The Core of the Czech Derivational Dictionary

Radek Sedláček

Faculty of Informatics, Masaryk University Brno
rsedlac@fi.muni.cz

Abstract

Amongst all available language resources for the Czech language one can find a lot of useful dictionaries, databases and corpora. There are machine readable dictionaries of literary Czech (Havránek, 1989; Filipec, 1998), the dictionary of Czech synonyms (Pala, 2000) and two encyclopaedia: Otto and Diderot. Moreover, Czech researchers have two morphological databases (Hajič, 2001; Sedláček and Smrž, 2001), the Prague Dependency Treebank (Hajič, 1998; Hajičová, 1998), the Czech version of WordNet (Balcanet, 2001) and the Czech National Corpus (Čermák, 1998) at their disposal. However, any kind of resources containing information about word-formation and the detailed morphemic structure of Czech words is still missing. In this paper we will introduce the core of such a resource: the Czech Derivational Dictionary (CDD) and the procedure of its creation.

1. Introduction

In spite of its pivotal importance for inflectional morphology, for syntax, and for lexicology, derivational morphology has remained one of the least-investigated subsystems of Czech.

The reasons for the relative neglect of Czech derivation and for the relatively unsatisfactory quality of the results obtained so far are both theoretical and practical. This statement is in no way intended to disparage previous work in this field (Dokulil, 1962; Klímová, 2001), much of which is both useful and stimulating. However, the fact remains that even the best work in Czech derivation does not attain the sophistication of works in inflectional morphology (Hajič, 2001; Osolsobě, 1996; Sedláček and Smrž, 2001).

The usual dictionaries provide us with extensive, although by no means exhaustive material for the study of anlaut affixes, backordered dictionaries give us similar material for terminal affixes (Slavíčková, 1975). However, there is no source of information which includes prefixes other than initial or suffixes which occur elsewhere than in auslaut, nor is there a single source to which we might turn for even a rough listing of Czech roots within their derivational families. The sources known to me are either not computer-readable (Slavíčková, 1975) or too pedagogically oriented to be useful for a project such as this (Šiška, 1998).

The distressing lack of such basic research tools has provided the impetus for the development of the present Czech Derivational Dictionary (CDD) and related works now in progress. It is important to emphasise that this paper is presented not as a scientific description of Czech derivational morphology, but as one of the research tools which may help make such a scientific description possible at some time in the future. Before discussing the organisational structure of the Derivational Dictionary, however, it may be helpful to describe briefly the history of the CDD project.

2. History of the Dictionary Project

The Czech Derivational Dictionary which forms the bulk of this volume is the result of some four years' work, with occasional interruptions due to competition for programming and computer time. The CDD has evolved through several intermediate stages. Both the strengths and the inevitable weaknesses of the final product can best be understood by following the project as it developed throughout the four-year period.

2.1. Choice of a Corpus

When it was decided to begin constructing the CDD, the most immediate problem was the selection of a suitable corpus of contemporary standard Czech (CSC) words.

Since the value of the CDD depends partly on the completeness and contemporaneity of the corpus on which it is based, it was important to find a source of as many CSC words as possible.

The best general dictionary then available was SSJC (Havránek, 1989), which contains some 100,000 words. This seemed to be rather antiquated and thus inadequate.

The only source which appeared both large enough and modern enough was the one-volume Retrograde Morphemic Dictionary (RMD) edited by Slavíčková (1975). In spite of some inadequacies, errors of both commission and omission, RMD has proved to be a generally adequate and reliable corpus for the study of CSC.

2.2. Transformation to Electronic Version

Once the corpus had been selected, the first task was the, as it then seemed, relatively simple one of getting it transformed into electronic version, in order to be able to manipulate it within the computer. The RMD contains some 60,000 words; these were to be scanned, recognised and stored into a file.

2.3. Division into Fields

As a working procedure it was decided to divide each Czech word into three fields:

- a root segment,
- a preroot segment, and
- a postroot segment.

The root segment is not necessarily a root in the etymological sense, although the synchronic roots of word families more often than not correspond to their etymological roots. The preroot segment includes all prefixes and also the first elements of compounds; strictly speaking, the preroot field contains not merely the first, but rather, all but the last element of compounds, regardless of the internal 'immediate constituent' structure of these compounds, e.g., *radio foto tele /graf/* where slashes enclose the root segment. The postroot segment contains all suffixes and the flexional ending if any. Preroot and postroot fields can be and frequently are void, and the same is true, although but rarely, for the root field.

