Morphological Resources of Derivational Word-Formation Relations

Lukáš Kyjánek
kyjanek@ufal.mff.cuni.cz
Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

November 2018

Abstract

The report focuses on existing morphological resources containing derivational word-formation relations. For each resource, the report describes history, licence, format, data structure and some other basic statistics to compare. Therefore, it means the first step to review and harmonize these resources in a similar way as it has already been done with resources of syntactic trees.

Introduction

Derivational morphology has gained increasing attention in recent years by more and more research groups across Europe. As a result, various computational linguistic resources dealing with the derivational morphology of several languages have been developed. These resources are potential for research of both the NLP and the linguistic typology. However, so far, there was no list or project that would bring and describe these resources together.

The report focuses on resources (simply called as lexicons) which publish and distribute data as digital datasets and which connect lemmas and/or morphemes with derivational relations.

The goal of these lexicons is to provide some kind of derivational word-formation networks. Despite the fact that these lexicons have the same area of interest for many languages (Section 1), they differ a lot, especially in used data structures (Section 2). Moreover, the derivational morphology is sometimes just the addition in a lexicon covering largely other linguistic phenomena, e.g. WordNet. From this point of view, resources of derivational relations were divided into: lexicons of clearly derivational word-formation relations (Section 3), lexicons of other linguistic phenomena (Section 4), corpora (see Section 5), digital explanatory dictionaries (see Section 6).

There also exist automatic analyzers segmenting lemmas to morphemes (including to derivational ones), e.g. French DériF (Namer, 2003), or analysers connecting lemmas, e.g. Czech Derivancze (Pala & Šmerk, 2015), but they are not reviewed here because they present tools, not datasets.

The report is one part of an ongoing Master Thesis of the author. The motivation for earlier publishing of this part is both to provide work covering the current situation of derivational word-formation resources and to find and discuss other potentially existing resources.

Contents

1 Languages overview 3

2 Data structures 4

3 Lexicons of clearly derivational word-formation relations 6
- 3.1 CatVar . . . 6
- 3.2 CELEX . . . 6
- 3.3 CroDeriV . . . 9
- 3.4 Démonette . . . 9
- 3.5 DeriNet . . . 10
- 3.6 DerIvaTario . . . 10
- 3.7 DErivBase . . . 11
- 3.8 DerivBase.hr . . . 12
- 3.9 DErivCelex . . . 12
- 3.10 Framorpho-FR . . . 12
- 3.11 Morphological Treebank . . . 13
- 3.12 Morphonette . . . 13
- 3.13 Nomage . . . 14
- 3.14 NomBank . . . 15
- 3.15 Nomlex-PT . . . 16
- 3.16 Polish WFN . . . 17
- 3.17 POLYMOTS . . . 17
- 3.18 Spanish WFN . . . 18
- 3.19 Unimorph . . . 18
- 3.20 VerbAction . . . 18
- 3.21 WFL . . . 19

4 Lexicons of other linguistics phenomena 20
- 4.1 E-Dictionary . . . 20
- 4.2 E-Lex . . . 20
- 4.3 Sloleks . . . 22
- 4.4 WiktiWF . . . 23
- 4.5 WordNet . . . 23

5 Corpora 30
- 5.1 Russian National Corpus . . . 30
- 5.2 Prague Depencency Treebank . . . 30
- 5.3 Universal Dependencies . . . 32

6 Digital explanatory dictionaries 34

7 Conclusion 36

Appendix A Where to view derivational data online

Appendix B Formats and Licenses of discovered resources

Appendix C Where to find reviewed resources

1 Languages overview

Lexicons connecting lemmas according to their derivational word-formation relations which are reviewed in this report cover 22 languages. They all belong to the Indo-European language family, are used in Europe and fit Standard Average European (Haspelmath, 2001), except 4 languages (Estonian, Finnish, Komi-Zyrian, and Turkish).

There were discovered 51 resources (the author obtained 39 to calculate some basic statistics) and 12 digital explanatory dictionaries containing derivational relations. Table 1 provides a more detailed overview.

Two ongoing projects are very promising and could increase the number of resources and languages. Project OSLIN (Section 6) builds digital explanatory dictionaries. It has published some for Asturian, Catalan and Portuguese, and now it is preparing new ones for Dutch, English, and Mirandes. Unfortunately, they all are available only online, not as datasets. The second one promising project is WiktiWF (Section 4.4). It has an ambition to provide datasets of derivational relations extracted from Wiktionary.org. It has published 5 datasets, and it is preparing 20 more, each for one language individually (it would add 10 new languages to Table 1).

Table 1. An overview of the languages for which resources containing derivational word-formation relations have been found. Numbers of found resources are represented in column Resources (in the format: clearly derivational ones + other phenomena + corpora + explanatory dictionaries = total count). Column Obtained (in the same order) shows how many of them the author has obtained.

Language	Family	Genus	Resources	Obtained
Asturian	Indo-European	Romance	0+0+0+1 = 1	0+0+0+0 = 0
Bulgarian	Indo-European	Slavic	0+1+0+0 = 1	0+0+0+0 = 0
Catalan	Indo-European	Romance	0+0+0+1 = 1	0+0+0+0 = 0
Croatian	Indo-European	Slavic	2+1+0+1 = 4	1+0+0+0 = 1
Czech	Indo-European	Slavic	1+2+1+2 = 6	1+1+1+0 = 3
Dutch	Indo-European	Germanic	1+1+0+1 = 3	1+1+0+0 = 2
English	Indo-European	Germanic	6+2+0+0 = 8	6+2+0+0 = 8
Estonian	Uralic	Finnic	0+1+0+0 = 1	0+1+0+0 = 1
Finnish	Uralic	Finnic	0+1+2+0 = 3	0+1+2+0 = 3
French	Indo-European	Romance	6+1+0+0 = 7	5+1+0+0 = 6
German	Indo-European	Germanic	4+2+0+2 = 8	3+1+0+0 = 4
Italian	Indo-European	Romance	1+0+0+0 = 1	1+0+0+0 = 1
Komi-Zyrian	Uralic	Permic	0+0+1+0 = 1	0+0+1+0 = 1
Latin	Indo-European	Italic	1+0+0+0 = 1	1+0+0+0 = 1
Polish	Indo-European	Slavic	1+2+0+1 = 4	1+2+0+0 = 3
Portuguese	Indo-European	Romance	1+1+0+1 = 3	1+1+0+0 = 2
Romanian	Indo-European	Romance	0+1+0+0 = 1	0+0+0+0 = 0
Russian	Indo-European	Slavic	1+0+1+1 = 3	0+0+0+0 = 0
Serbian	Indo-European	Slavic	0+2+0+0 = 2	0+0+0+0 = 0
Slovene	Indo-European	Slavic	0+1+0+0 = 1	0+1+0+0 = 1
Spanish	Indo-European	Romance	1+0+0+1 = 2	1+0+0+0 = 1
Turkish	Altaic	Turkic	0+1+0+0 = 1	0+1+0+0 = 1

2 Data structures

A selected data structure of lexicon is, as well as its content, the most important parameter. Six main ways how to organize derivational word-formation relations were found. However, two of them (TL, GL) are more format than the data structure, and one of them (RTM) is a variant of another (RTL). Some lexicons combine or offer more than one data structure. In most cases, one or more structures could be automatically converted to another. However, a deeper linguistic insight would be needed to make a complete harmonization of all data structures to one. Following paragraphs describe each discovered data structure separately.

