Word-Formation Network for Czech

Magda Ševčíková, Zdeněk Žabokrtský

Charles University in Prague, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
sevcikova@ufal.mff.cuni.cz, zabokrtsky@ufal.mff.cuni.cz

Abstract

In the present paper, we describe the development of the lexical network DeriNet, which captures core word-formation relations on the set of around 266 thousand Czech lexemes. The network is currently limited to derivational relations because derivation is the most frequent and most productive word-formation process in Czech. This limitation is reflected in the architecture of the network: each lexeme is allowed to be linked up with just a single base word; composition as well as combined processes (composition with derivation) are thus not included. After a brief summarization of theoretical descriptions of Czech derivation and the state of the art of NLP approaches to Czech derivation, we discuss the linguistic background of the network and introduce the formal structure of the network and the semi-automatic annotation procedure. The network was initialized with a set of lexemes whose existence was supported by corpus evidence. Derivational links were created using three sources of information: links delivered by a tool for morphological analysis, links based on an automatically discovered set of derivation rules, and on a grammar-based set of rules. Finally, we propose some research topics which could profit from the existence of such lexical network.

Keywords: word-formation, derivation, derivational morphology, lexical network

1. Introduction

In Natural Language Processing, considerable effort has been invested in the development of resources of morphological data (lemmatization, morphological analysis / synthesis, tagging). Nevertheless, the main focus has been laid on inflectional morphology while derivational morphology is underresourced in most languages; cf. (Zeller et al., 2013). Derivational data seem to have an application potential in Machine Translation (Weller et al., 2013), paraphrasing etc. on the one hand, and are highly required for theoretical research into word-formation on the other.

In the present paper, we describe the development of the lexical network DeriNet which includes the derivational relations in Czech. After a summarization of the state of the art of theoretical research and NLP approaches to Czech derivation (Sect. 2.), we introduce the theoretical background and formal structure of the network and the semi-automatic annotation procedure in Sections 3. and 4. Finally, we propose some research topics which could profit from the existence of such lexical network in Sect. 5.

2. Related work

2.1. Linguistic descriptions

Czech is a Slavic language with both rich inflectional and derivational morphology. Derivation is the most frequent and the most productive word-formation process in Czech. Other word-formation processes (conversion, compounding, and reflexivization of verbs (Dokulil et al., 1986)) correspond to a minor part of new word coinages in Czech.

An elaborate theoretical approach to derivation in Czech was carried out by (Dokulil, 1962). Dokulil discerned four onomasiological categories according to how the extralinguistic content is organized in language: category of substance (typically corresponding to the part of speech of nouns), quality (corresponding to adjectives), circumstance (corresponding to adverbs) and action (corresponding to verbs). Three types of shifts between and within these highly abstract categories were identified: (i) so-called transposition[1] is understood just as a change of the onomasiological category, without any accompanying shifts in meaning (for instance, a quality "being *objective*" is expressed as a substance by the noun *objectivity*), (ii) modification covers semantic shifts within an onomasiological category (cf. diminution of *židle* 'chair' to *židlička* 'small chair'), (iii) mutation is a shift across boundaries of onomasiological categories or within an onomasiological category, always accompanied with a change in meaning (e.g. *bloudit* 'stray' – *bludiště* 'labyrinth', *galerie* 'gallery' – *galerista* 'gallerist').

Onomasiological categories are subdivided into word-formation categories, which include words with the same categorical meaning that are derived from the same part of speech; in Czech, for instance, there is a numerous category of deadjectival names with the meaning of a quality. Within word-formation categories, word-formation types are further distinguished according to affixes used in the derivatives; e.g. deadjectival names of quality with the suffix *-ost* (roughly corresponding to English '-ness' or '-ity') constitute a word-formation type, another word-formation type includes deadjectival names of quality with the suffix *-ství* ('-ship' or '-ness' in English, according to the particular word) etc.

