PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 115 OCTOBER 2020 5–30

Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources

Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, Jonáš Vidra

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics

Abstract

The paper deals with harmonisation of existing data resources containing word-formation features by converting them into a common file format and partially aligning their annotation schemas. We summarise (dis)similarities between the resources and describe individual steps of the harmonisation procedure, including manual annotations and application of Machine Learning techniques. The resulting "Universal Derivations 1.0" collection contains 27 harmonised resources covering 20 languages. It is publicly available in the LINDAT/CLARIAH CZ repository and can be queried via the DeriSearch tool.

1. Introduction

There are several dozens of language resources which focus specifically on derivational morphology or capture some word-formation features in addition to other types of annotation. However, the resources differ greatly in many aspects, which complicates usability of the data in multilingual projects, including potential data-oriented research in word-formation across languages.

Being inspired by the recent developments in treebanking (cf. Buchholz and Marsi, 2006, McDonald et al., 2013, Zeman et al., 2014, Nivre et al., 2016b, and others), a harmonisation procedure was proposed to unify annotation schemas of word-formation resources. The harmonised resources were released under the title *Universal Derivations* (hereafter, UDer), with eleven resources covering eleven different languages in UDer version 0.5 (Kyjánek et al., 2019; Kyjánek et al., 2019). Kyjánek (2020) elaborated on the procedure to cover resources with other data structures. The present

 Corresponding author: kyjanek@ufal.mff.cuni.cz
Cite as: Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková, Jonáš Vidra. Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources. The Prague Bulletin of Mathematical Linguistics No. 115, 2020, pp. 5–30. doi: 10.14712/00326585.003.

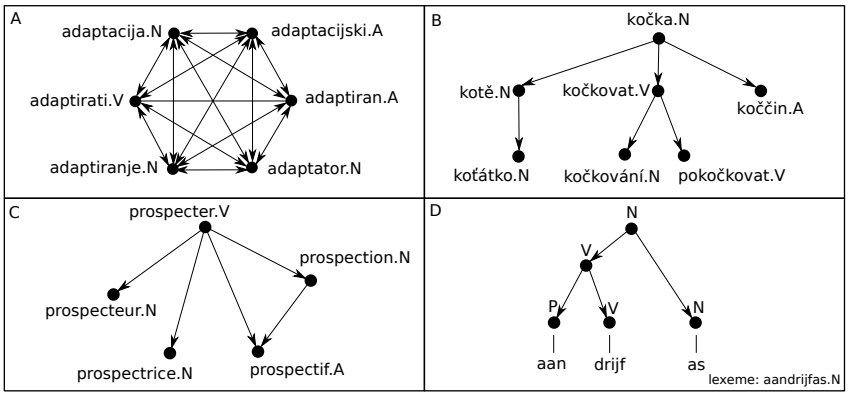

Figure 1. Data structures in available derivational resources: A. complete directed subgraph, B. rooted tree, C. weakly connected subgraph, D. derivation tree.

paper summarises the extended harmonisation procedure and introduces a new version of the UDer collection, which contains 27 harmonised resources for 20 languages (UDer 1.0, Kyjánek et al. 2020).

The paper is organised as follows: A brief overview of existing data resources, their underlying data structures, and more details on the resources selected for the harmonisation can be found in Section 2. The harmonisation is described step by step in Section 3, followed by some quantitative and qualitative features of UDer 1.0 and a description of a user query interface (Section 4).

2. Existing data resources and resources selected for harmonisation

Kyjánek (2018, 2020) listed about fifty machine-tractable resources where information related to word-formation of individual languages can be found. The resources differ in many aspects; specifically, in the data structure, in the file format, in the size in terms of both lexemes and derivational relations, and in the licenses under which the resources were released.

In what follows, the resources are compared using terms from the graph theory terminology (cf. Matoušek and Nešetřil 2009). In the first three types (see Figure 1), lexemes are represented as nodes and derivational relations as directed edges. The edges point from the base lexemes to the derived ones. In contrast, the basic building unit in the fourth type is a morpheme.

A. Some resources only group derivationally related lexemes together, i.e., lexemes that share a common root morpheme (hereafter, a derivational family). Individual derivational relations between lexemes are unspecified. Such families could be represented as complete directed subgraphs. Given that the structure mod-

els linguistic derivation, we represent such families rather by *complete directed subgraphs* (see A in Figure 1; adopted from DerivBase.hr for Croatian, Šnajder, 2014).[1]

B. If at most one base lexeme is identified for any derived lexeme, then the derivational family can be naturally represented as a *rooted tree* (B in Figure 1; from DeriNet for Czech, Vidra et al., 2019a). The tree root represents prototypically the simplest (unmotivated) lexeme, while leaf nodes contain the most complex lexemes (in terms of both morphological structure and derivational meaning) in a particular derivational family. The rooted tree data structure cannot capture compounding relations.

C. We represent derivational families as *weakly connected subgraphs* in resources that allow capturing more than one base lexeme for any derivative, e.g. compounds and double motivation (C in Figure 1; from Démonette for French, Hathout and Namer, 2014). Thus, the rooted-tree constraint does not hold in those resources.

D. Some resources focus on morphological segmentation of lexemes rather than derivational relations between lexemes. 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 (D in Figure 1; from Dutch section of CELEX2, Baayen et al., 1995); derivational relations between lexemes are then present only implicitly (based on shared sequences of morphemes).

The following criteria were applied to determine which of the existing resources will be harmonised: we wanted to cover all four data structures presented above; and we preferred resources specialised in capturing word-formation, covering languages not yet included in the collection, and published under an open license.

Bellow we briefly comment on each of the 27 resources selected for harmonisation (in alphabetical order).

[R1-en] CatVar (Habash and Dorr, 2003) is an automatically constructed Categorial Variation Database containing derivational families of English lexemes. The families were created by using the morphological segmentation obtained from several morphological segmenters and resources (including the English part of CELEX). Complete directed subgraphs are used to represent the data.

[R2-nl, R3-en, R4-de] CELEX is a large, manually created resource of comprehensive annotations for Dutch, English, and German. The three language parts were developed separately for psycholinguistic research. Word-formation features are inferred from three types of morphological segmentation provided by the resources: (a) segmentation of lexemes into bases and affixes, e.g. *collaboration* is segmented into *collaborate+ion*, (b) hierarchical segmentation of lexemes into morphemes organised into a derivation tree structure, and (c) flat segmentation of lexemes into morphemes

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

obtainable from the last tree level (hierarchical and flat segmentation are illustrated in example D in Figure 1).

[R5-fr] 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 is paradigm-oriented, i.e., it organises lexemes into (*partial*/*complete*) *derivational* (*sub*)*paradigms* using so-called *indirect relations*, and captures derivational series among lexemes. Derivational families are represented by weakly connected subgraphs.

[R6-cs] DeriNet is a lexical database of Czech that connects derivationally related lexemes. The data format used since version 2.0 (Vidra et al., 2019b) allows to represent compounding and other features, such as morphological categories, morphological segmentation, semantic labels, etc. Each derivational family is represented as a rooted tree.

[R7-es] DeriNet.ES (Faryad, 2019) is a DeriNet-like lexical database for Spanish. Its derivational relations were created by using substitution rules covering Spanish affixation. Resulting derivational families are organised into rooted trees.

[R8-fa] DeriNet.FA (Haghdoost et al., 2019) is a lexical database capturing derivations in Persian. It was created on top of the manually compiled Persian Morphologically Segmented Lexicon (Ansari et al., 2019). Derivationally related lexemes were identified and organised into DeriNet-like rooted trees by using automatic methods.

[R9-it] DerIvaTario (Talamo et al., 2016) is a database of manually morphologically segmented Italian lexemes. Each lexeme is assigned a unique ID, which interconnects lexemes across several existing resources to provide various pieces of information such as morphological categories and phonetic transcriptions. The data is processed as derivation tree structures.

[R10-de] DErivBase (Zeller et al., 2013) is a large-coverage lexicon for German, in which derivational relations were identified by more than 190 derivational rules, i.e., string substitutions, extracted from German reference grammar books. The resulting derivational families were automatically split into semantically consistent clusters by Zeller et al. (2014), forming weakly connected subgraphs.

[R11-hr] DerivBase.Hr is a database containing Croatian derivational families. Inspired by German DErivBase and DErivCELEX (Shafaei et al., 2017), DerivBase.Hr was created by using a set of derivational rules. Since the resource lists derivational families without specifying individual derivational relations, we represent the data as complete directed subgraphs.

[R12-ru] DerivBase.Ru (Vodolazsky, 2020) is a data resource of Russian derivationally related lexemes. While its lexemes came from Russian Wikipedia and Wiktionary, the relations were identified by a set of derivational rules extracted from Russian grammar books. Derivational families are represented as weakly connected subgraphs.

