Behaviour of the Czech Suffix *-ák* – A Case Study

Dana Hlaváčková, Karel Pala

NLP Centre
Faculty of Informatics, Masaryk University
Brno, Czech Republic
{hlavack,pala}@fi.muni.cz

Abstract. New techniques in Czech derivational morphology are discussed. They are based on the exploitation of the tool Deriv with integrated access to the main Czech dictionaries and corpora (SYN2000c and the new large Czech corpus CzTenTen12). The case study deals especially with the Czech suffix *-ák* – we describe its behaviour as completely as possible. The paper brings some new results in comparison with standard Czech grammars which do not rely on large data and software tools.

Key words: Czech morphology, Czech suffixes

1 Introduction

In the paper we report on an exploration of the derivational behaviour of the selected noun Czech suffixes *-ák, -ec, -ík,-ník*. We have used the tool Deriv (see `http://deb.fi.muni.cz/deriv`), which allows us to have a look at possibly all Czech nouns with the selected suffixes – here we will pay attention especially to the nouns with the suffix *-ák* that can be derived by the standard derivational processes. As far as we can say the previous exploration of the mentioned suffixes has been limited in number – because the authors of the standard Czech grammars ([1,2], further MČ and PMČ) did not have large corpora and machine dictionaries at their time, thus they were able to handle only a small number of examples.This inevitably meant that their results had to be only partial. The same can be said with regard to the Dokulil's [3] important work in which he laid the theoretical foundation of Czech word derivation but he did have sufficient data at his disposal.

2 Motivation

Our main purpose is to investigate the derivational relations in a more detailed and deeper way using larger and more complete Czech data. We are seeking to reach better coverage of the studied phenomena together with better precision. We start with the results that have been obtained for Czech formal morphology, namely with the morphological dictionary that is a part of Czech morphological analyzer Ajka [4]. Then we continue investigating derivational relations and

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 9–13, 2012.

behaviour of the selected suffixes in particular. As we use computer tools we need to make our description as formal as possible. We are convinced that better knowledge of derivational behaviour of the individual suffixes in Czech (and prefixes as well) is a necessary prerequisite for the development of the more inteligent searching tools that can be used in various applications, e.g. in searching engines for Web.

3 Computer processing of Czech word derivation

Obtaining better knowledge about the derivational relations in Czech requires larger data than was used so far in the mentioned standard Czech grammars [1,2]. Such data cannot be, however, reasonably processed manually, that would be too time consuming and would contain errors. At the moment we have at our disposal a large machine dictionary of Czech stems (approx. 400 000 items) whose coverage of Czech is about 95 %. To explore the behaviour of suffixes the tool Deriv has been developed, which makes it possible to formally describe the morphematic segmentation on the derivational level. This is handled with looking up the possible combinations of stems and suffixes (prefixes as well) also using regular expressions. In this way we obtain the required lists of nouns containing the particular suffixes as they occur in Czech. The tool Deriv [5] is integrated with two Czech corpora (SYN2000c [6] and CzTenTen12) as well as with the dictionary browser DEBDict, so the user (linguist) can see the behaviour of an explored suffix as completely as possible, namely its frequencies.

4 Starting data – morphological dictionary, corpora, ...

As we have said we work with the large machine dictionary of Czech stems (approx. 400 000 items) whose coverage of Czech is about 95 % (for comparison, the size of the SSJČ is twice smaller). It is a part of the Czech morphological analyzer Ajka [4] and the Deriv tool works with the lists of Czech stems generated appropriately for this purpose. The coverage 95 % means that the analyzer Majka is able to process any Czech text and recognize the word forms in it. Those 5 % are expressions consisting of numbers (dates, telephone numbers, etc.), e-mail addresses, URL's, words from other languages than Czech (most frequently Slovak and English) and others. Since the tool Deriv is linked with the two Czech corpora and six Czech dictionaries in Debdict the obtained results can be compared with them, especially with regard to the frequencies. The comparison of numbers obtained from the corpus SYN2000c with 114,363,813 and corpus CzTenTen12 with 5,414,437,666 tokens (see `http://ske.fi.muni.cz/auth/corpora/`) is then quite expressive. Thanks to links of the Deriv to the Word Sketch Engine also collocational data can be obtained. The data also contain the respective lists of Czech suffixes and prefixes.

