Derivancze — ***Deriv***ational ***An***alyzer of ***Cze***ch

Karel Pala and Pavel Šmerk(✉)

Faculty of Informatics, Masaryk University,
Botanická 68a, 60200 Brno, Czech Republic
{pala,smerk}@fi.muni.cz

Abstract. The paper describes a new tool Derivancze, which provides an information on derivational relations between Czech words. After a summary of linguistic descriptions of Czech derivation we present a structure of our data and types of derivational relations we use. We compare our approach and results with Czech lexical network DeriNet, in particular, we discuss many differences between the two approaches. Our tool presently works with Czech data only, but the solution is general and can be used also for other languages.

Keywords: Derivational morphology · Derivational analysis · Semantics of the derivational relations

1 Introduction

Standard morphological analyzers typically provide for an input word its corresponding basic form but as a rule they do not offer (or in a limited way only) information about derivational relations between words such as, for example, in Czech *otec – otcův* (*father – father's*), *řezat – řezání* (*cut – cutting*), *učit – učitel* (*teach – teacher*), etc. This information can be very useful for text indexation in searching or in the course of the syntactic analysis of the natural language and also for other applications.

In the highly inflectional languages like Czech derivational relations (further D-relations) represent a system of both formal and semantic relations that definitely reflects cognitive structures related to what may be characterized as a language ontology. For language users derivational affixes function as formal means by which they express semantic relations necessary for using language as a vehicle of communication. The affixes denote several sorts of meanings which we will try to classify in this paper. We will deal here primarily with Czech language but presented results can be applied with the necessary modifications to all Slavonic languages, see e. g. [1] or [2].

This work was supported by the Ministry of Education of CR within the Lindat Clarin Center.

P. Král and V. Matoušek (Eds.): TSD 2015, LNAI 9302, pp. 515–523, 2015.
DOI: 10.1007/978-3-319-24033-6_58

2 Motivation

The first important reason for doing all this is a belief that D-relations and derivational nests created by them reflect basic cognitive structures existing in natural language. These cognitive structures can be partly traced down in the standard Czech grammars [3] where they can be found under the term of the onomasiological categories. The semantics of the D-relations will be in the focus of our attention in the paper.

The second good reason for paying attention to the Czech derivational morphology is a need to describe the derivational relations as formally as possible and on this ground to develop software tools allowing to handle automatically D-relations between lexemes in Czech. The obtained results can be useful for various applications such as information extraction, indexing for searching engines, textual entailment, machine translation, etc.

The third inspiring reason is to confront the traditional description of the Czech derivational morphology as it can be found in the standard Czech grammars with its formal counterpart necessary for a computer treatment.

The last reason is to present a software tool, derivational analyzer called Derivancze, and to compare it partially with an existing similar derivational tool DeriNet [4].

3 Related Work

There is a well developed theoretical description of Czech derivational morphology by Dokulil [3,5]. It has served as an excellent starting point for a further work in this area (see, for example [6,7]). Dokulil in his explanation of the D-relations intertwines both semantic aspects of the Czech word derivation and its formal aspects in an interesting but also a complicated way.

It has to be remarked that Dokulil's theory and also other derivational descriptions in standard Czech grammars adopted from it are based on the partial data containing just typical well selected examples. The situation becomes different now when we have access to almost all relevant Czech data (relatively complete lists of affixes, stems, word lists obtained from corpora) and can process them with the appropriate software tools.

In particular, Dokulil works with what he calls onomasiological categories: modifications (smaller change of the meaning within the same POS: *učitel* – *učitelka*; *teacher*$_{\text{MASC}}$ – *teacher*$_{\text{FEM}}$), transpositions (change of POS without change of the meaning: *dobrý* – *dobře*; *good* – *well*, *padat* – *pád*; *to fall* – *a fall*), mutations (with a substantial change of the meaning: *slepý* – *slepec*; *blind* – *blind man*)) and reproductions (*bác* – *bácnout* (*squab* – *to do squab*) which include onomatopoic derivations) — within this framework he treats most of the derivational processes in Czech. It should be remarked that Dokulil's treatment of the D-relations is rather extensive, it takes 259 pages in [3], so it is not possible to mention all the relevant points in this paper. Thus here we are trying to follow just the main and most transparent derivational processes in Czech.