[R13-et] EstWordNet (Kerner et al., 2010) is an Estonian WordNet-like lexical database, into which derivational relations were added by Kahusk et al. (2010). The resulting families are represented as weakly connected subgraphs.

[R14-ca, R15-cs, R16-gd, R17-pl, R18-pt, R19-ru, R20-sh, R21-sv, R22-tr] Etymological WordNet (Gerard, 2014) is a lexical resource containing data extracted from the English section of Wiktionary. The Etymological WordNet aims, differently from other wordnets, at identifying lexemes linked by etymology. Besides etymological features, the Etymological WordNet also captures derivational relations between lexemes for almost 180 languages; however, only a few relations are captured for many languages. The data is mostly represented as weakly connected subgraphs.

[R23-fi] FinnWordNet (Lindén and Carlson, 2010) is a WordNet-like database created by translating English WordNet into Finnish. Derivational relations were added by Lindén et al. (2012). Derivational families are represented as weakly connected subgraphs.

[R24-pt] NomLex-PT (De Paiva et al., 2014) is a lexicon of Brazilian Portuguese verbs and deverbative nouns, which were extracted from already existing resources. Resulting derivational families are represented as weakly connected subgraphs.

[R25-en] The Morpho-Semantic Database (Fellbaum et al., 2007) is a stand-off database linking morphologically related nouns and verbs from English (Princeton) WordNet version 3.0 (Miller, 1998). Derivational relations were identified automatically and assigned 14 semantic labels. The data is represented as weakly connected subgraphs.

[R26-pl] The Polish Word-Formation Network (Lango et al., 2018) is a DeriNet-like lexical network for Polish. It was created by using pattern-mining techniques and a machine-learned ranking model. The Polish WFN was also enlarged with derivational relations extracted from the Polish WordNet (Maziarz et al., 2016). Derivational families are represented as rooted trees.

[R27-la] Word Formation Latin (Litta et al., 2016) is a manually-annotated resource specialised in capturing word-formation of Latin. Its lexeme set is based on the Oxford Latin Dictionary (Glare, 1968). In the Word Formation Latin database, the majority of derivational families is represented as rooted trees, but weakly connected subgraphs are used to capture compounds.

3. Harmonisation procedure

Once the resources were selected and their data structures identified, the harmonisation procedure started, focusing on both the data (i.e., lexeme set, derivational relations, and annotated features) and the annotation schemas (i.e., data structure, file format, feature-value pairs) of the resources.

Concerning the data, we have decided to make as few changes as possible. Thus, (i) the original sets of (derivationally related) lexemes are neither enlarged nor reduced; (ii) all derivational relations from the input resources are still preserved in

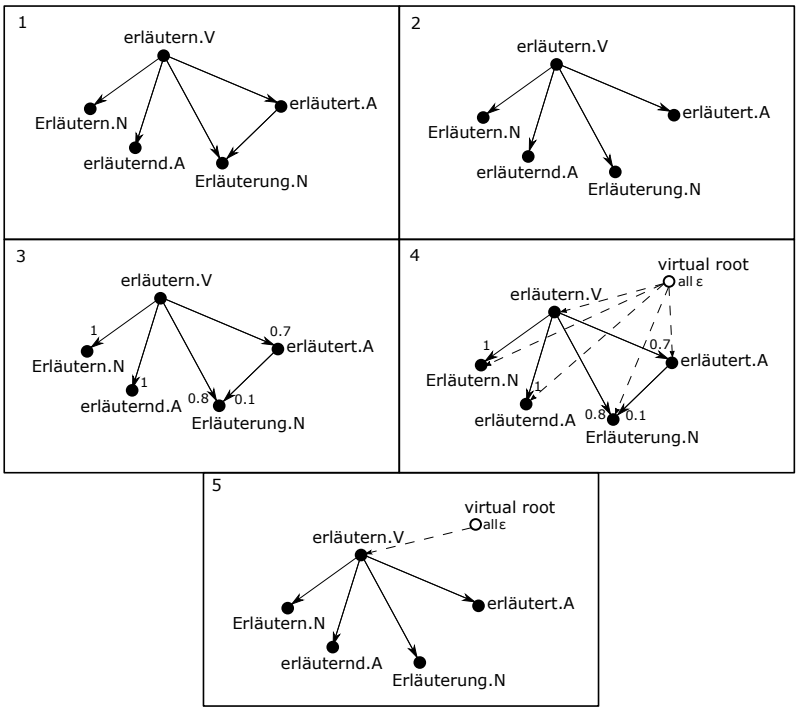

Figure 2. Five steps of the harmonisation procedure (illustrated on DErivBase data).

the resulting data, although they are restructured to fit the selected target annotation schema; and (iii) no new features or pieces of annotations are added to the data.

As for the annotation schema, we have selected the rooted-tree data structure and the file format used in DeriNet 2.0 (Vidra et al., 2019b) as the target data representation for all the resulting harmonised resources. In DeriNet 2.0, each derivational family is represented as a rooted tree (e.g. example B in Figure 1), which is internally organised according to the morphemic complexity of the lexemes involved, from the morphematically simplest lexeme in the root of the tree to the most complex ones in the leaves. Thus, it concurs 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 rooted-tree data structure makes it possible to store massive amounts of language data in a unified way, but it is not sufficient for capturing compounding and other more intricate phenomena, such as double motivation. These issues have been solved by introducing secondary edges allowing to specify any number of base lexemes and

derivatives in the target data structure.[2] We believe that such a representation is a reasonable compromise between expressiveness and uniformity. In addition, 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 and visualisation.

The target file format, in which the target data structure is stored, is a textual lexeme-based format consisting of ten tab-separated columns (Vidra et al., 2019b, pp. 86-88), inspired by the CoNLL-U format (Nivre et al., 2016a) used in Universal Dependencies treebanks. It has been designed to be as universal as possible to allow preserving key-value pairs specifying most of the annotated features relevant for studying word-formation, such as part-of-speech or any morphological categories of lexemes. The list of features can be extended as needed for any language, and lexeme features which cannot be easily expressed by a single value can also be stored in JSON format in the last column of the file.

The harmonisation procedure is illustrated in steps 1 through 5 in Figure 2 and further described in Subsections 3.1 to 3.5, respectively. Each of the steps is exemplified on two German resources (G-CELEX and DErivBase) in order to provide a better insight.

3.1. Importing data from the existing resources

At the beginning of the procedure, we import as much information as possible from the original, resource-specific file formats of the input resources. For instance, the DErivBase file format lists all lexeme pairs within the same derivational family and the shortest path between any two lexemes in the family (the top of Figure 3). The paths consist of derivational relations to which so-called *derivational rules* are assigned, e.g. `dVN09*>`.[3] In the case of G-CELEX, its file format stores individual lexemes with three types of morphological segmentation without specifying derivational relations between lexemes (the bottom of Figure 3).

First, we import lexemes. In most of the resources, a lexeme is represented as its lemma accompanied with its part-of-speech tag. In addition to lemma and part-of-speech tag, gender is used for representing nouns in DErivBase, while only a unique numeric ID is used in G-CELEX. We import only derivationally related lexemes from Estonian, Finnish, and Etymological WordNets, disregarding synonymy relations and the hyponymic/hyperonymic architecture completely.

After obtaining the lexeme sets, other pieces of annotations are imported, e.g. derivational and compounding relations between lexemes, morphological categories and

[2] This aspect resembles the case of Universal Dependencies, 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 that for each sentence, there is a core tree-shaped structure, possibly accompanied with a set of secondary (non-tree) edges.

[3] The asterisk (*) indicates that the rule is applied inversely.

```
Erläutern_Nn erläutern_V 1 Erläutern_Nn dVN09*> erläutern_V
Erläutern_Nn erläuternd_A 2 Erläutern_Nn dVN09*> erläutern_V dVA12> erläuternd_A
Erläutern_Nn erläutert_A 2 Erläutern_Nn dVN09*> erläutern_V dVA13> erläutert_A
Erläutern_Nn Erläuterung_Nf 2 Erläutern_Nn dVN09*> erläutern_V dVN07> Erläuterung_Nf
erläutern_V erläuternd_A 1 erläutern_V dVA12> erläuternd_A
erläutern_V erläutert_A 1 erläutern_V dVA13> erläutert_A
erläutern_V Erläuterung_Nf 1 erläutern_V dVN07> Erläuterung_Nf
erläuternd_A erläutert_A 2 erläuternd_A dVA12*> erläutern_V dVA13> erläutert_A
erläuternd_A Erläuterung_Nf 2 erläuternd_A dVA12*> erläutern_V dVN07> Erläuterung_Nf
erläutert_A Erläuterung_Nf 1 erläutert_A dNA25> Erläuterung_Nf
```

```
1\Tourenschi\...\Tour+en+Schi\NxN\...\((Tour)[N],(en)[N|N.N],(Schi)[N])[N]\...
2\Tourenwagen\...\Tour+en+Wagen\NxN\...\((Tour)[N],(en)[N|N.N],(Wagen)[N])[N]\...
3\tourenweise\...\Tour+en+weise\Nxx\...\((Tour)[N],(en)[B|N.x],(weise)[B|Nx.])[B]\...
4\Tourenzahl\...\Tour+en+zaehl\NxV\...\((Tour)[N],(en)[N|N.V],(zaehl)[V])[N]\...
5\Tourenzaehler\...\Tour+en+zaehl+er\NxVx\...\((Tour)[N],(en)[N|N.Vx],(zaehl)[V],(er)[N|NxV.])[N]\...
6\Tourismus\...\Tour+ismus\Nx\...\((Tour)[N],(ismus)[N|N.])[N]\...
7\Tourist\...\Tour+ist\Nx\...\((Tour)[N],(ist)[N|N.])[N]\...
8\Touristik\...\Tour+istik\Nx\...\((Tour)[N],(istik)[N|N.])[N]\...
9\touristisch\...\Tour+istisch\Nx\...\((Tour)[N],(istisch)[A|N.])[A]\...
```

Figure 3. Original file formats of DErivBase (top) and G-CELEX (bottom).