Dokulil's approach has become a widely respected and, in fact, the only common ground of word-formation descriptions in Czech linguistics during the last 50 years, cf. (Daneš et al., 1967), (Šmilauer, 1971), (Hauser, 1986), (Dokulil et al., 1986), (Karlík et al., 2000), (Čermák, 2012). Derivational affixes are usually listed according to

[1] Dokulil defined transposition in a specific way, closely to the syntactic derivation introduced by (Kuryłowicz, 1936).

the word-formation type, which however leads to redundancy in description in some cases on the one hand and to underspecification on the other (see Sect. 4.7.).

2.2. NLP approaches to Czech derivation

In flagrant opposition to the solid theoretical research, there are only few NLP approaches to Czech derivational morphology. A limited derivational information is provided by available morphological analyzers; immediate derivational history of a word (its base word) is encoded directly in the morphological lemma for selected groups of derived words (with a transparent word-formation structure) (Hajič, 2004), or the derivational relations are processed upon the morphological analysis ((Sedláček and Smrž, 2001), (Pala et al., 2003)). A basic set of just 14 derivational relations was added to the Czech WordNet database ((Pala and Smrž, 2004), (Pala and Hlaváčková, 2007)).

Possessive adjectives and deadjectival adverbs were converted into their base words within the deep-syntactic (tectogrammatical) annotation of Prague Dependency Treebank 2.0 ((Hajič et al., 2006), (Razímová and Žabokrtský, 2006)). These highly regular derivatives (so-called syntactic derivatives (Kuryłowicz, 1936)) are considered to have the same meaning as their base words (which is correspondingly captured by representing both the derivative and its base word by the same lemma at the tectogrammatical layer) but different syntactic functions (which is sufficiently represented by different semantic role labels). Besides this theoretical aspect, the number of lemmas was reduced significantly by this account.

Some of the regular types are incorporated in the Machine Translation system implemented within the Treex NLP framework (Popel and Žabokrtský, 2010). In spite of these, rather preliminary accounts, a publicly accessible, large-coverage derivational resource has not yet been available for Czech.

3. Linguistic background of the word-formation network

3.1. Derivation in Czech

Let us illustrate the complexity of derivation in Czech. There are several hundreds of affixes (in particular, suffixes) used for derivation of new coinages. Many affixes have several meanings and/or can be applied to base words from more than one part of speech. The suffixation is often accompanied by consonant and vowel changes of the word base, by vowel insertion or deletion and/or by decapitalization; these and other factors influence productivity of the particular suffixes.

For instance, about 400 suffixes are used to derive nouns (Štícha, 2012), approximately 15 out of them are involved in nouns with the meaning of a quality. Applying Baayen's hapax-based productivity measures ((Baayen, 1992), (Baayen, 1993)) on the data of representative subcorpora of the Czech National Corpus (CNC, 2000), (CNC, 2010), the suffixes *-ost* ('-ness'/'-ity') and *-ismus* ('-ism') are the most productive ones. The former suffix is compatible with adjectives of both Czech and foreign origin; besides the qualitative meaning (e.g. *hloupost* 'stupidity'), it expresses also objects, statements, actions etc. characterized by the quality (*řekl/udělal hloupost* 'he said/did a stupid thing').

The suffix *-ismus* is primarily combined with adjectives of foreign origin, however, besides the meaning of quality and corresponding objects, statements, actions etc. (like in the case of *-ost*), it also derives names of art movements, philosophical theories, ideologies etc. not only from adjectives but also from proper names (e.g. *klausismus* 'klausism' "ideology based on Czech ex-president Klaus thoughts") and even from sentence-like sequences (cf. *jetřebismus* 'it-is-necessary-ism' derived from *je třeba* '(it) is necessary'), which seems to considerably extend the derivational potential of the suffix.

Generating all sentence-like sequences as potential bases would necessarily lead to combinatorial explosion, so it is obvious that certain conditions must be imposed for inserting a word into the network.

3.2. Linguistic decisions on the design of the network

At the beginning of the network-creating procedure, a decision was made that the network should consist of lexemes that are to be extracted from an existing corpus of contemporary Czech. The alternative solution to generate lexemes (semi-)automatically would obviously lead to a huge number of still well-formed but not attested coinages. Even if there is an obviously derived word in the network but its base word is not part of the network, the base word was not generated. However, these cases are "serious" candidates for a future extension of the network.

