Merging Data Resources for Inflectional and Derivational Morphology in Czech

Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra, Adéla Limburská
Charles University in Prague, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
{zabokrtsky,sevcikova,straka,vidra}@ufal.mff.cuni.cz, limburska@seznam.cz

Abstract

The paper deals with merging two complementary resources of morphological data previously existing for Czech, namely the inflectional dictionary MorfFlex CZ and the recently developed lexical network DeriNet. The MorfFlex CZ dictionary has been used by a morphological analyzer capable of analyzing/generating several million Czech word forms according to the rules of Czech inflection. The DeriNet network contains several hundred thousand Czech lemmas interconnected with links corresponding to derivational relations (relations between base words and words derived from them). After summarizing basic characteristics of both resources, the process of merging is described, focusing on both rather technical aspects (growth of the data, measuring the quality of newly added derivational relations) and linguistic issues (treating lexical homonymy and vowel/consonant alternations). The resulting resource contains 970 thousand lemmas connected with 715 thousand derivational relations and is publicly available on the web under the CC-BY-NC-SA license. The data were incorporated in the MorphoDiTa library version 2.0 (which provides morphological analysis, generation, tagging and lemmatization for Czech) and can be browsed and searched by two web tools (DeriNet Viewer and DeriNet Search tool).

Keywords: derivation, inflection, morphology

1. Introduction

The present paper deals with merging two complementary resources of morphological data previously existing for Czech: first, the inflectional dictionary MorfFlex CZ used by a morphological analyzer capable of analyzing/generating several million Czech word forms according to the rules of Czech inflection (Hajič and Hlaváčová, 2013), and second, the recently developed lexical network DeriNet that stores derivational relations among several hundred thousand Czech lemmas (Ševčíková and Žabokrtský, 2014). The resulting publicly available data resource interconnects both types of information and thus allows exploiting the two morphological phenomena in NLP applications in a unified way. In addition, using the full inflectional dictionary led to a considerable growth of both recall and precision of the captured derivational relations.

2. Related work

Czech is a language with both a rich inflectional system and a complex derivational morphology. In the theoretical description, inflectional and derivational morphology have been traditionally separated from each other, derivations being described as a part of the word-formation system of Czech (in addition to compounding and combined word-formation processes); e.g. Dokulil et al. (1986), Karlík et al. (1995), Štícha (2013). The roots of the separation are traced back by Bednaříková (2009), who states that in the representative, "academic" grammars of Czech (Dokulil et al., 1986; Komárek et al., 1986) "word formation is seen as a sort of transition zone between morphology and lexicon" and "word formation means are considered non-morphological means, along with lexical means, means of word order and intonation etc."

Although the linguistic tradition influenced also developers of morphological software tools for Czech, who mostly approached inflection and derivation in different ways, some attempts to merge inflectional morphology with derivations have been already done. Besides the morphological analyzer which is based on the MorfFlex CZ dictionary (used in our present approach; see Section 3.), derivations were handled in the morphological analyzer Ajka/Majka and derived software tools (Sedláček and Smrž, 2001; Šmerk et al., 2007). The dictionary of the morphological analyzer can be searched for derivational related pairs (or n-tuples) by the Deriv tool (Osolsobě et al., 2009) using rules based on regular expressions, not on grammatical features. To reveal a list of pairs of verbs and derived nouns (such as *žehlit* 'to iron' > *žehlička* 'iron', *šít* 'to sew' > *šička* 'seamstress'), a set of rules has to be written that includes a separate rule for every vowel or consonant alternation involved. There is a more recently developed successor of Deriv called Derivancze (Pala and Šmerk, 2015), which is a web application that finds the base lemma for a given lemma as well as the list of its derived lemmas. Derivancze captures more than 255 thousand derivational links, classified into 17 semantically motivated types.

Another tool, Morfio, has been developed to search for pairs (or n-tuples) with a formally identical base and different formants in the Czech National Corpus (Cvrček and Vondřička, 2013). The tool makes it possible to include the most common alternations into the queries; however, it suffers from overgeneration as it is not possible to condition the alternations by the context.

The DeriNet network, which is – as a part of our approach – described in Section 4., has been developed as a publicly accessible, large-coverage resource of Czech derivational data and is, to the best of our knowledge, one of only few specialized resources of derivational data even in a broader context of different langauges. Most of the resources are very recent such as DerivBase for German

(Zeller et al., 2013), DerivBase.Hr for Croatian (Šnajder, 2014), Démonette network for French (Hathout and Namer, 2014), or the language-independent approach by Baranes and Sagot (2014).

