Universal Derivations Kickoff: A Collection of Harmonized Derivational Resources for Eleven Languages

Lukáš Kyjánek **Zdeněk Žabokrtský** **Magda Ševčíková** **Jonáš Vidra**
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
{kyjanek,zabokrtsky,sevcikova,vidra}@ufal.mff.cuni.cz

Abstract

The aim of this paper is to open a discussion on harmonization of existing data resources related to derivational morphology. We present a newly assembled collection of eleven harmonized resources named "Universal Derivations" (clearly being inspired by the success story of the Universal Dependencies initiative in treebanking), as well as the harmonization process that brings the individual resources under a unified annotation scheme.

1 Introduction

There are several dozen of language resources that either focus specifically on derivational morphology, or capture some derivational features in addition to other types of annotation. Being rooted in different approaches, the language resources differ greatly in many aspects. This fact complicates usability of the data in multilingual projects, including a potential data-oriented research in derivational morphology across languages. Last but not least, for developers of new data, it can be highly time-consuming to deal with various technical and other issues that somebody else may have already successfully solved.

The current situation with derivational resources is sort of similar to recent developments in treebanking. Efforts have been made to harmonize syntactic treebanks, for instance, in the CoNLL Shared Task 2006 (Buchholz and Marsi, 2006), in the HamleDT treebank collection (Zeman et al., 2014), or in Google Universal Treebanks (McDonald et al., 2013), converging into the Universal Dependencies project (Nivre et al., 2016), and that has become a significant milestone in the applicability of the treebanks.[1]

Being inspired by the harmonization of syntactic treebanks, we harmonized eleven selected derivational resources to a unified scheme in order to verify the feasibility of such undertaking, and to open a discussion on this topic, so far without any specific NLP application in mind. The collection is introduced under the, admittedly imitative, title *Universal Derivations* (UDer).

A brief overview of existing derivational resources and underlying data structures is given in Section 2; some details on the eleven resources to harmonize can be found in Section 3. The harmonization process is described in Section 4, followed by basic quantitative characteristics of the resulting UDer collection (Section 5).

2 Existing data resources for individual languages

Kyjánek (2018) listed 51 resources that capture information on derivational morphology of 22 different languages. The resources differ in many aspects, out of which the most important for us is the data structure, but other essential characteristics include the file format, the size in terms of both lexemes and derivational relations, and the licence under which the data were released.

To be able to compare the resources, we describe the content of the derivational resources for various languages using graph theory terminology. Such interpretation leads to a typology, dividing the resources

[1] Similarly to the evolution of multilingual syntactic datasets, we hope that the existence of our harmonized collection may lead to a snowball effect, as it could facilitate annotating word-formation resources for other languages, performing cross-lingual transfer experiments, allowing typological studies etc. On the other hand, the analogy is limited by the different nature of the two types of resources since, for instance, parsers trained on syntactic annotations can be applied on astronomical amounts of unseen texts, while vocabulary of a language whose word-formation is studied is growing only very slowly.

Proceedings of the 2nd Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019), pages 101–110,
Prague, Czechia, 19-20 September 2019.

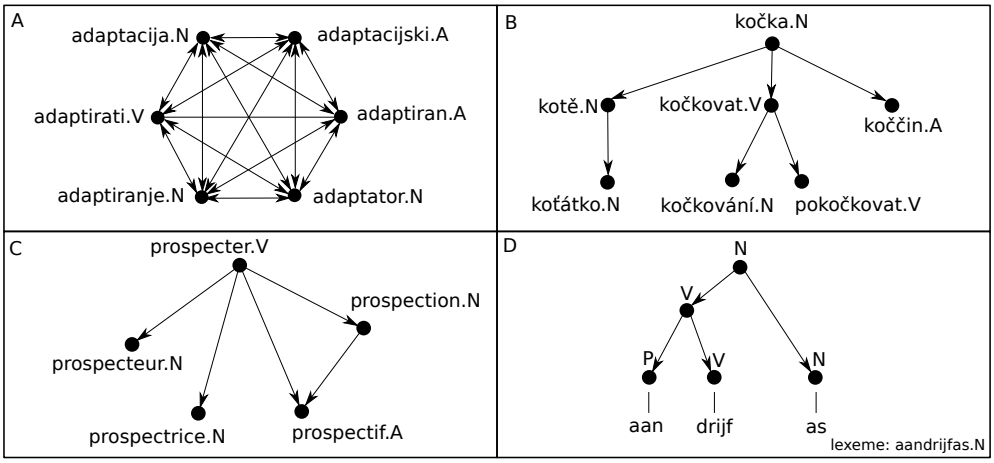

Figure 1: Data structures in available derivational resources: A. complete directed subgraph, B. rooted tree, C. weakly connected subgraph, D. derivation tree.

into four types listed below. In the first three types, lexemes are represented as nodes and derivational relations as directed edges, pointing to a derived lexeme from its base lexeme, while in the fourth type the basic building unit is the morpheme.