Segmentation rules were prepared for both the preroot and the postroot fields. For the former, the distribution of postfixes in RMD was inspected in detail and the rules based on this inspection; for the preroot segment, the SSJC dictionary served as equivalent source material. In both cases, tentative rules were written, and the dictionary was then inspected to determine the probable effect of these rules; the rules were then modified as necessary in the light of their probable effect.

2.4. The Second Stage of the CDD

The next stage in preparing the CDD consisted of two operations:

1. the results of the first segmentation had to be proofread and corrected;
2. the 60,000 words of RMD had to be reordered into families such that all words formed on the same synchronic root were grouped together.

Since proofreading of the computer routines for segmenting the preroot and postroot fields are of little interest, they need not be discussed here.

2.4.1. Reordering into Word-Families

The second part of the task of producing Stage II of the CDD consisted of reordering the 60,000 words of RMD into families.

This was a relatively simple affair, requiring only that the corrected version of CDD-I be alphabetised left-to-right within the root field, rather than from the left margin of the word. All words formed on */děd/* were thus grouped together, preceding the set formed on */děl/*, which itself preceded the set formed on */děloh/*, etc. Within each such group, entries were alphabetised left-to-right from the left margin of the word, zero being counted as the first letter of the alphabet. For example, all words formed on the root */děd/* which were prefixless preceded those which had prefixes, and the prefixed forms were themselves in alphabetical order. Finally, entries with the same prefix were alphabetised left-to-right within the postroot field. The results of these procedures were clusters of word-families such as that of *děd* '*grandfather*':

```
        /děd/
        /děd/ a
        /děd/ eč ek
        /děd/ ek
          ...
     pra /děd/
     pra /děd/ eč ek
pra pra /děd/
pra pra /děd/ eč ek
          ...
```

Each such word-family was assigned a label identical to the root on which it was built. For reasons which will become apparent below, this first label was termed the *occurrence root*. The entire corpus, corrected and ordered into these families, comprised Stage II of the CDD. This was by no means, however, a finished product.

2.5. The Stage III of the CDD

Although CDD-II was beginning to resemble the type of derivational dictionary of CSC envisaged at the outset of the project, it still had essential shortcomings. Specifically, there were two serious flaws in CDD-II.

2.5.1. Roots Homography

For one thing, the routines operating on strings of graphs were obviously incapable of distinguishing between homographic roots. This resulted in word-families which were actually mixtures of two or more separate families. The RMD contains some 203 words the graphic root of which is *vod*; two-thirds of these contain the root *vod* '*to lead*' as in *vodit* and one-third the root *vod* '*water*', e. g. in *voda*, but all of them were alphabetised together as if they constituted one large family.

The only solution to the problem caused by the existence of such homographic or multiply homographic groups was to proofread the entire CDD again and to assign distinct numeric identification tags to otherwise homographic roots. Thus, the homographic *vod* was split into *vod-1* '*to lead*' and *vod-2* '*water*', which permitted the words built on each of these two roots to be grouped separately.

2.5.2. Basic Roots

The second major flaw of CDD-II is following: due to the number and complexity of the morphophonemic alternations of CSC which are still more complicated in derivational families than they are in flexion, words of one and the same etymological family are scattered in small clusters throughout the dictionary. Words built on the root meaning 'take', for example, occur with the graphic root variants *ber*, *br*, *bír*, *běr*, and *bor*.

It was clearly necessary to assign all root allographs to an arbitrarily chosen canonical root form labelled the *basic root*; thus, all variants of 'take' were assigned to the basic root *br* '*take*'. The particular form chosen for the canonical root was to some extent arbitrary in any case, and decisions as to the shape of basic roots were made largely with a view to avoiding homography among these basic roots, i.e., to obviate the need for numeric tags on basic root forms. Therefore where two basic roots would serve equally well, that which appeared in the larger number of Czech words was usually chosen.

2.5.3. Final Ordering of the CDD-III

When all cases of homography in both occurrence and basic roots had been resolved, and when root allographs had been assigned to the appropriate basic root, Stage III of the CDD was ready to print out. CDD-III represented the 60,000 words of RMD, alphabetised by basic root; within each basic-root family, all words were grouped into sets according to the occurrence root, the latter also being in alphabetical order. Within the set of any given occurrence root, individual entries were alphabetised as described. CDD-III looked roughly as follows:

```
vod-1 | vod |    /vod/ i t
      |     |      ...
      |     | od /vod/ i t
      |     |      ...
      | voz | od /voz/ e n ý
      |     |      ...
      | vůd |    /vůd/ c e
      |     |      ...
vod-2 | vod |    /vod/ a
      |     |      ...
      |     |    /vod/ n í
      |     |      ...
```

2.6. The Stage IV of the CDD

As a tool for research on Czech derivational morphology, CDD-III was already unique, in comparison with the ordinary dictionary and a tergo dictionary, since it contained a large corpus of CSC words organised into derivational families. Once this major task had been accomplished, further changes were in the nature of refinements designed to get the individual words within each family into something more closely approximating derivational order, rather than the largely irrelevant left-to-right alphabetical order of CDD-III. To this end two further steps were undertaken:

1. compounds were reordered under their derivational bases, and
2. preroot segments other than compound elements were realphabetised.