We have to mention the attempts to handle Czech derivational morphology in a more formal way which have appeared recently. One of them is a tool developed by Ševčíková and Žabokrtský (2014) called DeriNet [4]. Another tool is a modified version of the Derivational Ajka developed at NLP Centre FI MU (Sedláček et al., 2005, it was not published and exists only as a computer program).

Apart from them there are two other tools. The first one is Deriv [8][1] developed at the NLP Centre FI MU and the second one is a tool called Morfio [9][2] built in ÚČNK FF UK. It has to be remarked, however, that both Deriv and Morfio are different from DeriNet and Derivancze. Particularly, Deriv is a web tool for exploring derivational relations among word forms from a morphological analyser of Czech using regular expressions. The results are linked to Czech corpora (CzTenTen, SYN2000) in order to make its manual post-editing easier. Morfio is a web interface as well allowing users to search the corpus by series of parallel queries which specify a chosen derivational model. It also analyses obtained results for the morphological productivity of affixes and estimates the completeness of the derivational model.

4 Design of Derivancze: in Constrast with DeriNet

We take advantage of the fact that there are publicly available data of the DeriNet network together with detailed description of its internals. It allows us to describe our decisions on the design of the Derivancze data by means of comparison of our approach with the approach of the DeriNet authors. The substantial differences can be drawn up in three parts.

4.1 Semantically Labelled Relations Instead of Purely Derivational Relations

The most prominent difference between the two approaches consists in our effort to classify somehow the relations between words. We work with semantically labelled relations whereas in DeriNet one finds just simple derivational relations without any explicit labels, at most they vary in their members' POS. In our view, this can be sufficient e. g. for relation adjective–adverb mentioned as a potential practical application of the network (subsection 5.1 of [4]) but for more sophisticated applications (text generation, condensation, paraphrasing or textual entailment) the more detailed information on the type of the link will be needed. This seems to be confirmed by the DErivBase [10] derivational lexicon, which is a German analog of the DeriNet, as its authors expect that for the derivationally close words also their semantic proximity will have to be captured because all existing applications assume strong correlation between derivational and sematic proximity.

[1] http://deb.fi.muni.cz/deriv/

[2] https://morfio.korpus.cz/

Therefore, in our data we aim at the D-relations for which a regular and transparent semantics can be found. So we are not interested in base words for *komunismus*, *rusismus* and *revmatismus*, i. e. words *komuna*, *Rus* and *revma*[3] because while from the formal point of view the derivational process is fully regular, the semantic relations inside the three pairs differ from one another. Similarly, for particular D-relations, we are not interested in word pairs which do not correspond with the semantics of the given D-relation. For example, all three contemporary Czech gramars ([3,6,7] mention *mdloba* (*faints*) as an example of a quality/property name derived by means of suffix -*oba*, but the relation to the base adjective *mdlý* (*bland*) is only formal and a "regular" *mdlost* (*blandness*) is semantically much more proper (unlike e. g. *chudoba* > *chudost* for *chudý*, similarly to English *poverty* > *poorness* for *poor*).

Moreover, in some cases we have to abandon purely formal approach and for the sake of completeness and consistency ignore the direction of the derivational process. For example, nouns describing actions or states denoted by verbs are regularly derived by means of suffix -*ní*: *pracovat*–*pracování* (*to work*–*a work* / *working*). But in some cases also other words can be used, e. g. *práce* in this case. From the formal point of view, *pracovat* is derived from *práce*, not conversely, but the information on direction is not interesting for real world applications: they need to give a verb and get the corresponding noun. A similar example are inhabitant names. Most of them are derived from the name of an area, but there are also many exceptions: *Vietnam*–*Vietnamec*, *Polsko*–*Polák* (*Poland*), *Rusko*–*Rus* (*Russia*). Clearly, the name of inhabitant is derived from the name of the area in the first case, both names are derived from some common base in the second case and the name of the area is derived from the name of the inhabitant (nation) in the third case. But from the practical point of view this is not relevant, a potential application will ask for an inhabitant name corresponding to the given area.

