Derivational Relations in Flectional Languages - Czech Case

Jaroslava Hlaváčová

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, 118 00 Praha 1, Czech republic
hlava@ufal.mff.cuni.cz

Jana Klímová

Institute of Czech Language, Czech Academy of Sciences
Letenská 4, 118 51 Praha 1, Czech republic
klimova@ujc.cas.cz

Abstract

When a text in any language is submitted to a morphological analysis, there always rest some unrecognized words. We can lower their number by adding new words into the dictionary used by the morphological analyzer but we can never gather the whole of the language. The system described in this paper (we call it "derivation module") deals with the unknown derived words. It aims not only at analyzing but also at synthesizing Czech derived words. Such a system is of particular value for automatic processing of languages where derivational morphology plays an important role in regular word formation.

Introduction

In the Czech National Corpus of 100 million word forms there is 2.3% of unrecognized words. These words can be classified into 3 main categories:

1. misspelled words and other rubbish (e.g. typos or unusual abbreviations),
2. proper names not included in the dictionary (geographical or personal names not commonly used throughout the language),
3. words derived from the words present in the dictionary (e.g. *brzditel* derived from the verb *brzdit = to brake*, could be translated as *braker* - someone who brakes).

Great majority of unknown words (68.1%) belongs to the category 2. The other two categories represent roughly 730 000 word forms. The way of treating unrecognized words was described in another paper (Hlaváčová, 2001). The method was strictly automatic and was based on observing ends of the words that were recognized properly by a morphological analyzer. There are some heuristics that can be added for better performance. However, there is a better possibility that takes into account deeper linguistic knowledge of the language.

In Czech, the most common way of derivation is adding affixes to a base word. While the set of possible suffixes is limited and very stable, new prefixes can be created very easily and are changing quickly. Although there is a great difference between using prefixes and suffixes for creating new words, we take the lists of affixes - both prefixes and suffixes - as constant for our purpose.

Formal Description

For the description of the derivation module we introduce following formal notation.

Every word w consists of individual letters $a_1, a_2,...,a_n$;

$$w = a_1a_2...a_n.$$

Let **V** be the morphological dictionary - the list of word forms. All the word forms from **V** are successfully morphologically analyzed, it means that the morphological analyzer M assigns them a couple $[L,T]$, where L is lemma and T morphological tag (see Hajič, 2001):

$$M: w \rightarrow [L,T].$$

If we say that a word w is unrecognized, it means that w does not belong to **V**, in other words:

$$M: w \rightarrow \varnothing.$$

Let **P** be list of prefixes, **S** list of suffixes.

We consider the term suffix in a broader sense than grammarians usually do. In our approach, the suffix consists of any possible combination of real suffix and ending. Let us call these strings flectional suffixes (FS). For instance in the word *podobnosti,* there the suffix is *ost* and ending *i* but the entry in the list **S** - the flectional suffix is *osti*. FS gives us grammatical information of the analyzed word, in our example the suffix *ost* represents a noun, its gender is feminine, the flectional suffix *osti* informs that the word form is in singular genitive or singular dative or in plural nominative or plural accusative.

We want the analyzed word to be assigned with lemma and tag and to find its base lemma B - the word from which w was derived. We are looking for a mapping (derivation mapping) D that assigns a triplet to our word:

$$D: w \rightarrow [L, T, B],$$

where L is lemma of w, T its tag and B its base lemma. We say that a word w is derived from the dictionary **V**, if the lemma B, from which w was derived, belongs to **V**.

Prefix derivation

The word w was derived from a base lemma B by means of a prefix, if there exists $i \in \langle 1, Mp \rangle$ such that $a_1...a_i$ belongs to **P** and $a_{i+1}...a_n$ belongs to **V**.

If M: $a_{i+1}...a_n \rightarrow [L_i T_i]$, then

$$L=a_1...a_iL_i,\ T=T_i \text{ and } B=L_i.$$

Mp is maximal length for which it is plausible to try to tear off the letters from the beginning of the word w. We assume that a stem is at least 3 characters long. That's why we set

Mp = min (length of the longest prefix, *n-3*).

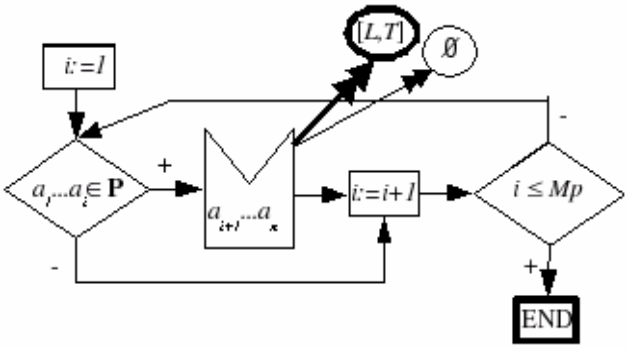

Figure 1: Algorithm of prefix derivation

Figure 1 shows the algorithm schema of the prefix recognition. The strange shape resembling upper case letter M represents morphological analyzer. Double arrows show its possible outputs - either a double [*L, T*] (thick one) or an empty set (thin one). From the non-empty output we derive the triplet [a_1...a_i*L,T,L*], where the first member means the lemma of the derived word, *L* is the base lemma and the tag is the same both for the base and derived lemma. If the output is empty, we still have to remember the prefix for the further analysis (see later). We process through the schema until we tear off the longest possible prefix from the set **P** or until the rest of the word is longer than 3, whichever case is reached first. Generally, it is possible to have more outputs. All of them are the result of the algorithm.

