The Morphology of the Czech Verb and Verb Derived Nouns and Adjectives as a Problem of the Formal Description and Automatic Analysis of the Czech Language

The text pursues the boundaries and possibilities of the automation of Czech derivational morphology (e.g. several types of nouns and adjectives derived from verbs).

The aim of the presented text is to recapitulate the results of the formal description of Czech verb inflexion and regular derivation from the *form of the verb* and to extend it to regular nominal derivation from the *stem* and *root of the verb* (e. g. nouns and adjectives derived from a given verb).

In the introductory chapter the impact on the field of the automatic morphological analysis of Czech are summarised. The formal description of Czech inflectional paradigms (Osolsobě 1996) was adopted as the base of the morphological analyzer *ajka* (Sedláček 2004), the tool developed and regularly used in the Centre for NLP at the Faculty of Informatics, Masaryk University in Brno. The morphological analyzer is able to process Czech word forms (both recognition and generating) as well as some regular derivations. The basis of the analyser is the morphological dictionary of Czech (Osolsobě 1996) in that the stems with inflection patterns (information about inflective forms and their morphological meanings e.g. person, number, etc.) are stored. The analyser generates and analyses data automatically; this means that to each simple word form all possible lemmas (basic forms e.g. nominative or infinitive) and all possible morphological meanings (of part of speech, gender, number, person, etc.) – morphological tags – could be automatically generated.

The inflectional paradigm of the verb is much more complicated than the nominal one. Besides paradigms of personal synthetic forms (e.g. indicative of present active and imperative) there are infinitive forms (infinitive, participles, gerundives). From the formal point of view the system of inflexional endings (personal endings, infinitive endings etc.) is nearly undifferentiated for all Czech verbs. The differences exist in the thematic vowel (sometimes combination of consonant + vowel) attached to the verb root forming the verb stem. Furthermore the root vowel often alternates both by inflection and derivation from the verb stem or from the verb root.

Word formation (inflection) and word derivation, i.e. forming new words from the corresponding word bases (root, stem), can be formally described as an operation on the strings of the characters (word form, lemma, morphological interpretation – tag). The derivational relations are typically regular and can be described as formal rules

The formal rules for both inflection and derivation work on the assumption that the unit defined as the verb root (formally, the string of character for written language analysis) is combined:

1) With the thematic suffixes and inflective endings (inflection);
2) With the thematic suffixes and derivation suffix (stem derivation);
3) With the derivation suffix (root derivation).

The unit defined as verb root is the smallest sign unit, e.g. the morpheme, that has more realizations/variants e.g. allomorphs. The allomorphs of the root differ in root vowel realisation and in root final consonant realisation. The grammar books designate these as *morphological alternations*.

What we consider to be an essential benefit of our work is that we have designed and tested our formal description of word-formation through derivation linked with a set of morphological alternations at different levels. The systematic description of rules for Czech allomorph occurrence in a verbal morphology based on the analysis of a large machine readable dictionary of Czech stems served us as point of departure. Following these rules, we monitored the usage of the allomorph of verbal forms and investigated further possibilities of variants in verb derived

word units. The rules mentioned above have been implemented in the formal description of derivational rules of selected types of nouns and adjectives derived from verbs and have been tested with the tool *Deriv*, the web interface (Šmerk 2009). This is an interactive tool able to process derivational relations in Czech according to the defined rules.

The formal rules, published and tested in our work, are formulated as substitution rules. Its format and the way of its presentation do not much differ from the programming language in which we often put a query to our corpus managers. For readers accustomed to working with language corpora it should not be a difficult task to draw on this exposition and to apply the suggested procedure (method) presented here for description in the overall field of word-formation systems in the frame of a computational linguistics which is oriented towards the analysis of word-formation. The corpora employed have been tagged at the morphologic level. Tagging should not be the final task, but merely a practical tool for work with the linguistic mega data-base. In morphology and word-formation we find a whole cluster of regularities and exceptions, which I have attempted to formulate in chapters 5 and 7. These can be of value to those working with language corpora, especially in cases where it is not desirable, or even possible, to use tagging.

The core of the Derivational Dictionary (DD), whose description is given in the 9th chapter, forms an integral part of our work. The data are accessible from http://deb.fi.muni.cz/deriv/ in the relevant directories. The DD consists of more than 96 000 twosome verb/derived nouns or adjectives. The frequency of each unit in corpus SYN2000 (100 million word forms) is joint and it is sufficient to switch on the corpus-concordance and dictionary browser, and than the usage in the corpus and in the definition in several Czech dictionaries is evident.