Even further, in some cases something similar to suppletion in inflectional morphology appears. As well as plural forms of almost all Czech words are created regularly and only a few exceptions have irregular (*přítel*–*přátelé*; *friend*–*friends*) or suppletive (*člověk*–*lidé*; *man*–*people*) forms, we can see that, for example, almost all masculine → feminine changes are expressed by a respective suffix (*učitel*–*učitelka* above, *dělník*–*dělnice*; *worker MASC*–*FEM*) and only a few exceptions display some irregularities (*tchán*–*tchyně*; *father-in-law*–*mother-in-law*) — or they are entirely "suppletive" (*syn*–*dcera*; *son*–*daughter*). But then, because it is hard to find any reasonable argument why the application should get an answer if it requests for a feminine form of "regular" nouns *vnuk* (*grandson*) or *medvěd* (*bear*) (in Czech *vnuč-ka* and *medvěd-ice*) and should not in case of "suppletive" nouns *syn* (*son*) or *kůň* (*horse*) (in Czech *dcera* and *kobyla*), aiming at completness and consistency, we have to admit that even word pairs like *syn*–*dcera* should be counted as derivational, albeit "suppletive".

On the whole, the relations in our approach are based on the formal derivational relations, but as the semantics is what matters after all, they do not fully agree with the D-relations as they are treated in the standard Czech grammars

[3] *communism*, *russism*, *rheumatism*, *commune*, *Russian*, and *rheuma*

— these, as we hinted above, are not suitable for use in practical applications. It has to be remarked that we are trying to follow just the main and somehow “fuzzy” tendencies as there are not any objective criteria for decision what is semantically transparent and what is disguised (and to what extent).

4.2 More than One Base Word and Semantic Equivalence

In the DeriNet network, every word is allowed to have at most one base word, but this constraint seems to be too restrictive in some cases. For instance, *virový* (*viral*) is a relational adjective derived either from *vir* or *virus* (two shapes of the same foreign word, *virus*). Another nice example offers the DeriNet network itself: the base word of *antikomunista* is *antikomunismus* (*anticommunist*, *anticommunism*), but the base word of *antikomunistův* is *komunistův* (*anticommunist's*, *communist's*) which is clearly inconsistent. But even if the authors would prefer suffixation over prefixation (or vice versa), there would be no obvious reason for such decision. In our view, much better solution is to admit that *antikomunistův* can be derived both from *antikomunista* (with suffix -*ův*) and *komunistův* (with prefix *anti*-).

From the previous explanation it immediately follows that we also need some concept of semantic equivalence, at least to be able to distinguish cases like *virový*, where the two possible base words are semantically equal, from cases like *antikomunistův*, where they are different. This equivalence is going to cover also orthographical variants (*socialismus*/*socializmus*, but only *socialistický*; *socialism–socialist(ic)*) and synonymic suffixes (*normal*ita/*normál*nost, but only *normalizace* or *normální*; *normality*, *normalization*, *normal*).

4.3 Overgeneration Followed by Filtering through Language Corpora

For an initialization of the DeriNet network, only lemmata with SYN corpus [11] frequency ≥ 2 were used. In Derivancze, we prefer to acquire as many “correctly” derived pairs as possible[4] and then add frequencies from corpus or corpora. Doing this we should be able not only to obtain the same results as DeriNet, but also to make a distinction between impossible and infrequent. Moreover, we can offer this information also to the user or application to let them decide between synonymic suffixes. For example, name of the quality/property expressed by an adjective *hluchý* (*deaf*) can be both *hluchost* and *hluchota* — althought the suffix -*ost* is much more productive in general, *hluchota* is around two orders of magnitude more frequent than *hluchost*.