Suffix derivation

For every suffix there is a set of derivation rules of three types:

RS - solving the separation of suffixes,
RA - for the letter alternations in stems,
RE - for the addition of an appropriate flectional suffix creating the base lemma.

Generally, a sequence of derivation rules is applied so that the analyzed word *w* is assigned with its relevant derivation mapping D.

The word *w* was derived from a base lemma *B* by means of a suffix, if there exists $j \in \langle n\text{-}Ms, n \rangle$ such that $a_j...a_n$ belongs to **S** and there is a sequence of derivation rules **RS, RA, RE** that creates the base lemma *B*. Again, similarly as for prefix derivation, we set

Ms = min (maximal length of suffixes, *n-3*)

because we assume that the length of the stem is at least 3 characters.

The application of derivation rules is then as follows:

1. **RS** cuts off a flectional suffix and converts it to a basic form, i.e. singular nominative for nouns.

Example:

w = *normiček*
iček ∈ **S**
RS: *iček* (masc. sg. nom.) --> *iček*
iček (fem. pl. gen.) --> *ička*
iček (neutr. pl. gen.) --> *ičko*

2. **RA** checks if there is, for the given flectional suffix, a letter alternation in the stem. If there is, **RA** makes appropriate changes in the word stem, if not, the stem remains unchanged. There is always one identical rule **RA** not changing the stem.

Example:

RA: *norm* --> *nůrm* (prolongation of the stem vowel)
norm --> *norm* (no change)

3. **RE** adds a new suffix to all the potential (possibly changed) stems forming base lemmas. Now, we have to check if the base lemmas really exist, in other words, if *B* ∈ **V**. In that case we have solved the unknown word. If not, the unknown word rests unknown and it is necessary to find another method how to recognize it (in the worst case - manually).

Example:

RE: *iček* --> *ek* (*normek* ∉ **V**, *nůrmek* ∉ **V**)
ička --> *a* (***norma*** ∈ **V**, *nůrma* ∉ **V**)
ičko --> *o* (*normo* ∉ **V**, *nůrmo* ∉ **V**)

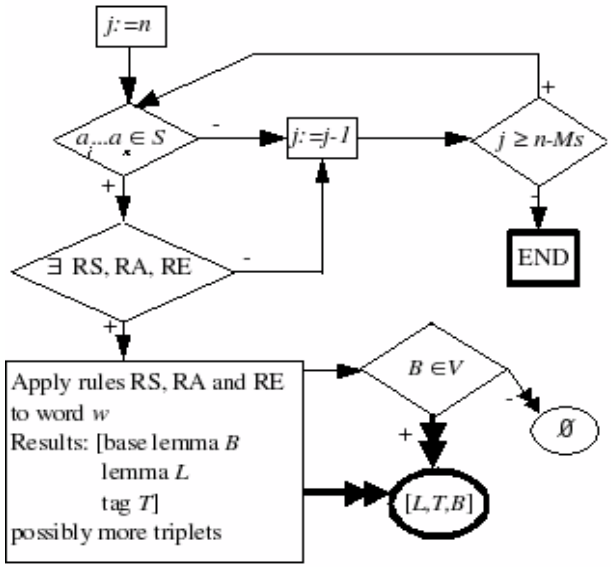

Figure 2: Algorithm of suffix derivation

In Figure 2, the algorithm proceeds until we tear off the longest possible suffix or until the rest of the word is longer than 3. Outputs of the second algorithm are triplets [lemma, tag, base lemma] which are accepted only when the base lemma *B* is included in the vocabulary **V** of word forms. Again, there can be more than one output.

Analysis of unrecognized derived words

The analysis of an unknown word consists of 2 steps:

1. prefix analysis,
2. suffix analysis.

If the prefix analysis outputs an empty set for a prefix, we tear the prefix off and past the rest to the suffix analysis. Therefore, in those cases the suffix analysis is processed twice - with the whole unknown word and with its rest after tearing a possible prefix off.

Example: unknown word *w* = *euronormiček* (lit. *small euronorm*, gen. pl.)

