A Czech Morphological Lexicon

Hana Skoumalová
Institute of Theoretical and Computational Linguistics
Charles University
Celetná 13, Praha 1
Czech Republic
hana.skoumalova@ff.cuni.cz

Abstract

In this paper, a treatment of Czech phonological rules in two-level morphology approach is described. First the possible phonological alternations in Czech are listed and then their treatment in a practical application of a Czech morphological lexicon.

1 Motivation

In this paper I want to describe the way in which I treated the phonological changes that occur in Czech conjugation, declension and derivation. My work concerned the written language, but as spelling of Czech is based on phonological principles, most statements will be true about phonology, too.

My task was to encode an existing Czech morphological dictionary (Hajič, 1994) as a finite state transducer. The existing lexicon was originally designed for simple C programs that only attach "endings" to the "stems". The quotation marks in the previous sentence mean that the terms are not used in the linguistic meaning but rather technically: *Stem* means any part of a word that is not changed in declension/conjugation. *Ending* means the real ending and possibly also another part of the word that is changed. When I started the work on converting this lexicon to a two-level morphology system, the first idea was that it should be linguistically more elegant and accurate. This required me to redesign the set of patterns and their corresponding endings. From the original number of 219 paradigms I got 159 that use 116 sets of endings. Under the term paradigm I mean the set of endings that belong to one lemma (e.g. noun endings for all seven cases in both numbers) and possible derivations with their corresponding endings (e.g. possessive adjectives derived from nouns in all possible forms). That is why the number of paradigms is higher then the number of endings.

In this approach, it is necessary to deal with the phonological changes that occur at boundaries between the stem and the suffix/ending or between the suffix and the ending. There are also changes inside the stem (e.g. *přítel* 'friend' × *přátelé* 'friends', or *hnát* 'to chase' × *ženu* 'I chase'), but I will not deal with them, as they are rather rare and irregular. They are treated in the lexicon as exceptions. I also will not deal with all the changes that may occur in a verb stem—this would require reconstructing the forms of the verbs back in the 14th century, which is outside the scope of my work. Instead, I work with several stems of these irregular verbs. For example the verb *hnát* ('to chase') has three different stems, *hná-* for infinitive, *žen-* for the present tense, imperative and present participles, and *hna-* for the past participles. The verb *vést* ('to lead') has two stems, *vés-* for the infinitive and *ved-* for all finite forms and participles. The verb *tít* ('to cut') has the stem *tn-* in the present tense, and the stem *ťa-* in the past tense; the participles can be formed both from the present and the past stem. For practical reasons we work either with one verb stem (for regular verbs) or with six stems (for irregular verbs). These six stems are stems for

infinitive, present indicative, imperative, past participle, transgressive and passive participle. In fact, there is no verb in Czech with six different stems, but this division is made because of various combinations of endings with the stems.

2 Types of phonological alternations in Czech

We will deal with three types of phonological alternations: palatalization, assimilation and epenthesis. Palatalization occurs mainly in declension and partly also in conjugation. Assimilation occurs mainly in conjugation. Epenthesis occurs both in declension and in conjugation.

2.1 Epenthesis

An epenthetic *e* occurs in a group of consonants before a Ø-ending. The final group of consonants can consist of a suffix (e.g. *-k* or *-b*) and a part of the stem; in this case the epenthesis is obligatory (e.g. *kousek* × *kousku* 'piece', *malba* × *maleb* 'painting'). In cases when the group is morphologically unseparable, the application of epenthesis depends on whether the group of consonants is phonetically admissable at word end. In loan words, the epenthetic *e* may occur if the final group of consonants reminds a Czech suffix (e.g. *korek* × *korku* 'cork', but *alba* × *alb* 'alb'). In declension, two situations can occur:

- The base form contains an epenthetic *e*; the rule has to remove it, if the form has a non-Ø ending, e.g. *chlapec* 'boy', *chlapci* dative/locative sg or nominative pl.
- The base form has a non-Ø ending; the rule has to insert an epenthetic *e*, if the ending is Ø, e.g. *chodba* 'corridor', *chodeb* genitive pl.

In conjugation, an epenthetic *e* occurs in the past participle, masculine sg of the verb *jít* 'to go' (and its prefixed derivations): *šel* 'he-gone', *šla* 'she-gone', *šlo* 'it-gone'. The rule has to insert an epenthetic *e* if the form has a Ø-ending.