[4] We are not able to clarify what exactly means this “correctly” as it always be questionable in cases of rare word forms. It should be noted there, that contemporary Czech grammars cannot be trusted concerning statements what is possible. For example for passive verbal adjectives the grammars mention only transitive verbs or intransitive verbs with indirect object, but one can recall, e. g. *padaná jablka* (literally *fallen apples*), where the verb *padat* has no object at all. That is why we prefer to try to generate such forms for all or almost all verbs, even where they may seem “incorrect” (because not used).

5 Results

The Derivancze itself is implemented in the same way as the morphological analyzer `majka` [12], i. e. the data are represented as a simple list of `query:response`, which is converted to a minimal finite state automaton. Derivancze does not do any real analysis, but it only looks up all possible responses (derived forms and D-relations) for a given query (input word). It means that all possible analyses for all known inputs are precomputed in a compilation phase, thus the tool itself remains very simple and fast.

The current version of data comprises the following D-relations[5]:

- `k1verb`, `k2pas`, `k2proc`, `k2rakt`, `k2rpas`, and `k2ucel` from verbs, where:
  - `k1verb` derives nouns describing process, action or state denoted by the verb (*kropit–kropení*; *sprinkle–sprinkling*),
  - `k2pas` and `k2rpas` are passive participle and past passive adjectival participle, i. e. two forms of adjectives which describe the patient or object of the action (*kropit–kropen/kropený*; *sprinkle–sprinkled*),
  - `k2proc` derives present active adjectival participles, i. e. adjectives decribing a subject doing the action (*kropit–kropící*; *sprinkle–sprinkling* (man)),
  - `k2rakt` are past active adjectival participles, i. e. adjectives describing subjects which have completed the action (*pokropit–pokropivší*; *sprinkle–who has springled st.*), and
  - `k2ucel` derives adjectives which describe an object used for the action (*kropit–kropicí*; *sprinkle–sprinkling* (machine)),
- verb → agent noun relation `k1ag` (*bádat–badatel*; *research–researcher*),
- adjective → name of the property relation `k1prop` (*rychlý–rychlost*; *fast–speed*),
- adjective → adverb relation `k6a` (*dobrý–dobře*; *good–well*),
- noun → possessive adjective relation `k2pos` (*otec–otcův*; *father–father's*),
- noun → relational adjective relation `k2rel` (*virus–virový*; *virus–viral/virus*), semantically perhaps the most heterogenous relation among the Derivancze relations,
- relations `k1f`, `k1jmf`, and `k1jmr` express changes in gramatical gender:
  - `k1f` derives feminines from general masculines (*doktor –doktorka*; *doctor*$_{\text{MASC}}$ –*doctor*$_{\text{FEM}}$),
  - `k1jmf` derives feminine forms of surnames (*Novák–Nováková*), and
  - `k1jmr` derives family forms of surnames (*Novák–Novákovi*) — it should be noted that `k1f` and `k1jmf` cannot be joined because of names of nationalities, which can also act as surnames, but the derived forms differ (*Rus–Ruska* X *Rusová*, i. e. *Russian*$_{\text{FEM}}$ X *Mrs. Rus*),

[5] The names of D-relations can be seen as completely arbitrary, but in fact they are slightly based on the morphological analyser `majka` tagset [13]: `k1`, `k2` and `k6` denote that the derived word is a noun, adjective and adverb respectively.

- area or city → inhabitant name relation `k1obyv` (*Kanada–Kanaďan*; *Canada–Canadian*), formally the most heterogenous relation in Derivancze,
- noun → deminutive relation `k1dem` (*dům–domek*; *house–little house*).