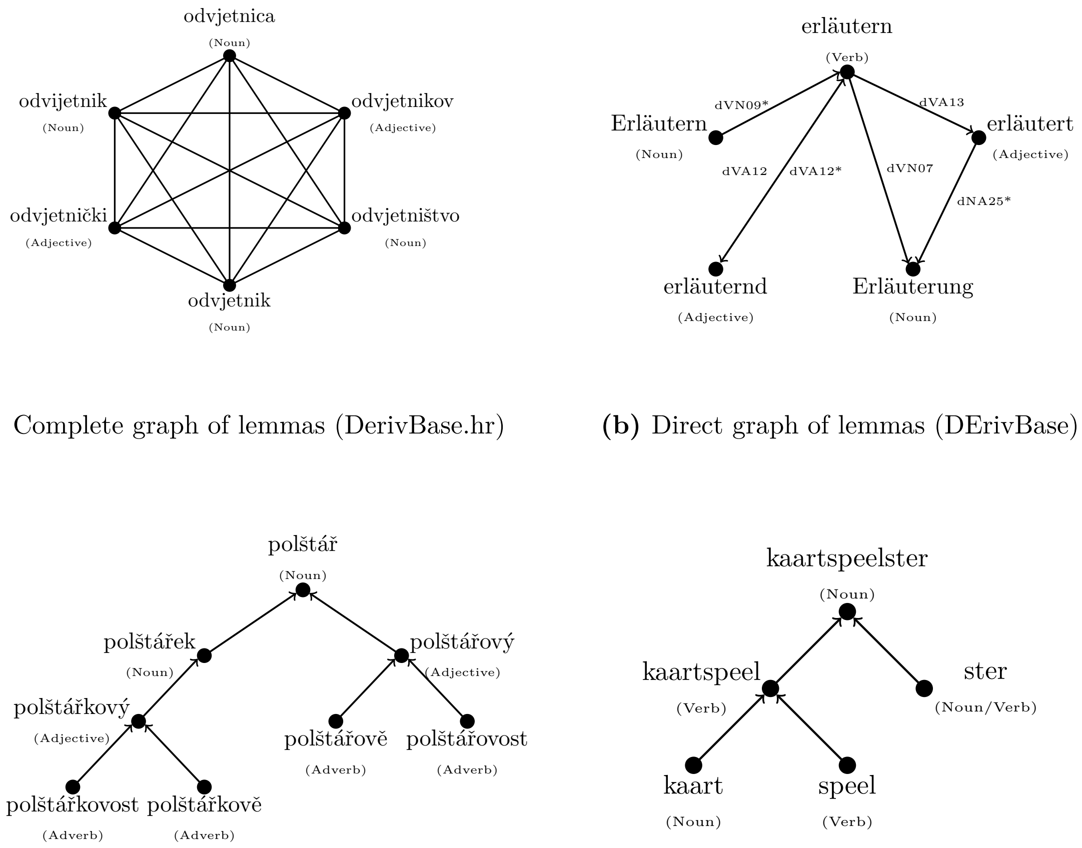
A complete graph of lemmas (CGL) is one of the often used data structures. Each derivational family is represented by the complete graph in which the nodes represent lemmas, and the edges represent derivational relations. There is no more specification of the relations because every node is connected to each other in the derivational family. Figure 1a shows an example of one derivational family from Croatian DerivBase.hr (Šnajder, 2014). This data structure can easily be generated from most other structures described below.

A general graph of lemmas (DGL or IGL) is the most often used data structure. Each derivational family is represented by the general graph in which the nodes represent lemmas, and the edges represent derivational relations. In general graphs of lemmas, each edge connects two specific nodes (ascendant, i.e. derivational parent, and descendant, i.e. derivational child). Two types of general graphs of lemmas can be distinguished: **a direct graph of lemmas (DGL)**, and **an indirect graph of lemmas (IGL)**, however, the second one did not appear among the resources found. Direct graphs of lemmas specify which node is the derivational parent/child (according to the edge direction), whereas indirect graphs of lemmas do not. In both types, some resources label the edges by used derivational rule or semantic category (or both). An important feature of this structure is that it can contain cycles. Figure 1b shows an example of one derivations family from German DErivBase (Zeller, Šnajder, & Padó, 2013).

A rooted tree of lemmas (RTL) is a fairly complex data structure which represents each derivational family as the rooted tree. The nodes represent lemmas, the edges represent derivational relations between two specific nodes (ascendant, i.e. derivational parent, and descendant, i.e. derivational child), and the root represents an unmotivated lemma, i.e. base word, of the derivational family. The main feature of this structure is that it cannot contain cycles. Every node must have only one derivational parent. There is still some space to label relations by used derivational rule or semantic category (or both). This data structure tries to respect linguistic theories about derivational morphology (introduced for example by Dokulil (1962, 1967)). Figure 1c shows an example of one derivational family from Czech DeriNet (Ševčíková & Žabokrtský, 2014).

A rooted tree of morphemes (RTM) is the data structure very similar to the rooted tree of lemmas. However, the nodes represent the individual morphemes of the lemmas (segmented automatically or manually), the edges represent connections between morphemes, and the root represents a whole lemma. Cycles are not allowed and each descendant, i.e. child morpheme, must have only one ascendant, i.e. parent morpheme. If the root of the lemma (lexical morpheme) is not marked, it is difficult to automatically create a complete derivational family precisely. Figure 1d shows an example of one lemma from Dutch D-CELEX (Baayen, Piepenbrock, & Gulikers, 1996) using the rooted tree of morphemes.