A. In some resources, derivationally related lexemes (i.e. lexemes that share a common root morpheme; hereafter, a derivational family) are simply grouped together, leaving particular derivational relations within the groups underspecified (cf. DerivBase.hr for Croatian, Šnajder, 2014). Such derivational families could be represented as complete subgraphs. However, given that the structure models linguistic derivation, we should represent such derivational families rather by *complete directed subgraphs* (see A in Figure 1).[2]

B. If at most one base lexeme is captured for any derived lexeme, then the derivational family can be naturally represented as a *rooted tree* with a designated root node representing a lexeme that is considered as further unmotivated (cf. DeriNet for Czech, Vidra et al., 2019a; B in Figure 1).

C. A *weakly connected subgraph* (in which any lexeme can have more than one base lexeme) is used for representing derivational families in resources in which the rooted-tree constraint does not hold, e.g. in Démonette for French (Hathout and Namer, 2014; C in Figure 1).

D. A *derivation tree* (in the terminology of Context Free Grammars), with morphemes in its leaf nodes and artificial symbols in non-terminal nodes, can be used for describing how a lexeme is composed of individual morphemes (cf. Dutch section of CELEX2, Baayen et al., 1995, D in Figure 1); derivational relations between lexemes are then present only implicitly (based on shared sequences of morphemes).

3 Data resources selected for harmonization

For the pilot stage of the harmonization project, we selected 11 data resources, all of them based either on rooted trees or weakly connected subgraphs (see B and C in Figure 1). The original resources (in alphabetical order) are briefly described below in this section.

Démonette is a network containing lexemes assigned with morphological and semantic features. It was created by merging existing derivational resources for French (cf. Morphonette, Hathout, 2010; VerbAction, Tanguy and Hathout, 2002; and DériF, Namer, 2003). Démonette focuses on suffixation and captures also so-called *indirect relations* (representing *sub-paradigms*) and derivational series among lexemes. Derivational families are represented by weakly connected subgraphs.

[2]Keeping the quadratic number of edges in the data might seem rather artificial at the beginning, however, it is a good starting point as it allows for applying graph algorithms analogously to other types.

DeriNet is a lexical database of Czech that captures derivational relations between lexemes. Each derivational family is represented as a rooted tree.

DeriNet.ES is a DeriNet-like lexical database for Spanish which is based on a substantially revised lexeme set used originally in the Spanish Word-Formation Network (Lango et al., 2018). In DeriNet.ES, derivational relations were created using substitution rules covering Spanish affixation (Faryad, 2019). Resulting derivational families are organized into rooted trees.

DeriNet.FA is a lexical database capturing derivations in Persian, which was created on top of manually compiled Persian Morphologically Segmented Lexicon (Ansari et al., 2019). By using automatic methods, derivationally related lexemes were identified and organized into DeriNet-like rooted trees (Haghdoost et al., 2019).

DErivBase is a large-coverage lexicon for German (Zeller et al., 2013) in which derivational relations were created by using more than 190 derivational rules extracted from reference grammars of German. The resulting derivational families were automatically split into semantically consistent clusters, forming weakly connected subgraphs.

The Morphosemantic Database from English WordNet 3.0 (hereafter, English WordNet) is a stand-off database linking morphologically related nouns and verbs from English WordNet (Miller, 1995) in which synonymous lexemes are grouped into so-called *synsets*, which are further organized according to the hyponymy/hyperonymy relations. Derivational relations were identified and assigned 14 semantic labels (Fellbaum et al., 2007). Derivational families are represented by weakly connected subgraphs.

EstWordNet (Kerner et al., 2010) is a WordNet-like lexical database for Estonian, which did not cover derivational morphology originally. Derivational relations were added by Kahusk et al. (2010); derivational families are represented by weakly connected subgraphs.

FinnWordNet is another WordNet-like database; it is based on the English database which was translated into Finnish (Lindén and Carlson, 2010). Derivational relations were added later by Lindén et al. (2012). Derivational families are represented by weakly connected subgraphs.

NomLex-PT is a lexicon of nominalizations in Portuguese (De Paiva et al., 2014), which were extracted from existing resources. Resulting derivational families are represented by weakly connected subgraphs.

The Polish Word-Formation Network is a DeriNet-like lexical network for Polish created by using pattern-mining techniques and a machine-learned ranking model (Lango et al., 2018). The network was enlarged with the derivational relations extracted from the Polish WordNet (Maziarz et al., 2016). Each derivational family is represented as a rooted tree.