3. Inflectional resource: MorfFlex CZ

MorfFlex CZ is a Czech morphological dictionary developed originally by Jan Hajič as a spelling checker and lemmatization dictionary; it is a plain list of lemma-tag-form triples and the latest version contains nearly 985 thousand unique lemmas and more than 120 million word forms (Hajič and Hlaváčová, 2013). For each word form, full inflectional information is available, using the positional tagging scheme proposed by Hajič (2004). In addition, lemmas can also contain basic derivational, semantic and named entity information.

The dictionary deals with lexical homonymy to a limited extent but it tries to avoid lexical distinctions that are not morphologically based. It does, for instance, distinguish between formally identical (homonymous) words of different part-of-speech categories. Each of the homonymous lemmas is identified by a base form (lemma) followed by a number; cf. the lemmas *podle-1* and *podle-2* assigned in MorfFlex CZ to the string *podle* which can be either an adverb (meaning 'meanly') or a preposition ('along').

3.1. Internal representation of MorfFlex CZ

The MorfFlex CZ itself is a flat list of lemma-tag-form triples (see Fig. 1), so that it can be used easily in applications requiring Czech morphology.

However, MorfFlex CZ is generated from a structured representation, which is based on paradigm system described by Hajič (2004). A paradigm represents several inflectional forms of a lemma. It represents either all inflectional forms of a lemma (a full paradigm), or only some of them (a partial paradigm). Partial paradigms are needed in Czech because many lemmas do not belong to a full paradigm, but have a group of regular inflections.

Technically, a paradigm is a set of ending-tag pairs. A lemma then includes several entries, each consisting of a (technical) root and a paradigm. For each root and paradigm, a set of form-tag pairs is created by catenating the root with the paradigm endings and using its tags. For completely irregular words, there exists a special paradigm `0` with one empty ending, which is used to represent one word form with a specified tag. See an example of a paradigm in Fig. 2 and entries for the above mentioned lemmas *podle-1* and *podle-2* in Fig. 3.

Paradigms and lemma entries use a different tag set than the MorfFlex CZ entries. Instead of positional tags consisting of 15 characters, so-called compact tags described by Hajič (2004) are used. The compact tags are isomorphic to the positional tags but drop unused positions for the given part-of-speech category, and are easier to work with once you become familiar with them.

```
podle-1_^(*3ý-1) Dg-------3N---6 nejnepodlejc
podle-1_^(*3ý-1) Dg-------3N---- nejnepodleji
podle-1_^(*3ý-1) Dg-------3A---6 nejpodlejc
podle-1_^(*3ý-1) Dg-------3A---- nejpodleji
podle-1_^(*3ý-1) Dg-------1N---- nepodle
podle-1_^(*3ý-1) Dg-------2N---6 nepodlejc
podle-1_^(*3ý-1) Dg-------2N---- nepodleji
podle-1_^(*3ý-1) Dg-------1A---- podle
podle-1_^(*3ý-1) Dg-------2A---6 podlejc
podle-1_^(*3ý-1) Dg-------2A---- podleji
podle-2 RR--2---------- podle
```

Figure 1: Entries for lemmas *podle-1* ('meanly') and *podle-2* ('along') in the MorfFlex CZ dictionary, each lemma-tag-form triple on a separate line. The technical suffix `_^(*3ý-1)` ("remove 3 characters and add `ý-1`") encodes the base lemma (i.e. adjective *podlý-1*) of the adverb.

```
ev
  e[DG1@], eji[DG#@], ejc[DG#@-6]
```

Figure 2: The paradigm `ev` with three endings in the MorfFlex CZ dictionary.

```
podl ev =podle-1
podle 0 =podle-2+R2
```

Figure 3: Paradigm entries for the lemmas *podle-1* ('meanly') and *podle-2* ('along') in the MorfFlex CZ dictionary.

In Czech, negation is often formed by the *ne-* prefix (as in *nepodle* 'unmeanly'), and superlative forms are often derived from comparative forms by the prefix *nej-* (e.g., *nejpodleji* 'most meanly' derived from *podleji* 'more meanly'). Therefore, every paradigm ending has two additional boolean flags: whether it can form a negation and whether it can form a superlative (if it can form both, *nejne-* prefix is used). These flags are indicated by using `@` and `#` characters in the tag, which are then replaced by `A`/`N` for positive/negative variant, and `2`/`3` for comparative/superlative, respectively. That is why 10 entries for the lemma *podle-1* in Fig. 1 are generated based on three endings of the `ev` paradigm shown in Fig. 2.

3.2. Derivational information in MorfFlex CZ

A substantial part of the lexicon of Czech could be mapped onto so-called derivational paradigms in the MorfFlex CZ dictionary. If a lemma belongs to a derivational paradigm, several other lemmas can be derived from it. See the derivational paradigm `ye` in Fig. 4. Using the `ye` derivational paradigm, the representation of the lemmas *podle-1* and *podle-2* in the MorfFlex CZ dictionary is displayed in Fig. 5.