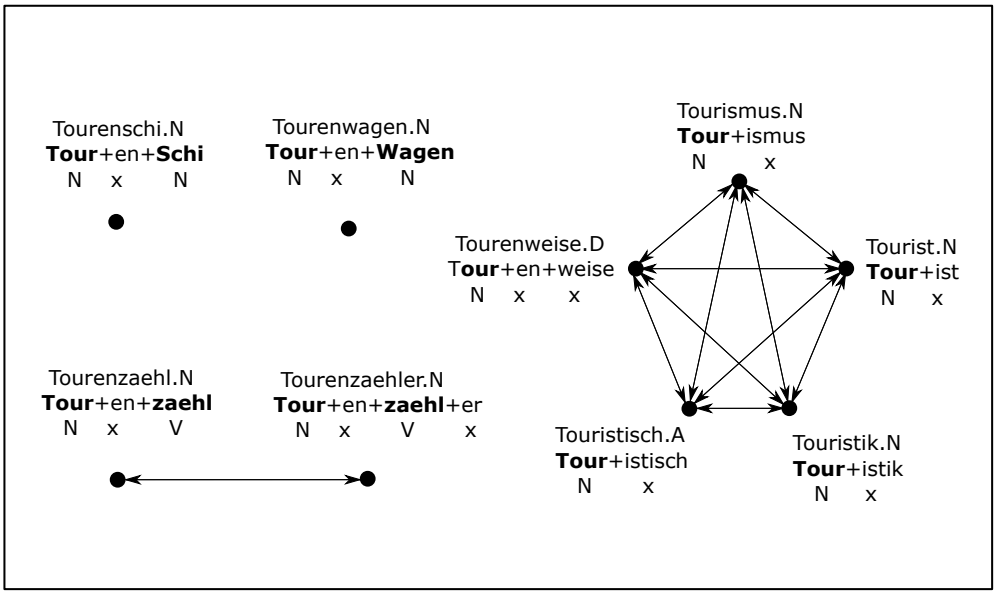

Figure 4. Constructing derivational subgraphs from morphological segmentation.

Input resource	Imported features from the input data resources									
	DER	COM	POS	MCG	SEG	SEM	HSG	PAR	TID	DRL
[R1-en] CatVar	✓	–	✓	–	–	–	–	–	–	–
[R2-nl] D-CELEX	–	✓	✓	–	✓	–	✓	–	✓	–
[R3-en] E-CELEX	–	✓	✓	–	✓	–	✓	–	✓	–
[R4-de] G-CELEX	–	✓	✓	–	✓	–	✓	–	✓	–
[R5-fr] Démonette	✓	–	✓	✓	✓	✓	–	✓	–	–
[R6-cs] DeriNet	✓	✓	✓	✓	✓	✓	–	–	✓	–
[R7-es] DeriNet.ES	✓	–	–	–	–	–	–	–	–	–
[R8-fa] DeriNet.FA	✓	–	–	–	–	–	–	–	–	–
[R9-it] DerIvaTario	–	–	✓	–	✓	–	–	–	✓	–
[R10-de] DErivBase	✓	–	✓	✓	–	–	–	–	–	✓
[R11-hr] DerivBase.Hr	✓	–	✓	–	–	–	–	–	–	–
[R12-ru] DerivBase.Ru	✓	–	✓	–	–	–	–	–	–	✓
[R13-et] EstWordNet	✓	–	✓	–	–	–	–	–	–	–
[R14-ca] EtymWordNet-cat	✓	–	–	–	–	–	–	–	–	–
[R15-cs] EtymWordNet-ces	✓	–	–	–	–	–	–	–	–	–
[R16-gd] EtymWordNet-gla	✓	–	–	–	–	–	–	–	–	–
[R17-pl] EtymWordNet-pol	✓	–	–	–	–	–	–	–	–	–
[R18-pt] EtymWordNet-por	✓	–	–	–	–	–	–	–	–	–
[R19-ru] EtymWordNet-rus	✓	–	–	–	–	–	–	–	–	–
[R20-sh] EtymWordNet-hbs	✓	–	–	–	–	–	–	–	–	–
[R21-sv] EtymWordNet-swe	✓	–	–	–	–	–	–	–	–	–
[R22-tr] EtymWordNet-tur	✓	–	–	–	–	–	–	–	–	–
[R23-fi] FinnWordNet	✓	–	✓	–	–	–	–	–	–	–
[R24-pt] NomLex-PT	✓	–	✓	–	–	–	–	–	–	–
[R25-en] The M-S Database	✓	–	✓	–	–	✓	–	–	–	–
[R26-pl] The Polish WFN	✓	–	–	–	–	–	–	–	–	–
[R27-la] Word Formation Latin	✓	✓	✓	✓	✓	–	–	–	✓	–

Table 1. Features imported from the individual data resources: derivational relations (DER), compounding relations (COM), part-of-speech tags (POS), morphological categories (MCG), morphological segmentation (SEG), semantic labels (SEM), hierarchical segmentation (HSG), subparadigmatic relations (PAR), unique technical IDs of lexemes (TID), derivational rules (DRL).

morphological segmentation, semantic labels, etc. We also extract custom features, such as derivational rules from DErivBase and hierarchical morphological segmentation from G-CELEX; see Table 1.

Based on the imported relations and features, we identify the original data structure type for each family; see step 1 in Figure 2. The original data for the particular family is presented at the top of Figure 3. In derivation tree structures, where the relations between lexemes are not captured, e.g. in G-CELEX, we generate relations on the basis of the shared (root) morphemes and longer subsequences of morphemes in the lexemes.[4] For instance, derivational relations are generated for *Tour+ist* '*tourist*' on the basis of `Nx` representing suffixation of the base *Tour* (cf. line 5 and 7 in the bottom of Figure 3). We generate compounding relations too, e.g. for *Tour+en+Wagen* '*tour-*

[4]Homonymy of morphemes is a difficult problem to solve here.

ing car' segmented as `NxN` (cf. line 2 in the bottom of Figure 3). However, we do not apply further harmonisation steps to them. In the case of generated derivational relations, we obtain derivational families represented as complete or weakly connected subgraphs (see Figure 4), in which the target rooted-tree data structures have to be identified, if the particular family is not already tree-shaped.

For DeriNet, DeriNet.ES, DeriNet.FA, and the Polish WFN, which contain rooted trees as their original data structure, the following steps 2, 3, and 4 are unnecessary, and the resources are included into the resulting collection by skipping to the last step of the whole procedure (Section 3.5).

3.2. Annotating derivational families

As the next step in harmonisation of non-tree resources, we have identified rooted trees of non-tree-shaped families. Most resources contain only a handful of such families (see Table 2), making it possible to identify the rooted tree manually in all of them. However, CatVar, D-CELEX, E-CELEX, G-CELEX, DErivBase, DerivBase.Hr, DerivBase.Ru and FinnWordNet contain too many non-trees to be handled by hand. In such resources, we have manually annotated a uniformly random sample of 400 to 600 derivational families, which served as training and testing data for development of supervised Machine Learning models.

In all non-tree-shaped derivational families, the annotators' task was to choose derivational relations which form a tree-shaped structure (see step 2 in Figure 2). During the annotations, annotators[5] were not allowed to add any new lexemes and relations. The phenomena on which the annotators decided are exemplified in Figure 5. In tree A (from DErivBase), both the adjective *stehend* '*standing*' and the verb *nachstehen* '*lag behind*' were considered as base lexemes for the adjective *nachstehend* '*lagging behind*' because they share a long common substring, but the verb was chosen as the linguistically adequate solution as *nachstehend* is a present participle form of this verb and is not assumed to be a prefixation of another participle (*stehend*). In contrast, in B in Figure 5 two representations seem to be equally acceptable: either two parallel subtrees are constructed (one for affirmatives *gelenkig* '*flexible*' and *Gelenkigkeit* '*flexibility*', the second one for negatives *ungelenkig* '*inflexible*' and *Ungelenkigkeit* '*inflexibility*'), or negated lexemes are directly linked with their affirmative forms, i.e., *gelenkig* → *ungelenkig* and *Gelenkigkeit* → *Ungelenkigkeit*. We chose the latter solution because it keeps the trees more compact and is insensitive to missing lexemes as compared to the former option, and applied it across the harmonised resources.