Word Formation Latin is a resource specialized in word-formation of Latin (Litta et al., 2016). The lexeme set is based on the Oxford Latin Dictionary (Glare, 1968). In the Word Formation Latin database, the majority of derivational families is represented by rooted trees but weakly connected subgraphs are used to capture compounds.

4 Harmonization process

4.1 Target representation

The data structure of the DeriNet database is used as the target representation for the remaining ten resources to harmonize. In DeriNet, each tree corresponds to a derivational family. In each tree, the derivational family is internally organized according to the morphemic complexity of the lexemes, from the morphematically simplest lexeme in the root of the tree to the most complex ones in the leaves of the structure, concurring thus with the linguistic account of derivation as a process of adding an affix to a base in order to create a new lexeme (Dokulil, 1962; Iacobini, 2000; Lieber and Štekauer, 2014).

This simple but, at the same time, highly constrained data structure makes it possible to organize massive amounts of language data in a unified way, but it is not sufficient for modelling compounding and other more intricate phenomena, such as double motivation. In the DeriNet 2.0 format, which was released recently (Vidra et al., 2019b) and which is used as the target representation in the presented harmonization process, some of the issues have been solved by introducing *multi-node relations* and other features modelling the language phenomena in a more adequate way.

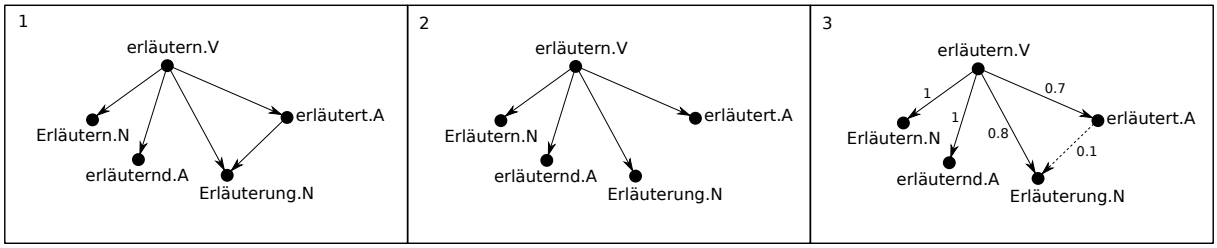

Figure 2: The process of harmonization of a weakly connected graph (an example from DErivBase).

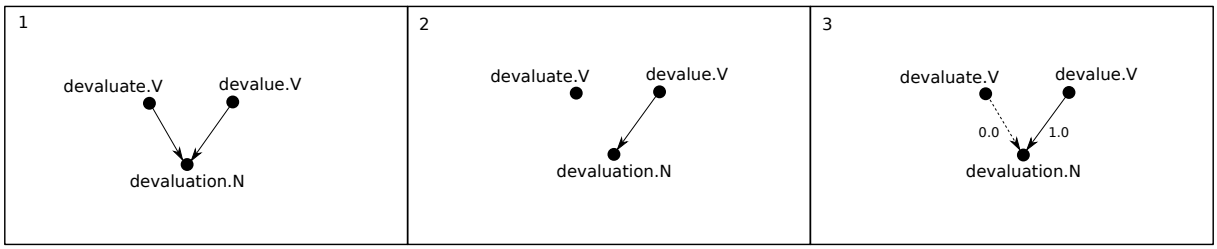

Figure 3: The process of harmonization of a weakly connected graph (an example from English WordNet) that leads to the splitting of the graph (due to spelling variants).

During any harmonization effort, one faces the trade-off between expressiveness and uniformity. A target framework with high expressiveness and flexibility might be able to subsume information exactly as it is present in any input data resource (preserving the annotation scheme with all its linguistic and technical decisions), but it would be just a mere file format conversion without offering any new or more general insights. On the other hand, if the target framework pushes too much on simplicity and uniformity, it could leads to ignoring some features that are important in a particular language. We have to search for something between these two extremes, as we really cannot keep both perfect flexibility and generalization at the same time. We believe that choosing rooted trees is a reasonable compromise: we keep selected word-formation relations in a tree-shaped skeleton (and we hope that multilingual analogies will be enlightened this way), while non-tree edges from the original resources are memorized too in the resulting collection, however, on a less prominent place. Last but not least, choosing the tree approach is hard to resist from the practical perspective: it simplifies many technical aspects (compared to less constrained graphs), such as data traversing, visualization, and evaluating annotator agreement.[3]

4.2 Importing data from existing resources

Derivational resources differ in the formats in which they are distributed. Therefore, as the first step of the harmonization process, the data files were converted into a common file format.