All lemmas created by a derivational paradigm have the derivational information stored as a technical suffix of the lemma (see Fig. 1). Even though the derivational information is available for 665 thousand lemmas in the MorfFlex CZ dictionary, only the most regular types of derivatives with a transparent derivational structure are assigned a respective technical suffix.

If a lemma has a technical suffix encoding derivational information, it refers to a lemma from which the current lemma was derived. In the latest released version of MorfFlex CZ, any such lemma can be referenced. However, following the needs of proper representation of derivational

```
ye 0,ye,0,0,0,0,*
ye 0,ev,r0,e,0,0,-
ye ost,kt1n,r0,ost,0,0,-
```

Figure 4: The derivational paradigm `ye`, from which three full (non-derivational) paradigms are derived: `ye`, `ev` (with an unmodified root and a lemma created from the original root by adding the suffix *-e*) and `kt1n` (with the suffix *-ost* added to the original root and a lemma being the same as the new root).

```
podl ye =podlý-1
podle 0 =podle-2+R2
```

Figure 5: Paradigm entries for the lemmas *podle-1* ('meanly') and *podle-2* ('along') based on the derivational paradigm `ye` from Fig. 4.

morphology, the newest version of MorfFlex CZ has been changed to refer to the direct predecessor of the given lemma.

4. Derivational resource: DeriNet

The DeriNet network was used as the source of information on derivational morphology in Czech (Ševčíková and Žabokrtský, 2014). In DeriNet, relations between derived words and their base words are modeled as an oriented graph. Nodes of the graph correspond to lemmas (including the information on their parts of speech). Edges represent derivational steps between lemmas. The orientation of edges reflects the process of derivation: the edge points from the base lemma to the derived one. Each lemma has at most one base lemma.

DeriNet grows gradually; the most important steps can be summarized as follows:

- The first version of the network of derivational relations (DeriNet 0.1) was created as a result of a pilot study which focused on certain types of deadjectival nouns only.
- DeriNet 0.5 already contained lemmas of four basic parts of speech: nouns, adjectives, verbs and adverbs. The main selection criterion for choosing lemmas was their frequency in the Czech National Corpus (CNC, 2014). As the repertoire of morphological means used for derivation in Czech is very broad, the derivational links were generated using a pipeline of various methods, from applying highly reliable suffix substitution patterns, through automatically detected substitution patterns accompanied with a few manually identified exceptions, to lists of derivational pairs that were mostly assembled (or at least confirmed) manually.
- DeriNet 0.9 was not extended substantially, but rather improved in terms of quality. This version was used as the input for merging with the inflectional data for Czech.
- DeriNet 1.0 is the result of merging the DeriNet version 0.9 with the MorfFlex CZ dictionary; the process is described in Sections 5.1. to 5.3.

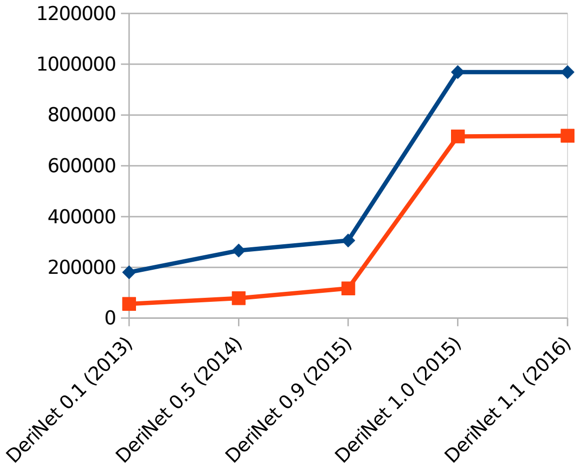


Figure 6: Growth of the DeriNet network in terms of the number of lemmas (nodes; the upper line) and derivational relations (edges; the lower line).

- DeriNet 1.1, the most recent version of the network, adds derivational links between lemmas in which consonant and vowel alternations are identical with those appearing in their inflectional paradigms; see Sect. 5.4. for more details.

The development of DeriNet in terms of its size is documented in Fig. 6.

5. The process of merging

5.1. Unifying the list of lemmas

The biggest change resulting from merging the inflectional dictionary MorfFlex CZ with the derivational network was the large number of lemmas not previously processed in DeriNet. MorfFlex CZ contained 985 thousand lemmas, while DeriNet in version 0.9 was roughly three times smaller (around 310 thousand lemmas) and — with the exception of around 20 lemmas — was a subset of the MorfFlex CZ dictionary.