Figure 1. Visualization of derivational families according to used data structure.



(a) Complete graph of lemmas (DerivBase.hr)

(b) Direct graph of lemmas (DErivBase)

(c) Rooted tree of lemmas (DeriNet)

(d) Rooted tree of morph. (D-CELEX)

Tagged lemmas (TL) is not a data structure in common sense. This format is usually used in corpora in the form of some tag for the lemma. The tag contains derivational information, e.g. used affix or semantic category. It is quite difficult to create a different structure from this format (used tags assume wider linguistic insight, and then using some data structure).

Glossa for lemmas (GL) is also not a data structure in common sense. This format is usually used in the digital explanatory dictionaries. Glossa means more or less structured text containing derivational information, e.g. used affix or semantic category. It is quite difficult to create a different structure from this format (wide extraction and structuring of data would be necessary, and then using some data structure).

3 Lexicons of clearly derivational word-formation relations

Lexicons containing clearly derivational word-formation relations have been developed in the few past years. There were discovered 26 datasets of this type covering 12 languages (usually Indo-European languages). The author of the report obtained 22 datasets to do more detailed statistics and observations, see Table 2 on page 7.

As was already said above, datasets differ in most properties. For example, different lemma size can be seen, even for the datasets that were created as part of the same project (cf. CELEX, and Polish and Spanish WFN). They also differ in their format (`txt`, `tsv`, `xml`; all data in many files, or in one file) and licenses, see Appendix B.

Following subsections describe each resource individually. Some of them can be queried online. The list of URL links for all discovered resources is listed in Appendix A.

3.1 CatVar

CatVar (Habash & Dorr, 2003), the Categorial-Variation Database, is an automatically created lexicon of English. It focuses on the categorical variations of English lemmas, i.e. derivationally related (especially by suffixation), nouns, adjectives, verbs, and adverbs. The main motivation to build CatVar was to enable the improving results of Information Retrieval (IR), Natural Language Generation (NLG) and Machine Translation (MT).

Lemmas and relations of CatVar come from a combination of resources and algorithms, e.g. *Lexical Conceptual Structure Verb and Preposition Databases*, *Brown Corpus*, morphological analysis lexicon *Englex*, *Nomlex*, *Dictionary of Contemporary English 3*, *Princeton WordNet 1.6*, and *Porter stemmer*. Used method for creating CatVar clusters lemmas based on results of *Porter stemmer*.

Listing 1 shows a format of CatVar. Each line contains a whole derivational family consisting of lemmas with their part-of-speech classification and id of its source separated by a hash sign. Catvar organizes its derivational families to the complete graph of lemmas. The dataset is available for queries on the web (see Appendix A).

Listing 1: An example of format of English CatVar.

```
$ invite_N%3#invite_V%63#invitee_N%35#invited_AJ%1#inviting_AJ%3#invitation_N%11#
    invitation_AJ%1#invitational_AJ%3
$ corrupt_V%63#corrupt_AJ%7#corruption_N%11#corrupted_AJ%1#corrupting_AJ%1#corruptive_AJ
    %1#corruptness_N%33#corruptible_AJ%3#corruptibility_N%1
```

3.2 CELEX

Celex (Baayen et al., 1996) is a large, manually constructed psycholinguistic lexical database available for English, German, and Dutch that was carefully verified by experts and is widely used in psycholinguistics. It involves much linguistic information, including derivational morphology. Derivational relations are represented in the rooted tree of morphemes, however, the root of each lemma is marked, so it would be possible to convert this structure to all others described in Section 2.

Table 2. Statistics for lexicons containing clearly derivational word-formation relations. Abbreviations for Structure coresponds to abbreviations in Section 2. Relations means edges connecting lemmas. Families means non-singleton families of lemmas. Singletons (singleton families) means derivational families consisting of only one lemma. Part-of-speech: N for noun, A for adjective, V for verb, D for adverb, O for others. Minuses in table means that information could not be obtained.

Language	Resource	Ver	Structure	Lemmas	Relations	Families	Singletons	Part-of-speech [%]				
								N	A	V	D	O
Croatian	CroDeriV[1]	1.6	RTM	14,493	-	-	-	0	0	100	0	0
Croatian	DerivBase.hr	1.0	CGL	99,606	1,528,481	14,818	40,733	59	29	12	0	0
Czech	DeriNet	1.6	RTL	1,027,758	803,404	125,799	98,629	44	35	6	15	0
Dutch	D-CELEX[2]	3.1	RTM	121,787	137,161	5,702	35,429	64	8	8	1	19
English	ADJADV	1.0	DGL	5,005	2,581	2,424	0	0	51	1	48	0
English	CatVar	2.1	CGL	82,675	77,532	13,368	38,604	60	24	11	5	0
English	E-CELEX[2]	2.5	RTM	43,650	56,245	10,535	3,165	56	18	16	9	1
English	NOMLEX	2001	DGL	1,964	1,025	941	0	52	0	48	0	0
English	NOMLEXPlus	1.0	DGL	7,756	4,450	3,298	5	57	6	37	0	0
English	NOMADV	1.0	DGL	318	161	158	0	50	0	0	50	0
French	Démonette	1.2	DGL, RTL	22,620	96,027	7,542	0	64	33	3	0	0
French	Famorpho-FR	1.0	CGL	635	2,228	119	54	62	24	11	3	0
French	Morphonette	0.1	DGL	29,310	96,107	8,607	0	58	25	14	3	0
French	Nomage	1.0	DGL	1,298	667	656	11	51	0	49	0	0
French	POLYMOTS	2.0	DGL	-	-	-	-	-	-	-	-	-
French	VerbAction	1.0	DGL	15,885	9,393	6,513	0	58	0	42	0	0
German	DErivBase	2.0	CGL, DGL, RTL	280,336	50,774	20,371	214,916	85	10	5	0	0
German	DErivCelex	2.0	CGL	46,644	189,265	5,422	20,774	58	19	19	3	1
German	G-CELEX[2]	2.0	RTM	51,338	64,372	6,142	4,263	58	19	19	3	1
German	Morph. Treebank[3]	1.0	RTM	-	-	-	-	-	-	-	-	-
Italian	DerIvaTario	1.0	RTM	11,147	14,044	4,872	1,348	51	27	13	9	0
Latin	WFL	1.0.1	RTL	29,504	26,181	4,248	0	46	27	24	2	1
Polish	Polish WFN[4]	0.5	RTL	262,887	189,226	32,333	41,328	-	-	-	-	-
Portuguese	Nomlex-PT	1.0	DGL	7,024	4,238	2,787	0	60	0	40	0	0
Russian	Unimorph[5]	-	RTM	-	-	-	-	-	-	-	-	-
Spanish	Spanish WFN	0.5	RTL	162,751	18,441	11,322	132,988	-	-	-	-	-

[1] The dataset is not available. However, the author of this technical report contacted the authors of CroDeriV and they specified some numbers (in table).