5 Case study – behaviour of the Czech suffix *-ák*

The obtained list contains the derived nouns with the suffixem *-ák* which can be characterized in the following way:

- expressive and slang nouns, also obsolete ones
- nouns productive in deriving one word expressions

The number of the Czech nouns ending with the string *-ák* is 1351, from them there are 724 lemmata in masculine animate and 627 lemmata in masculine inanimate (they include also substandard lemmata not occurring in SSJČ and proper (family) names.

6 Derivational categories and their classification

For the nouns in the lists we propose the classification comprising the following categories with reference to the classifications that can be found in MČ and PMČ.

- Nouns derived from nouns:
  - agentive nouns - *dudák* (bagpiper), *koňák* (horseman), *sedlák* (farmer), *tramvaják* (tram driver)
  - nouns denoting inhabitants - *Brňák* (inhabitant of Brno), *Hanák* (inhabitant of Haná), *Malostraňák* (dweller of the Prague quarter Malá Strana), *Pražák* (inhabitant of Prague
  - nouns expressing membership in a group of people - *devětsilák* (member of the group Devětsil), *esenbák* (cop), *tatrovák* (worker in the Tatra factory)
  - nouns denoting animals (derived from feminines) - *lišák* (male fox), *myšák* (male mouse), *opičák* (male monkey)
  - augmentatives - *hnusák* (scumbag), *sviňák* (goat), *úchylák* (pervert)
- Nouns derived from adjectives:
  - nouns denoting bearers of properties - *blonďák* (blond man), *dobrák* (brick), *chudák* (poor fellow), *silák* (strongman)
- Nouns derived from numerals
  - nouns denoting order - *prvák* (first-year student), *druhák* (second-year student
  - nouns denoting roe deers and deers with regard to their antlers - *desaterák* (ten pointer), *dvanácterák* (royal stag)
- Nouns derived from verbs
  - agentive nouns - *divák* (viewer), *honák* (cowboy), *pašerák* (smuggler), *zpěvák* (singer
  - nouns denoting instruments - *bodák* (bayonet), *drapák* (grab), *hasák* (pipe-wrench), *naviják* (reel)

It is necessary to remark that number of the nouns in Table 1 is smaller than the numbers given above. This is due to the fact that we have put aside the nouns for which the reasonable derivational relation between the basic and derived form cannot be established or it is very difficult.

Table 1. The coverage of the nouns belonging to the individual categories

Category	Masculine animate	Masculine inanimate
agentive	98	-
inhabitants	78	-
groups	62	-
animals	16	-
augmentatives	47	-
bearers of property	128	
order	5	-
antlers	7	-
agentive1	105	-
instruments		-
Total	**546**	

7 Results and Conclusions

In the paper we paid attention to some selected Czech noun suffixes for which we describe their derivational behaviour. It has to be stressed that we have concentrated on just one suffix, namely *-ák*, which serves as a pattern that can be applied to other mentioned suffixes as well. The concrete results (numbers) are, however, brought only for the *-ák*. On the other hand, it is obvious that this kind of description can be applied to all mentioned suffix.

The main result is Table 1 which shows what meanings *-ák* have, the basic analysis standing behind Table 1 has been performed manually. The classification offered in Table 1 in fact removes ambiguity of *-ák*, which a human user resolves easily but computer applications cannot work without it.

We would like to stress that more can be said about the behaviour of *-ák* and suffixes of the same type – the paper is one of the first exercises in this respect. It also has to be noted that the use of the tool Deriv and large data made it possible to offer the results which display a reasonable coverage.

Acknowledgments

This work has been partly supported by the Ministry of Education of the Czech Republic project No. LM2010013 (Lindat–Clarin – Centre for Language Research Infrastructure in the Czech Republic), and by the Czech Science Foundation under the project P401/10/0792.

References

1. Komárek, M.: Mluvnice češtiny I (Grammar of Czech I). Academia, Praha (1986)
2. Karlík, P., Grepl, M., Nekula, M., Rusínová, Z.: Příruční mluvnice češtiny. Lidové noviny (1995)

3. Dokulil, M.: Tvoření slov v češtině I (Word Derivation in Czech I). Nakladatelství ČSAV, Praha (1962)
4. Šmerk, P.: K počítačové morfologické analỳze češtiny. (2010)
5. Šmerk, P.: Deriv. (2009) Web application interface (in Czech), accessible at: `http://deb.fi.muni.cz/deriv`.
6. ICNC: Czech National Corpus – SYN2000. Institute of the Czech National Corpus, Praha (2000) Accessible at: `http://www.korpus.cz`.