Another decision that we have faced concerned the relations to be captured within the network. Derivation is a dominant, nevertheless, an integral part of word-formation system, it often combines with composition. However, in order to make a clear delimitation, only derivational relations are captured within the network, neither products of combined word-formation processes (composition with derivation) nor compounds are linked up with their base word even if they are part of the network. This decision is reflected in a definition feature of the network architecture, namely, that a single base word is allowed to be specified for each word.

3.3. Minimal approach to polysemy

As it is almost always the case in lexicography, assigning a single network node to a single lemma (canonical form of a word) is not enough. However, choosing the right level of polysemy granularity is a notorious (and unsolved) problem. We did not want to couple the network with any of the already existing lexical resources (such as Czech WordNet), since their granularity is too fine-grained for our purpose. Instead, we decided to follow a minimal approach that does not lead to generating false derivational links: a lemma should be ideally split into two (or more) nodes if and only if

- it was coincidentally generated from two formally different base words as in the case of the adverb *tržně* with which one lexeme comes from the adjective *tržní* 'market (price)' and the other one from *tržný* 'lacerated (wound)', the noun *dlaždička* is either a diminu-

tive from the noun *dlaždice* 'paving tile', or a female variant of the male profession name *dlaždič* 'paver', or

- two senses of the lemma lead to different (sets of) derived words.

3.4. Treatment of orthographic variants

In Czech, there are vowel and consonant alternations in words of Czech as well as foreign origin, which can be traced back to various factors such as dialects, differences in style, variability of the language in time etc. These alternations lead to different types of orthographic variants; for instance, *vyndávat*/*vyndavat* 'to take out', *diskuze*/*diskuse* 'discussion', *citron*/*citrón* 'lemon'. All types of orthographic variants are handled as single lexemes within the network and coupled with the base words in accordance with the particular variation, i.e, *vyndávat – vyndávání* vs. *vyndavat – vyndavání* ('to take out' – 'taking out'), *diskuze – diskuzní* vs. *diskuse – diskusní* ('discussion' – 'discussion (group)').
However, if the variation arises just in the particular word (and is not present in the base words), both orthographic variants are linked up with the same base word as being two different derivatives of it; cf. both *nakladač*/*nakládač* 'loader' are linked up with the verb *nakládat* 'to load'.

4. Building the network

4.1. Network representation

In our approach, the relations between derived words and their base words are modeled as an oriented graph. Nodes of the graph correspond to lexemes (a lexeme represents a lemma, possibly only with a subset of its senses, as discussed above). Edges represent derivational steps between lexemes. The orientation of edges reflect the word-formative process: the edge points from a base lexeme to a derived lexeme. Each lexeme can have at most one base lexeme. Therefore the whole derivations graph is a forest composed of tree-shaped clusters consisting of derivationally related lexemes, see Figure 1.

4.2. Annotation process

When filling the derivation network with actual lexemes (nodes) and derivational relations (edges) between them, we combine three sources of information: existing data resources (corpora and lexical resources), linguistically motivated heuristics, and manually written rules. We try to automatize the process as much as possible; however, as we prefer precision to coverage, candidate pairs of a derivative and its base word were checked manually before creating an edge in the network, unless they came from a highly reliable resource. The main steps of the network construction are addressed in the following paragraphs.

4.3. Network initialization

At the beginning, the network is initialized with a set of lexemes, without any mutual connections. Given that some of the derivational processes in Czech are extremely productive, it was hard to find any linguistically justified boundary for the amount of lexemes to include. As already explained, we decided to use only lexemes whose existence is supported by corpus evidence. We extracted all noun, adjective, verb and adverb lemmas from the SYN subcorpus of the Czech National Corpus (CNC, 2014), which contains around 2.7 billion tokens. We used only lemmas that fulfilled the following conditions: they occurred at least twice in the corpus, they don't contain any digit or punctuation symbol, they contain at least two letters and at least one of them is lowercased (to suppress abbreviations). This filter led to around 260,000 lexemes inserted into the network.