From all resources, we imported as much information as possible about lexemes and word-formation, e.g. morphological features, semantic labels, segmentation, compounding etc., however, we could not preserve all the information present in the original data. For instance, we did not import explicit information about the origin of each feature in Démonette. From Estonian and Finnish WordNet, we extracted all lexemes but processed only derivationally related ones, disregarding synonymy relations and the hyponymic/hyperonymic architecture completely.

4.3 Identifying rooted trees in weakly connected graphs

According to the typology sketched in Section 2, the DeriNet, DeriNet.ES and DeriNet.FA databases, and the Polish Word-Formation Network contain rooted trees, while all the other selected resources consist of weakly connected graphs, in which the spanning tree (tree-shaped skeleton) has to be identified.

[3]Again, this resembles the case of UD, where it was also clear from the very beginning that trees are insufficient for capturing all syntactic relations (e.g. with more complex coordination expressions). The recent UD solution is similar to ours: for each sentence there is a core tree-shaped structure, possibly accompanied with a set of secondary (non-tree) edges.

Therefore, a procedure of selecting rooted tree edges out of a weakly connected graph was applied to the French Démonette, German DErivBase, English WordNet, Estonian EstWordNet, Finnish FinnWordNet, Portuguese NomLex-PT, and Word Formation Latin; see Figure 2 for individual steps. In these resources, a lexeme was allowed to refer to two or more base lexemes, for example, due to compounding, double motivation, or spelling variants (see step 1 in Figure 2).

The data of Démonette, English WordNet, EstWordNet, NomLex-PT, and Word Formation Latin contained a small number of derivational families represented by non-tree structures, therefore, we could select the most appropriate incoming link manually for all those families. In the case of DErivBase and FinnWordNet, there were many such non-tree edges, so we decided to apply Machine Learning. We annotated a small sample of both resources (see step 2 in Figure 2) to train classifiers that predict scores estimating a chance of a derivational relation between two lexemes to be present, or absent, respectively.

Our feature set employed in the classifiers consisted of part-of-speech categories, Levenshtein distance (Levenshtein, 1966), length difference and character n-grams of both the base lexeme and the derived lexeme, and boolean features manifesting whether the initial and final unigrams and bigrams of the base lexeme and the derivative were identical. We tested a number of classification techniques and evaluated them using held-out data in terms of F-score. Logistic Regression performed best for FinnWordNet (F-score = 76.13 %), while Decision Trees achieved the highest F-score for DErivBase (F-score = 82.71 %). Using the classification models, we assigned estimated edge-presence scores to all edges except for leaf nodes, for which no decision-making was needed (see step 3 in Figure 2).

We chose resulting trees by maximizing the sum of scores using the Maximum Spanning Tree algorithm introduced by Chu and Liu (1965) and Edmonds (1967) implemented in a Python package NetworkX (Hagberg et al., 2008). The resulting tree-shaped skeleton is drawn with solid lines and the non-tree edges are drawn with dashed lines in the step 3 in Figure 2. In the harmonized data, we saved both types of edges, but the non-tree ones were processed as secondary ones.

However, some derivational families did not contain a rooted tree in the respective weakly connected graphs. This situation can be caused, for example, by spelling variants, that is illustrated on the data from English WordNet in Figure 3. English verbs "devalue" and "devaluate" are proposed as base lexemes for lexeme "devaluation" (see step 1 in Figure 3) but only one is allowed in the rooted tree (see step 2 in Figure 3), which leads to splitting the family into two (see step 3 in Figure 3). One of the families contains the lexemes "devalue" and "devaluation", the second one has a single lexeme "devaluate". Using links between the new roots ("devalue" and "devaluate"), we kept information about splitting the family in the harmonized data.

4.4 Converting the data into the DeriNet 2.0 format

Using the application interface developed for DeriNet 2.0,[4] we stored the trees resulting from the previous steps into the DeriNet 2.0 format, which was designed to be as language agnostic as possible; see Vidra et al. (2019b) in this volume.

The lemma set of each resource and all features assigned to the lexemes (e.g. morphological features) were converted first. It was also necessary to create a unique identifier for each lexeme to prevent technical problems caused by the same string form or homonymy of lexemes. An identifier pattern consisting of the string and the part-of-speech category of the lexeme was sufficient for all harmonized resources except for Démonette, DeriNet, DErivBase and Word Formation Latin.