The prefix analysis outputs one empty set for the prefix *euro* because the rest of the word *normiček* is also

unknown. The suffix analysis with the whole word *euronormiček* does not yield any output but the second suffix analysis, with the rest *normiček*, outputs a meaningful base lemma *norma*. In this case the unknown word was derived from the base word *norma* using prefix derivation with the prefix *euro* and the suffix derivation with the diminutive suffix *ička*. It is not possible to decide which of the derivation came first.
The result of the analysis is then as follows:
w = *euronormiček*
flectional suffix = *iček*
T = noun, feminine, plural, genitive
L = *euronormička*
B = *norma*
Other complications are multiple derivations - derived words on the basis of words already derived. The algorithm that was presented above should be slightly modified in order to allow this possibility. As the recursion cannot be too long, the algorithm always ends in a reasonable time. An example of a complicated derived word to analyze is *ne-od-děl-itel-ného*. The stem is *děl*, prefixes are *od* and *ne* (prefix of negation), suffixes are *itel* (creating nouns of action from verbs) and *ný* (creating adjectives from nouns) and ending *ého* (expressing an adjective masculine or neuter singular genitive). The word was derived in steps:
děl (stem) --> *dělit* (verb) --> *dělitel* (noun of action) --> *dělitelný* (adjective) --> *oddělitelný* (adjective with the prefix *od*) --> *neoddělitelný* (adjective with the prefix *ne*)
Results of analysis:
flectional suffix = *ného*
B = *dělitel*
L = *neoddělitelný*

Examples

The following examples are from the category of diminutives. All have the same flectional suffix *ček* (preceded with an epenthetic vowel) that represents different genders.

- *w* = *foťáček* (literally *a small camera*)

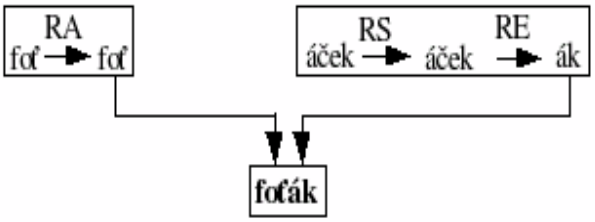

Figure 3: Analysis of the word *foťáček*

Result of the analysis:
stem: *foť*
flectional suffix: *áček*
T = noun masculine singular nominative
B = *foťák*
L = *foťáček*

- *w* = *plechovčiček* (literally *a small can*)

Result of the analysis:
T= noun feminine plural genitive
stem: *plechovč*
flectional suffix: *iček*
B = *plechovka*
L = *plechovčička*

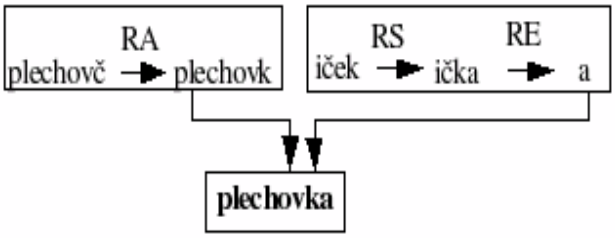

Figure 4: Analysis of the word *plechovčička*

- *w* = *políneček* (lit. *a very small log*)

stem: *polín*
flectional suffix: *eček*
T = noun neuter plural genitive
B = *poleno*
L = *polínečko*

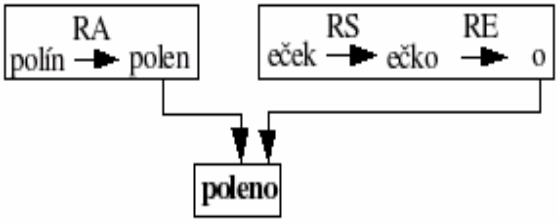

Figure 5: Analysis of the word *políneček*

Conclusion

The derivation module is an opened system, it means that for every suffix a special module can be added. This module is able to analyse words derived with that suffix by applying derivation rules. The derivation module is also able to generate new words according to the derivation rules. Words in the vocabulary **V** are assigned with derivation patterns that define the derivation rules for given suffixes (i.e. the POS of the base and derived lemma and the sets of RS, RA and RE rules). It is necessary to offer a root and type of derivation we like to apply. However, such generation can create a word that does exist only potentially, even if it was generated according to the valid rules for the affixes.
The derivation module works for the Czech language, but as the principles of word formation are similar for some other flectional languages, it can be easily modified for them.

Acknoledgements

The research was supported by the Grant Agency of the Czech Republic No. 405/03/0913

References

Daneš F., Dokulil M., Kuchař J. (1967): Tvoření slov v češtině 2 (Odvozování podstatných jmen), Academia Praha

Hajič, J. (2001): Disambiguation of Rich Inflection (Computational Morphology of Czech). The Karolinum Press.

Hlaváčová, J.(2001): Morphological Guesser of Czech Words. In Proceedings TSD 2001 (pp. 70 – 75). Springer-Verlag Berlin Heidelberg.

Klímová J., Pala K. (2000): Application of WordNet ILR in Czech Word-formation, In Proceedings of the Second International Conference on Language Resources and Evaluation (pp. 987-992). Athens.

Klímová J., Kocek J. (2000): Derivation in the Czech National Corpus, In Proceedings of the Second International Conference on Language Resources and Evaluation (pp. 1463-1468). Athens.

Klímová J. (2001): Počítačové zpracování vybraných slovotvorných typů v češtině, PhD. dissertation, Faculty of Mathematics and Physics, Charles University Prague.