4.4. Derivational links delivered by the tool for morphological analysis

The lemmatization and morphological tags which we rely on were described in (Hajič, 2004); the corpus texts were tagged using the Morce tagger (Spoustová et al., 2007). Some of the lemmas contain so called technical suffixes, which might—among other things—contain an information about the derivation of the current lemma (the technical suffix sometimes encodes the stringwise difference with respect to the original lemma). Using this information led to around 46,015 derivational links.

4.5. Specialized rules

We employed the fact that there are some highly regular derivational relations in Czech which can be described by simple rules. For instance, adverbs derived from adjectives can be often obtained by a very simple suffix substitution. In this way we obtained 5,586 derivational links from adjectives to adverbs. Similarly, we generated 5,863 nouns derived from adjectives by the suffix *-ost*, and 930 derivational links that are based on a Latin or Greek prefix, such as *super-* or *meta-*. However, in all the cases hand-coded lists of exceptions were needed.

4.6. Automatically discovered set of derivation rules

As it was already mentioned, there are literally hundreds of derivation types in Czech. We tried to extract some of the transformation rules automatically. The procedure was based on the assumption that if two words share a sufficiently long sequence of characters, they are more likely to be derivationally related. Instances of such pairs were extracted from the list of lemmas and used for acquiring more general suffix substitution rules. 35 most reliable rules were selected manually out of around 400 automatically inferred derivation rules, equipped with lists of exceptions and applied on the network.
This round was limited only to replacing suffixes, while consonant and vowel changes, vowel insertion/deletion and decaptalization were not included. Thus, only regularly derived words of very frequent word-formation types were covered; for instance, regularly derived deverbal nouns with the suffix *-ní*/*-tí*, possessive adjectives, or female variants of surnames (suffix *-ová* in Czech).
This approach proved to be very efficient, as it brought 13,957 new derivational links.

4.7. Grammar-based set of derivation rules

A list of substitution rules was compiled manually from the description of word-formation in a representative grammar book of Czech (Karlík et al., 2000). The grammatical

description is organized, as a first-level criterion, according to the part of speech of the derivatives (nouns, adjectives, verbs, pronouns, and adverbs), followed by the part of speech of the base words (nouns from nouns, nouns from adjectives etc.) and, finally, by the meaning of the derivatives (i.e. in accordance with Dokulil's concept of word-formation types); within each word-formation type, individual affixes were listed from the most productive to peripheral ones. In the grammar book, base words for each affix were delimited by general semantic features (and/or morphological and/or phonological features) and exemplified with several examples, which were the only source of specific information on changes between the base and derived word in many cases.

Thus, on the one hand, the original grammatical description seems to be rather overspecified since some affixes are described several times due to semantic nuances that do not mirror in formal changes during the derivation (for instance, the suffix *-ák* is listed twice as a means of deverbal derivation: among agent nouns, as in *honák* 'cattle drover' derived from *honit* 'to drive', and among instruments of activities, e.g. *sušák* 'dryer' from *sušit* 'to dry'), and rather underspecified on the other hand, especially with regard to coinages that undergo individual changes with regard to the base word (e.g. the suffix *-enka* is recorded as deriving denominal names of various cards and permits, cf. the changes in *jízda* 'driving' – *jízdenka* 'ticket', *místo* 'seat' – *místenka* 'seat reservation ticket', *účet* 'invoice'– *účtenka* 'bill', *povolení* 'permission' – *povolenka* 'permit' etc. that all must be formalized as separate substitution rules in order to find the corresponding candidate pairs).