2.2 Palatalization and assimilation

Palatalization or assimilation at the morpheme boundaries occurs when an ending/suffix starts with a soft vowel. The alternations are different for different types of consonants. The types of consonants and vowels are as follows:

- hard consonants—*d, (g,) h, ch, k, n, r, t*
- soft consonants—*c, č, ď, j, ň, ř, š, ť, ž*
- neutral consonants—*b, l, m, p, s, v, z*
- hard vowels—*a, á, e, é, o, ó, u, ů, y, ý* and the diphthong *ou*
- soft vowels—*ě, i, í*

The vowel *ú* cannot occur in the ending/suffix so it will not be interesting for us. I also will not discuss what happens with 'foreign' consonants *f, q, w* and *x*—they would be treated as *v, k, v* and *s*, respectively. The only borrowing from foreign languages that I included to the above lists is *g*: This sound existed in Old Slavonic but in Czech it changed into *h*. However, when later new words with *g* were adopted from other languages, this sound behaved phonologically as *h* (e.g. *hloh, hlozích*—from Common Slavonic *glog* 'hawthorn', and *katalog, katalozích* 'catalog').

The phonological alternations are reflected in writing, with one exception—if the consonants *d, n* and *t* are followed by a soft vowel, they are palatalized, but the spelling is not changed:

spelling:		phonology:	
dě, di		/*ďe*/, /*ďi*/	
ně, ni		/*ňe*/, /*ňi*/	
tě, ti		/*ťe*/, /*ťi*/	

In other cases the spelling reflects the phonology. In the further text I will use { } for the morpho-phonological level, / / for the phonological level and no brackets for the orthographical level. In the cases where the orthography and phonology are the same I will only use the orthographical level. Let us look at the possible types of alternation of consonants:

- Soft consonant and *ě* — The soft consonant is not changed, the soft *ě* is changed to *e*.
  {*číčě*} → *číče* 'pussycat' dative sg
- Soft or neutral consonant and *i/í* — No alternations occur.
  {*číči*} → *číči* 'pussycat' genitive sg

- Hard consonant and a soft vowel — The alternations differ depending on when and how the soft vowel originated.

  Assimilation:

  - {*kj*} → *č*
    tlak 'pressure' → *tlačen* 'pressed'
  - {*hj*} → *ž*
    mnoho 'much, many' → *množení* 'multiplying'
  - {*gj*} → *ž*
    It is not easy to find an example of this sort of alternation, as *g* only occurs in loan words that do not use the old types of derivation. In colloquial speech it would be perhaps possible to create the following form:
    pedagog 'teacher' → *pedagožení* 'working as a teacher'
  - {*dj*} → *z*
    sladit 'to sweeten' → *slazení* 'sweetening'
    This sort of alternation is not productive any more—in newer words palatalization applies:
    sladit 'to tune up' → *sladění* 'tuning up'
    In some cases both variants are possible, or the different variants exist in different dialects—the east (Moravian) dialects tend to keep this phonological alternation, while the west (Bohemian) dialects often abandoned it.
  - {*tje*} → *ce*
    platit 'to pay' → *placení* 'paying'
    This alternation is also not productive any more. The newest word that I found which shows this sort of phonological alternation is the word *fotit* 'to take a photo' → *focení* 'taking a photo'.

  Palatalization:
  During the historical development of the language several sorts of palatalization occured—the first and second Slavonic palatalization and further Czech palatalizations.

  - {*kě/ki*} → *če/či* (1st palat.)
    matka 'mother' → *matčin* possesive adjective
  - {*kě/ki*} → *ce/ci* (2nd palat.)
    matka → *matce* dative/locative sg
  - {*hě/hi*} → *že/ži* (1st palat.)
    Bůh 'God' → *Bože* vocative sg
  - {*hě/hi*} → *ze/zi* (2nd palat.)
    Bůh → *Bozi* nominative/vocative pl
  - {*gě/gi*} → *že/ži* (1st palat.)
    Jaga a witch from Russian tales → *Jažin* possesive adjective
  - {*gě/gi*} → *ze/zi* (2nd palat.)
    Jaga → *Jaze* dative/locative sg
  - {*dě*} → /*ďe*/ → *dě*
    rada 'council' → *radě* dative/locative sg
  - {*tě*} → /*ťe*/ → *tě*
    teta 'aunt' → *tetě* dative/locative sg

  Both palatalization and assimilation yields the same result:

  - {*ch*} → *š*
    moucha 'fly' → *mouše* dative/locative sg, *mušší* derived adjective
  - {*n*} → /*ň*/ → *n*
    hon 'chase' → *honit* 'to chase', *honěný* 'chased'
  - {*r*} → *ř*
    var 'boil' → *vařit* 'to cook', *vaření* 'cooking'

- Neutral consonant and *ě* — The alternations differ depending on when and how *ě* originated.