2.6.1. Compound reordering

In CDD-III all preroot entities were treated identically. No distinction was made between prefixes *v, na, pro*, prefixoids *anti, poly*, and the first elements of true compounds *tele* in *telegraf*, etc.; all such preroot entries were simply listed in left-to-right alphabetical order. Consequently, compounds were often far removed from their derivational bases, and when a base had several compound derivatives, the latter were separated from each other by intervening prefixed derivatives.

The first step in compound reordering was to separate prefixes from compounding elements, all prefixoids were treated as compounding elements for this purpose; this was accomplished by providing a list of prefixes, which is a closed set, and classifying all items not found on this list as compounding elements.

In this revised version of CDD, the compounds formed on the bases *odběratel, odběratelka, odběratelský* and *odběratelsky* are listed in alphabetical order immediately after each of these bases.

The result of the reordering just described was a decided advance over the original form of CDD-III. However, the reordering itself brought to light a problem which has only rarely been noted in the literature, for example, in Dokulil's (1962) study, but which is more widespread and of greater theoretical implications than generally recognised.

It is clear even from the limited material that the bulk of compounds in Czech are also suffixal derivatives: *velkoodběratelka* and *velkoodběratelský* are not simply compounds of *odběratelka* and *odběratelský* respectively, but are suffixal derivatives of *velkoodběratel* as well, and to treat such words as simple compounds is to ignore an essential derivational relation of CSC.

It is obvious that no linear printout can render simultaneous dual derivation satisfactorily, and unless decisions are to be made on an ad hoc basis from one word-family to the next, there are only two possibilities: either all such suffixal compounds are to be treated as compounds of suffixed bases, or they are to be considered as suffixal derivatives of compound bases.

Inspection of several hundred sample suffixal compounds led to the conclusion that the second of these two solutions was preferable. To illustrate with a few small sets of words:

```
       od /běr/ a tel
mal  o od /běr/ a tel         compound
mal  o od /běr/ a tel k  a    derivative
mal  o od /běr/ a tel sk ý    derivative
mal  o od /běr/ a tel sk y    derivative
velk o od /běr/ a tel         compound
velk o od /běr/ a tel k  a    derivative
velk o od /běr/ a tel sk ý    derivative
velk o od /běr/ a tel sk y    derivative
       od /běr/ a tel k  a    derivative
       od /běr/ a tel sk ý    derivative
       od /běr/ a tel sk y    derivative
```

2.6.2. Prefix Realphabetisation

The compound reordering routines described above left all regular, i.e., noncompounding, preroot entries as they were in CDD-III, i.e., in straight left-to-right alphabetical order. That this order is indeed far removed from derivational order is clear from the following example: *vézt*, from which all other infinitives such as *odvézt, přivézt, svézt*, etc. are derived, should stand before all of them, but does not.

The inconvenience of such order is obvious. In order to bring such words into an order which more closely approximates their real derivational relations, it was necessary to reorder the prefixes 'from inside out', that is, to ensure that *odvézt* follows rather than precedes *vézt*, from which it is derived, etc. This was a relatively simple procedure, which consisted in effect of eliminating the original left-to-right order in favour of small subgroups, each of which was headed by the verbal prefix closest to the root field and contained all further entries with this closest prefix plus secondary prefixes. The fragment of the */véz/* family then appeared as:

```
      /véz/ t
   do /véz/ t
   na /véz/ t
   od /véz/ t
po po /véz/ t
  pro /véz/ t
  pře /véz/ t
        ...
   za /véz/ t
```

which is obviously much closer to genuine derivational order than was the version of CDD-III.

CDD-IV was the last of the major revisions undertaken in the development of the CDD. This version, with compounds reordered and prefixes rearranged in something closer to derivational order, provided the final format of the Czech Derivational Dictionary.

3. The Present Format of the Dictionary

The basic outlines of the format of the CDD have already been given in the historical sketch of the project.