[5]We annotated most of the resources ourselves using several monolingual and multilingual dictionaries. In the case of DerivBase.Ru, the annotator was a Russian native speaker and a linguist at the same time.

Input resource	Tree-shaped families	Tree-shaped relations	Non-tree-shaped families	Non-tree-shaped relations	Manually annotated families	Manually annotated relations
[R1-en] CatVar	0	0	13,367	155,064	600	7,618
[R2-nl] D-CELEX	0	0	5,449	1,733,364	419	6,596
[R3-en] E-CELEX	0	0	6,725	109,002	411	6,654
[R4-de] G-CELEX	0	0	5,615	145,936	449	5,720
[R5-fr] Démonette	7,050	12,849	286	1,303	286	1,303
[R9-it] DerIvaTario	0	0	1,992	28,088	440	5,454
[R10-de] DErivBase	15,831	21,795	3,962	33,215	431	5,226
[R11-hr] DerivBase.Hr	0	0	14,818	3,056,962	610	7,548
[R12-ru] DerivBase.Ru	7,653	10,076	10,293	279,817	455	10,754
[R13-et] EstWordNet	428	470	28	65	28	65
[R14-ca] EtymWordNet-cat	2,879	4,422	40	191	40	191
[R15-cs] EtymWordNet-ces	2,284	4,788	70	543	70	543
[R16-gd] EtymWordNet-gla	2,412	4,688	57	403	57	403
[R17-pl] EtymWordNet-pol	2,822	24,106	59	879	59	879
[R18-pt] EtymWordNet-por	1,166	1,586	15	41	15	41
[R19-ru] EtymWordNet-rus	715	2,926	36	474	36	474
[R20-sh] EtymWordNet-hbs	1,694	6,111	20	238	20	238
[R21-sv] EtymWordNet-swe	2,865	4,075	20	376	20	376
[R22-tr] EtymWordNet-tur	1,837	5,188	84	769	84	769
[R23-fi] FinnWordNet	2	2	6,345	29,781	377	2,432
[R24-pt] NomLex-PT	2,751	4,124	34	111	34	111
[R25-en] The M-S Database	5,690	7,580	128	420	128	420
[R27-la] Word Formation Latin	5,230	21,946	43	741	43	741

Table 2. Number of tree-shaped and non-tree-shaped families in the input resources and the size of manually annotated samples. Structures consisting of a single lexeme (so-called singletons), and relations explicitly labelled as compounding are not considered.

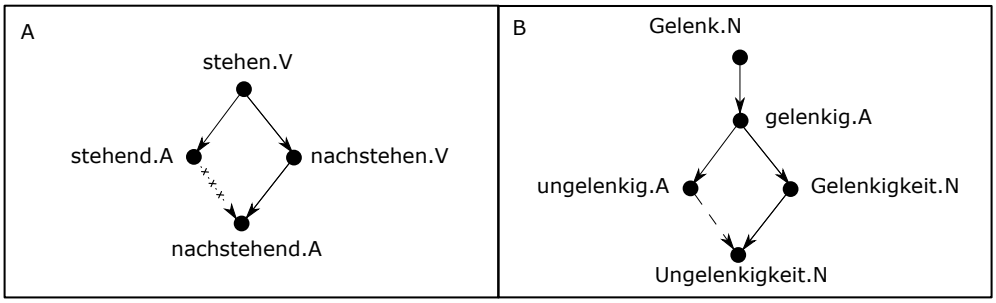

*Figure 5. Manual annotation (DErivBase): A. The prefixed adjective (*nachstehend*) captured as derived from the prefixed verb (vs. the rejected relation represented by the dotted line with tiny crosses). B. The noun with a negative prefix (*Ungelenkigkeit*) can be seen as a deadjectival derivative (from* ungelenkig*; cf. the dashed line) or a denominal derivative (*Gelenkigkeit*); the latter representation is preferred in UDer.*

3.3. Scoring derivational relations

The above-mentioned manual annotation was aimed at selecting a tree-shaped subgraph out of the original resource. Given the annotated data, we want to automatise this task for all families using Machine Learning.

From the Machine Learning perspective, the task can be formalised in various ways. We choose an approach consisting of two phases:

1. We train a scoring model that assigns a numerical score to each edge; the higher the score, the higher the chance that a given edge belongs to the rooted tree.
2. We choose the rooted tree with the maximum sum of edge scores.

We tackle only the first phase using Machine Learning, as described in the following paragraphs. Once the edge scores are given, the globally optimal rooted tree can be found deterministically in the second phase, as described in more detail in Section 3.4.

Manually annotated data does not provide us with any numerical values to train a scorer directly. What we have in each annotated family are relations that were manually marked to be included into the tree (positive examples), while all the other relations from the original data resources are considered as rejected (negative examples). Using this view, we can reformulate the scoring task as a classification task: we train a binary classifier that predicts each relation to be accepted or rejected. Then we use the classifier's confidence about the positive class as the score.

The classification data was prepared as follows: we split the annotated data into the training (65 %), validation (15 %), and hold-out (20 %) sections, and provided all positive and negative instances with the following one-hot encoded features: (a) part-of-speech categories, (b) morphological categories, such as gender, aspect, etc., if present in the original resource, (c) initial and final character n-grams of both the base lexeme and derivative, (d) custom features included in particular resources, e.g. derivational rules in DerivBase.Ru; and of the following numeric features: (e) Levenshtein distance (Levenshtein, 1966), (f) Jaro-Winkler distance (Jaro, 1989; Winkler, 1990), (g) Jaccard distance (Jaccard, 1912), and (h) length of the longest common substring.

Table 3 summarises performance of the following classification methods: Naive Bayes, Logistic Regression, Decision Tree, Random Forest, Perceptron, and K-Nearest Neighbour. Clearly, it was necessary to train a separate classification model for each data resource. If hyper-parameter settings were needed, the values were set using grid-search on the training and validation sections. The standard classification methods are compared with a simple probabilistic baseline whose score is a maximum-likelihood conditional probability estimation conditioned only by the pair of POS values of related lexemes.

Finally, two things should be emphasised. First, the achieved performances are not directly comparable across different data resources, as the complexity of the particular classification tasks might be highly different. Second, the classification performance

Resource	ML method	ε	Scoring relations VALIDATION	Scoring relations HOLDOUT	Identifying trees VALIDATION	Identifying trees HOLDOUT
[R1-en] CatVar	Decision Tree	0.5	44.6 / 82.4	44.9 / 80.7	51.6 / 83.1	53.3 / **81.0**
[R2-nl] D-CELEX	Decision Tree	0.3	47.2 / 81.1	47.7 / 77.1	54.2 / 81.1	53.0 / **79.5**
[R3-en] E-CELEX	Decision Tree	0.5	47.1 / 74.0	47.1 / 74.0	59.7 / 74.9	59.4 / **73.8**
[R4-de] G-CELEX	Decision Tree	0.5	45.8 / 75.6	46.1 / 76.8	57.5 / 79.5	57.5 / **77.4**
[R9-it] DerIvaTario	Decision Tree	0.6	47.7 / 77.5	47.5 / 76.0	48.7 / 78.1	50.0 / **75.1**
[R10-de] DErivBase	Logistic Regression	0.1	24.9 / 88.6	25.4 / 85.8	75.1 / 93.4	78.9 / **92.1**
[R11-hr] DerivBase.Hr	Decision Tree	0.2	45.2 / 77.2	45.4 / 80.7	56.4 / 81.1	58.3 / **81.0**
[R12-ru] DerivBase.Ru	Logistic Regression	0.0	35.1 / 83.0	34.1 / 83.1	49.3 / 84.4	45.0 / **85.5**
[R23-fi] FinnWordNet	Random Forest	0.3	38.2 / 74.0	37.8 / 70.1	62.0 / 80.2	62.9 / **76.9**

Table 3. Evaluation of F-scores calculated for harmonisation procedure that uses simple baseline vs. Machine Learning model (in form: `simple_baseline / ml_model`).

should be considered only a proxy measure and cannot be assumed to correlate perfectly with the quality of induced rooted trees.

3.4. Constructing rooted trees

We construct the resulting score-optimal rooted tree on top of a derivational family with scored relations using the Maximum Spanning Tree (MST) algorithm (Chu and Liu, 1965; Edmonds, 1967). If any rooted tree exists in the input graph, then this algorithm is guaranteed to find the rooted tree with maximum sum of scores, see step 4 in Figure 2.

