The condition can describe either the stem end, or the tag, or both (then the two conditions are separated by a slash '/'). On the right side, there is the list of paradigms which is prepended to the list of paradigms from the lexicon—if they are present, they specify some unusual, non-typical behaviour of the stem.

Then, for example, the following entries in the lexicon

```
Klaus k1gMqJOP
houslista:-i,+-é k1gM
```

can stand for a markedly longer definitions

```
Klaus:muž,pán_nPc[67],+pán_nPc4 k1gMqJOP
houslista:žena_nS,-ovi,pán_nP,-ové,-i,+-é k1gM
```

3 Case Study: Masculine Animate Nouns

As a case study we use the new format for a description of masculine animate nouns. In the old data format, these nouns are described by 217 different paradigms.[1] The Table 1 lists all lexical descriptions which are shared by at least 10 lemmata (the representative is chosen arbitrarily).

Taken as a whole, the figures in the table show that more than 92.3 % of masculine animate nouns can be described only by means of part-of-speech specification, some pieces of semantic information and/or internal (morphotactic) structure.[2]

The description of the new paradigms is 13 times smaller than the equivalent 217 old paradigms — and it is even 24 times smaller, if one does not count definitions shared among different genders or parts-of-speech.

[1] For the sake of completeness it should be stated that one of these old paradigms has not assigned any lemma and the most of the paradigms for surnames are duplicates, i. e. there exists an identical paradigm for non-surnames.

[2] E. g. `=an` in `Severo+evrop=an` is a suffix which fully determines the inflection of the word, see [5] for more details.

Table 1. The most frequent descriptions in the dictionary

13,871	69.17	`gaučo k1gM`
2,207	11.01	`Ionesc[ko k1gMqJOP`
1,654	8.25	`Severo+evrop=an`
683	3.41	`Mario k1gMqJO`
440	2.19	`kok.eš:-ové k1gM`
321	1.60	`sob.ěk:-i k1gM`
146	0.73	`uniat:-é k1gM`
90	0.45	`invalida:-é,+-i k1gM`
90	0.45	`košer:+-ové k1gM`
58	0.29	`dutoroh=ý k1gMnP`
52	0.26	`tatí:neskl k1gM`
41	0.20	`pterosaur-us:+-i k1gM`
35	0.17	`v%ol k1gMqA`
22	0.11	`příchoz:muž,-ové`
17	0.08	`Ferrari:neskl k1gMqJOP`

16	0.08	`pán k1gM`
		`pane nSc5`

12	0.06	`Řek k1gMqJN`

10	0.05	`Ciceron k1gMqJO`
		`Cicero nSc1`

4 Conclusions and Future Work

The primary goal of the new format was to significantly reduce the redundancy of the current descriptions of the morphological data, but it has several more advantages:

- the words can be filed under the paradigms found in the traditional grammar books;
- it is easy to handle the graphemic and phonologocal alternations;
- the format allows for much more linguistically acceptable and interpretable description of the data and if the same phenomenon can be described in more than one way, the format even allows us to interpret these descriptions differently;
- it is possible to describe markedness and it is possible to distinguish what is regular or at least typical and what is idiosyncratic or peripheral—and more than that: the idiosyncracy can be described only by means of departures from the rules.

The new format even allows for a description of word formation relations and therefore the morpheme structure of the words, but it is not discussed in this paper (see [5] for more details).

Within the future work, the rest of data is and will be converted to the new format. Then the tougher part of the task will follow: a description of a word formation relations.

Acknowledgements

This work has been partly supported by the Ministry of Education, Youth and Sports of Czech Republic under the project LC536 and within the National Research Programme II project 2C06009, and by the Czech Grant Agency under the project GA 407/07/0679.

References

1. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). Karolinum, Praha (2004).
2. Hlaváčová, J.: Formalizace systému české morfologie s ohledem na automatické zpracování českých textů (Formalization of the Czech Morphology System with Respect to the Automatic Processing of Czech Texts). Ph.D. thesis, Faculty of Arts, Charles University, Praha, Czech Republic (2009).
3. Osolsobě, K.: Algoritmický popis české formální morfologie a strojový slovník češtiny (Algorithmical Description of the Czech Formal Morphology and Machine Dictionary of Czech). Ph.D. thesis, Faculty of Arts, Masaryk University, Brno, Czech Republic (1996).
4. Sedláček, R.: Morphemic Analyser for Czech. Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno, Czech Republic (2005).
5. Šmerk, P.: K počítačové morfologické analýze češtiny (On Computational Morphological Analysis of Czech). Ph.D. thesis, Faculty of Informatics, Masaryk University, Brno, Czech Republic (2010).