5.2. Discovering new derivations

As mentioned above, the DeriNet network has been generated by a pipeline of tools in which various kinds of rules (especially regular-expression based suffix substitution rules) were combined with lists of positive and negative exceptions and other manual annotations. Thus, it was technically easy to apply the pipeline with only minor modifications on the larger set of MorfFlex CZ lemmas, but the lists of exceptions had to be updated manually. About 27 thousand derivational links were changed in this way. The resulting derivational network was labeled as DeriNet 1.0.

Linguistically, we hypothesized that the density of exceptions is considerably lower among lemmas that were not present in DeriNet 0.9 due to general language economy principles – exceptions are worth remembering rather for more frequent language units. Given that the selection criterion for lemmas in DeriNet 0.9 was based on their frequency in the Czech National Corpus, most irregularly derived words were probably captured already in DeriNet 0.9, and the complement lemmas can be relatively reliably detected by automatic rules. In other words, we expected that the less frequent a lemma is, the more regularly it is derived from the viewpoint of derivational patterns. However, this was a working hypothesis and had to be proved empirically first.

5.3. Quality measurement

In order to measure the quality of DeriNet 1.0 in comparison with DeriNet 0.9, we annotated two gold-standard sets of 1,000 lemmas each. For every lemma, a set of base lemmas (usually a single one, but there can be several, in case of ambiguity, or none may exist) was manually specified.

The first set, utilized as development data during creation of DeriNet 1.0, was randomly selected from DeriNet 0.9. The second set, used for evaluation only, was randomly selected from DeriNet 1.0. This poses a problem with interpretation of results: the development set is directly applicable to DeriNet 0.9, but it cannot be used for unbiased evaluation of DeriNet 1.0; while the evaluation set contains many lemmas not present in the old version and therefore is not directly applicable to it.

However, when ignoring nonexistent lemmas, both sets produce similar results on DeriNet 0.9, even though the evaluation set is restricted to just 339 lemmas. We have therefore used the evaluation set for measuring precision and recall on both versions.

We measured the quality in terms of edge-wise precision and recall. Although few manual annotations were added specifically for the version 1.0, precision did not decrease and stayed in the 98–99% range. Recall increased from 75 % to 85 %.

There were two main sources of this improvement. First, our observations confirmed that less frequent words are derived regularly. Second, automatized rules used for building both DeriNet 0.9 and 1.0 predicted some false derivations (a wrong base lemma for a derived lemma) in the version 0.9, because the correct base lemma was simply not present in the network.

5.4. Mapping vowel/consonant alternations from inflections onto derivations

Derivation in Czech, especially suffixation, is often accompanied by consonant and vowel changes at the root-suffix boundary (e.g. *hřích* 'sin' > *hříšný* 'sinful'), or within the root (*list* 'leaf' > *lístek* 'small leaf'), or even in both positions simultaneously, cf. the alternations *h–ž* and *í–ě* in *sníh* 'snow' > *sněžný* 'snowy'.

As experienced during the creation of the DeriNet network, alternations can neither be omitted since they affect a substantial part of derivations in Czech, nor applied mechanically. For instance, there are vowel alternations *a–á* and *á–a* that are applied e.g. in derivations *vrata* 'gate' > *vrátný* 'porter' and *vrátit* 'return' > *vratný* 'returnable' respectively, but cannot be applied to the nouns *slavistika* 'Slavic studies' and *slávista* 'supporter of the sport club Slávie' (cf. the derivatives *slavistický* 'Slavic' and *slávistický* 'belonging to supporters of Slávie'). Another example is the *k–č* alternation at the root-suffix boundary (e.g. *matrika* 'register' > *matriční* 'registry (book)') which is not present in the adjective *matiční* that was derived from *matice* 'foundation', not from *matika* 'math'. This shows that implementing alternations just as simple substitution rules would lead to serious overgeneration of derivational links.

The interconnection of the MorfFlex CZ dictionary with the DeriNet network made it possible to employ the fact that some of the root alternations that occur during derivation are present in the inflectional paradigms of the respective words as well; for instance, the alternation *í–ě* in *sníh* 'snow' > *sněžný* 'snowy' can be observed in the genitive singular form of the noun (*sníh*.nom-sg – *sněhu*.gen-sg).

First of all, a list of lemmas with alternating vowels and/or consonants in the inflectional paradigm was automatically extracted from MorfFlex CZ. Second, the list was used to search for pairs of base-target words which correspond to suffix substitution patterns (which were part of the pipeline used for compilation of the network). Only manually confirmed pairs have been included into DeriNet; this extension was published as DeriNet 1.1. Although the number of newly added links was relatively small (roughly 3,000 new links), it enriches the diversity of DeriNet with previously unnoticed types of derivations.