DeriNet uses so-called *tag masks*[5] instead of part-of-speech category. In Démonette and DErivBase, the identifier contains also a gender (for nouns only) of the lexeme, and Word Formation Latin needs to use the ID from its original version due to the subtle differentiation of lexeme meanings. For example, there are three meanings of the lexeme "gallus" captured in the Word Formation Latin resource ("a farmyard cock", "an inhabitant of Gaul", and "an emasculated priest of Cybele"; Glare 1968), i.e. three entries with the same graphemic form and morphological features but with the different derivational families.

[4] https://github.com/vidraj/derinet

[5] The tag mask represents the intersection of the set of part-of-speech tags of all inflected forms of a particular lexeme. By comparing positions of values in each tag, the tag mask consists of values (whether the value was the same across all tags) or question marks (otherwise). For more details, see Vidra et al. (2019b).

Resource	Language	Extracted from original Lexemes	Relations	Families	After harmonization Lexemes	Relations	Families	License
Démonette 1.2	French	21,290	14,152	7,336	21,290	13,808	7,482	CC BY-NC-SA 3.0
DeriNet 2.0	Czech	1,027,665	808,682	218,383	1,027,665	808,682	218,383	CC BY-NC-SA 3.0
DeriNet.ES	Spanish	151,173	36,935	114,238	151,173	36,935	114,238	CC BY-NC-SA 3.0
DeriNet.FA	Persian	43,357	35,745	7,612	43,357	35,745	7,612	CC BY-NC-SA 4.0
DErivBase 2.0	German	280,775	55,010	235,287	280,775	44,830	235,945	CC BY-SA 3.0
English WordNet 3.0	English	13,813	8,000	5,818	13,813	7,855	5,958	CC BY-NC-SA 3.0
EstWordNet 2.1	Estonian	115,318	535	456	988	507	481	CC BY-SA 3.0
FinnWordNet 2.0	Finnish	44,173	29,783	6,347	20,035	13,687	6,348	CC BY 3.0
Nomlex-PT 2017	Portuguese	7,020	4,235	2,785	7,020	4,201	2,819	CC BY 4.0
Polish WFN 0.5	Polish	262,887	189,217	73,670	262,887	189,217	73,670	CC BY-NC-SA 3.0
Word Formation Latin	Latin	29,708	22,687	5,273	29,708	22,641	5,320	CC BY-NC-SA 4.0

Resource	Singleton nodes	#Nodes	Tree depth	Tree out-degree	Part-of-speech distribution [%] Noun	Adj	Verb	Adv	Other
Démonette 1.2	69	2.8 / 12	1.1 / 4	1.8 / 8	63.0	2.5	34.5	–	–
DeriNet 2.0	96,208	4.7 / 1638	0.8 / 10	1.1 / 40	44.0	34.8	5.5	15.7	–
DeriNet.ES	98,325	1.3 / 35	0.2 / 5	0.3 / 14	–	–	–	–	–
DeriNet.FA	0	5.7 / 180	1.5 / 6	3.3 / 114	–	–	–	–	–
DErivBase 2.0	215,823	1.2 / 51	0.1 / 7	0.1 / 13	85.5	9.9	4.6	–	–
English WordNet 3.0	65	2.3 / 6	1.0 / 1	1.3 / 6	56.9	–	43.1	–	–
EstWordNet 2.1	21	2.1 / 3	1.0 / 2	1.0 / 3	15.9	29.0	7.9	47.2	–
FinnWordNet 2.0	3	3.2 / 36	1.5 / 9	1.5 / 13	55.3	29.2	15.5	–	–
Nomlex-PT 2017	17	2.5 / 7	1.0 / 1	1.5 / 7	59.8	–	40.2	–	–
Polish WFN 0.5	41,332	3.6 / 214	1.0 / 8	1.1 / 38	–	–	–	–	–
Word Formation Latin	63	5.6 / 130	1.5 / 6	3.0 / 42	46.0	27.4	23.8	–	2.8

Table 1: Ten language resources of the UDer collection before and after the harmonization, and some basic quantitative features of the UDer collection. Columns #Nodes, Tree depth, and Tree outdegree are presented in average / maximum value format.

In the second step, we converted tree-shaped derivational relations and added details about each relation, e.g. semantic label, type of relation, affix, depending on the annotation in the original resource. Because Word Formation Latin captures also compounding, these relations were included too.

The rest of the imported data and some by-products of the harmonization process (esp. the non-tree derivational relations, links between roots in the case of splitting the original family, and resource-specific annotation, e.g. indirect relations in Démonette) were converted for each resource in the last step.