In this round, vowel insertion/deletion (*e* is mostly inserted as in *léčba* 'treatment'– *léčebna* 'sanatorium'; *k* is deleted during the derivation of the adjective *vimperský* from the town name *Vimperk*), decaptalization (cf. the adjective *vimperský*) and vowel and consonant changes were involved. There were 18 vowel changes (mostly vowel lengthening/shortening, e.g. *žít* 'to live' – *žití* 'living', *list* 'leaf' – *lístek* 'leaflet') and 11 consonant changes (mostly into palatal counterparts, e.g. *stuha* 'ribbon' – *stužka* 'small ribbon') integrated into the process of generation of the base-target pairs. Vowel changes typically occur within the stem of the word whereas consonant changes mostly appear on the stem-suffix boundary. Those changes that affect very frequent word-formation types were specified within substitution rules so that, for instance, specialized rules such as *V-at→N-ání* (*adaptovat* 'adapt' – *adaptování* 'adapting') and *V-át→N-aní* (*brát* 'to take' – *braní* 'taking') were preferred to a general rule *V-t→N-ní*.

4.8. Application of the grammar-based rules

The compiled list of substitution rules was checked for uniqueness; for instance, nearly 50 rules (that occur more than once in the grammatical description since they belong to several word-formation types) were deleted from the starting list of more than 500 substitution rules describing derivation of nouns. The resulting list of rules was automatically mapped on the lexemes of the network. Though a careful inclusion of vowel and consonant changes, it was necessary to attach a list of irregularly derived words.

Nodes	
N	148 296
A	80 037
V	22 186
D	15 660
Total	266 179
Derivation relations	
N2A	21 646
V2A	16 140
V2N	13 523
N2N	9 007
A2N	8 999
A2D	7 587
V2V	1 449
A2A	353
D2D	13
A2V	10
V2D	8
Total	78,735

Table 1: Distribution of nodes according to the part of speech of lexemes (N - noun, A - adjective, V - verbs, D - adverb), and distribution of derivational links according to the part of speech of base and derived lexemes.

The generated output data (base-target pairs) was divided into two lists. The first of them contained derivatives for which a single base word was suggested. The second list contained derivatives for which more than one base word was generated. Both lists were manually checked; in the first of them, the base-target pairs were either confirmed or rejected, in the second list incorrect pairs were marked. The manual process resulted in a new (shorter) list of rules and exceptions.

The application of the hand-coded derivation rules led to 4,253 derivational links.

4.9. Increasing inter-cluster consistency

Having gathered a sufficiently large amount of derivational links, we could have started exploring some regular patterns frequently appearing in derivation clusters. For instance, if two clusters contain two sets of lexemes derived by the same set of affixes, then the clusters are likely to have the same internal structure. This observation does not bring new derivational links, but makes the network more consistent, not only in the sense of removing errors, but also in systematizing the annotation in places in which linguistic intuition allows multiple interpretations of the derivation process. Using this approach, 760 clusters were internally restructured.

4.10. Statistical properties of the network

The dataset resulting from the procedure described above is called DeriNet Version 0.5. In this version, DeriNet contains 266,179 Czech lexemes and 78,735 derivational links that connect base and derived lexemes. Selected quantitative properties of the network data are shown in Table 1.

The size of the clusters in the network varies greatly. Obviously, most clusters contain just one lexeme, but one can