However, rooted trees are not guaranteed to exist in derivation families imported from the input resources. In order to make sure that the MST algorithm will not fail, we add a temporary virtual root into each family and connect it with all lexemes in the family (see Figure 6.). The score of such newly added relations is equal to ε; the optimal value of ε is grid-searched using the validation sections for each resource separately.

Besides guaranteeing that the MST algorithm will not fail, adding a virtual root with ε-scored relations makes it possible to effectively split a family into two or more disconnected components, as the virtual root is deleted at the end. However, all roots of the divided family are still interlinked in the last JSON-encoded column in the target file format of the resulting harmonised data, e.g. roots *betreffen* '*to affect*' and *Betreff* '*subject*' in Figure 6.

The performances of identification of rooted trees applied to data predicted by both the simple baseline models and the Machine Learning models are presented in Table 3; the bold numbers show the final performances of the whole harmonisation procedure.

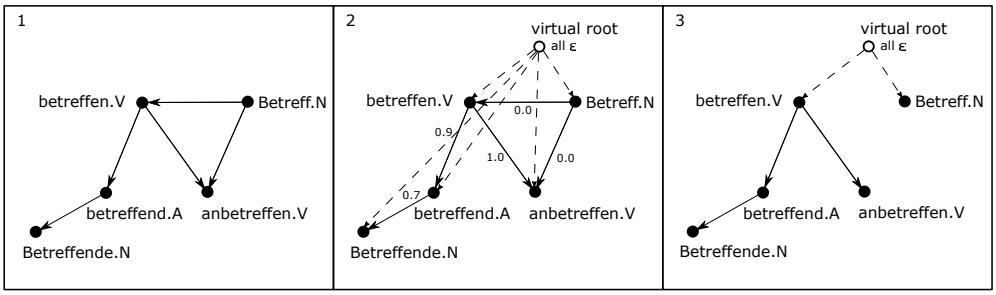

Figure 6. Identification of rooted trees by maximising the sum of scores (DErivBase).

3.5. Converting data into the DeriNet 2.0 format

Finally, we convert the identified rooted trees (except for the virtual root and its relations; cf. step 5 in Figure 2) into the target DeriNet 2.0 file format using the application interface developed for DeriNet 2.0.[6]

We preserve all relations from the original data resources, including compounding relations and relations not chosen by the MST algorithm, just that these additional relations are stored in a less prominent place in the target file format.[7]

We also convert all (custom) features assigned to the lexemes (e.g. part-of-speech categories, morphological categories, morphological segmentation, etc.) and relations, such as semantic labels or derivational rules. Part-of-speech values and morphological categories are also harmonised using the Universal Features annotation schema (Nivre et al., 2016a); however, values of semantic labels are kept in their original forms because they differ significantly across the resources.

We convert the lemma set of each resource and all features assigned to the lexemes first. A unique identifier for each lexeme is created to prevent technical problems caused by the same string form. For example, DErivBase uses a combination of the lemma, its part-of-speech tag, and its gender (for nouns), but G-CELEX combines the lemma and its part-of-speech tag with the original numeric ID to disambiguate lexemes.

After that, we convert tree-shaped derivational relations and add their annotations, for instance, derivational rules from DErivBase. They are stored under the key `Rule=`*x*, where *x* is the original rule identifier. In G-CELEX, a complete (hierarchical) morphological segmentation as well as compounding relations are also included.

[6] `https://github.com/vidraj/derinet/tree/master/tools/data-api/derinet2`

[7] However, we do not preserve non-tree relations in the harmonised versions of CatVar and DerivBase.Hr. It would be too redundant, as we represent their derivational families as complete directed subgraphs initially.

```
1.0 erläutern#VERB  erläutern VERB  _ _ _ _ _ {}
1.1 Erläutern#NOUN#Neut Erläutern NOUN  Gender=Neut _ 1.0 Rule=dVN09&Type=Deriv _ {}
1.2 Erläuterung#NOUN#Fem  Erläuterung NOUN  Gender=Fem  _ 1.0 Rule=dVN07&Type=Deriv _ {}
1.3 erläuternd#ADJ  erläuternd  ADJ _ _ 1.0 Rule=dVA12&Type=Deriv _ {}
1.4 erläutert#ADJ erläutert ADJ _ _ 1.0 Rule=dVA13&Type=Deriv _ {"other_parents": "1.2&Rule=dNA25"}
```

Figure 7. A derivational family from DErivBase harmonised in the target file format.

Finally, we add some other information, such as the original non-tree derivational relations excluded during the harmonisation and links between tree roots if an original family was divided after the identification of rooted trees. [8] These annotations are stored in the last JSON-encoded column in the target file format.

Figure 7 presents a derivational family from DErivBase harmonised to the final file format. The meaning of individual columns is as follows: (i) internal ID consisting of the word-formation family number and the lexeme number separated by a dot, (ii) unique identifier for each lexeme, (iii) lemma, (iv) part-of-speech tag, (v) morphological features, (vi) surface morphological segmentation, (vii) ID of the base lexeme, (viii) annotations of the relation referenced to by the internal ID, (ix) column reserved for other potential relations, (x) JSON-encoded data.

4. Universal Derivations collection

The resulting collection, Universal Derivations version 1.0 (UDer 1.0), contains 27 resources covering 20 languages; see Table 4 summarising basic characteristics of the collection and Figure 8 with examples of harmonised trees. If a particular language is covered by more resources in the collection, the same lexeme was chosen, cf. the English verb *to abandon* in Catvar, E-CELEX, and the Morpho-Semantic Database, or the Russian noun ***весна*** '*spring*' with different derivatives and different relations in DerivBase.Ru and EtymWordNet-rus. The tree of the Polish verb *chcieć* '*to want*' in the Polish WFN differs from the EtymWordNet-pol tree in that it also includes (inflected) word forms.

4.1. Selected quantitative and qualitative properties

Selected quantitative and qualitative details on the UDer 1.0 collection are documented in Table 4 and commented in the following subsections.

Lex(emes). The lexeme sets are adopted from the original data resources, except for EstWordNet, FinnWordNet, and Etymological WordNets from which only derivationally related lexemes are taken. Multi-word lexemes (used mostly for phrasal verbs

[8] They are not preserved for the harmonised versions of CatVar and DerivBase.Hr, because we represented their families as complete (directed) subgraphs at the beginning of the procedure.

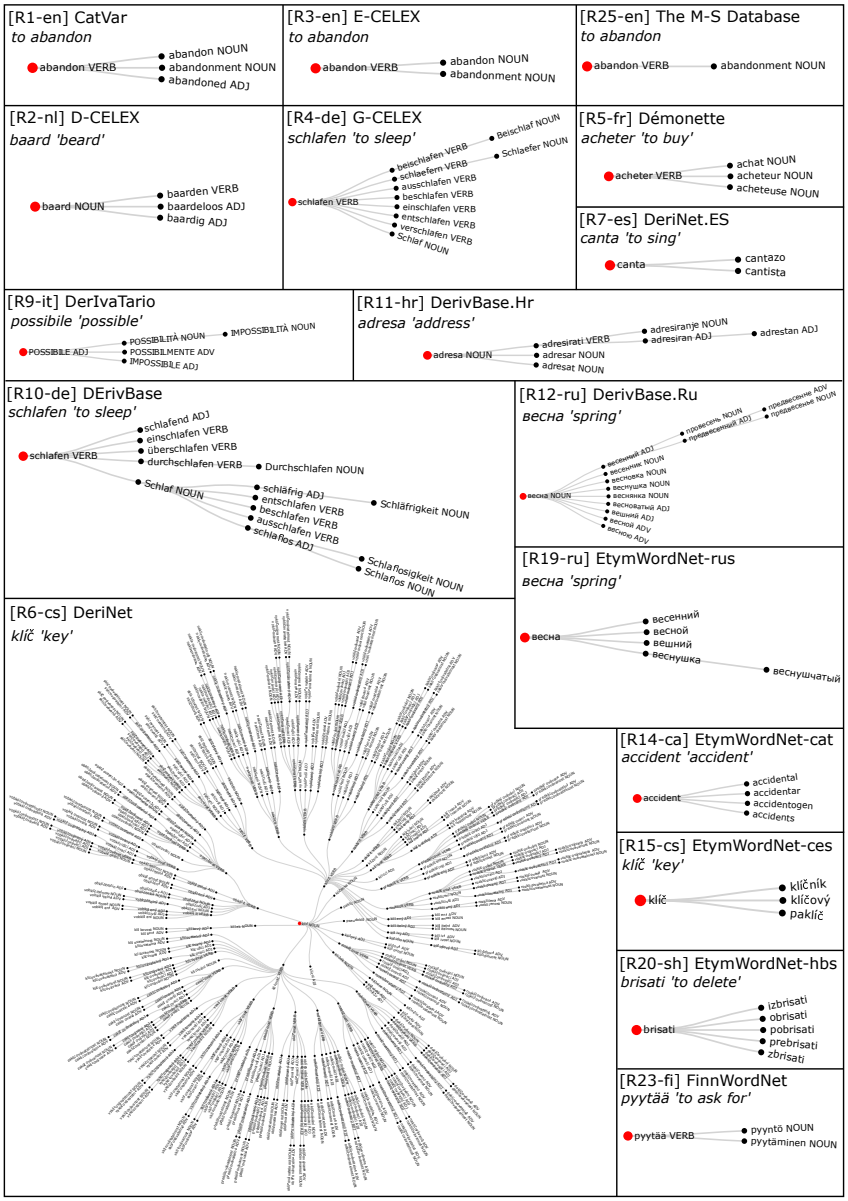

Figure 8. Harmonised rooted trees in UDer 1.0 (part 1).