  Assimilation:

  - {*bje*} → *be*
    zlobit 'to irritate' → {*zlobjení*} → *zlobení* 'irritating'
  - {*mje*} → *me*
    zlomit 'to break' → {*zlomjený*} → *zlomený* 'broken'
  - {*pje*} → *pe*
    kropit 'to sprinkle' → {*kropjení*} → *kropení* 'sprinkling'

- $\{vje\} \rightarrow ve$
  lovit 'to hunt' $\rightarrow$ $\{lovjení\}$ $\rightarrow$ *lovení* 'hunting'
- $\{sje\} \rightarrow še$
  prosit 'to ask' $\rightarrow$ $\{prosjení\}$ $\rightarrow$ *prošení* 'asking'
  This type of assimilation is not productive any more. In newer derivations $\{sje\} \rightarrow se$ (e.g. *kosit* 'to mow' $\rightarrow$ *kosení* 'mowing').
- $\{zje\} \rightarrow že$
  kazit 'to spoil' $\rightarrow$ $\{kazjení\}$ $\rightarrow$ *kažení* 'spoiling'
  This type of assimilation is also not productive any more. In newer derivations $\{zje\} \rightarrow ze$ (e.g. *řetězit* 'to concatenate' $\rightarrow$ *řetězení* 'concatenating').

Palatalization:
With *b*, *m*, *p* and *v* no alternation occurs ($\{vrbě\}$ 'willow' dative/locative sg $\rightarrow$ *vrbě*).

- $\{sě\} \rightarrow se$
  vosa 'wasp' $\rightarrow$ $\{vosě\}$ $\rightarrow$ *vose* dative/locative sg
- $\{zě\} \rightarrow ze$
  koza 'goat' $\rightarrow$ $\{kozě\}$ $\rightarrow$ *koze* dative/locative sg

Both palatalization and assimilation yields the same result:

- $\{lje\} \rightarrow le$
  školit 'to school' $\rightarrow$ $\{školjení\}$ $\rightarrow$ *školení* 'schooling'
- $\{lě\} \rightarrow le$
  škola 'school' $\rightarrow$ $\{školě\}$ $\rightarrow$ *škole* dative/locative sg

- Group of hard consonants and a soft vowel. Here again either palatalization or assimilation occurs.

  Assimilation:

  - $\{stj\} \rightarrow /šť/$
    čistit 'to clean' $\rightarrow$ *čištění* 'cleaning'
  - $\{slj\} \rightarrow šl$
    myslit 'to think' $\rightarrow$ *myšlení* 'thinking'

  Palatalization:

  - $\{sk\} \rightarrow /šť/$
    kamarádský 'friendly' $\rightarrow$ *kamarádští* masculine animate, nominative pl, *kamarádštější* 'more friendly'
  - $\{ck\} \rightarrow /čť/$
    čacký 'brave' $\rightarrow$ *čačtí* masculine animate, nominative pl, *čačtější* 'braver'
  - $\{čk\} \rightarrow /čť/$
    žluťoučký 'yellowish' $\rightarrow$ *žluťoučtější* 'more yellowish', but *žluťoučcí* masculine animate, nominative pl

The alternations affect also the vowel *ě*. When it causes palatalization or assimilation of the previous consonant, it looses its 'softness', i.e. *ě* $\rightarrow$ *e*:
$\{matkě\} \rightarrow matce$
$\{sestrě\} \rightarrow sestře$
$\{školě\} \rightarrow škole$

3 Phenomena treated by two-level rules in the Czech lexicon

As the Czech lexicon should serve practical applications I did not try to solve all the problems that occur in Czech phonology. I concentrated on dealing with the alternations that occur in declension and regular conjugation, and the most productive derivations. The rest of alternations occurring in conjugation are treated by inserting several verb stems in the lexicon. The list of alternations and other changes covered by the rules:

- epenthesis
- palatalization in declension
- palatalization in conjugation
- palatalization in derivation of feminine nouns from masculines
- palatalization in derivation of possessive adjectives
- palatalization in derivation of adverbs
- palatalization in derivation of comparatives of adjectives and adverbs

- palatalization or assimilation in derivation of passive participles
- shortening of the vowel in suffixes *-ík* (in derivation of feminine noun from masculine) and *-ův* (in declension of possesive adjectives)

For the Czech lexicon I used the software tools for two-level morphology developed at Xerox (Karttunen and Beesley, 1992; Karttunen, 1993). The lexical forms are created by attaching the proper ending/suffix to the base form in a separate program. To help the two-level rules to find where they should operate, I also marked morpheme boundaries by special markers. These markers have two further functions:

- They bear the information about the length of ending (or suffix and ending) of the base form, i.e. how many characters should be removed before attaching the ending.
- They bear the information about the kind of alternation.