5.5. Resolving homonymy

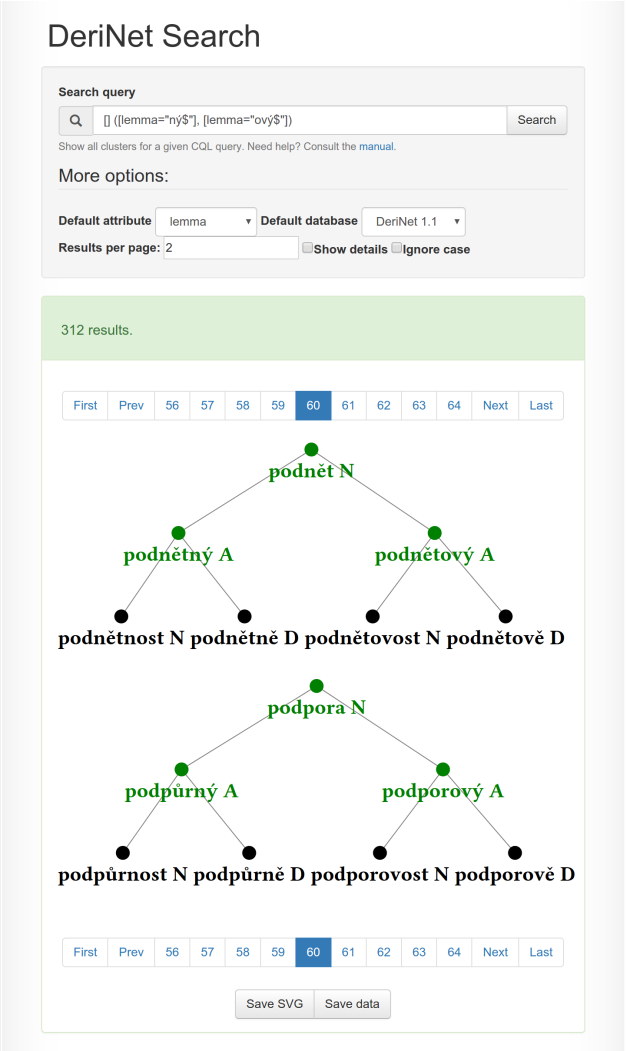
Representation of derivational relations in a natural language cannot dispense with a consistent approach to lexical ambiguity, especially of polysemy and homonymy. In the network, a polysemous lemma is represented by a single node and all derivatives of it are listed as its child nodes. By contrast, homonymous lemmas originate in different base words and have different derivatives, so they have to be discerned and represented by separate nodes in the DeriNet network.

The approach to homonymy implemented in MorfFlex CZ, which was very broad without a clear-cut boundary to polysemy, has been revised automatically in the first step, followed by a manual annotation. The resulting set of app. 170 homonymous lemmas, which are represented by two or more nodes in the DeriNet network, includes the following types:

1. formally identical lemmas that are composed of different morphemes; for instance, the verb *proudit* which can either be analyzed as derived from the noun *proud* 'flow' (*proud-it* 'to flow'), or from the verb *udit* 'to smoke' (*pro-udit* 'to smoke thoroughly');

2. lemmas formed from different bases by semantically broad affixes, e.g. adjectival suffixes *-ový* or *-ný* expressing a broad relation to a noun (the adjective *masový* relates both to the noun *masa* 'mass' and *maso* 'meat' in Czech, *vinný* both to *víno* 'wine' and *vina* 'blame').

3. lemmas derived from the same base by a polysemous affix; for instance, the suffix *-ič* occurring regularly


DeriNet Search

Search query: [] ([lemma="ný$"], [lemma="ový$"])

Show all clusters for a given CQL query. Need help? Consult the manual.

More options:

Default attribute: lemma Default database: DeriNet 1.1

Results per page: 2 □ Show details □ Ignore case

312 results.

Figure 7: A screenshot of the DeriNet Search web-based application. The query `[] ([lemma="ný$"], [lemma="ový$"])` searches for adjectives which were derived by two different, but productive suffixes.

both in agentive nouns and in nouns denoting instruments (e.g. *čistič* ‘cleaner’ as a person or an instrument noun which have different inflectional paradigms in Czech and only the former one is the base word for the possessive adjective *čističův* ‘cleaner’s’).

6. The resulting data resources and applicational and user interfaces

The set of derivational links resulting from merging the two data resources was published in fall 2015 as DeriNet 1.0, and is available on the web in the LINDAT/Clarin repository under the CC-BY-NC-SA license (Vidra et al., 2015). The most current version of the data, DeriNet 1.1, and future versions will be published in the LINDAT/Clarin repository as well.

The data of the DeriNet network can also be accessed online by using several recently developed tools.

6.1. Application interfaces

DeriNet data have been incorporated to the 2.0 version of MorphoDiTa (Straková et al., 2014),[1] which is a library providing morphological analysis, morphological generation, tagging and lemmatization with state-of-the-art performance for Czech. The version 2.0 allows traversing derivational trees by providing two methods for climbing the trees up and down (i.e., for a given lemma its derivational antecedent and the list of descendants are returned, respectively).