[2] Numbers of Relations, Families and Singletons are for reference only. Used script for linking derivationally related lemmas was quite naive, which affected those numbers.

[3] Statistics could not be calculated for German Morphological Treebank because of the license of GermaNet (the author of this technical report does not have it).

[4] Part-of-speech tags could not be calculated because this resource does not contain part-of-speech tags.

[5] The author of Russian Unimorph do not provide data other than on the web.

3.2.1 D-CELEX

Lemmas of Dutch CELEX (Baayen et al., 1996) come from *Van Dale's Comprehensive Dictionary of Contemporary Dutch*, *Word List of the Dutch Language ('Groene Boekje')*, and text corpus of the Institute for Dutch Lexicology.

Listing 2 shows a format of Dutch CELEX. Each line consists of the lemma and various linguistic information separated by a slash. The most important are positions for the lemma (position 2), morphological status (position 4), simple segmentation (position 9), a marked root of the lemma (position 10), morphological segmentation and part-of-speech classification (position 13).

Listing 2: An example of format of Dutch D-CELEX.

```
$ 33883\gestuntel\11\C\1\Y\Y\Y\ge+stuntel\x1\N\N\((ge)[N|.V],(stuntel)[V])[N]\N\N\N
$ 100013\stuntel\2\U\0\Y\Y\Y\\\\\\\\N
$ 100015\stuntelig\108\C\1\Y\Y\Y\stuntel+ig\Nx\N\N\((stuntel)[N],(ig)[A|N.])[A]\N\N\N
$ 61827\melkhaar\0\C\1\Y\Y\Y\melk+haar\NN\N\N\((melk)[N],(haar)[N])[N]\N\N\N
```

3.2.2 E-CELEX

Lemmas of English CELEX (Baayen et al., 1996) come from *Oxford Advanced Learner's Dictionary* and *Longman Dictionary of Contemporary English.*

Listing 3 shows a format of English CELEX. Each line consists of the lemma and various linguistic information separated by a slash. The most important are positions for the lemma (position 2), morphological status (position 4), simple segmentation (position 12), a marked root of the lemma (position 13 and 21), morphological segmentation and part-of-speech classification (position 22).

Listing 3: An example of format of English E-CELEX.

```
$ 8333\collaborate\72\C\\1\N\N\N\N\Y\col+labour+ate\xNx\ASA\N\N\N\#-ur+r#\N\N\ASA\((col)[
    V|.Nx],((labour)[V])[N],(ate)[V|xN.])[V]\N\N\N
$ 8334\collaboration\102\C\\1\N\N\N\N\Y\collaborate+ion\1x\SA\N\N\N\-e#\N\N\ASAA\(((col)[
    V|.Nx],((labour)[V])[N],(ate)[V|xN.])[V],(ion)[N|V.])[N]\N\N\N
$ 8335\collaborationism\0\C\\1\N\N\N\N\Y\collaboration+ism\Nx\SA\N\N\N\#\N\N\ASAAA\((((
    col)[V|.Nx],((labour)[V])[N],(ate)[V|xN.])[V],(ion)[N|V.])[N],(ism)[N|N.])[N]\N\N\N
```

3.2.3 G-CELEX

Lemmas of German CELEX (Baayen et al., 1996) come from *Bonnlex*, *Molex*, and *Noetic Circle Services*. It should be noticed that German CELEX uses an old orthographical standard. However, Steiner (2016) invented and implemented Perl scripts for spelling correction.

Listing 4 shows a format of German CELEX. Each line consists of the lemma and various linguistic information separated by a slash. The most important are positions for the lemma (position 2), morphological status (position 4), simple segmentation (position 9), a marked root of the lemma (position 10), morphological segmentation and part-of-speech classification (position 14).

Listing 4: An example of format of German G-CELEX.

```
$ 17\abarbeiten\3\C\1\Y\Y\Y\ab+arbeit\xV\N\N\N\((ab)[V|.V],(arbeit)[V])[V]\N\N\N\Y\r2\N
$ 65\Abbrucharbeit\4\C\1\Y\Y\Y\Abbruch+Arbeit\NN\N\N\N\((((ab)[V|.V],(brech)[V])[V])[N
    ],((arbeit)[V])[N])[N]\Y\N\N\N\S3/P3\N
$ 4578\ausarbeiten\139\C\1\Y\Y\Y\aus+arbeit\xV\N\N\N\((aus)[V|.V],(arbeit)[V])[V]\N\N\N\Y
    \r2\N
```

3.3 CroDeriV

CroDeriV (Šojat, Srebačić, Pavelić, & Tadić, 2014) is a manually created lexicon of Croatian. It focuses exclusively on the morphological structure and derivational relatedness of verbs. The main motivation to build CroDeriV was to enrich of the *Croatian WordNet* and the *Croatian Morphological Lexicon*, however, CroDeriV finds a good use also separately as the lexicon of derivational networks.

Lemmas of CroDeriV come from different sources: machine-readable and paper dictionaries of Croatian language, *Croatian National Corpus v3.0*, and *hrWaC*. Used method for creating this lexicon is based on manual analyzing and segmenting lemmas into lexical, inflectional and derivational morphemes. After that, all results were manually checked and corrected.

The dataset is available for queries on the web (see Appendix A), but it is not downloadable. According to information from authors of CroDeriV, it could organize data into the simplified rooted tree of morphemes ("simplified" in this case means that all morphemes are linked to the lexical morpheme).

3.4 Démonette

Démonette (Hathout & Namer, 2014) is the lexicon of the derivational morphology of French. It merges so far existed similar French lexicons. Démonette is a big step for French derivational morphological resources because it has merged all French derivational data for nouns, adjectives and verbs. It focuses on suffixation.

Lemmas of Démonette come from *TLFnome lexicon*, *VerbAction*, and *Lexeur*. Relations and labels are extracted from *DériF*, *Morphonette* and *VerbAction*.

Démonette symmetrically connects derived lemmas to their ascendant, i.e. derivational parent, by direct relations, and to the other lemmas in the derivational family by indirect relations. These relations, as well as the semantic category of lemmas, are labelled. Beside the derivational families, Démonette provides so-called derivational series, i.e. derivational paradigms, similar to these in Morphonette.