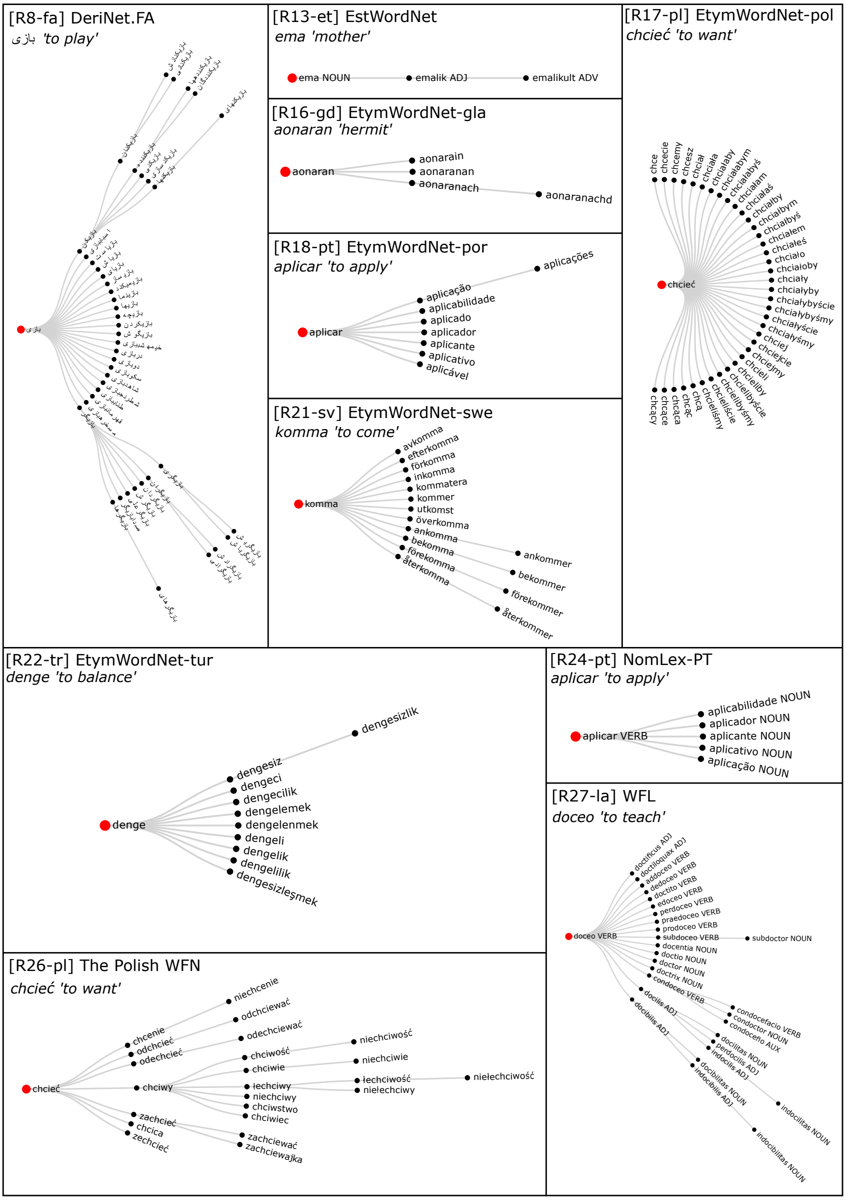

Figure 8. Harmonised rooted trees in UDer 1.0 (part 2).

Lang & Resource	Lex	Rel	Fam	Singl	Size	Depth	Out-deg	POS distrib.	License
[R1-en] CatVar	82,675	24,873	57,802	45,954	1.4/18	0.3/7	0.3/10	60/24/11/5/0	OSL-1.1
[R2-nl] D-CELEX	125,611	13,435	112,176	107,112	1.1/301	0.1/11	0.1/73	63/8/8/1/21	–
[R3-en] E-CELEX	53,103	9,826	43,277	37,951	1.2/51	0.2/8	0.2/33	47/15/13/7/18	–
[R4-de] G-CELEX	53,282	13,553	39,729	34,156	1.3/39	0.2/11	0.3/35	52/17/17/2/12	–
[R5-fr] Démonette	21,290	13,808	7,482	69	2.8/12	1.1/4	1.8/8	63/2/34/0/0	C +NC 3.0
[R6-cs] DeriNet	1,027,665	809,282	218,383	96,208	4.7/1638	0.8/10	1.1/40	44/35/5/16/0	C +NC 3.0
[R7-es] DeriNet.ES	151,173	36,935	114,238	98,325	1.3/35	0.2/5	0.3/14	0/0/0/0/0	C +NC 3.0
[R8-fa] DeriNet.FA	43,357	35,745	7,612	0	5.7/180	1.5/6	3.3/114	0/0/0/0/0	C +NC 4.0
[R9-it] DerIvaTario	8,267	1,787	6,480	5,255	1.3/13	0.2/5	0.2/6	51/26/14/9/0	C 4.0
[R10-de] DErivBase	280,775	43,368	237,407	216,982	1.2/46	0.1/5	0.1/13	86/10/5/0/0	C 3.0
[R11-hr] DerivBase.Hr	99,606	35,289	64,317	50,100	1.5/945	0.3/21	0.4/863	59/30/12/0/0	C 3.0
[R12-ru] DerivBase.Ru	270,473	133,759	136,714	116,037	2.0/1142	0.3/13	0.4/36	62/18/17/3/0	Apache 2.0
[R13-et] EstWordNet	988	507	481	22	2.1/3	1.0/2	1.0/3	16/29/8/47/0	C 3.0
[R14-ca] EtymWordNet-cat	7,496	4,568	2,928	8	2.6/13	1.1/4	1.5/13	0/0/0/0/0	C 3.0
[R15-cs] EtymWordNet-ces	7,633	5,237	2,396	14	3.2/48	1.1/4	2.0/42	0/0/0/0/0	C 3.0
[R16-gd] EtymWordNet-gla	7,524	5,013	2,511	15	3.0/15	1.1/3	1.8/13	0/0/0/0/0	C 3.0
[R17-pl] EtymWordNet-pol	27,797	24,876	2,921	19	9.5/75	1.1/3	8.3/66	0/0/0/0/0	C 3.0
[R18-pt] EtymWordNet-por	2,797	1,610	1,187	9	2.4/57	1.0/3	1.3/57	0/0/0/0/0	C 3.0
[R19-ru] EtymWordNet-rus	4,005	3,227	778	15	5.1/44	1.0/3	4.0/44	0/0/0/0/0	C 3.0
[R20-sh] EtymWordNet-hbs	8,033	6,303	1,730	6	4.6/108	1.0/3	3.6/107	0/0/0/0/0	C 3.0
[R21-sv] EtymWordNet-swe	7,333	4,423	2,910	3	2.5/116	1.0/3	1.5/116	0/0/0/0/0	C 3.0
[R22-tr] EtymWordNet-tur	7,774	5,837	1,937	11	4.0/42	1.1/4	2.8/22	0/0/0/0/0	C 3.0
[R23-fi] FinnWordNet	20,035	11,922	8,113	1,461	2.5/20	1.0/5	1.3/14	55/29/15/0/0	C 4.0
[R24-pt] Nomlex-PT	7,020	4,201	2,819	17	2.5/7	1.0/1	1.5/7	60/0/40/0/0	C 4.0
[R25-en] The M-S Database	13,813	7,855	5,958	65	2.3/6	1.0/1	1.3/6	57/0/43/0/0	C +NC 3.0
[R26-pl] The Polish WFN	262,887	189,217	73,670	41,332	3.6/214	1.0/8	1.1/38	0/0/0/0/0	C +NC 3.0
[R27-la] WFL	36,417	32,414	4,003	121	9.1/524	1.7/6	4.3/236	46/29/21/0/4	C +NC 4.0

Table 4. Language resources harmonised in the UDer collection and their basic quantitative features. Columns Tree size, Tree depth, and Tree out-degree are presented in average / maximum value format. Part-of-speech distribution is ordered as follows: nouns, adjectives, verbs, adverbs, and other categories. (C is abbreviation for CC BY-SA in License)

and named entities) occur in E-CELEX (6,600), FinnWordNet (1,297), the Morpho-Semantic Database (105), DerivBase.Ru (60), EstWordNet (14), DerIvaTario (6), and Démonette (2).