5 UDer Collection

The resulting collection, Universal Derivations version 0.5 (UDer 0.5), includes eleven resources covering eleven different languages listed in Table 1. Using the DeriNet 2.0 file format, UDer provides derivational data in the same annotation scheme, in which a rooted tree is the backbone of each derivational family, however, other original derivational relations that are not involved in the trees due to the harmonization process are also included as the secondary relations to the harmonized data. Tree-shaped derivational families with the verb "evaluate" (and their equivalents in particular languages) in all harmonized resources are displayed in Figure 4. Basic quantitative properties of the collection are summarized in the following subsections; information about the availability of the collection is provided, too.

5.1 Selected quantitative properties

Selected quantitative characteristics of the resources involved in the harmonized collection can be compared in Table 1. The lexeme sets were adopted from the original data resources, except for WordNets. From FinnWordNet and EstWordNet, only derivationally related lexemes were admitted.

After the harmonization process, the number of derivational relations decreased in resources capturing derivational families in weakly connected graphs, however, the number of relations after the harmonization as given in the table includes only tree-shaped relations. Non-tree relations are also stored but on a less prominent place (the number of them can be calculated as a difference between extracted relations and

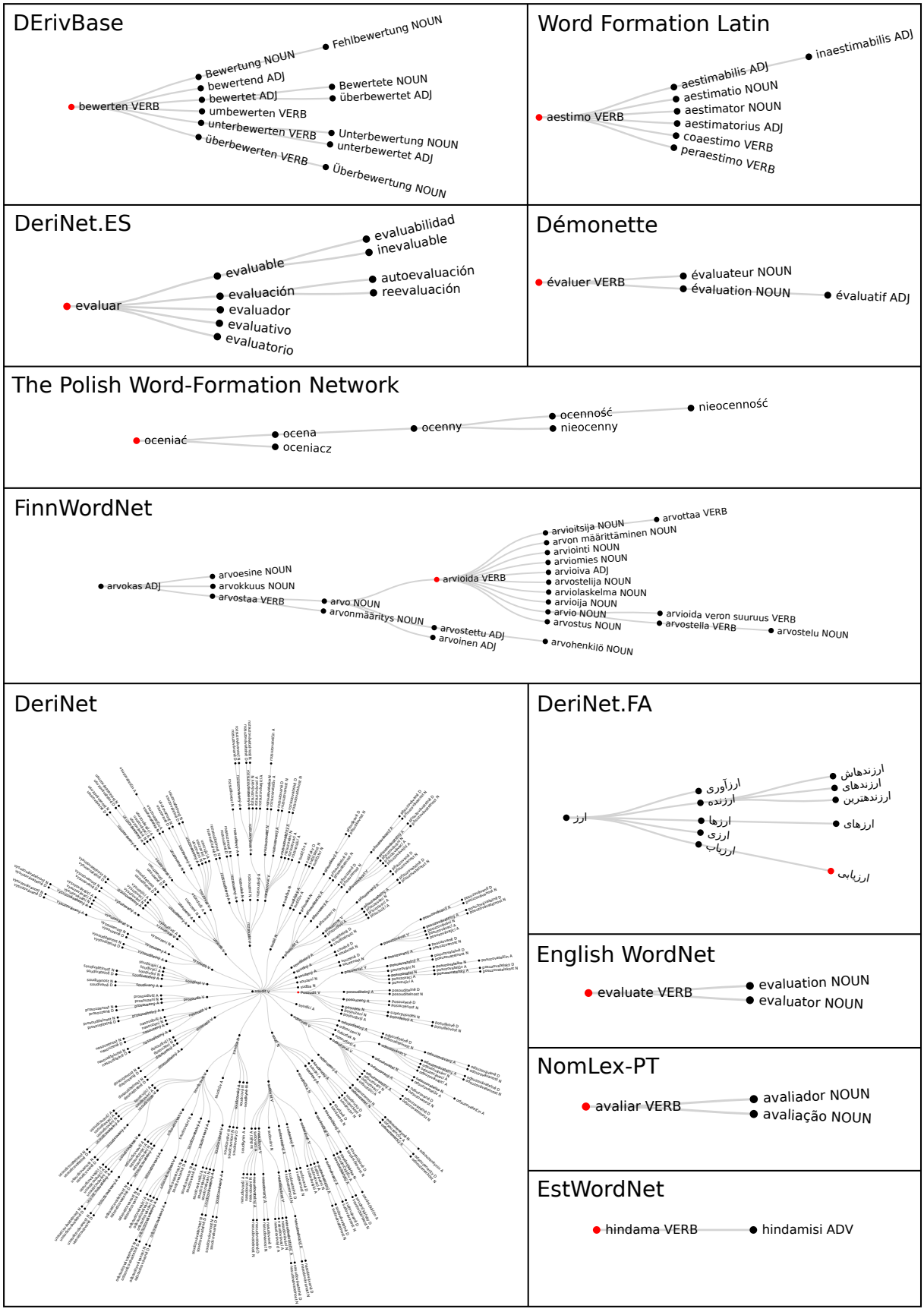

Figure 4: Harmonized rooted trees for verb “evaluate” in UDer v0.5 collection.

relations after the harmonization).