Listing 5 shows a format of Démonette. It is an easy-to-read `xml` format. Lemmas and their morphological and semantic labels are inside tags <targetWord> and <sourceWord>. The label of their relation is inside tag <relationType>. Used word-formation process is inside tags <sourceFormConstrucion> and <targetFormConstrucion>. Tag <targetMeaningConstruction> provides the definition for the target word. Démonette organizes its derivational families to the directed graph of lemmas (so-called indirect relations) and also to the rooted tree of lemmas (so-called direct relations).

Listing 5: An example of (`xml`) format of French Démonette.

```
<morphologicalRelation origin="derif">
  <targetWord>
    <writtenForm origin="tlfnome">abaissement</writtenForm>
    <morphoSyntacticTag origin="tlfnome">Ncms</morphoSyntacticTag>
    <morphoSemanticType origin="demonette">@ACT</morphoSemanticType>
  </targetWord>
  <sourceWord>
    <writtenForm origin="tlfnome">abaisser</writtenForm>
    <morphoSyntacticTag origin="tlfnome">Vmn----</morphoSyntacticTag>
    <morphoSemanticType origin="demonette">@</morphoSemanticType>
  </sourceWord>
  <relationType origin="derif">
    <direction>descendant</direction>
    <complexity>simple</complexity>
  </relationType>
  <targetFormConstruction><constructionalProcess origin="derif">suf</
      constructionalProcess>
    <constructionalExponent origin="derif">ment</constructionalExponent>
    <constructionalTheme origin="derif">abaiss</constructionalTheme>
  </targetFormConstruction>
  <sourceFormConstruction>
  </sourceFormConstruction>
  <targetMeaningConstruction>
    <concreteDefinition origin="derif">action de abaisser</concreteDefinition>
    <abstractDefinition origin="demonette">action de @</abstractDefinition>
  </targetMeaningConstruction>
</morphologicalRelation>
```

3.5 DeriNet

DeriNet (Ševčíková & Žabokrtský, 2014) is a semi-automatically created lexicon of Czech. It focuses on all derivational relations between and within nouns, adjectives, verbs, and adverbs.

Lemmas of DeriNet come from the *SYN subcorpus* of the Czech National Corpus and from the Czech inflectional dictionary *MorfFlex CZ*. Used method for creating DeriNet is based on a semi-automatic annotation procedure. Derivational relations are searched and processed automatically from tools for morphological analysis, from a discovered set of derivational rules, and from a grammar-based set of rules, however, before including to DeriNet, they are mostly manually checked.

DeriNet organizes data as the rooted tree of lemmas, so it is currently limited to the derivational relations because they are the most frequent and most productive word-formation process in the Czech language. Nevertheless, DeriNet already marks compounds and processed derivational compounds, but soon (since version 2.0) it will include edges for compounds relations, and it will semantically label some lemmas. In all cases, DeriNet tries to get as close as possible to Dokulil's 1962 approach to the derivational morphology which has become widely respected.

Listing 6 shows a format of DeriNet. Each line contains an id of the lemma, the lemma, its so-called techlemma (for connection to *MorfFlex CZ*, *MorhoDiTa*, *PDT* etc.), part-of-speech classification, and id of the ascendant, i.e. derivational parent for the lemma, all separated by tabulators. The dataset is available for queries on the web (see Appendix A).

Listing 6: An example of format of Czech DeriNet.

```
$ 140365  drženě    drženě_^(*1ý)    D  140368
$ 140366  držení    držení_^(*2t)    N  140369
$ 140367  drženost  drženost_^(*3ý)  N  140368
$ 140368  držený    držený_^(*2t)    A  140369
$ 140369  držet     držet_:T         V
```

3.6 DerIvaTario

DerIvaTario (Talamo, Celata, & Bertinetto, 2016) is a manually created lexicon of Italian. It focuses on analyzing and segmenting nouns, adjectives, verbs, and adverbs into lexical and derivational morphemes.

Lemmas of DerIvaTario come from *COLFIS corpus* (Corpus e Lessico di Frequenza dell'Italiano Scritto). Used method for creating this lexicon is based on manual annotation. After that, all annotations were validated through an inter-annotator agreement experiment.

DerIvaTario is still linked to *COLFIS*, therefore, it provides various information for each lemma. In addition, DerIvaTario is also linked to *Phonitalia*, so the phonetic transcription, syllabification and stress position can be recovered for each lemma, too.

Listing 7 shows a format of DerIvaTario. Each line contains an id of the lemma, wordform of the lemma, and its morphemes, all separated by semicolons. The lexical morpheme is always before derivational ones. Therefore, Derivatario organizes its derivational families to the rooted tree of morphemes, however, it can be easily automatically converted to the complete graph of lemmas. The dataset is available for queries on the web (see Appendix A).

Listing 7: An example of format of Italian DerIvaTario.

```
$ 36937;GOMMISTA;GOMMA:root;ISTA:ista:mt1:ms1;;;;;
$ 36940;GOMMOSO;GOMMA:root;OSO:oso:mt1:ms1;;;;;
$ 46953;LEGALIZZAZIONE;LEGGE:suppl;ALE:ale:mt7:ms1;IZZARE:izzare:mt1:ms1;ZIONE:zione:mt1:
    ms1;;;
$ 49878;MANIERISMO;MANIERA:root;ISMO:ismo:mt1:ms2a;;;;;
$ 49879;MANIERISTA;MANIERA:root;ISMO:ismo:mt1:ms2a;ISTA:ista:mt6:ms1;;;;
```

3.7 DErivBase

DErivBase (Zeller et al., 2013) is a semi-automatically created lexicon of German. It focuses on the derivationally related nouns, adjectives, and verbs. This lexicon and its constructional method inspired the building of several another lexicon, such as *DerivBase.hr* and *DErivCelex*.

Lemmas of Derivbase come from a large German-language web corpus *SDEWAC*. Lemmatization and part-of-speech classification are done by *TreeTagger*. Used method for creating DErivBase is heuristical, based on the rule-based framework for clustering lemmas into derivational families. Rules are manually extracted from grammar books and include zero derivation, prefixation, suffixation, circumfixation, and stem changes. In version 2.0, Zeller, Padó, and Šnajder (2014) automatically split the derivational families into semantically consistent clusters. DErivBase was evaluated on the manually classified sample.

