SBORNÍK PRACÍ FILOZOFICKÉ FAKULTY BRNENSKÉ UNIVERZITY
STUDIA MINORA FACULTATIS PHILOSOPHICAE UNIVERSITATIS BRUNENSIS
A 41, 1993

KAREL PALA, KLÁRA OSOLSOBĚ

# CZECH STEM DICTIONARY

## 1. INTRODUCTION

In our contribution we would like to present the results of the research performed at the Dept. of Czech Language, Philosophical Faculty of Masaryk University, Brno in the course of the last three years (Pala, Osolsobě, Franc, 1987, Halasová-Osolsobě, 1989-90, Osolsobě, Pala, 1990).

We have set the following tasks:

1. to prepare as complete a machine dictionary of Czech stems as possible (at present containing about 175 000 entries);

2. to work out an algorithmic description of Czech morphology based on a complete list of pattern (model) words now containing 510 pattern words (Osolsobě, 1991);

3. to prepare a complete machine glossary of Czech words containing 190 000 entries and based on the glossary of the representative Czech dictionary *(Dictionary of Literary Czech Language)* published in1960 and reprinted in 1989;

4. to build a collection of computer programs for the automatic morphological parsing and generating Czech word forms containing also a Czech spelling checker and other programs for creating, updating and mantaining machine dictionaries on personal computers (Pala, Osolsobě, Franc, 1987, Osolsobě, Pala, 1990).

## 2. CZECH MORPHOLOGY AND DICTIONARIES

As our first step we have decided to develop a dictionary of Czech word stems.

The basic structure of the Czech stem dictionary can be divided into three parts:

1. those words that cannot be inflected, i. e. adverbs, prepositions, conjunctions, particles and interjections;

2. forms of words with irregular inflection, i. e. personal, demonstrative, negative and indefinite pronouns and also a few irregular verbs, as eg. *být (to be), mít (to have), chtít (to want), jít (to go), ...;*

3. a dictionary of stems of regularly inflected parts of speech, i. e. nouns, adjectives, numerals and all regular verbs.

Uninflected words and irregularly inflected words represent the part of our dictionary that can be called the „fixed dictionary". It is a full list of the noninflected and irregularly inflected word forms that are found in Czech, together with the information about their part of speech and respective grammatical categories associated to them.

As a result we now have a dictionary of stems having the following structure:

1. on each line there is an entry consisting of a basic noun, adjective or verb form segmented as a stem and an ending, particularly,
   – as a stem without a final consonant, final consonant and ending,
   – as a root, word forming suffix and ending,
   – as a stem, word forming suffix and ending,
   – as a stem, stem forming suffix and ending.

From a formal point of view a stem can be defined as a string of characters (lower case letters) ending in a space or nonalphabetic character (eg. by hyphen „-"), followed either by an ending or a final consonant, word forming suffix, stem forming suffix and ending or followed by a space.

Examples:

*pán-O (lord)*
*žen-a (woman)*
*vl-k-O (wolf)*
*mat-k-a (mother)*
*sídl-išt-ě (settlement, housing estate)*
*pěkn-ý (nice)*
*děl-a-t (to do)*

2. a symbol denoting which part of speech the basic form belongs to. With nouns we also associate the respective four genders (see also Sgall, 1960), eg.
   @1Z - animate masculine
   @1W - inanimate masculine
   @1F - feminine
   @1N - neuter.

Further we distinguish hard adjectives – @2T,
and soft adjectives – @2M.

Verbs as a whole are denoted as @5A.

For all uninflected parts of speech we use the symbol @NO and they have unified inflection pattern denoted as >dummy.