In addition to traversing a derivational tree directly, MorphoDiTa incorporates derivational information in more high-level way. Every operation producing a lemma (notably morphological analysis and lemmatization) can instead produce either the root of its derivational tree, the whole derivation path from the lemma to the root, or the whole derivational tree containing the lemma.

One can easily exploit this functionality in various NLP tasks. For instance, for query expansion in an information retrieval system, a lemma from a query can be replaced by a bag of all lemmas from the derivational tree, possibly weighted by the length of their path to the original lemma.

In applications in which lexical sparsity is an issue even after lemmatization, the lemmatization step can be “prolonged” effectively by replacing a lemma by the lemma of the root node of the associated derivational tree.

6.2. User interfaces

The functionality described in Sect. 6.1. is available also in the MorphoDiTa web service,[2] both as an online demo and as REST API.

Additionally, the DeriNet data can be browsed and searched online using two recently developed tools, DeriNet Viewer and DeriNet Search tool.

DeriNet Viewer[3] is a simple viewer of derivational trees for given lemmas. It also provides grouping of derivational trees according to their shape and showing simple statistics for the whole derivational network.

DeriNet Search tool[4] is a web-browser application that enables searching for derivational clusters based on their structure and attributes they contain. See Figure 7 for a screenshot of the user interface.

The DeriNet Search tool makes it possible to search the data for a specific lemma by specifying a set of regular expression constraints in square brackets, like so:

```
[attribute1="regex1"
attribute2="regex2" ...]
```

Only lemmas matching all the conditions are selected.

In the DeriNet Search tool, the following node attributes are available:

lemma and *techlemma* (techlemma is the lemma plus technical suffixes as provided by the MorfFlex CZ dictionary),

pos (part-of-speech tag),

[1] http://ufal.mff.cuni.cz/morphodita

[2] http://lindat.mff.cuni.cz/services/morphodita/

[3] http://ufal.mff.cuni.cz/derinet/viewer

[4] http://ufal.mff.cuni.cz/derinet/search

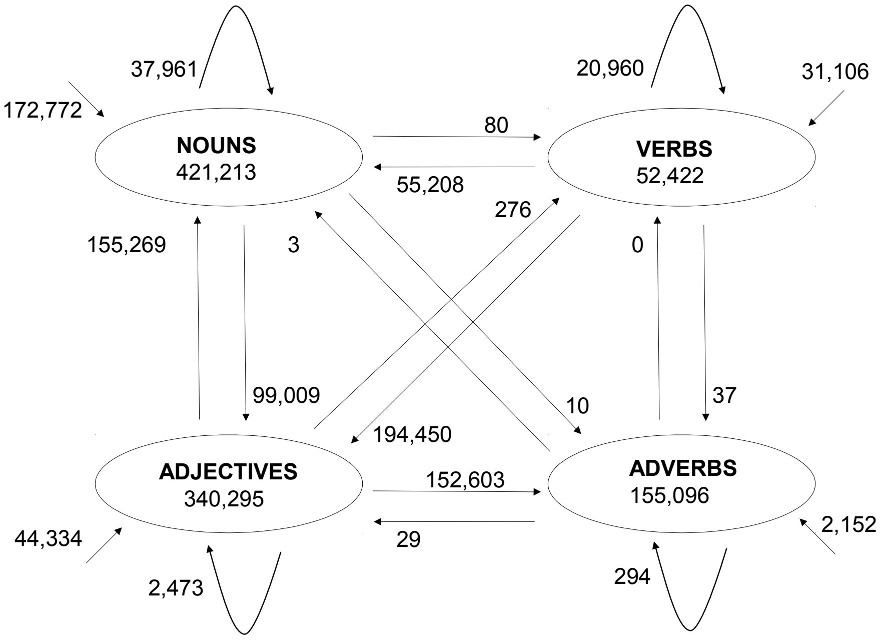


Figure 8: Absolute frequency of DeriNet 1.1 nodes with respect to their part-of-speech category (nodes in the diagram) and absolute frequency of derivational links with respect to part of speech of their base and derived lemmas (arrows in the diagram). The four arrows leading from out of the diagram represent lemmas that are not considered derived in DeriNet 1.1 (i.e., roots of derivational clusters).

id (its internal numerical ID) and

parent (ID of the derivational parent).

Here are several examples of different types of queries which could be formulated in the DeriNet Search tool:

1. The query `[pos="N" lemma=".*ová$"]` selects all clusters containing a noun ending in *-ová* – these are usually female names derived from male variants.