The relations of the first group, verbal derivatives, are useful for tagging and syntactic analysis as the derived words somehow retain valences of the base verbs, i. e. they retain a syntactically relevant behavior (perhaps except for `k2ucel`), the other relations were either requested by our commercial partners (Czech search engine Seznam.cz) or useful for various kinds of text generation (e. g. [14]). The data itself are partially taken from data of morphological analyzer [15] and Czech WordNet [16,17], other relations and data were added from various sources, e. g. `k1obyv` is from [18], but in the most cases the Deriv tool [8] was utilized.

The Table 1 shows a distribution of the derivational pairs in Derivancze data according to the D-relation. The row `var` is the semantic equivalence introduced in the Section 4.2. To make a comparison with DeriNet easier, we added another two columns CzTenTen and SYN with numbers of pairs whose both members occur in respective corpus CzTenTen [19] or SYN [11] more than once (the criterion for inclusion a lemma to DeriNet was the same).

Table 1. Distribution of derivational pairs according to D-relation

Relation	**# of pairs**	**CzTenTen**	**SYN**
k1ag	703	588	447
k1dem	6342	5250	3342
k1f	3170	2343	1854
k1jmf	2230	2049	2114
k1jmr	2212	1786	19
k1obyv	262	241	209
k1prop	9886	7503	5975
k1verb	34781	20466	15097
k2pas	34847	11273	192
k2pos	30953	11879	6861
k2proc	15765	7040	5539
k2rakt	18106	1150	600
k2rel	20023	16782	13257
k2rpas	35017	17844	12343
k2ucel	1672	1582	1390
k6a	39065	17281	11678
var	565	406	98
total	255599	125463	81015

Obviously, the numbers of pairs occurring in corpora are rather approximative as they depend on particular lemmatization of the respective corpora (cf. [20] for CzTenTen and [21,22] for SYN). For instance, lemma of *Novákovi* in the SYN corpus is *Novák*, not *Novákovi* as in CzTenTen, thus no `k1jmr` pair should be found in SYN. But if the name is unknown to the morphological analysis component of the tagger, the lemma is retained equal to the word form, i. e.

Varmužovi is lemmatized as *Varmužovi*. This is the cause of the very low, but still non-zero count of `k1jmr` (and also `k2pas` and `var`) pairs in SYN.

Presently, Derivancze cannot be freely downloaded, but complete data is accessible through a web interface http://nlp.fi.muni.cz/projects/derivancze.

6 Conclusions and Future Work

In the paper we have presented the results of the computational analysis of basic and most regular D-relations in Czech exploiting the older unpublished derivational version of the morphological analyzer D-Ajka and re-designed it as a new derivational analyzer for Czech with the name Derivancze. Though the whole project is a "work-in-progress" and the analysis is far from complete, the number of the captured D-relations and generated derivational pairs is reasonable and covers the basic D-relations in Czech.

We have compared Derivancze with the DeriNet network: the most important differences are that Derivancze covers only semantically transparent D-relations, assigns explicit labels to them and namely prefers semantic consistency if the semantical and formal aspect of D-relations diverge from each other. It should be noted that the last seems to be novel not only for Czech derivational morphology tools and descriptions.

No evaluation has been done yet, but we plan to measure an "added value" of our data for applications which are able to exploit them.

We have also data for some other D-relations, but as they are semantically less transparent, we prefer to work on more regular relations, namely between verbs (aspectual changes, iterativity, etc.) at first. In the future we also aim to precisely describe productive derivational paradigms to be able to predict and recognize word forms derived from loanwords and other new words emerging in Czech texts.

Finally, we would like to note that the Derivancze is not only a theoretical result in Czech derivational morphology: one of its versions has been used as a concrete application in the Czech search engine Seznam.cz and it also serves as an instrument for various kinds of text generation ([14]).

