# Computational Processing of Czech Derived Words

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

**Abstract**

The system presented in this paper is concerned with the computational processing of the selected types of Czech word-formation. The developed programming tool (word-formation module) aims at analysing and synthesising Czech derived words. Such a system is of particular value for automatic processing of Czech language where derivational morphology plays an important role in regular word-formation due to which new words come continuously into being.

The analytic function of this module is used in the process of recognition of unknown words. The word-formation module processes words that cannot be previewed in static sources of computer lexicons and are not identified in the process of lemmatisation. The generative (syntactic) function of this tool plays its role in several NLP applications and is used e.g. for enlarging the static part of the lexicon.

One of the productive word-formation types – diminutives – was selected for the illustration of these functions. Derivation relations between the basic words (stored in the lexicon) and derived words serve for the formulation of derivation rules that form the basis for the definition of derivation patterns and word-formation module algorithms.

## 1 Word-formation module

The word-formation module is an efficient tool for recognising unknown words from texts that were not identified during the process of lemmatisation (see Figure 1 in 2.2).

Taken most generally, the aim of the programming tool is to perform the following two tasks:

(a) to generate a potential word from a given stem and affix(es),

(b) to analyse a potential word and determine its stem and affix(es).

The vocabulary of a language is not a fixed list of words but a growing and developing, potentially infinite set. The words in this set can be split into two classes: basic words and potential words.

The basic words are those which are - at least synchronically - felt to be unmotivated as to derivation, while the potential words are those where the motivation for the derivation of their form is still intuitively strongly felt. The basic words are stored in the static part of the lexicon and all other words that are still processable by the programming tool (word-formation module) [Klímová 2001] are considered potential, i.e. derived. Derived word could serve as a base for the creation of a new word, too.

### 1.1 Sources of data

In the course of the work, the lemmatised part of the Czech National Corpus [SYN 2000] (containing 100 million current words) served as an important source of contemporary Czech language data and for illustrating the productivity of suffixes.

Backordered dictionaries [Slavíčková 1975; Králík et al. 1986] provide material for the information about suffixes and about the possible alternations in stem caused by these suffixes.

The dictionaries [SSČ 1994] and [SSJČ 1971] accessible in electronic form enriched the entity of data used for the definition of derivation rules and illustrated the language used several decades ago in comparison with the language of the 1990's provided by the SYN2000 corpus. Some differences in formation of words by chosen suffixes in the given time periods are mentioned.

### 1.2 Definition of derivation patterns

The functional variability of stems and possible combinations of stems and affixes was studied on the basis of all data available in electronic form. The derivation relations functioning in the process of creating new words were defined by using all the information stored in the database system (see in 2.1).

The derivation pattern represents the entity of all words derived by the given suffix; it defines the ability of combining the given suffix with the respective set of basic words. Every pattern is expressing the possible changes of the stem vowel and/or the final stem consonant depending on the given suffix. The paradigmatic properties of the suffix are given by the respective derivation pattern; the semantic point of view is not taken into account here.

The word-formation paradigm as the basic term of the word-formation system is expressing the properties of the given suffix, i.e. the POS of the basic and derived words and all the alternations caused by the given suffix. This paradigm sums all the properties of the derivation patterns for the respective suffix (examples of derivation patterns for diminutives see in 2.4).

### 1.3 Analysis of derived words

Every analysed textual word form first enters the lemmatising module. A word that has not been identified during the process of lemmatisation can be either an unknown basic word or a derived word that is not stored in the lexicon. This word is transferred to the word-formation module.

Generally, in the process of analysing derived words, all means used in the formation of the given word have to be determined, i.e. the paradigmatic and semantic features of the derived word, the grammatical categories of the basic word and affixes used, the type(s) of alternations applied. The changes in the stem of basic words may be of the following kinds:

softening of final stem consonant, change of quality of vocal, embedding/adding of a consonant.

The analysis can be unsuccessful when

(a) the respective basic word is not included in the lexicon (see in 2.1),

(b) the processed word is a basic word ended by the same string of characters as the respective suffix and is not included in the list of exceptions,

(c) a linguistic phenomenon was omitted during the formulation of the algorithm serving as the basis for the programming of the respective module.

### 1.4 Generation of words

From the point of view of generation of new words there are two main limits in word-formation [Dokulil 1962]:

(a) Each suffix has its function (semantic properties) and gives it to the newly created word. This statement supposes that every affix has its meaning and could be combined only with certain stems from the semantic point of view. The database of semantic codes (see in 2.1) has to serve as a basis for the semantic description of affixes and stems.

The automatically generated words are correct from the point of view of alternation but from the semantic point of view they can be classified in the following way:

(1) mentioned in one of the Czech language dictionaries [SSČ 1994] or [SSJČ 1971],
(2) usable,
(3) strange,
(4) unusable.