Rel(ations). Table 4 counts only relations involved in tree-shaped families; non-tree relations (compounding, etc.) are not included, although they are present in the harmonised data too. Compounds are captured and connected to their base lexemes in D-CELEX (3,949), G-CELEX (2,563), Word Formation Latin (1,747), E-CELEX (621), and DeriNet (600; other 32,479 compound lexemes are identified by a label without (yet) being connected to their base lexemes).

Fam(ilies) and Singl(etons). The column of derivational families counts only families which have more than one lexeme, including families created by splitting the original families into more parts during the identification of rooted trees. Links between the new roots of the divided families are also stored in the harmonised data. As for the amount of singletons (one-node trees), it seems to correlate with the ways

the original resources have been created. Many singleton trees occur in resources that were built by finding derivational relations within the lexeme set (bottom-up approach), i.e., the CELEXes, DeriNet, DeriNet.ES, DErivBase, and the Polish Word-Formation Network, whereas the lower number of singletons is documented for resources that included lexemes depending on whether the lexeme was derivationally linked to any other lexeme (top-down approach). The number of singleton trees could increase during the harmonisation as a result of splitting the original families.

Tree size, depth, and out-deg(ree). The columns describe the average and maximum size of derivational families, their average and maximum depth (i.e., the distance from the tree root to the furthest node), and out-degree (i.e., the highest number of direct children of a single node). In average, the biggest derivational families can be found in EtymWordNet-pol,[9] Word Formation Latin, DeriNet.FA, and DeriNet, while the smallest families are in the CELEXes and DErivBase, as their data is made up mostly of singletons; a similar tendency can also be seen for the maximum tree sizes. The biggest tree with more than 1.6 thousand lexemes is captured in Derinet, namely for the Czech verb *dát* '*to give*'. The tree depths and out-degrees document that NomLex-PT and the Morpho-Semantic Database contain nouns derived from verbs only. The small absolute depths combined with high absolute out-degrees indicate that derivational families are relatively spread in Etymological WordNets.

POS distrib(ution). Lexemes are assigned part-of-speech tags only in less than a half of the harmonised resources. Nouns, adjectives, verbs, and adverbs are captured in CatVar, the CELEXes, DeriNet, DerivBase.Ru, and EstWordNet. Démonette, DErivBase, DerivBase.HR, and FinnWordNet lack adverbs whereas Démonette and DErivBase have a low number of adjectives. Word Formation Latin contains pronouns, auxiliaries, and lexemes unspecified for the part of speech. The Morpho-Semantic Database and NomLex-PT are limited to nouns and verbs only.

Semantic labels. Derivational meanings are labelled in Démonette, DeriNet, and the Morpho-Semantic Database. However, the labels cannot be merged as they have different interpretations; harmonisation of these labels is one of the future tasks. While the Morpho-Semantic Database assigns labels based on WordNet semantic types, i.e., Agent, Body, By, Destination, Event, Instrument, Location, Material, Property, Result, State, Undergoer, Uses, and Vehicle; morpho-syntactic tags are used in Démonette, i.e., ACT, RES, AGF, AGM, and PRP; and labels in DeriNet are rooted in comparative semantic concepts, namely DIMINUTIVE, POSSESSIVE, FEMALE, ITERATIVE (inspired by Bagasheva 2017), and ASPECT.

Morphological segmentation. A partial or complete morphological segmentation is included in the CELEXes, Démonette, DeriNet, DerIvaTario, DerivBase, DerivBase.Ru, and Word Formation Latin, though the scope of annotation differs largely. While a complete morphological segmentation occurs in the CELEXes and DerIvaTario,

[9] It should be mentioned that Etymological WordNets often represent inflectionally related lexemes as derivation.

only those morphemes which are part of a particular derivational relation are segmented in Démonette and Word Formation Latin. Morhological segmentation in DeriNet is currently limited to identification of root morphemes.

4.2. Publishing and licensing

The presented collection UDer 1.0 is freely available in a single data package in the LINDAT/CLARIAH CZ repository[10] under the licenses listed in Table 4. Relevant scripts for harmonising the original resources and releasing the UDer collection are available in the GitHub repository.[11] The UDer data can be processed using software developed within the DeriNet project, especially the DeriSearch tool,[12] and the Python application interface for DeriNet 2.0;[13] see UDer web page for more information and updates.[14]

4.3. Query interface

As illustrated in Figure 8, UDer trees can grow quite big. Even if the file format is text-based, it has been optimised rather for data exchange, and it is difficult to read by a naked eye (especially when patterns composed of multiple nodes are considered). Thus a specialised interface is needed for human users to browse the UDer data.

Currently we use an updated version of DeriSearch (Vidra and Žabokrtský, 2017) for searching and visualisation purposes. The query language of the tool was recently extended to support querying non-tree graph structures such as compounding, as well as specific node and relation features such as morphological segmentation and semantic labels (Vidra and Žabokrtský, 2020).

The query language supports searching for individual lexemes by imposing regular expression conditions (possibly more of them, combined using logical operators) on their properties. At the same time, it supports searching for contiguous subgraphs composed of multiple lexemes connected by word-formation relations.

Figure 9 shows results for two sample queries. The first query searches for a single specific node whose lemma is *betreffen*, while the second query expresses a more general pattern that matches any node which is a verb and from which at least three child nodes are derived (without conditioning their properties).

Quantitative results of another set of DeriSearch queries, this time applied across several languages, are summarised in Table 5. Given that POS is the only node attribute conditioned in the four queries, we evaluated the queries on all UDer datasets

[10] http://hdl.handle.net/11234/1-3236

[11] https://github.com/lukyjanek/universal-derivations

[12] https://quest.ms.mff.cuni.cz/derisearch2/v2/databases/

[13] https://github.com/vidraj/derinet/tree/master/tools/data-api/derinet2

[14] http://ufal.cz/universal-derivations

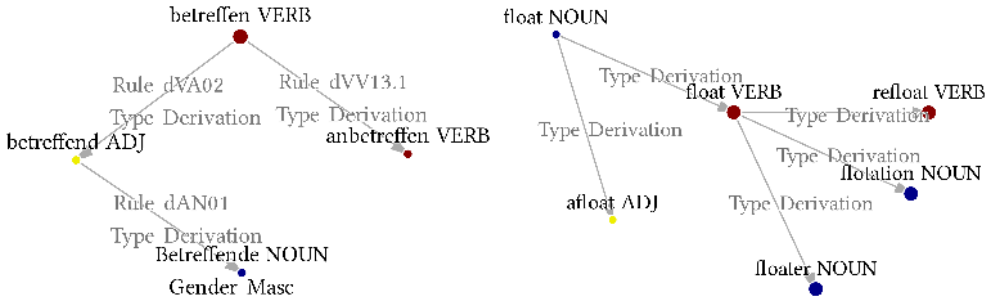

Figure 9. Result of searching for `[lemma="betreffen"]` in DErivBase using DeriSearch (left). Notice that the noun Betreff *is not present, as explained in Section 3.4 and illustrated in Figure 6. One of the results for `[pos="VERB"] ([], [], [])` in the E-CELEX database, visualised by DeriSearch (right).*

in which POS values are available. Please note that the columns are described using a shortened notation, for instance `V(N,N,N)` corresponds to query:

```
[pos="VERB"]([pos="NOUN"],[pos="NOUN"],[pos="NOUN"])
```

Let us use, for example, French, German, Croatian, and Czech to illustrate the subgraphs found by DeriSearch:

- `V(N,N,N)` represents three different nouns derived from the same verb, such as in the case of the French nouns *armeteur* '*armeteer*', *armeur* '*armorer*', and *armement* '*armament*', all derived from the verb *armer* '*to arm*'.
- `V(A,A)` represents two adjectives derived from the same verb, such as the German adjectives *heimatlich* '*native*' and *heimatlos* '*without homeland*' derived from *Heimat* '*homeland*'.
- `V N A` represents patterns in which an adjective is derived from a noun which is derived from a verb, such as in the case of Croatian triple *obujmiti* '*to embrace*', *obujam* '*volume*', *obujamski* '*volumetric*'.
- `N(A,A)` represents two adjectives derived from the same noun, such as *Aristotelův* '*Aristotle's*' and *aristotelský* '*Aristotelian*' derived from *Aristoteles* '*Aristotle*' in Czech.

