PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 99 APRIL 2013 101–105

REVIEWS

Review of the book "Morfologie českého slovesa a tvoření deverbativ jako problém strojové analýzy češtiny"

(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)

Klára Osolsobě

Spisy Masarykovy univerzity v Brně, Filozofická fakulta, č. 401. MUNI Press 2011, 220 pp., ISBN 978-80-210-5565-0.

Reviewed by Jaroslava Hlaváčová

Czech deverbatives are usually derived from verbs by means of mostly very regular patterns. The rules for deverbative derivations have been described by many linguists in many previous works. Nonetheless, those descriptions are not possible to convert directly into a system usable in the field of natural language processing (NLP). They are not detailed enough to cover the whole diversity of morphological alternations in the particular verb forms taken as the basis for the derivations. For automatic processing, it is necessary to present the description down to the slightest detail. Moreover, the rules have to be expressed within a formalism that ensures their exactness and unambiguity.

The only widely used system utilizing the regularity of deverbative generation in NLP is the Prague morphological tool, in which the generation rules constitute an intrinsic part. However, the rules are very general, the result of which is a massive overgeneration. In morphological analysis, it does not matter as much, but in other fields of usage, it is naturally better to generate only those words that (may) really exist.

Thus, the presented treating deverbatives with respect to their application in NLP is a meritorious achievement.

All the linguistic data were taken from the Czech morphological dictionary which forms the basis for the automatic morphological analyzer "ajka" used for morphological tagging Czech corpora at Masaryk University in Brno. Until now, the dictionary has served mainly for analysis and synthesis of inflected word forms. The book brings an impulse for its usage in the domain of word formation. The first step was made with deverbatives, as they are one of the most productive parts of the Czech word-formation.

For automatic finding candidates of deverbatives, the software tool Deriv developed at the Faculty of informatics at MU Brno was used. It searches the dictionary for pairs of words, that meet special substitution rules transforming the first word into the second one. In that sense, the "derivation" is in reality no real derivation. It is "mere" checking, if there are pairs of words present in the dictionary, fulfilling certain (substitution) rules. The author prepared tens of such rules. The resulting pairs had to be checked manually because of great homonymy. The final lists are not presented in the book. They rest on the server where they were processed. The author gives (in Section IX) only a list of directories and files where the results are stored. The results are not accessible directly, it is necessary to ask an administrator for permission. I used an anonymous access, so I was able to use the tool for several trials, but did not manage to look into the author's results. I would expect that the author make her findings, or at least several representative examples, accessible.

The book is organized "procedurally". The Introduction (Section I) contains a short historical overview. The theoretical background and terminology is covered by the next three sections. Section V contains observations of the data with consequences in the form of rules that are later (in Section VII) specified more precisely for using in the automatic tool Deriv which is described in Section VI. Having a brief user introduction to Deriv, the Section VII brings the final rules used for searching the dictionary with the tool. Sections VIII and IX are devoted to the results. The Conclusion (Section X) summarizes the book, together with a short English résumé. The book ends with a detailed bibliography and three appendices: A – description of morphological tags, B – web interface of Deriv, C – examples of results.

Let us have a more detailed look at the essential parts of the book.

In the first three sections, there is a very brief overview of the morphological tools used in Brno. There is also stated the objective of the work – the formal description of "realized" derivations of selected types of Czech deverbatives. The word in quotes is important (and as such should not be explained in the book only as a footnote), because it restricts the linguistic material only to the morphological dictionary itself. The author admits that it would be more valuable to use a larger basis, for instance a big language corpus. However, the dictionary size is still quite large. Another aim of the work was to investigate the possibility of automatic language analysis with respect to word-formation and the dictionary served as a natural word reserve ready to use. Moreover, the dictionary offers additional pieces of information, namely the morphological tags. If a corpus was utilized for a similar research, there would be

necessary to choose a slightly different approach, as there appear unknown words, not recognized by the morphological dictionary. As such, the research cannot rely on their morphological tags.

The fourth section presents the terminology used in word segmentation, with the special attention to deverbatives.

Section VI contains a short user manual of the software tool Deriv. There is also a very brief introduction into the syntax of regular expressions (table on p. 46). The tool itself was probably designed to fit the needs of the author and as such it works surely very well. However, its user interface is not very intuitive and it would demand certain time to get used to it. Nevertheless, it is not the subject of this review.

The heart of the book are the sections V, VII and VIII where the main fiddly work of the author is presented.

Section V summarizes morphological alternations occurring in verbs and deverbatives. They are presented in the form of "observations", from which the author derives "rules".[1] The section is divided into several subsections, first of them being introductory, others are sorted according to the place where the alternation occurs.

Section VII brings the list of all analyzed types of deverbatives, each of them having its own subsection. They have always the same structure: At first there is a description of the derivation, including alternations relevant to the given type. Then comes a formulation of substitution rules for the software tool Deriv. The rules are presented in the form of a table, together with an example and the number of resulting pairs [verb, deverbative]. The last two columns express the number of real pairs and over-generated ones.[2] Each subsection ends with a discussion about observed overgeneration illustrated by a bar graph. My notice concerning graphs in the whole book see below.