The classification was done by hand and is user dependent. This ability of the programming tool is theoretically utilisable for some word-formation types, as *e.g.* deverbative nouns, diminutives, and numeral derivatives. In any case it is necessary to take into account the problem of overgeneration and all kinds of possible exceptions. The semantic codes assigned to stems and affixes could serve for solving the problem of computerised generation of words. These codes determine the theoretical combination of the given elements of word-formation. Practically this work is to be done by hand and is very time consuming. This problem is illustrated by the suffix *-tel* that is marked by the semantic code expressing an intellectual activity and is combinable with verbs having the same semantic property, *e.g. učit (to teach), bádat (do research).*

(b) The suffix could be combined only with a certain form of the stem. This statement says that only certain form of the stem (*e.g.* the verbal present or past tense stem) could serve as the basic stem for the creation of a new word by given suffix. The database of affixes

describes the most productive affixes and gives the basic information for defining the derivation relations.

## 2 Computer Implementation

### 2.1 Structure of the system

The whole system of processing words from texts is a modular system consisting of two main (lemmatising and word-formation) modules (see Figure 1) and the lexicon standing in the core of the system. The database system provides the programming tool with necessary information.

The program fulfilling the functions of the word-formation module is an opened modular system. It means that for every word-formation type (*e.g.* diminutives, nouns of agents, names of locations, properties, etc.) a unique program module is created.

Lexicon is the basic computerised list of words serving as the base for various natural language processing systems. It contains not only static information about words but it is able to express the relations between words, too. It describes the formal properties of words and their lexical decomposition and therefore serves as a basic lexical entity for the dynamic word-formation module. The word-formation processes will be incorporated in the hierarchically organised lexicon. Apart from all information about words the lexicon will include a set of prototypes of derived lexical items with the definitions of derivation relations; this will enable the lexicon to cover also words (word forms) whose presence in the main lexicon is unforeseeable.

The data arrived at during the linguistic research were first stored in a FoxPro database system, where two databases were created: the database of suffixes and the database of semantic codes.

(a) Database of suffixes describes the most productive suffixes and gives the basic information for defining the derivation relations.

(b) Database of semantic codes serves for the semantic description of affixes and stems.

### 2.2 Word-formation control module

Words that are not identified in the lemmatising module (derived words unknown by the lemmatiser, *e.g. tygřík (little tiger), chatička (little cottage), náměstíčko (little square)*) are passed over to the word-formation module.

The word form is first converted to its potential basic form – lemma (*e.g. chatičkou –> chatička, tygříkem –> tygřík, náměstíčka –> náměstíčko, etc.*) in the control module and then transferred to the respective submodule according to its ending. This is only a supposed form of the lemma. The potential lemma is then classified according to its paradigmatic and semantic properties and passed over to the respective word-formation submodule (*e.g.* words ended by *-tel* are transferred to the submodule of nouns of agents).

Diminutives are first classified according to its gender, one suffix form can represent different