When comparing individual lines in Table 5, one quickly notices striking quantitative discrepancies among the resources. The counts seem to be far from correlated, and hence the variability can be hardly attributed only to different sizes of the input resources (though their sizes differ in the order of magnitude). The existence of genuine linguistic differences among the languages cannot serve as a sole explanation either, as resources for related languages (or even two resources for a same language) differ substantially too. The most viable explanation is that—is spite of our harmon-

	V(N,N,N)	V(A,A)	V N A	N(A,A)
[R1-en] CatVar	558	863	508	156
[R2-nl] D-CELEX	3	0	1	326
[R3-en] E-CELEX	96	49	125	50
[R4-de] G-CELEX	160	123	273	166
[R5-fr] Démonette	1664	0	408	2
[R6-cs] DeriNet	1510	54874	3655	9124
[R9-it] DerIvaTario	6	7	11	1
[R10-de] DErivBase	332	1363	369	283
[R11-hr] DerivBase.Hr	86	445	385	1062
[R12-ru] DerivBase.Ru	1166	265	2342	2511
[R13-et] EstWordNet	0	0	0	0
[R23-fi] FinnWordNet	559	79	263	634
[R24-pt] NomLex-PT	303	0	0	0
[R25-en] The M-S Database	192	0	0	0
[R27-la] WFL	1010	654	468	995

Table 5. Counts of results found for four sample queries (shortened notation) across different resources.

isation efforts—there is still a long way to overcome the diversity of design decisions petrified in the original data resources and to reach fully comparable networks.

5. Conclusions

This paper presented a procedure for unifying annotation schemas of resources capturing word-formation. Twenty-seven resources covering 20 (mostly European) languages were harmonised using a semiautomatic procedure, and their harmonised versions were publicly released under the title Universal Derivations (UDer 1.0). The harmonised resources allow processing data of multiple languages by the same software tools. The DeriSearch engine was presented here as a tool for visualisation and querying the data.

One of the goals of our harmonisation efforts is to initiate a discussion about the design decisions made, including the choice of the target schema and particular features to harmonise. Harmonisation is necessarily a compromise in that it is impossible to keep all information and allow processing it in an efficient unified way at the same time. Still, we hope that the benefits of the presented efforts outweigh the negatives.

Acknowledgements

We would like to thank all researchers who made their derivational resources publicly available under open licenses. Special thanks also go to Anna Nedoluzhko for manual annotations of Russian data and for valuable comments on the draft of this article.

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 260 575. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIAH CZ project (LM2015071, LM2018101).

Bibliography

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

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

Bagasheva, Alexandra. Comparative Semantic Concepts in Affixation. In Santana-Lario, Juan and Salvador Valera, editors, *Competing Patterns in English Affixation*, pages 33–65. Peter Lang, Bern, 2017. ISBN 978-30-3432-701-5.

Buchholz, Sabine and Erwin Marsi. CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning*, pages 149–164. ACL, 2006. doi: 10.3115/1596276.1596305.

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

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

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

Edmonds, Jack. Optimum Branchings. *Journal of Research of the national Bureau of Standards*, 71B (4):233–240, 1967. doi: 10.6028/jres.071B.032.

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

Fellbaum, Christiane, Anne Osherson, and Peter E Clark. Putting Semantics into WordNet's "Morphosemantic" Links. In *Language and Technology Conference*, pages 350–358. Springer, 2007. doi: 10.1007/978-3-642-04235-5_30.

Gerard, de Melo. Etymological Wordnet: Tracing The History of Words. In *Proceedings of the Language Resources and Evaluation (LREC-2014)*, pages 1148–1154, Reykjavik, Iceland, 5 2014. ELRA. ISBN 978-2-9517408-8-4.

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

Habash, Nizar and Bonnie Dorr. A Categorial Variation Database for English. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, volume 1, pages 17–23, Stroudsburg, 2003. Association for Computational Linguistics. doi: 10.3115/1073445.1073458.

Haghdoost, Hamid, Ebrahim Ansari, Zdeněk Žabokrtský, and Mahshid Nikravesh. 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, 2019.

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

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

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

Jaccard, Paul. The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2):37–50, 1912. doi: 10.1111/j.1469-8137.1912.tb05611.x.

Jaro, Matthew A. Advances in Record-linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989. doi: 10.1080/01621459.1989.10478785.

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

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

Kyjánek, Lukáš, Zdeněk Žabokrtský, Jonáš Vidra, and Magda Ševčíková. Universal Derivations v0.5, 2019. URL `http://hdl.handle.net/11234/1-3041`. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kyjánek, Lukáš, Zdeněk Žabokrtský, Jonáš Vidra, and Magda Ševčíková. Universal Derivations v1.0, 2020. URL `http://hdl.handle.net/11234/1-3236`. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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

Kyjánek, Lukáš. Harmonisation of Language Resources for Word-Formation of Multiple Languages. Master's thesis, Charles University, Faculty of Mathematics and Physics, 2020.

Kyjánek, Lukáš, Zdeněk Žabokrtský, Magda Ševčíková, and Jonáš Vidra. Universal Derivations Kickoff: A Collection of Harmonized Derivational Resources for Eleven Languages. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 101–110, Prague, 2019. ISBN 978-80-88132-08-0.

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

Levenshtein, Vladimir I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

Lieber, Rochelle and Pavol Štekauer. *The Oxford handbook of derivational morphology*. Oxford University Press, Oxford, 2014. doi: 10.1093/oxfordhb/9780199641642.001.0001.

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

Lindén, Krister, Jyrki Niemi, and Mirka Hyvärinen. Extending and updating the Finnish Wordnet. In *Shall We Play the Festschrift Game?*, pages 67–98. Springer, 2012. doi: 10.1007/978-3-642-30773-7_7.

Litta, Eleonora, Marco Passarotti, and Chris Culy. Formatio Formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the 3rd Italian Conference on Computational Linguistics*, pages 185–189, 2016. doi: 10.4000/books.aaccademia.1799.

Matoušek, Jiří and Jaroslav Nešetřil. *Invitation to Discrete Mathematics*. Oxford University Press, Oxford, 2009. ISBN 978-0-19-857043-1.

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

McDonald, Ryan, 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. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 92–97. ACL, 2013.

Miller, George. *WordNet: An Electronic Lexical Database*. MIT press, 1998. ISBN 9780262561167.

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

Nivre, Joakim, 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. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Language Resources and Evaluation (LREC-2016)*, pages 1659–1666, Portorož, Slovenia, 2016a. ELRA. ISBN 978-2-9517408-9-1.

Nivre, Joakim, 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. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1659–1666. ELRA, 2016b.

Shafaei, Elnaz, Diego Frassinelli, Gabriella Lapesa, and Sebastian Padó. DErivCELEX: Development and Evaluation of a German Derivational Morphology Lexicon based on CELEX. In *Proceedings of the Workshop on Resources and Tools for Derivational Morphology*, pages 117–127, Milan, Italy, 2017. ISBN 978-88-9335-225-3.

Šnajder, Jan. 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. ELRA, 2014.

Talamo, Luigi, Chiara Celata, and Pier Marco Bertinetto. DerIvaTario: An Annotated Lexicon of Italian Derivatives. *Word Structure*, 9(1):72–102, 2016. doi: 10.3366/word.2016.0087.

Tanguy, Ludovic and Nabil Hathout. 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)*, Nancy, France, 2002. ATALA.

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

Vidra, Jonáš, Zdeněk Žabokrtský, Magda Ševčíková, and Lukáš Kyjánek. 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, 2019b.

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

Vidra, Jonáš and Zdeněk Žabokrtský. Next step in online querying and visualization of word-formation networks. In *Proceedings of the 23rd International Conference on Text, Speech and Dialogue* (*TSD 2020*). Springer, 2020. doi: 10.1007/978-3-030-58323-1_15. In print.

Vodolazsky, Daniil. DerivBase.Ru: A Derivational Morphology Resource for Russian. In *Proceedings of the Language Resources and Evaluation* (*LREC-2020*), volume 20, pages 3930–3936, Marseille, France, 2020.

Winkler, William E. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research*, pages 354–359. ERIC, 1990.

Zeller, Britta, Jan Šnajder, and Sebastian Padó. DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1201–1211. ACL, 2013.

Zeller, Britta, Sebastian Padó, and Jan Šnajder. Towards Semantic Validation of a Derivational Lexicon. In *Proceedings of COLING 2014*, pages 1728–1739, Dublin, Ireland, 8 2014. Dublin City University and Association for Computational Linguistics. ISBN 978-1-941643-26-6.

Zeman, Daniel, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. HamleDT: Harmonized Multi-Language Dependency Treebank. *Language Resources and Evaluation*, 48(4):601–637, 2014. doi: 10.1007/s10579-014-9275-2.

Address for correspondence:
Lukáš Kyjánek
kyjanek@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University
Malostranské náměstí 25
118 00 Praha 1, Czech Republic