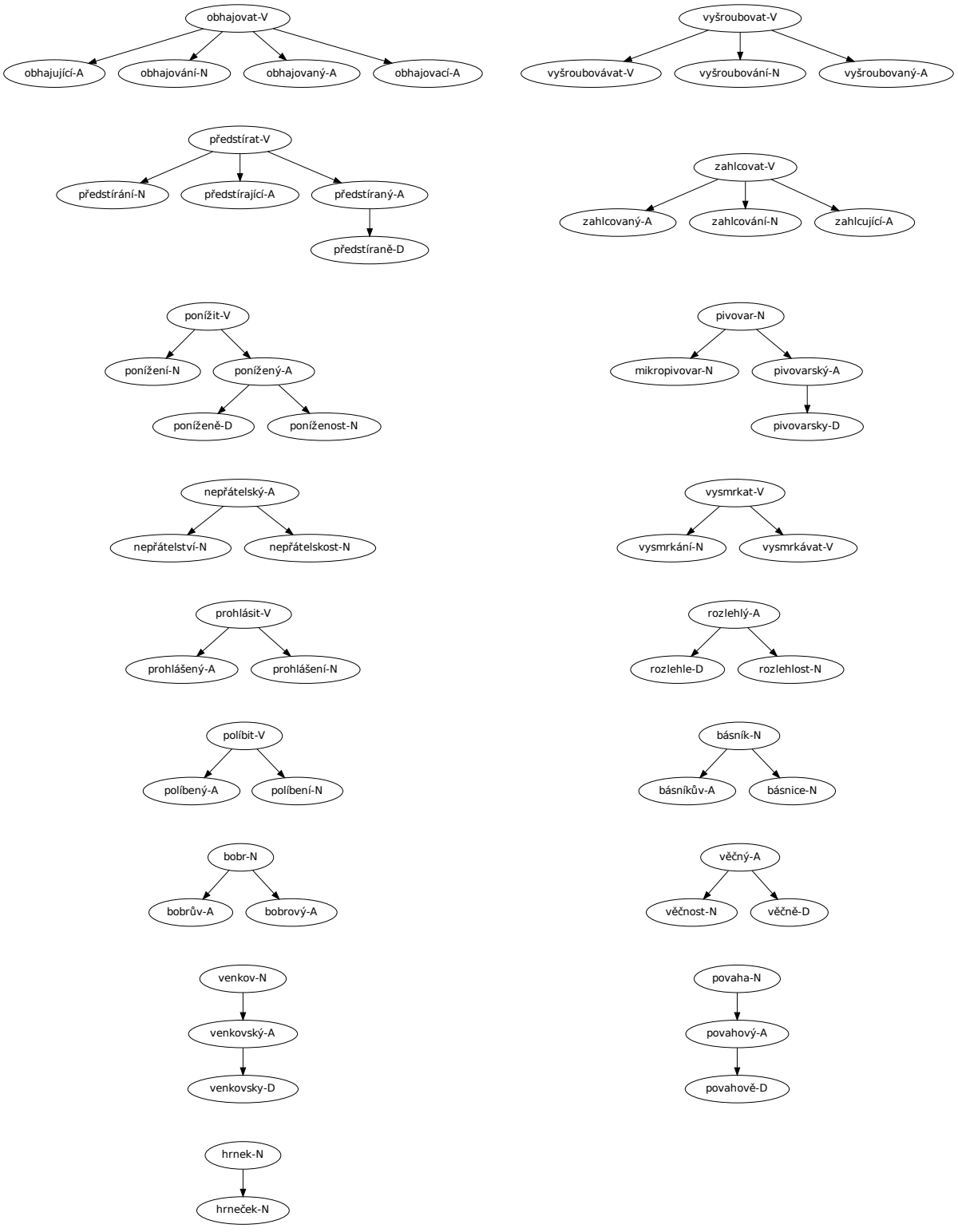

Figure 1: Examples of diverse shapes of derivation clusters. Nodes in the clusters are labeled with their lemmas and parts of speech.

find clusters containing more than 20 lexemes. The same hold for internal structure of clusters in the sense of cluster depth and width: there are many clusters containing five or more derived lexemes on the same level, while chains with depth of five or more can be found too. The diversity of cluster shapes is illustrated by Figure 1.

It should be emphasized that the given statistics mirror rather the current state of the building-up process of the

network rather than characteristics of Czech lexicon. There are many types of derivational relations which we have not captured in the network so far, and thus it is very likely that the network will become more densely connected in the future. For example, verb prefixation and aspectual pairs are expected to govern tens of thousands of derivational links which are not yet covered in the current version of the network.

4.11. Data distribution

DeriNet Version 0.5 is publicly available on the Internet at `http://ufal.mff.cuni.cz/derinet`. It can be used under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License (CC-BY-NC-SA). The data is available in a simple line-oriented format as well as in a self-documenting XML-based format.

5. Envisaged usage of the network

5.1. Derivational information within Machine Translation

The primary motivation of our effort lies in NLP applications. Let us illustrate it on two potential usages of the network in Machine Translation.