3. An inflection pattern (see Sect. 3) is introduced by the character >, e.g.

| | | |
|---|---|---|
| *pán* | *@1Z* | *>pán* |
| *klá-r-a* | *@1F* | *>sá* |
| *sídl-išt-ě* | *@1N* | *>sídl* |
| *pěkn-ý* | *@2T* | *>nov* |
| *děl-a-t* | *@5A* | *>děl* |

The following components of the entry are not obligatory:

4. the character ~ expresses that the word contained in the entry can have a negative form obtained by adding the prefix *ne–*, *(not)*,

5. the character " is used with adjectives to express the fact they may form the superlative with the prefix *nej–*, *(most)*,

6. the character ^ shows that the word contained in the entry starts in uppercase (e.g., proper noun),

7. all uninflected words contain information about the word category they belong to in the form „/kat:V, /kat:Y, /kat:X, /kat:W, /kat:Z", eg.,

| | | |
|---|---|---|
| *pán* | *@1Z* | *>pán* |
| *klá-r-a* | *@1F* | *>sá ^* |
| *sídl-išt-ě* | *@1N* | *>sídl* |
| *pěkn-ý* | *@2T* | *>nov ~"* |
| *děl-a-t* | *@5A* | *>děl ~* |
| *tam* | *@NO* | *>dummy /kat:V* |

8. an entry also can contain „switches", i.e. strings of characters beginning with / and followed by a string of lowercase letters of the Czech alphabet. A switch defines the word forming structure of a word or a word deriving suffix. Thus a dictionary where entries contain these switches yields a classification of the entries according to word formation classes. In this way the switches represent a base for more detailed classification of the words according to the respective word formation types. An example:

*anarchist-k-a @1F >mat /istkaa (type a))*

| | | |
|---|---|---|
| *koč-k-a* | *@1F* | *>mat /kaa (type a))* |
| *pojist-k-a* | *@1F* | *>aljaš /kab (type b))* |
| *učitel-k-a* | *@1F* | *>mat /telkaa (type a))* |

## 3. THE LIST OF PATTERNS AND ITS STRUCTURE

The list of patterns displays a two level structure:

1. definitions of the sets of endings,
2. definitions of the patterns.

### 3.1. DEFINITIONS OF THE SETS OF ENDINGS

A set of endings is a collection of endings, and there are 166 of them in our system, that for all words belonging to one inflection type expresses the same grammatical meaning. At the same time the requirement has to be fulfilled that all the endings belonging to one set combine with just one variant of the stem. This, of course, means that the endings that combine with the forms in which alternations of stem occur belong to a special set.

An example:

The endings connected with the typical pattern *pán* are:

*-O, -a, -u, -ovi, -em, -i, -ové, -é, -ů, -ům, -y, -ech, -ích, -e.*
The can be divided into the following subsets:

```
    =V1
[Qsm]
        (a,2)
        (u,3)
        (ovi,3)
        (a,4)
        (u,6)
        (ovi,6)
        (em,7)
[Qpm]
        (ů,2)
        (ům,3)
        (y,4)
        (y,7)
```

```
   =VI
[Qpm]
        (i,1)
        (i,5)

   =VOVE
[Qpm]
        (ové,1)
        (ové,5)

   =VE
[Qpm]
        (é,1)
        (é,5)

   =V13X
[Qsm]
        ( _,1)

   =VVE
[Qsy]
        (e,5)

   =VVU
[Qsy]
        (u,5)
```

The definition of the set of endings has a fixed format. It begins with a character „=", followed by the name of the defined set consisting of uppercase letters and digits. On the folowing line (in the third column) there is a string of characters in square brackets denoting the part of speech, number, and gender for words that arise as a combination of the stems of a certain type and the ending(s) of the defined set. The next line(s) contains the definition of single endings belonging to the set. In the brackets one can gradualy find the respective endings separated by a comma „," from a digit (1, 2, 3, 4, 5, 6, 7) denoting case or person. The following symbols are used:

| | | | |
|---|---|---|---|
| Q – | noun | V – | adverb |
| R – | adjective | W – | preposition |
| S – | pronoun | X – | conjunction |
| T – | numeral | Y – | particle |
| U – | verb | Z – | interjection |
| s – | singular | p – | plural |

1....7 case

| | |
|---|---|
| m | gender, animate masculine |
| w | gender, inanimate masculine |
| f | genden, feminine |
| n | genden, neuter |
| z | m + w (animate masculine + inanimate masculine) |
| y | m + w + n (animate masculine + inanimate masculine + neuter) |
| q | w + f (inanimate masculine + feminine) |
| u | w + n (inanimate masculine + neuter) |
| x | m + w + f + n (animate masculine + inanimate masculine + feminine + neuter) |
| a | positive of adverbs |
| b | comparative of adverbs |

| | |
|---|---|
| A 1st person | C 3rd person |
| B 2nd person | G infinitive |

The above mentioned combinations of genders and other categories are quite freguent, therefere it is very useful to have the presented abbreviations for them.