The dataset of DErivBase provides several files. One of them contains a list of derivational families in the same format as *DerivBase.hr* and *CatVar* do (see Listing 8). Another one file contains a list of individual derivational relations labelled by used derivational rules and by the length of the path (see Listing 9). Based on this second format, it is quite well to obtain the rooted tree of lemmas structure. Thus it organizes data to the complete graph, direct graph, and to the rooted tree of lemmas. Derivational rules are separated in another file (see Listing 10).

Listing 8: An example of format of German DErivBase (list of derivatioanl families).

```
$ überbelegen_V Überbelegung_Nf überbelegt_A belegen_V belegend_A belegt_A Belegung_Nf
    belegbar_A Beleg_Nm Beleger_Nm unterbelegen_V Unterbelegung_Nf unterbelegt_A
    Fehlbelegung_Nf fehlbelegt_A
```

Listing 9: An example of format of German DErivBase (list of derivational relations/paths).

```
$ Beleg_Nm Beleger_Nm 1 Beleg_Nm dNN05> Beleger_Nm
$ Beleg_Nm Unterbelegung_Nf 2 Beleg_Nm dNV21> unterbelegen_Ven dVN07> Unterbelegung_Nf
```

Listing 10: An example of format of German DErivBase (list of derivational rules).

```
$ -- rechnen -> Rechnung, entleeren -> Entleerung
$ dVN07 = dPattern "dVN07"
$   (sfx "ung") verbs fNouns
$ -- Tunnel -> untertunneln, Keller -> unterkellern
$ dNV21 = dPatternSS "dNV21"
$   (pfx "unter") nouns verbs
```

3.8 DerivBase.hr

DerivBase.hr (Šnajder, 2014) is an automatically created lexicon of Croatian. It focuses on suffixal derivation between and within nouns, verbs, and adjectives which is a very productive derivational process in Croatian.

Lemmas of DerivBase.hr come from large web corpus *hrWaC*, and they are clustered to derivational families. Used method for creating this lexicon is inspired by Zeller et al. (2013), the authors of *DerivBase*, a similar lexicon for German.

The dataset contains two versions according to the used approach of creating DerivBase.hr: an unsupervised and knowledge-based induction. The unsupervised induction clusters lemmas based on a string distance. The knowledge-based induction uses an inflectional lexicon and a set of derivational patterns. Authors of this resource recommend the knowledge-based version because of its higher quality.

Listing 11 shows a format of Derivbasehr. Each line contains a whole derivational family consisting of lemmas with their part-of-speech classification separated by space. DerivBase.hr organizes derivational families to the complete graph of lemmas.

Listing 11: An example of format of Croatian DerivBase.hr.

```
$ bojovnik_N bojić_N bojev_A bojo_N bojovan_A bojati_V bojište_N bojenje_N bojen_A bojani
    č_N bojanje_N bojan_N bojan_A bojnik_N bojnica_N bojani_A bojano_N bojanov_A
    bojanka_N boj_A bojica_N bojilo_N bojil_N bojiti_V
```

3.9 DErivCelex

DErivCelex (Shafaei, Frassinelli, Lapesa, & Padó, 2017) is an automatically created lexicon of German. It groups derivationally related nouns, adjectives, verbs and adverbs into the derivational families. It includes suffixation, prefixation and also composition and derivational composition.

Lemmas of DErivCelex come from *German CELEX*, a large, manually constructed lexicon resource. Used method for creating DErivCelex is based on processing the manual annotation of *CELEX* format. Shafaei et al. (2017) describe the algorithm, which could be also used on *Dutch CELEX* and *English CELEX* but, to our best knowledge, it so far was not realized. Compared to *DErivBase*, a coverage of DErivCelex is lower, but DErivCelex contains fewer false positives because it is built on the basis of a cleaner resource. However, it also should be noticed that *German CELEX* is a quite old resource, so it contains an old orthographical standard.

Listing 12 shows a format of DErivCelex. Each line contains an id and a whole derivational family consisting of lemmas with their part-of-speech classification separated by space. Derivcelex organizes its derivational families to the complete graph of lemmas.

Listing 12: An example of format of German DErivCelex.

```
$ 10 unabänderlich_A unveränderlich_A veränderbar_A abändern_V Veränderlichkeit_N Ä
    nderung_N umändern_V änderbar_A abänderlich_A ändern_V veränderlich_A Abänderung_N
    verändern_V Unveränderlichkeit_N Umänderung_N Veränderung_N
```

3.10 Framorpho-FR

Framorpho-FR (Hathout, 2005) is a semi-automatically created lexicon of French derivational families. It groups derivationaly related nouns, adjectives, verbs, adverb, and one interjection. It focuses on prefixation, suffixation and conversion.

Lemmas of Framorpho-FR come from *Trésor de la Langue Française (TLF)* starting with FR-. Derivational relations were discovered using program DeClique (Hathout, 2005). After that, data was manually revised.

Listing 13 shows a format of Framorpho-FR. It is distributed in easy-to-read `xml`.

Listing 13: An example of format of French Famorpho-FR.

```
<family>
<entry><written_form>fraise</written_form><cat>noun</cat></entry>
<entry><written_form>fraiser</written_form><cat>verb</cat></entry>
<entry><written_form>fraisé</written_form><cat>adjective</cat></entry>
</family>
```

3.11 Morphological Treebank

Morphological Treebank for German (Steiner, 2017) is a semi-automatically created lexicon of German. It merges derivational data from *German CELEX* and *GermaNet*, so it contains nouns, adjectives, verbs and adverbs.

Lemmas (95k) and derivational relations of Morphological Treebank come from *German CELEX* and *GermaNet*. *German CELEX* is a large, manually constructed psycholinguistic lexical database, and *GermaNet* is a German WordNet, lexical-semantic database. Resulting Morphological Treebank merges these two resources, corrects an old orthographical standard of *German CELEX* (Steiner, 2016), and adds 18 additional word-formation rules. It covers all German word-formation processes, including composition.

Due to the license agreement of *GermaNet* and *CELEX*, the dataset of Morphological Treebank for German is not available, but the scripts for building this resource are. However, the *GermaNet* and *German CELEX* are needed.

3.12 Morphonette

Morphonette (Hathout, 2010) is an automatically created lexicon of French. It focuses on the derivationally related nouns, adjectives, verbs and adverbs. This lexicon has become a part of *Démonette*.

Lemmas of Morphonette come from *TLFnome* and *TLFindex lexica*. Used method for creating Morphonette is based on measurement of morphological similarity between lemmas and on the formal analogy of discovered paradigms.