First, in syntactically informed Machine Translation systems (such as TectoMT, (Popel and Žabokrtský, 2010)), sometimes part of speech of available translation equivalents does not fit the target sentence syntactic context. For instance, an adjectival attribute of an English gerund is usually translated as a Czech adverb if the gerund phrase is translated as a subordinating clause. If the adjective-adverb pair is not available in the translation component (which happens often for rare words, no matter how large the training parallel corpus is), then our network can come to help. In other words, the network can be used for adding previously unseen translation pairs.[2]

Second, partitioning of the lexical space induced by the transitive closure of the derivational relation can be used for generating less sparse lexical features for discriminative translation models, such as for the one described in (Mareček et al., 2010).

5.2. Linguistic research in productivity in word-formation

As already mentioned, the network is expected to be a reliable resource of data for theoretical research into word-formation, in particular for research into productivity, which has become one of the central issues of word-formation research at least in the last three decades.

Besides the quantitative features of derivatives with a certain affix, which can be extracted from a corpus (number of hapax legomena, token and type frequency of the derivatives), information on base words is required in many approaches in order to obtain a more complex picture of productivity of an affix (cf. Aronoff's concept of possible words (Aronoff, 1976), systemic productivity in (Dokulil, 1962) or qualitative productivity in (Lüdeling and Evert, 2005)). Such information can be acquired neither from corpora nor from dictionaries since the base words of a particular suffix cannot be delimited without a specialized lexical resource.

Using the word-formation network, the following (linguistic) questions can be easily answered for Czech derived words:

- What are the base words of words with a particular affix? (Which part of speech do the base words belong to? Are they non-derived (primary) vs. derived words?)
- What are the derivatives of a word / of a group of words? (Are there any at all? – e.g. possessive adjectives are terminal nodes in the network. Which part of speech do the derived words belong to? Are they derived by suffixes, prefixes?)
- What is the type frequency of derivatives of a certain type?

6. Conclusion

In this paper, we have introduced a new language data resource focused on derivational morphology in Czech. The DeriNet network subsumes many different types of derivational relations and is being further extended.

Besides many "standard" lexicographic problems (such as the question which words should be included into the network, how polysemy and existence of orthographic variants should be captured, what should be done with word capitalization), we were facing several derivation-specific design questions, too. For instance, in some cases it was hard to choose derivation direction for a pair of related words (even worse, sometimes there is no correct answer, e.g., if they both are created from a base word that does not exist in the contemporary language any more).

Another question arises whether and how the current network model could be extended to combined or compounding word-formation, in which a new word is created from two or more words.

7. Acknowledgements

This research has been supported by GA ČR P406/12/P175. The work has been using language resources developed and/or stored and/or distributed by the LINDAT-Clarin project of the Ministry of Education of the Czech Republic (project LM2010013).

8. References

Mark Aronoff. 1976. *Word Formation in Generative Grammar*. The MIT Press, Massachussetts.

Harald Baayen. 1992. Quantitative aspects of morphological productivity. In G. Booij and J. van Marle, editors, *Yearbook of Morphology 1991*, pages 109–149, Dordrecht. Kluwer.

Harald Baayen. 1993. On frequency, transparency, and productivity. In G. Booij and J. van Marle, editors, *Yearbook of Morphology 1992*, pages 181–208, Dordrecht. Kluwer.

[2] Of course, such translation pairs are not fully equivalent to those seen in the parallel data, as they come without any statistics; the needed redistribution of probability mass can be only approximated.

CNC. 2000. Czech National Corpus – SYN2000, Institute of Czech National Corpus, Faculty of Arts, Charles University in Prague.

CNC. 2010. Czech National Corpus – SYN2010, Institute of Czech National Corpus, Faculty of Arts, Charles University in Prague.

CNC. 2014. Czech National Corpus – SYN, Institute of Czech National Corpus, Faculty of Arts, Charles University in Prague.

František Daneš, Miloš Dokulil, and Jaroslav Kuchař. 1967. *Tvoření slov v češtině 2: Odvozování podstatných jmen*. Academia, Prague.