The number of derivational families after the harmonization process remained the same for resources representing derivational families as rooted trees, but it increased in resources that organized derivational families in the weakly connected graphs. The growth is caused by splitting the original family because some weakly connected graphs did not contain the rooted tree (cf. Figure 3 and Section 4.3). Nevertheless, the information about splitting the original family is stored in form of links between roots of the rooted trees in the harmonized data.

The second part of Table 1 indicates the number of singleton nodes (some derivational families contain just a single lexeme). The number of singleton nodes correlates with the way the resource was created. The high number of singleton nodes occurs in resources that were built-up from lexeme set to finding derivational relations within them, i.e. DeriNet, DeriNet.ES, DErivBase, and The Polish Word-Formation Network, whereas the lower number of singleton nodes is documented in resources that included lexemes depending on whether the lexeme was derivationally related to another lexeme. The number of singleton nodes could increase due to splitting the original family during the harmonization of these resources.

As for the average and maximum size of derivational families, their average and maximum depth (i.e. the distance of the furthest node from the tree root) and out-degree (i.e. the highest number of direct children of a single node) is compared across the harmonized resources which illustrate a general condition of the resources after the harmonization process. On average, the biggest derivational families can be found in DeriNet.FA, Word Formation Latin, and DeriNet, while the smallest families are in DErivBase and DeriNet.ES, as their data are made up mostly of singletons. A similar tendency can also be seen for the maximum size of nodes (lexemes) in the trees (families). DeriNet contains the biggest tree with the root "dát" ("give") having more than 1.6 thousand nodes. On the other hand, in the small and sparse data of EstWordNet, all trees contain three or even fewer nodes.

As for the part-of-speech categories, DeriNet and EstWordNet cover nouns, adjectives, verbs, and adverbs. Word Formation Latin lacks adverbs but it contains pronouns, auxiliaries and lexemes unspecified for the part of speech. Démonette, DErivBase and FinnWordNet also lacks adverbs, and both Démonette and DErivBase have a low number of adjectives. English WordNet and NomLex-PT are limited to nouns and verbs. The part-of-speech categories are not available for DeriNet.ES, DeriNet.FA, and the Polish Word-Formation Network.

5.2 Publishing and licensing

The presented UDer 0.5 collection is freely available in a single data package in the LINDAT/CLARIAH CZ repository[6] under the licenses listed in Table 1. The UDer data can be also queried using DeriSearch tool[7] (Vidra and Žabokrtský, 2017) and processed using other software developed within the DeriNet project, especially the Python application interface for DeriNet 2.0.

6 Conclusions and final remarks

This paper introduced a collection of derivational resources which have been harmonized to a common annotation scheme. The collection is publicly available. In the near future, we plan to evaluate the harmonization process in terms of consistency and adequacy across languages, and we are going to harmonize data resources for other languages and from other types of data structures, too.

The process of harmonization of linguistic data resources is always a compromise between expressiveness and uniformity. It is impossible to keep all the information stored in the diverse original resources and allow processing them all in an efficient unified way at the same time to allow multilingual or cross-lingual research. However, we believe that the benefits of the presented harmonization efforts outweigh the negatives and, above all, that it will open a (previously almost non-existent) discussion on the harmonization of derivational resources.

[6] http://hdl.handle.net/11234/1-3041
[7] http://ufal.mff.cuni.cz/derinet/derinet-search

Acknowledgments

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, by the Charles University Grant Agency (project No. 1176219) and by the SVV project number 260 453. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIAH CZ project (LM2015071, LM2018101).

References

Ebrahim Ansari, Zdeněk Žabokrtský, Hamid Haghdoost, and Mahshid Nikravesh. 2019. Persian Morphologically Segmented Lexicon 0.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-3011.

Harald R. Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. CELEX2. Linguistic Data Consortium, Catalogue No. LDC96L14.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning*. ACL, pages 149–164.

Yoeng-Jin Chu and T. H. Liu. 1965. On the Shortest Arborescence of a Directed Graph. *Scientia Sinica* 14:1396–1400.

Valeria De Paiva, Livy Real, Alexandre Rademaker, and Gerard de Melo. 2014. NomLex-PT: A Lexicon of Portuguese Nominalizations. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*. ELRA, pages 2851–2858.

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

Jack Edmonds. 1967. Optimum Branchings. *Journal of Research of the national Bureau of Standards* 71B(4):233–240.