Morphonette is grounded in a paradigmatic conception of derivational morphology. Morphonette differs between derivational families (family of lemmas with the same root) and derivational series (set of lemmas that forms same formal analogies). Derivational series is described as 'triplets of the form $(w, r, sr(w))$, where w is the entry, r is a member of the derivational family of w and $sr(w)$ is the derivational series of w with respect to r; in other words, $sr(w)$ is the set of words that participate in relations similar to the relation between w and r.' (Hathout & Namer, 2014)

Listing 14 shows a format of Morphonette. It consists of the relation between ascendant, i.e. derivational parent, and descendant, i.e. derivational child. Each this relation also includes lemmas of its derivational series. Morphonette organizes its derivational families to the directed graph.

Listing 14: An example of format of French Morphonette.

```
<filament>
<entry><written_form>frissonner</written_form><transcription>ffrriissoonnei</
    transcription><cat>Vmn----</cat></entry>
<parent><written_form>frisson</written_form><transcription>ffrriisson</transcription><cat
    >Ncms</cat></parent>
<sub_series>
<member><written_form>buissonner</written_form><transcription>bbuyiissoonnei</
    transcription><cat>Vmn----</cat></member>
<member><written_form>hérissonner</written_form><transcription>eirriissoonnei</
    transcription><cat>Vmn----</cat></member>
<member><written_form>friponner</written_form><transcription>ffrriippoonnei</
    transcription><cat>Vmn----</cat></member>
<member><written_form>palissonner</written_form><transcription>ppaalliissoonnei</
    transcription><cat>Vmn----</cat></member>
<member><written_form>polissonner</written_form><transcription>ppoolliissoonnei</
    transcription><cat>Vmn----</cat></member>
<member><written_form>saucissonner</written_form><transcription>ssaussiissoonnei</
    transcription><cat>Vmn----</cat></member>
<member><written_form>soupçonner</written_form><transcription>ssouppssoonnei</
    transcription><cat>Vmn----</cat></member>
</sub_series></filament>
```

3.13 Nomage

Nomage (Balvet, Barque, & Marín, 2010) is a semi-automatically created lexicon of French so-called nominalizations. Nominalization in this sense corresponds to nominalizations in generative grammar described, for example, by Chomsky et al. (1968). It tries to propose data representing an influence of verbs, adjectives or adverbs on the nominalised form. Nomage focuses on nouns derived from verbs.

Lemmas of Nomage come from *French Treebank*. The used method processes common nouns (excluding person names) with one of following affixes: -ade, -age, -ance, -ée, -ence, -ment, -sion, -tion, -ure, -xion. In a broader sense, it can be claimed that Nomage is semantically tagged. It uses 4 tags (and their combinations; e.g. etat = state, act = activity, acc = achievement, ach = perfective) for tagging verbs and 3 tags (hab = habit, objet = object, objetinfo = informational object) for tagging nouns. Tags are based on transformational tests of aspectual patterns (originated from valency frames) proposed by the authors of Nomage.

Listing 15 shows a format of Nomage. It is an easy-to-read `xml` format. Unfortunately, it is formatted incorrectly, so it needs to be processed with regular expressions as `txt` format. Derivational child and its part-of-speech classification can be found in a tag <feat att="POS" ...>. The next one tag specifies derivational process. Information about meaning and an aspectual pattern is placed between tags <Sense>. Derivational parent is written in tag <SenseRelation>.

Listing 15: An example of format of French Nomage.

```
<LexicalEntry>
  <Lemma>
    <feat att="POS" val="noun"/><feat att="writtenForm" val="abjuration"/>
    <feat att="affix" val="ion"/>
  </Lemma>
  <Sense id="abjuration1">
    <PredicativeRepresentation>
      <feat att="label" val="abjuration de Y par X"/>
      <feat att="patron" val="N de Y par X"/></PredicativeRepresentation>
      <AspectualClass><feat att="label" val="ACH"/></AspectualClass>
    <SenseExample>
      <val-list><feat att="label" val="Guerre ethnique larvée au Caucase, dialogue de
          sourds entre Gorbatchev et les Lituaniens, _*abjuration*_ du communisme par le
          PC polonais, spectaculaires valses - _*hésitations*_, en Roumanie et en RDA, de
           ce qu’ on hésite Ã  appeler encore pouvoir ; heurts, en Bulgarie, entre pro et
           anti-turcophones, risque grandissant d’_*implosion*_ de la Yougoslavie : 1990
          a démarré tellement en fanfare, dans les pays de l’Est, qu’on a le sentiment
          de n’avoir encore rien vu."/>
      </val-list>
    </SenseExample>
  </Sense>
  <SenseRelation target="abjurer1"/>
</LexicalEntry>
```

3.14 NomBank

NomBank (Meyers et al., 2004) is a semi-automatically created annotation project of English so-called nominalizations. Nominalization in this sense corresponds to nominalizations in generative grammar described, for example, by Chomsky et al. (1968). It tries to propose data representing an influence of verbs, adjectives or adverbs on the nominalised form. NomBank provides data for nominalizations of verbs, adjectives, and adverbs.

Lemmas and relations come from *PropBank Corpus* (the *Wall Street Journal Corpus* of the *Penn Treebank*). NomBank also includes and enlarges NOMLEX which was developed separately before the NomBank project.

3.14.1 NOMLEX

NOMLEX (Macleod, Grishman, Meyers, Barrett, & Reeves, 1998) is a manually created lexicon of English nominalizations of verbs (nouns derived from verbs). It started with noun affixes: -ion, -ment, -er, -ee, -al, -ing and with a conversion. First, it was developed separately from the NomBank project and focused on nouns derived from verbs. When the NomBank project started, NOMLEX was improved and included as *NOMLEXPlus* to the NomBank project.

Listing 16 shows a format of NOMLEX. It is well-structured `txt` format. Each entry starts with NOM and an orthographical form of a derivational child (ORTH) following by derivational parent (VERB) and valency patterns.

Listing 16: An example of format of English NOMLEX.

```
(NOM       :ORTH "abasement"     :VERB "abase"
                                 :PLURAL *NONE*
                                 :NOM-TYPE ((VERB-NOM))
                                 :VERB-SUBJ ((NOT-PP-BY)
                                             (DET-POSS))
                                 :SUBJ-ATTRIBUTE ((COMMUNICATOR))
                                 :OBJ-ATTRIBUTE ((COMMUNICATOR))
                                 :VERB-SUBC ((NOM-NP :OBJECT ((DET-POSS)
                                                              (N-N-MOD)
                                                              (PP-OF)))))
```

3.14.2 ADJADV

ACJADV is an automatically created lexicon of English adverbs (and 9 verbs) derived from adjectives. During the creating process, *CatVar* was used.