## 3.2. PATTERN DEFINITIONS

The number of patterns is quite large if cmpared with standard grammars of Czech (Havránek, Jedlička, 1981, Petr a kol, 1986). There are 510 patterns in our list, and one half are noun patterns, one third – verb patterns, one tenth – adjective patterns, and the rest – numeral patterns (Osolsobě, 1991). The reason for having this number of patterns is that we want to capture all the doublets and exceptions in Czech declension and conjugation.

A pattern is defined as a pattern word and sets of endings – their combinations capture all the word forms associated with a given pattern. Pattern words can be segmented in several ways:

- stem + <> + set (sets) of endings,
- stem + < word forming suffix > + set (sets) of endings,
- stem + < final consonant > + set (sets) of endings,
- stem + < final consonant + word forming suffix > + set (sets) of endings,
- stem + < comparative suffix > + set (sets) of endings,
- stem + < final consonant + comparative suffix > + set (sets) of endings,
- stem + < stem forming suffix > + set (sets) of endings,
- stem + < final consonant + stem forming suffix > + set (sets) of endings.

The strings in angle brackets can be called intersegments and there are 241 of them in our system. They include:

1. final consonants where the following alternations take place: *k-c-č, h-z-ž, g-z-ž, ch-š, r-ř,*
2. final consonants with the „graphic" alternations: *d–ď, t–ť, n–ň,*
3. final consonant groups with epenthetical *–e* that is omitted in the cases with full ending: *eb–b, el–l, em–m, en–n, er–r, et–t, ev–v,*
4. possessive adjectives with the word forming suffixes *–ův/–ov, –in,*
5. suffixes used for forming the comparative, e.g., -ejš-,
6. alternating word forming suffixes like *-ost, -ství,*
7. stem forming suffixes as e.g., *-ova-,*
8. alternating consonants of verb stems + stem forming suffixes (*-z-* > *-ž-*),
9. passive participle suffix for deriving verbal nouns and adjectival participles (*-ní, -ný, -ící*),
10. combinations of the types mentioned above.

Examples:

```
+slon
     <> V1,V13X,VOVE,VVE,VZ,VI
     <ův> PRIVL1X
     <ov> PRIVL1
+/medvíd
     <ek> V13X
     <k> V1,VOVE,VVU
     x <kův> PRIVL1X
     <kov> PRIVL1
     <c> VI, VQ
+/nov
     <> PRT1,PRT2,6B,6H,6C
     <ějš> PRK
+/uč
     <> W1E,W2A,W4A
     <í> W1D
     <i> W3A,W5A
     <en> V13
```

## 4. AN ALGORITHMIC DESCRIPTION OF CZECH MORPHOLOGY

It can be seen that the dictionary of stems and list of patterns together represent an algorithmic description of Czech morphology. For each stem we know the pattern according to which the stem is inflected, thus we are able not only to recognize any word form for which there is a stem in the dictionary, we can generate it as well. The links between the patterns and stems represent the core of the algorithm of Czech morphological analysis and synthesis.

The dictionary of stems and list of patterns exist in two forms:

1. as **text files** that are readable for a linguist and can be edited; here relations between the patterns and stems are visible and fully accessible and can be easily changed, checked and corrected.
2. as **program (exe) files** in an internal machine code; here all the relations are hidden in a machine code and their correctness can be tested only by analyzing or generating particular Czech word forms.

Preprocessing programs have also been developed that translate the dictionaries as text files into machine files. In this way a user can create new versions of dictionaries and update them. They do not represent a direct part of the algorithmical description of Czech morphology but we have found them useful for building, updating and maintaining the whole system.