Beside the markers for morpheme boundaries I also use markers for an epenthetic *e*. As I said before, *e* is inserted before the last consonat of a final consonant group, if the last consonant is a suffix, or if the consonant group is not phonetically admissable. However, as I do not generally deal with derivation nor with the phonetics, I am not able to recognize what is a suffix and what is phonetically admissable. That is why I need these special markers.

Another auxiliary marker is used for marking the suffix *-ík*, that needs a special treatment in derivation of feminine nouns and their possesive adjectives. The long vowel *í* must be shortened in the derivation, and the final *k* must be palatalized even if the Ø-ending follows. I need a special marker, as *-ík-* allows two realizations for both the sounds in same contexts:
Two realizations of *í*
úředník 'clerk' → *úřednice* 'she-clerk', but *rybník* 'pond' → *rybníce* locative sg
Two realizations of *k*
úředník × *úřednic* (genitive pl of the derived feminine)

In the previous section, I described all possible alternations concerning single consonants. When I work with the paradigms or with the derivations, it is necessary to specify the kind of the alternation for all consonants that can occur at the boundary. For this purpose I introduced four types of markers:

^1P — 1st palatalization for *g*, *h* and *k*, or the only possible (or no) palatalization for other consonants. I use this marker also for palatalization *c* → *č* in vocative sg of the paradigm *chlapec*. The final *c* is in fact a palatalized *k*, so there is even a linguistic motivation for this.

^2P — 2nd palatalization for *g*, *h* and *k*, or the only possible (or no) palatalization for other consonants.

^A — Assimilation (or nothing).

^N — No alternation.

These markers are followed by a number that denotes how many characters of the base form should be removed before attaching the ending/suffix. Thus there are markers ^1P0, ^2P0, ^1P1, etc. The markers starting with ^N only denote the length of the ending of the base form—and instead of using ^N0 I attach the suffix/ending directly to the base form. Fortunately, nearly all paradigms and derivations cause at most one type of alternation, so it is possible to use one marker for the whole paradigm.

The markers for an epenthetic *e* are ^E1 (for *e* that should be deleted) and ^E2 (for *e* that should be inserted). The marker for the suffix *-ík* in derivations is ^IK.

Here are some examples of lexical items and the rules that transduce them to the surface form:

(1) `doktorka^1P1in^2P0ých`
The base form is *doktorka* 'she-doctor'. The marker ^1P1 denotes that the possible alternation at this morpheme boundary is (first) palatalization and that the length of the ending of the base form is 1 (it means that *a* must

be removed from the word form and the possible alternation concerns *k*). The marker ^2P0 means that the derived possessive adjective has a Ø-ending and the possible alternation at this morpheme boundary is palatalization. If we rewrite this string to a sequence of morphemes we get the following string: *doktork-in-ých.* The sound *k* in front of *i* is palatalized, so the correct final form is *doktorčiných*, which is genitive plural of the possessive adjective derived from the word *doktorka.*

Let us look now at the two-level rules that transduce the lexical string to the surface string. We need four rules in this example: two for deleting the markers, one for deleting the ending *-a*, and one for palatalization. The rules for deleting auxiliary markers are very simple, as these markers should be deleted in any context. The rules can be included in the definition of the alphabet of symbols:

```
Alphabet
%^1P0:0 %^1P1:0
%^2P0:0 %^2P1:0 %^2P2:0 %^2P3:0
%^A2:0
%^N1:0 %^N2:0 %^N3:0 %^N4:0
%^E1:0 %^E2:0 %^IK:0
```

This notation means that the auxiliary markers are always realized as zeros on the surface level.