Section VIII presents a quantitative analysis of all the resulting pairs [verb, deverbative] with respect to the overgeneration. Derivation types are compared in groups that were created according to the basis of the generation. The tables given for every group repeat summed figures from the tables of Section VII, but some mistakes have crept in.[3] They are not serious mistakes, but their presence unfortunately decrease the credit of the whole, in reality very laborious and valuable work.

The main shortcoming of the book is lack of index. Also a list of all the abbreviations presented on a separate page at a special place of the book (at the beginning, or

[1] Their numbering, especially on pages 26 to 30, is a bit confusing.

[2] Here, a defect occurs in all the subsections – the tables presenting substitution rules do not have properly labeled the 3rd (no label) and 4th column (label "pair" is not instructive enough). Similar tables in Section VIII are labeled correctly.

[3] See for instance numbers at the top of page 76, where the percentage is calculated correctly, and the summarized table with a slightly different (and wrong) figure on page 171 for the suffix *-tel*. Another mistake occurs in the figures for the suffix *-dlo* on page 171 where right and over-generated pairs do not sum to the total given in the third column (wrong copy of figures from page 91).

at the end) would be appreciated. There are quite a lot of abbreviations used through the whole book, but their explanation is presented only inside the text, mainly on pages 20 and 24.[4] Also the system of references within the book is not adequate. The references "see above / below" are not possible to follow, if they point farther than several paragraphs.[5] I have also a critical remark on graphs. They are presented in Section VII for individual deverbative suffixes, and then in Section VIII, where they are summarized according to the types of their generation. All the graphs present the same – relation between number of correctly generated deverbatives and over-generated ones, expressed in percents. Thus, there are always pairs of numbers summing up to 100%. The graphs are not designed very well. Firstly, they all are too wide for the presented data – the right (and bigger) side of every graph is always empty. Secondly, the marks of the x-axis do not have any value – in fact, they do not represent anything and therefore their presence is very confusing. And thirdly, as all the graphs present the same relationship, it would be more illustrative if the y-axis was always calibrated from 1 to 100, not sometimes to another amount, according to the concrete numbers.[6] In my opinion, the graphs would be more clear, if there were only one bar, always of the same length, divided in the given ratio, not two of them. A horizontal bar, always in the same size, with numbers expressed directly in the graph, not above, would probably fit the given purpose the best.

Throughout the book, there is quite a lot of minor inaccuracies, typos, sometimes even minor mistakes. Inconsistencies are present also in formalized expressions of strings. Sometimes, the author uses regular expressions, sometimes she chooses possibly a "more readable" slash (/) for alternatives, sometimes even within one paragraph,[7] or one expression.[8] The slash is actually overused as a whole, which sometimes may cause problems with understanding, especially when it is used within one sentence in more meanings, as in the 3rd paragraph from the bottom on page 47 where the slashes mean at first alternatives and at the end of the sentence it serves as a sign for a pair.

My final critical remark concerns the names of sections and subsections. The main sections do not have their names, they are only numbered. It is naturally the question of the author's attitude, but the subsections should perhaps be more structured, probably also numbered for a better cross referencing. I would also number the graphs and tables.

---

[4]Moreover, the abbreviation VSB introduced there under item *d*) is modified as SBV in item *h*) and later in the table summarizing these abbreviations (page 26 above). In the footnote 97 on the page 42, there is again VSB.

[5]See for instance footnotes 476 and 477 on page 171.

[6]Compare for instance graphs on pages 72, 83 and 87.

[7]Rule 2 on p. 42.

[8]At the end of the Rule 1 on the same page.

The book contains three essential results: the lists of pairs [verb, deverbative] that can be used in various domains of NLP, such as machine translation, data mining, summarization and many others. The second substantial achievement is the formulation of the set of formalized rules describing the relationship between verbs and their deverbatives. Those rules may be utilized for detection of other similar pairs, that have not been present in the morphological dictionary, in another source – for instance in a language corpus. The third contribution is the method itself. Formulation of additional rules may detect other relationships among words of different parts of speech and their derivatives.

The author has investigated an immense amount of linguistic material, categorized it and stated formal rules for derivation of deverbative nouns and adjectives from verbal stems or word forms. The book represents an important contribution to the domain of Czech word-formation. My main critical remarks concerned mostly technical aspects of the book. Such a laborious work would definitely deserve a more careful technical preparation. However, the factual content of the book presents an important bridge between descriptive and computational linguistics and as such constitutes an essential achievement in the modern linguistic research.

Address for correspondence:
Jaroslava Hlaváčová
hlavacova@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics,
Charles University in Prague
Malostranské náměstí 25
118 00 Praha 1, Czech Republic