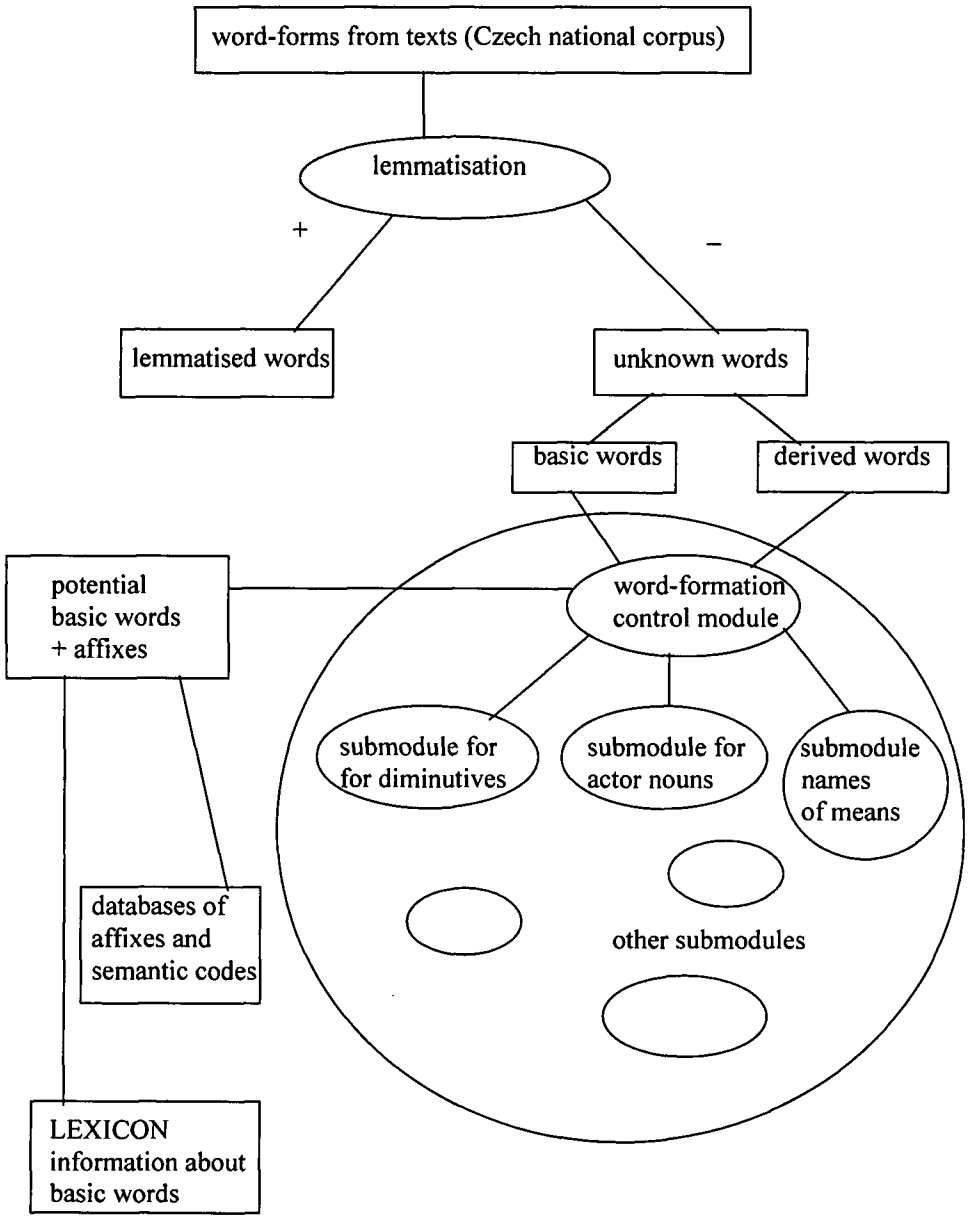


Figure 1: Processing of words from texts

grammatical categories, *e.g.* the suffix *-čka* can be tagged by five different tags (1. position: POS, 2. position: M - masculine, F - feminine, N - neuter, 3. position: S - singular, P -

plural, 4. position: case): NFS1, NMS2, NMS4, NNP1, NNP4. The analysed word-form is then transferred to the diminutive submodule according to its supposed gender.

If the word was analysed successfully in any submodule, the given word is assigned with its basic word (found in the lexicon) and all means used for the creation of this word (affixes, types of alternation applied) are determined.

The word is processed by all submodules till it is identified. If the analysed word form was not identified in any submodule, it has to be classified manually by the user. These unanalysable words can be unknown basic words that seem as to be derived words.

### 2.3 Word-formation submodules

Every submodule processes words derived by a certain suffix. First the exceptions have to be solved – words that have the same ending (string of characters) as the given suffix but that are not words derived by this suffix, e.g. *hotel (hotel), kotel (boiler), pytel (sack)* are words not created by the suffix *-tel* (used for the creation of agent nouns) and some words ended by the string *-čka* (female diminutive suffix) do not belong to the class of female diminutives, *e.g. akademička (she academician), fyzička (she physicist).*

All derivation rules assigned to the given suffix are continuously applied and the derivation pattern to which the word belongs is determined. If the respective basic word is found successfully in the lexicon, the derived word is classified with the relevant derivation paradigm. In the case of unsuccessful analysis the processed word form is returned back to the control module and then sent to another submodule.

### 2.4 Derivation of diminutives

Diminutives constitute a very numerous class (from almost every Czech noun several diminutives can be derived) and hence creating a module able to cope with them enables to enlarge the coverage of any system using this module.

| | Masculine nouns | feminine nouns | neuter nouns |
|---|---|---|---|
| primary suffixes | *-ík, -ek* | *-ka* | *-ko* |
| secondary suffixes | *-ček*<br>*-ínek*<br>*-oušek* | *-čka*<br>*-inka, -e/ěnka, unka*<br>*-w/ouška* | *-čko*<br>*-inko*<br>*-átečko* |
| tertiary suffixes | *-číček*<br>*-činek*<br>*-eneček, ineček*<br>*-ulínek* | *-čička*<br>*-čenka,-činka*<br>*-enečka, -inečka*<br>*-ulinka, -ulenka* | *-čičko*<br>*-činko*<br>*-enečko*<br>*-inečko,-ulinko* |

Table 1: Some Czech suffixes for the derivation of diminutives

It is useful (see in Table 1) to distinguish primary, secondary and tertiary diminutives. The difference between these three classes is in the degree of intensifying the diminution or the

emotional feature. The table and examples of diminutives below illustrate the rich productivity of Czech diminutive suffixes.