The dictionary consists of more than 60,000 Czech words, 43% of which are nouns, 27% adjectives, 25% verbs and 5% adverbs, divided into cca 1700 word-families. Each family has been assigned a basic or canonical root form and all families are alphabetically ordered according to these basic roots. Within each basic-root family, there are several occurrence-root subfamilies. Each such subfamily contains all the words built on the given allomorph of the basic root, and the subfamilies themselves occur in alphabetical order.

Within each occurrence-root subfamily, words are grouped into smaller nests on the basis of their preroot field structure. Nonprefixed words, i.e., words with zero prefix, zero being counted as the first letter of the alphabet, precede prefixed words, the latter being listed according to the alphabetical order of the prefixes.

Prefixed words themselves are grouped so as to reflect their immediate constituent structure as closely as possible. The primary ordering criterion is the prefix nearest to the root. Words with more than one prefix are listed as subgroups within the group whose place is defined by the first prefix, i.e. that nearest the root. Prefixes, in other words, are ordered ’from inside out’.

Words with identical preroot and root fields are alphabetised by suffix, i.e. by their postroot fields. Postroot alphabetisation is by segment, not across the entire field, so that, for example, the suffix *ov k a* precedes *ova t*, since *ov* precedes *ova*.

Compounds are listed immediately after the simplex upon which they are built, e.g. *doprava* is followed by *auto-doprava*, . . . , *rychlodoprava*, and only then does the next alphabetical suffixal derivative of *dopravce*, *dopravní* occur. All words with compounding elements in the preroot field and with suffixes in the postroot field are treated as suffixal derivatives of compounds, and not as compounds of suffixal derivatives.

Compounds of compounds are treated just as compounds themselves, i.e, the compound of a compound, together with all its own suffixal derivatives, follows immediately after the compound upon which it is based, thus preceding the latter’s suffixal derivatives.

To recapitulate: ordering follows from the basic root, then the occurrence root, then the preroot segment, then the postroot segment.

4. Conclusion

The foregoing sections have made it clear just what the CDD is and what it is not. It should not be approached as if it pretended to be a complete description of Czech derivational morphology; such a description is still far in the future. If taken at face value, however, as a large collection of material organised according to consistent internal logic, there is some reason to hope that it may prove a useful aid to the further study of the principles and procedures of word-formation in contemporary standard Czech.

Acknowledgement

This work was supported by Ministry of Education, Research Program CEZ:J07/98:143300003.

5. References

Balcanet, 2001. Balkanet project. http://dblab.upatras.gr.

Dokulil, M., 1962. *Theory of Word-Formation*. Praha: Nakladatelství ČSAV. In Czech.

Čermák, F., 1998. Czech National Corpus: Its Character, Goal and Background. In P. Sojka et al. (ed.), *Text, Speech, Dialog*. Brno, Czech Republic: Masaryk University Press.

Filipec, J., 1998. *Dictionary of Literary Czech for School and Public*. Praha: Academia, 2nd edition. In electronic version.

Hajič, J., 1998. Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In Eva Hajičová (ed.), *Issues of Valency and Meaning. Studies in Honor of Jarmila Panevová*. Prague Karolinum, Charles University Press, pages 12–19.

Hajič, Jan, 2001. *Disambiguation of Rich Inflection*, volume I (Computational Morphology of Czech). Prague Karolinum, Charles University Press. 334 pp.

Hajičová, E., 1998. Prague Dependency Treebank: From Analytic to Tectogrammatical Annotation. In *Proceedings of TSD 1998*. Brno, Czech Republic.

Havránek, B., 1989. *Dictionary of Literary Czech*, volume 1-8. Praha: Academia. In electronic version.

Šiška, Z., 1998. *Base Morphemic Dictionary*. Olomouc: Vydavatelství Univerzity Palackého, 1st edition.

Klímová, J., 2001. *Computational Processing of Selected Word-Formative Types in Czech*. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University, Praha. In Czech.

Osolsobě, K., 1996. *Algorithmic Description of Czech Formal Morphology*. Ph.D. thesis, Faculty of Arts, Masaryk University, Brno. In Czech.

Pala, K., 2000. *Dictionary of Czech Synonyms*. Praha: Nakladatelství Lidové noviny, 3rd edition.

Sedláček, R. and P. Smrž, 2001. A New Czech Morphological Analyser `ajka`. In *Proceedings of TSD 2001*. Berlin: Springer-Verlag.

Slavíčková, E., 1975. *Retrograde Morphemic Dictionary of Czech)*. Praha: Academia, 1st edition.