## 5. COMPUTER PROGRAMS

Presented linguistic description in the form of algorithms enabled us to develop linguistic software consisting of three groups of programs:

1. A set analyzers, i. e. programs that are able to recognize Czech word forms taken from an arbitrary Czech text file and in this way to check their correctness, i. e. a Czech spelling checkers. They are implemented as a RAM resident (screen) checker and also as a file checker working with complete Czech text files. Both programs perform full morphological analysis of Czech inflected word forms and they say *yes* if a checked word is correct or *no* if a checked word is wrong or if the respective stem cannot be found in the dictionary.

2. This group of programs can for each inflected word form occuring in a Czech text file find its basic form, i. e. nominative, singular for nouns and adjectives and infinitive for verbs (**a lemmatizer**). Each recognized word form can also be associated with its corresponding grammatical categories and if there is a homonymy all the existing possibilities are offered to a user. It is obvious that both these programs can serve as part of an integrated parser. There is also a possibility to generate all word forms that can be

derived from a selected stem. These programs in fact represent a machine dictionary of Czech and can become a module within a larger natural language understanding system.

3. Programs for building dictionaries from any text file:

   - dictionaries of word forms
   - freguency dictionaries of word forms
   - inverted (a tergo) dictionaries of word forms sorted according to their endings;

   Their limitation is that they are worked out just for Czech language and it is not possible to offer to them easily other East European alphabets (which can be done without trouble with eg. MICRO OCP (Oxford, 1991) or WORDCRUNCHER.

   Related to these are programs for editing and updating, maintaining and creating a machine stem dictionary and a list of patterns. It is also possible to merge dictionaries and assign patterns to the respective stems by means of a special editor developed just for this purpore.

## 6. CONCLUSIONS

The system of stems and patterns is designed in such a way that it captures not only Czech declension and conjugation but also basic word forming types in Czech. Within the system a user can get:

- all word forms that can be derived from a given stem by declension and conjugation and the respective grammatical information associated with them.

- all words that can be formed from a given stem by regular word formation processes, i. e. the system is able to produce verbal nouns from verbs, participles from verb stems and adverbs from adjectives plus the respective comparatives and superlatives.

In this sense our machine dictionaries are of some interest not only for a Czech user but also for foreign students of Czech and linguists interested in Slavonic languages.

In further research it is our aim to use the present system and programs for modelling more complicated word formation processes in Czech. For this, however, we also have to solve the problems with homonymy in Czech (e.g.*ženu* –

**verb, Ist pers. sg or noun, fem, acc. sg)** as well as the Czech verb prefixation. The machine dictionaries also have become a core of the Czech lemmatizer used as a part of the Czech computer thesaurus containing now about 20 000 entries (Pala, Ševeček, Všianský, 1992).

## BIBLIOGRAPHY

HALASOVÁ-OSOLSOBĚ, K.: Algoritmický popis české formální morfologie substantiv a adjektiv, SPFFBU, A 37-38, 1989-90, S. 83-97.

HAVRÁNEK, B., JEDLIČKA, A.: Česká mluvnice, SPN, Praha, 1981.

OSOLSOBĚ, K.: Popis systému českých substantivních a slovesných vzorů, rukopis, Brno, 1991.

OSOLSOBĚ, K., PALA, K.: Czech Stem Dictionary for IBM PC XT/AT, Conference on Computer Lexicography, Balatonfüred, September 1990.

PALA, K., OSOLSOBĚ, K., FRANC, S.: Česká morfologie a syntax v PROLOGU, sb. semináře SOFSEM'87, VUSEIAR Bratislava, 1987.

PALA, K., ŠEVEČEK, P., VŠIANSKÝ, J.: Český počítačový synonymický slovník, softwarový produkt a rukopis, Brno 1992.

PETR, J., A KOLEKTIV AUTORŮ. Akademická mluvnice češtiny 1, 2, Academia, Praha, 1986.

RUSÍNOVÁ, Z.: Současná česká morfologie, SPN, Praha, 1984.

SGALL, P.: Soustava pádových koncovek v češtině, AUC – Slavica Pragensia 2, s. 65-84, 1960.

SLOVNÍK SPISOVNÉHO JAZYKA ČESKÉHO. Academia, Praha, 1960, 1989.