2. Parent-child derivational relations are queried by concatenating several attribute expressions together. The lemma to the left is a parent of the lemma to the right, e.g.
`[lemma="^[a-z].*"] [lemma=".*ův$"]`
selects all clusters, in which a lemma ending in *-ův* (a typical masculine possessive suffix) is derived from a lemma that starts with a lowercase unaccented letter.

3. Multiple children are specified using parentheses and commas: `[] ([lemma=".*ův$"], [lemma=".*ová$"])` selects clusters, where a lemma ending in *-ův* and a lemma ending in *-ová* are derived from the same base. The base lemma is not constrained by any attributes.

7. Conclusions and future work

The present paper describes our recent work on merging the derivational lexical network DeriNet with the Czech inflectional dictionary MorfFlex CZ. The resulting resource, which contains 970 thousand lemmas connected with 715 thousand derivational relations, is currently the biggest publicly available data resource for Czech derivational morphology.

The fact that derivational and inflectional information is available within a single data resource now not only offers new experimental directions (e.g., using root lemmas as a sort of boosted lemmatization in NLP applications), but will bring a new perspective on the two original resources too and will enable new cross-tests of consistency of both that were not considered before.

In addition, we plan to extend the derivational component along several dimensions.

First, a coarse-grained semantic classification of derivational relations, which is currently being implemented into the data, will make it possible to base the search also on semantic criteria. The list of semantic types distinguished in the DeriNet network will involve agentive nouns, feminine nouns, diminutives, possessives, etc.

Second, we are planning to employ a few more data resources from which new derivation links can be extracted, such as the valency lexicon of Czech verbs VALLEX (Lopatková et al., 2015) in which aspectual counterparts (pairs of perfective and imperfective verbs) are connected.

Third, in the more distant future, we would like to go beyond the current one-antecedent limitation in order to be able to capture also compounding.

8. Acknowledgements

This work was supported by the Grant No. 16-18177S of the Grant Agency of the Czech Republic. It has been us-