The rule for deleting the ending *-a* looks as follows:

```
"Deletion of the ending -a-"
a:0 <=> _ [ %^N1: | %^1P1: | %^2P1: ] ;
        _ t: [ %^N2: | %^N4: ];
```

The first line of the rule describes the context of a one-letter nominal ending *a*, and the second line describes the context of an infinitive suffix with ending *-at* or *-ovat.*

The rule for palatalization *k* → *č* looks as follows:

```
"First palatalization k -> č"
k:č <=> _ (%^IK:) [ a: | ě: ] %^1P1: i ;
        NonČČS: _ (End) %^1P1: ě: ;
```

The first line describes two possible cases: either the derivation of a possesive adjective from a feminine noun (*doktorka* → *doktorčin*), or the derivation of a possesive adjective from a feminine noun derived from a masculine that ends with *-ík* (*úředník* → (*úřednice* →) *úředničin*).

The second context describes a comparative of an adjective, or a comparative of adverb derived from that adjective (*hořký* → *hořčejší*, *hořčeji*). The set NonČČS contains all character except *c*, *č* and *s* and it is defined in a special section. This context condition is introduced, because the groups of consonants *ck*, *čk* and *sk* have different 1st palatalization.

The label End denotes any character that can occur in an ending and that is removed from the base form.

(2) korek^2P0^E1em
The base form of this word form is *korek* 'cork'; the marker ^2P0 means that the possible alternation is (second) palatalization and that the length of ending of the base form is 0. The marker ^E1 means that the base form contains an epenthetic *e*, and *em* is the ending of instrumental singular. The correct final form is *korkem.* The rule for deleting an (epenthetic) *e* follows:

```
"Deletion of e"
e:0 <=> Cons _ c: %^N2:;
    _ [ %^1P1: | %^2P1: | %^N1: | %^N2: ];
    Cons _ Cons: ([%^1P0:|%^2P0:]) %^E1: Vowel:;
    _ t:0 [ %^2P2: | %^N2: ];
```

The first line describes the context for deletion of the suffix *-ec* in the derivation of the type *vědec* 'scientist' → *vědkyně* 'she-scientist'.

The second context is the context of the ending *-e* or the suffix *-ce.* This suffix must be removed in the derivation of the type *soudce* 'judge' → *soudkyně* 'she-judge'.

The third context is the context of an epenthetic *e* that is present in the base form and must be removed from a form with a non-Ø ending. The sets Cons and Vowel contain all consonants and all vowels, respectively.

The fourth line describes the context for deletion of the infinitive ending *-et.*

The whole program contains 35 rules. Some of the rules concern rather morphology than phonology; namely the rules that remove endings or suffixes. One rule is purely technical; it is one of the two rules for the alternation *ch* → *š*, as *c* and *h* must be treated separately

(though *ch* is considered one letter in Czech alphabet). Six rules are forced by the Czech spelling rules (e.g. rules for treating /ď/, /ť/ and /ň/ in various contexts, or a rule for rewriting $y \rightarrow i$ after soft consonants). 18 rules deal with the actual phonological alternations and they cover the whole productive phonological system of Czech language. The lexicon using these rules was tested on a newspaper text containing 2,978,320 word forms, with the result of more than 96% analyzed forms.

4 Acknowledgements

My thanks to Ken Beesley, who taught me how to work with the Xerox tools, and to my father, Jan Skoumal, for fruitful discussions on the draft of this paper.

References

Jan Hajič. 1994. *Unification Morphology Grammar*, Ph.D. dissertation, Faculty of Mathematics and Physics, Charles University, Prague.

Josef Holub, and Stanislav Lyer. 1978. *Stručný etymologický slovník jazyka českého (Concise etymological dictionary of Czech language)*, SPN, Prague.

Lauri Karttunen, and Kenneth R. Beesley. 1992. *Two-Level Rule Compiler*, Xerox Palo Alto Research Center, Palo Alto.

Lauri Karttunen. 1993. *Finite-State Lexicon Compiler*, Xerox Palo Alto Research Center, Palo Alto.

Kimmo Koskenniemi. 1983. *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*, Publication No. 11, University of Helsinki.

Arnošt Lamprecht, Dušan Šlosar, and Jaroslav Bauer. 1986. *Historická mluvnice češtiny (Historical Grammar of Czech)*, SPN, Prague.

Jan Petr et al. 1986. *Mluvnice češtiny (Grammar of Czech)*, Academia, Prague.

Jana Weisheitelová, Květa Králíková, and Petr Sgall. 1982. Morphemic Analysis of Czech. No. VII in *Explizite Beschreibung der Sprache und automatische Textbearbeitung*, Faculty of Mathematics and Physics, Charles University, Prague.