References

1. Šojat, K., Srebačić, M., Tadić, M., Pavelić, T.: CroDeriV: a New Resource for Processing Croatian Morphology. In: Calzolari, N., et al. (eds.) Proceedings of LREC 2014. ELRA, Reykjavik (2014)
2. Pala, K.: Derivational Relations in Slavonic Languages. In: Proceedings of the FASSBL 2008, pp. 21–28. Croatian Language Technologies Society, Zagreb (2008)
3. Dokulil, M., et al.: Mluvnice češtiny I (Grammar of Czech I). Academia, Praha (1986)
4. Ševčíková, M., Žabokrtský, Z.: Word-Formation Network for Czech. In: Calzolari, N., et al. (eds.) Proceedings of LREC 2014. ELRA, Reykjavik (2014)
5. Dokulil, M.: Teorie odvozování slov (Theory of the Word Derivation). Academia, Praha (1962)

6. Karlík, P., et al.: Příruční mluvnice češtiny (Reference Grammar of Czech). Nakladatelství Lidové noviny, Praha (1995)
7. Čechová, M., et al.: Čeština – řeč a jazyk (Czech – Speech and Language). ISV nakladatelství, Praha (2002)
8. Hlaváčková, D., Osolsobě, K., Pala, K., Šmerk, P.: Exploring Derivational Relations in Czech with the Deriv Tool. In: NLP, Corpus Linguistics, Corpus Based Grammar Research, Bratislava, Slovakia, Tribun, pp. 152–161 (2009)
9. Cvrček, V., Vondřička, P.: Nástroj pro slovotvornou analýzu jazykového korpusu (A Tool for Word-Formation Analysis of a Language Corpus). In: Gramatika a korpus, Hradec Králové, Gaudeamus (2013)
10. Zeller, B., Padó, S., Šnajder, J.: Towards Semantic Validation of a Derivational Lexicon. In: Proceedings of COLING 2014: Technical Papers, Dublin City University and ACL, pp. 1728–1739 (2014)
11. Ústav Českého národního korpusu FF UK: Český národní korpus – SYN (Czech National Corpus – SYN), Praha (2014). http://www.korpus.cz (cited April 1, 2015)
12. Šmerk, P.: Tools for Fast Morphological Analysis Based on Finite State Automata. In: Recent Advances in Slavonic Natural Language Processing 2014, Brno, Tribun EU, pp. 147–150 (2014)
13. Jakubíček, M., Kovář, V., Šmerk, P.: Czech Morphological Tagset Revisited. In: Recent Advances in Slavonic Natural Language Processing 2011, Brno, Tribun EU, pp. 29–42 (2011)
14. Nevěřilová, Z.: Paraphrase and Textual Entailment Generation in Czech. Computación y Sistemas 18 (2014)
15. Veber, M., Sedláček, R., Pala, K., Osolsobě, K.: A Procedure for Word Derivational Processes Concerning Lexicon Extension in Highly Inflected Languages. In: Proceedings of LREC 2002, Las Palmas de Gran Canaria, pp. 1254–1259. ELRA (2002)
16. Pala, K., Hlaváčková, D.: Derivational Relations in Czech WordNet. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, pp. 75–81. ACL, Praha (2007)
17. Horák, A., Smrž, P.: VisDic – Wordnet Browsing and Editing Tool. In: Proceedings of GWC 2004, Brno, Czech Republic, Masaryk University, pp. 136–141 (2003)
18. Filipec, J., et al.: Slovník spisovné češtiny. Academia, Praha (1994)
19. Suchomel, V.: Recent Czech Web Corpora. In: Recent Advances in Slavonic Natural Language Processing 2012, Brno, Tribun EU, pp. 77–83 (2012)
20. Šmerk, P.: Towards Morphological Disambiguation of Czech. PhD thesis proposals, Faculty of Informatics, Masaryk University, Brno (2007) (in Czech)
21. Hajič, J.: Disambiguation of Rich Inflection (Computational Morphology of Czech). Charles Univeristy Press, Prague, Czech Republic (2004)
22. Spoustová, D., Hajič, J., Votrubec, J., Krbec, P., Květoň, P.: The best of two worlds: Cooperation of statistical and rule-based taggers for czech. In: Proceedings of the Workshop on Balto-Slavonic Natural Language Processing, Prague, pp. 67–74. ACL (2007)