The diminutive suffixes are combinable with most of nouns, it means that there are nearly no restrictions from the semantic point of view but the rich alternation makes difficulties in the process of computerisation of this type of word-formation.

In Table 2 the derivation patterns for masculine diminutives, which illustrate derivation relations and express different kinds of alternation used in the process of derivation (EV - embedding of a vowel, N - no alternation, PSV - prolongation of stem vowel, SFC - softening of final consonant, (SFC) - possible softening of final consonant, SSV - shortening of stem vowel) of masculine diminutives, are given. These derivation patterns were formulated on the basis of all data accessible in electronic form (Czech national corpus, Czech dictionaries). The derivation pattern represents the entity of all words derived by the given suffix, it defines the ability of combining the given suffix with the respective set of basic words. Every pattern is expressing the possible changes of the stem vowel and/or the final stem consonant depending on the given suffix. The word-formation paradigm as the basic term of the word-formation system is expressing the properties of the given suffix, i.e. the POS of the basic and derived words and all the alternations caused by the given suffix.

| Derivation pattern | Alternation | Examples |
|---|---|---|
| *dům - domek - domeček* | SSV | *vůl - volek - voleček,*<br>*dvůr - dvorek - dvoreček,*<br>*stůl - stolek - stoleček* |
| *džbán - džbánek - džbáneček* | (SFC) | *práh - prážek, hrách - hrášek,*<br>*háj - hájek - háječek,*<br>*sloup - sloupek - sloupeček* |
| *chlapec - chlapeček* | SFC | *kopec- kopeček, měsíc-měsíček* |
| *kartáč - kartáček* | N | *míč - míček, kotouč - kotouček* |
| *krtek - krteček* | SFC | *zámek - zámeček, žebřík - žebříček,*<br>*hák - háček, krk - krček, býk - býček* |
| *list - lístek - lísteček* | PSV | *medvěd - medvídek-medvídeček,*<br>*sud - soudek - soudeček,*<br>*vlas - vlásek - vláseček* |
| *uzel - uzlík - uzlíček* | EV | *nehet - nehtík - nehtíček,*<br>*kotel – kotlík - kotlíček,*<br>*mazel - mazlík - mazlíček* |
| *vlak - vláček* | PSV, SFC | *drak - dráček, kluk - klouček,*<br>*potok - potůček* |
| *vůz - vozík - vozíček* | SSV, (SFC) | *kůl - kolík - kolíček,*<br>*vítr - větřík – větříček* |

Table 2: Derivation patterns for masculine diminutives

## 3 Conclusion

The aim of this paper was to describe some typical word-formation procedures in Czech and to present the tool developed for the generation and analysis of derived words. This instrument enables to assign the grammatical categories to the words derived by certain suffixes by the means of a set of derivational patterns and rules.

Seen from the practical viewpoint, the system enriches a static lexical framework with a dynamic derivation module able to process a rather huge set of lexical items that are predictable on the basis of derivation regularities. The level of explicitness achieved allows for applications in such systems as, e.g., grammatical tagging of derived words and spelling checkers, which are currently based on fixed lists of stems and their morphology. It is supposed that the word-formation module will enrich the functions of the lemmatizer and will be able to analyse words derived by the selected suffixes.

## References

[Čermák 1990] Čermák, F., 1990. Syntagmatika a paradigmatika tvoření slov II, Morfologie a tvoření slov. Univerzita Karlova Praha, Czech republic.

[Dokulil 1962] Dokulil, M., 1962. Tvoření slov v češtině, Academia Praha, Czech republic.

[Klímová 2001] Klímová J., 2001. Computational Processing of Selected Types of Czech Word-formation, Ph.D-thesis, Matematicko-fyzikální fakulta, Univerzita Karlova Praha, Czech republic.

[Králík et al. 1986] Králík J., J. Petr & M. Těšitelová, 1986. Retrográdní slovník současné češtiny, Academia Praha, Czech republic.

[Slavíčková 1975] Slavíčková, E., 1975: Retrográdní morfematický slovník češtiny Academia Praha, Czech republic.

[SSČ 1994] Slovník spisovné češtiny pro školu a veřejnost, 1978, Academia Praha, Czech republic.

[SSJČ 1971]Slovník spisovného jazyka českého 1-4, 1957-1971. Academia Praha, Czech republic.

[SYN 2000], Český národní korpus SYN2000, Ústav Českého národního korpusu, Filozofická fakulta, Univerzita Karlova Praha, 2000. http://ucnk.ff.cuni.cz.