Miloš Dokulil, Karel Horálek, Jiřina Hůrková, Miloslava Knappová, and Jan Petr. 1986. *Mluvnice češtiny 1*. Academia, Prague.

Miloš Dokulil. 1962. *Tvoření slov v češtině 1: Teorie odvozování slov*. Academia, Prague.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková-Razímová. 2006. *Prague Dependency Treebank 2.0*. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.

Jan Hajič. 2004. *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Karolinum, Prague.

Přemysl Hauser. 1986. *Nauka o slovní zásobě*. SPN, Prague.

Petr Karlík, Marek Nekula, and Zdeňka Rusínová. 2000. *Příruční mluvnice češtiny*. NLN, Prague.

Jerzy Kuryłowicz. 1936. Dérivation lexicale et dérivation syntaxique. *Bulletin de la Société de linguistique de Paris*, 37:79–92.

A. Lüdeling and S. Evert. 2005. The emergence of productive non-medical *-itis*. Corpus evidence and qualitative analysis. In S. Kepser and M. Reis, editors, *Linguistic Evidence. Empirical, Theoretical and Computational Perspectives*, pages 351–370, Mouton De Gruyter. Berlin – Boston.

David Mareček, Martin Popel, and Zdeněk Žabokrtský. 2010. Maximum entropy translation model in dependency-based MT framework. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 201–201, Uppsala, Sweden. Uppsala Universitet, Association for Computational Linguistics.

Karel Pala and Dana Hlaváčková. 2007. Derivational Relations in Czech WordNet. In *Proceedings of the 2007 ACL Workshop on Balto-Slavonic Natural Language Processing*, pages 75–81, Prague.

Karel Pala and Pavel Smrž. 2004. Building Czech Wordnet. *Romanian Journal of Information Science and Technology*, 7:79–88.

Karel Pala, Radek Sedláček, and Marek Veber. 2003. Relations between inflectional and derivation patterns. In *Proceedings of the 2003 EACL Workshop on Morphological Parsing of Slavic Languages*, pages 1–8, Budapest.

Martin Popel and Zdeněk Žabokrtský. 2010. TectoMT: Modular NLP framework. In Hrafn Loftsson, Eirikur Rögnvaldsson, and Sigrun Helgadottir, editors, *Lecture Notes in Artificial Intelligence, Proceedings of the 7th International Conference on Advances in Natural Language Processing (IceTAL 2010)*, volume 6233 of *Lecture Notes in Computer Science*, pages 293–304, Berlin / Heidelberg. Iceland Centre for Language Technology (ICLT), Springer.

Magda Razímová and Zdeněk Žabokrtský. 2006. Annotation of Grammatemes in the Prague Dependency Treebank 2.0. In *Proceedings of the LREC 2006 Workshop on Annotation Science*, pages 12–19.

Radek Sedláček and Pavel Smrž. 2001. A New Czech Morphological Analyzer *ajka*. In *Proceedings of the 4th International Conference Text, Speech and Dialogue (TSD 2001)*, pages 100–107, Železná Ruda.

Drahomíra Spoustová, Jan Hajič, Jan Votrubec, Pavel Krbec, and Pavel Květoň. 2007. 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, ACL 2007*, pages 67–74, Praha.

František Čermák. 2012. *Morfématika a slovotvorba češtiny*. NLN, Prague.

Vladimír Šmilauer. 1971. *Novočeské tvořní slov*. SPN, Prague.

František Štícha. 2012. Miloš Dokulil and his theory of productivity in word-formation. *Korpus – gramatika – axiologie*, 6:3–9.

Marion Weller, Max Kisselew, Svetlana Smekalova, Alexander Fraser, Helmut Schmid, Nadir Durrani, Hassan Sajjad, and Richrd Farkas. 2013. Munich-Edinburgh-Stuttgart Submissions at WMT13: Morphological and Syntactic Processing for SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 232–239, Sofia, Bulgaria.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 2013. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1201–1211, Sofia, Bulgaria.