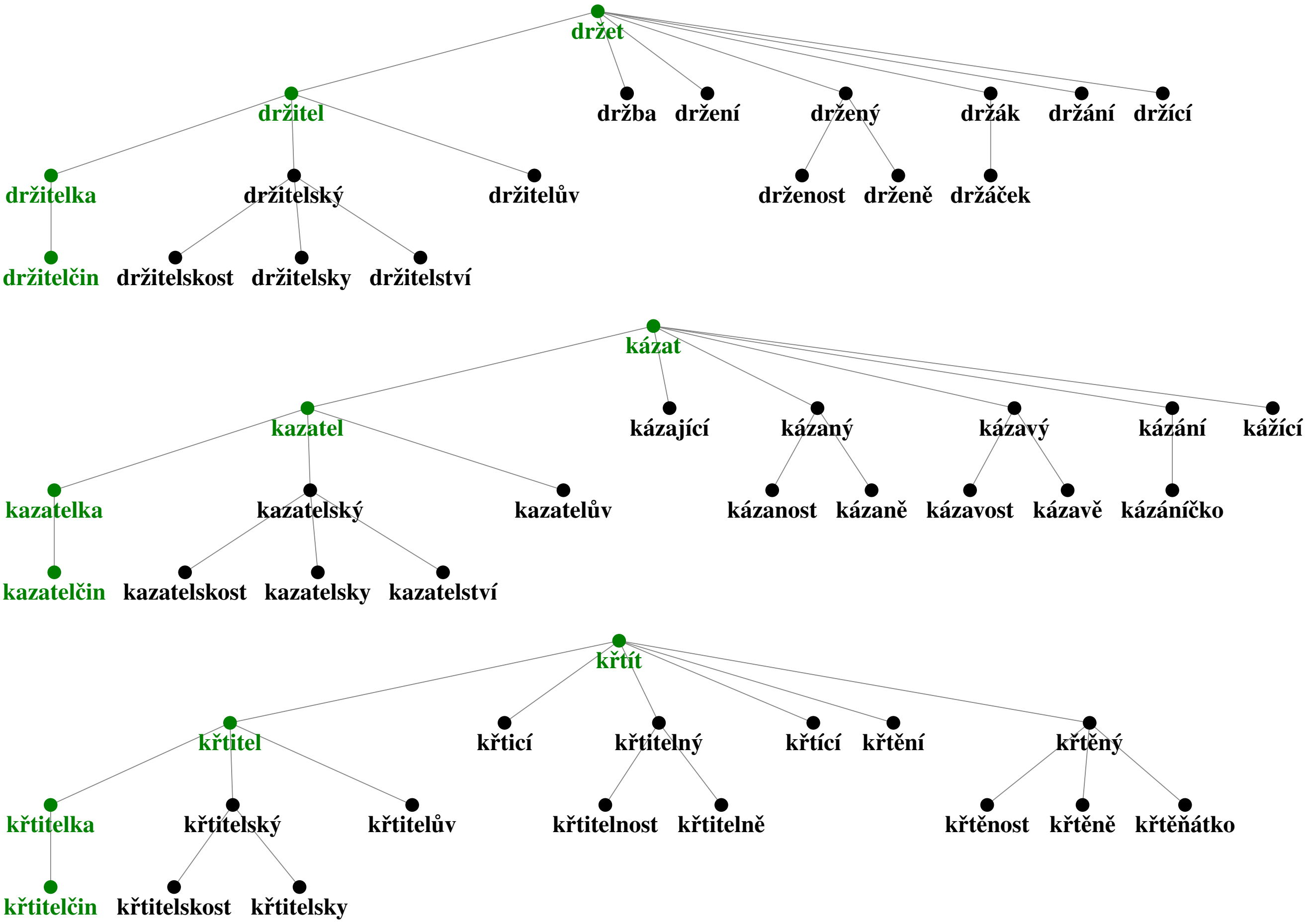


Figure 9: An example of the DeriNet Search output for the query `[pos="V"] [pos="N" lemma="tel$"] [pos="N" lemma="ka$"] [pos="A" lemma="in$"]`, which corresponds to a very productive pattern consisting of a four-node path: (1) a verb (e.g. *držet* ‘to hold’ in the first tree) from which (2) an agentive noun is derived using the *-tel* suffix (*držitel* ‘holder’) which is further turned (3) to a feminine noun by the *-ka* suffix (*držitelka* ‘female holder’) from which (4) a possessive adjective is derived by the *-in* suffix (*držitelčin* ‘female holder’s’).

ing language resources developed, stored, and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2015071).

9. Bibliographical References

Baranes, M. and Sagot, B. (2014). A Language-independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 2793–2799, Reykjavik, Iceland, May.

Bednaříková, B. (2009). *Slovo a jeho konverze*. UPOL, Olomouc.

Cvrček, V. and Vondřička, P. (2013). Nástroj pro slovotvornou analýzu jazykového korpusu. In *Gramatika a korpus*, Hradec Králové. Gaudeamus.

Dokulil, M., Horálek, K., Hůrková, J., Knappová, M., and Petr, J. (1986). *Mluvnice češtiny 1*. Academia, Prague.

Hajič, J. (2004). *Disambiguation of Rich Inflection: Computational Morphology of Czech*. Karolinum Press, Prague.

Hathout, N. and Namer, F. (2014). Démonette, a french derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11(5):125–168.

Karlík, P., Nekula, M., and Rusínová, Z. (1995). *Příruční mluvnice češtiny*. NLN, Prague.

Komárek, M., Kořenský, J., Petr, J., and Veselková, J. (1986). *Mluvnice češtiny 2*. Academia, Prague.

Osolsobě, K., Hlaváčková, D., Pala, K., and Šmerk, P. (2009). Exploring Derivational Relations in Czech with the Deriv Tool. In *NLP, Corpus Linguistics, Corpus Based Grammar Research*, pages 152–161, Bratislava, Slovakia. Tribun.

Pala, K. and Šmerk, P. (2015). Derivancze – Derivational Analyzer of Czech. In *Text, Speech, and Dialogue*, volume 9302 of *Lecture Notes in Computer Science*, pages 515–523. Springer International Publishing.

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

Ševčíková, M. and Žabokrtský, Z. (2014). Word-Formation Network for Czech. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 1087–1093, Reykjavik, Iceland, May.

Šmerk, P., Sojka, P., and Horák, A. (2007). Morphemic analysis: A dictionary lookup instead of real analysis. In *First Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2007*, pages 77–85, Brno. Masaryk University.

Šnajder, J. (2014). DerivBase.hr: A High-Coverage Derivational Morphology Resource for Croatian. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 3371–3377, Reykjavik, Iceland, May.

Štícha, F. (2013). *Akademická gramatika spisovné češtiny*. Academia, Praha.

Straková, J., Straka, M., and Hajič, J. (2014). Open-source tools for morphology, lemmatization, pos tagging and named entity recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June. Association for Computational Linguistics.

Zeller, B., Šnajder, J., and Padó, S. (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.

10. Language Resource References

CNC. (2014). *Czech National Corpus – SYN*. Institute of Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic.

Jan Hajič and Jaroslava Hlaváčová. (2013). *MorfFlex CZ*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Markéta Lopatková and Václava Kettnerová and Eduard Bejček and Anna Vernerová and Zdeněk Žabokrtský. (2015). *VALLEX 3.0 - Valency Lexicon of Czech Verbs*. Charles University in Prague.

Jonáš Vidra and Zdeněk Žabokrtský and Magda Ševčíková and Milan Straka. (2015). *DeriNet 1.0*. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.