Ján Faryad. 2019. Identifikace derivačních vztahů ve španělštině. Technical Report TR-2019-63, Faculty of Mathematics and Physics, Charles University.

Christiane Fellbaum, Anne Osherson, and Peter E Clark. 2007. Putting Semantics into WordNet's "Morphosemantic" Links. In *Language and Technology Conference*. Springer, pages 350–358.

P. G. W. Glare. 1968. *Oxford Latin dictionary*. Clarendon Press, Oxford.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*. pages 11–15.

Hamid Haghdoost, Ebrahim Ansari, Zdeněk Žabokrtský, and Mahshid Nikravesh. 2019. Building a Morphological Network for Persian on Top of a Morpheme-Segmented Lexicon. In *Proceedings of the 2nd Workshop on Resources and Tools for Derivational Morphology*. Charles University.

Nabil Hathout. 2010. Morphonette: A Morphological Network of French. *arXiv preprint arXiv:1005.3902* .

Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French Derivational Morpho-Semantic Network. *Linguistic Issues in Language Technology* 11:125–162.

Claudio Iacobini. 2000. Base and Direction of Derivation. In *Morphology. An International Handbook on Inflection and Word-formation*, Mouton de Gruyter, volume 1, pages 865–876.

Neeme Kahusk, Kadri Kerner, and Kadri Vider. 2010. Enriching Estonian WordNet with Derivations and Semantic Relations. In *Baltic hlt*. pages 195–200.

Kadri Kerner, Heili Orav, and Sirli Parm. 2010. Growth and Revision of Estonian WordNet. In *Principles, Construction and Application of Multilingual WordNets*. Narosa Publishing House, pages 198–202.

Lukáš Kyjánek. 2018. Morphological Resources of Derivational Word-Formation Relations. Technical Report TR-2018-61, Faculty of Mathematics and Physics, Charles University.

Mateusz Lango, Magda Ševčíková, and Zdeněk Žabokrtský. 2018. Semi-Automatic Construction of Word-Formation Networks (for Polish and Spanish). In *Proceedings of the 11th International Conference on Language Resources and Evaluation*. ELRA, pages 1853–1860.

Vladimir I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10:707.

Rochelle Lieber and Pavol Štekauer. 2014. *The Oxford handbook of derivational morphology*. Oxford University Press, Oxford.

Krister Lindén and Lauri Carlson. 2010. FinnWordNet–Finnish WordNet by Translation. *LexicoNordica – Nordic Journal of Lexicography* 17:119–140.

Krister Lindén, Jyrki Niemi, and Mirka Hyvärinen. 2012. Extending and updating the Finnish Wordnet. In *Shall We Play the Festschrift Game?*, Springer, pages 67–98.

Eleonora Litta, Marco Passarotti, and Chris Culy. 2016. Formatio Formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the 3rd Italian Conference on Computational Linguistics*. pages 185–189.

Marek Maziarz, Maciej Piasecki, Ewa Rudnicka, Stan Szpakowicz, and Paweł Kędzia. 2016. plWordNet 3.0 – a Comprehensive Lexical-Semantic Resource. In *Proceedings of the 26th International Conference on Computational Linguistics*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 2259–2268.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Täckström Oscar, Bedini Claudia, Castelló B. Núria, and Lee Jungmee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. ACL, volume 2, pages 92–97.

George A Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38(11):39–41.

Fiammetta Namer. 2003. Automatiser l'analyse morpho-sémantique non affixale: le système DériF. *Cahiers de grammaire* 28:31–48.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Tsarfaty Reut, and Zeman Daniel. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. ELRA, pages 1659–1666.

Ludovic Tanguy and Nabil Hathout. 2002. Webaffix: un outil d'acquisition morphologique dérivationnelle à partir du Web. In *Actes de la 9e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*. ATALA, Nancy, France.

Jonáš Vidra and Zdeněk Žabokrtský. 2017. Online Software Components for Accessing Derivational Networks. In *Proceedings of the Workshop on Resources and Tools for Derivational Morphology*. EDUCatt, pages 129–139.

Jonáš Vidra, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, and Šárka Dohnalová. 2019a. DeriNet 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2995.

Jonáš Vidra, Zdeněk Žabokrtský, Magda Ševčíková, and Lukáš Kyjánek. 2019b. DeriNet 2.0: Towards an All-in-One Word-Formation Resource. In *Proceedings of the 2nd Workshop on Resources and Tools for Derivational Morphology*. Charles University.

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*. ACL, volume 1, pages 1201–1211.

Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2014. HamleDT: Harmonized Multi-Language Dependency Treebank. *Language Resources and Evaluation* 48(4):601–637.

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