Listing 17 shows a format of ADJADV. It is well-structured `txt` format as well as NOMLEX. Each entry starts with ADJADV and an orthographical form of a derivational parent (ORTH) following by derivational child (ADV) and valency patterns.

Listing 17: An example of format of English ADJADV.

```
(ADJADV :ORTH "abject"
        :ADV "abjectly"
        :FEATURES ((MANNER-ADV))
        :SEMI-AUTOMATIC T)
```

3.14.3 NOMADV

NOMADV is an automatically created lexicon of English adverbs derived from nouns.

Listing 18 shows a format of NOMADV. It is well-structured `txt` format as well as NOMLEX. Each entry starts with NOMADV and an orthographical form of a derivational parent (ORTH) following by derivational child (ADV) and valency patterns.

Listing 18: An example of format of English NOMADV.

```
(NOMADV :ORTH "alternative"
        :ADV "alternatively"
        :FEATURES ((META-ADV :EPISTEMIC T))
        :SEMI-AUTOMATIC T)
```

3.14.4 NOMLEXPlus

NOMLEXPlus is based on a manualy created *NOMLEX*. It enlarges *NOMLEX* with nominalizations of adjectives.

Listing 19 shows a format of NOMLEXPlus. It is well-structured `txt` format as well as NOMLEX. Each entry starts with NOM or NOMADJ and an orthographical form of a derivational child (ORTH) following by derivational parent (VERB or ADJ) and valency patterns.

Listing 19: An example of format of English NOMLEXPlus.

```
(NOMADJ :ORTH "ability"
        :ADJ "able"
        :NOM-TYPE ((ADJ-NOM))
        :FEATURES ((GRADABLE))
        :SUBJ-ATTRIBUTE ((NHUMAN)
                         (ACTION)
                         (COMPANY)
                         (COMMUNICATOR))
        :OBJ-ATTRIBUTE ((PROPOSITION)
                        (ACTION))
        :ADJ-SUBC ((NOM-INTRANS :SUBJECT ((N-N-MOD)
                                          (DET-POSS)
                                          (PP :PVAL ("of"))))
                   (NOM-ADJ-TO-INF :SUBJECT ((N-N-MOD)
                                             (DET-POSS)
                                             (PP :PVAL ("of")))
                                   :NOM-SUBC ((TO-INF :SC T))))
        :SEMI-AUTOMATIC T)
```

3.15 Nomlex-PT

Nomlex-PT (De Paiva, Real, Rademaker, & De Melo, 2014) is a semi-automatically created lexicon of Portuguese nominalizations. Nominalization in this sense corresponds to nominalizations in generative grammar described, for example, by Chomsky et al. (1968). It tries to propose data representing an influence of verbs, adjectives or adverbs on the nominalised form. Nomlex-PT focuses on nouns derived from verbs, and it is also known as Nomlex-BR because it processes Brazilian Portuguese.

Lemmas and relations come from various resources, e.g. *Portuguease Wiktionary*, *FrameNet*, *AC/DC Corpus*, translated from English *NOMLEX* and French *NOMAGE* and Spanish *AnCora-Nom.* It started with nouns ending with affixes: -ção, -mento, ida, -ura, -or, -nte, -ada, and some others. Nomlex-PT was fully integrated into OpenWordNet-PT (Rademaker, De Paiva, de Melo, & Coelho, 2014).

Listing 20 shows a format of Nomlex-PT. It is `rdf` format. Plural form of derivational child is placed between tags <nomlex:plural ... >. Derivational parent can be found between tags <nomlex:verb ... > and derivational child between tags <nomlex:noun ... >. Tag <dc:provenance ... > contains information about the origin of described relation.

Listing 20: An example of format of Portuguese Nomlex-PT.

```
<Description rdf:about="http://arademaker.github.com/nomlex-br/instances/nomlex-beirar-
    beira">
 <nomlex:plural xml:lang="pt">beiras</nomlex:plural>
 <rdf:type rdf:resource="http://arademaker.github.com/nomlex/schema/Nominalization"/>
 <nomlex:verb rdf:resource="http://arademaker.github.com/wn30-br/instances/word-beirar"/>
 <nomlex:noun rdf:resource="http://arademaker.github.com/wn30-br/instances/word-beira"/>
 <dc:provenance xml:lang="pt">wiktionary-en</dc:provenance>
</Description>
```

3.16 Polish WFN

Polish Word-Formation Network (Lango, Ševčíková, & Žabokrtský, 2018) is a semi-automatically created lexicon of Polish. It focuses on the construction of derivational networks between lemmas, especially using an unsupervised method. It was constructed together with *Spanish WFN.*

Lemmas of Polish WFN come from the *Grammatical Dictionary of Polish*, and *Polish WordNet.* Used method for creating Polish WFN is based on a sequential pattern mining technique to construct useful morphological features in an unsupervised manner. For each lemma, the most probable base word is chosen from a resulting list of possible base words given by the model. Finally, the derivational relations from *Polish WordNet* were extracted and included to Polish WFN.

Listing 21 shows a format of Polish WFN. It keeps the same data structure (the rooted tree of lemmas) and format as *DeriNet*, although the techlemma does not make sense (because it does not serve to connect other linguistic resources, nor does it bring new information about lemma), and the part-of-speech classification is unspecified. However, it could be considered as a first step to harmonize lexicons of derivational networks. Each line contains an id of the lemma, the lemma, its so-called techlemma (it is same as lemma), part-of-speech classification (empty), and id of the ascendant, i.e. derivational parent for the lemma, all separated by tabulators. The dataset is available for queries on the web (see Appendix A).

Listing 21: An example of format of Polish WFN.

```
$ 125824  zatyrać  zatyrać      112583
$ 155298  natyrać  natyrać      112583
$ 70592   potyrać  potyrać      112583
$ 112583  tyrać    tyrać
```

3.17 POLYMOTS

Polymots (Gala, Rey, & Zock, 2010) is an automatically created lexicon which groups French lemmas into morpho-phonological families. These families represent a continuity of form and sense. Polymots includes derivational morphology and linguistic information (semantic features) similar to WordNets.

Unfortunately, the dataset of Polymots is available only for queries (see Appendix A). It is possible to query by lemmas, their form, derivations, senses, affixes, productivity etc.

3.18 Spanish WFN

Spanish Word-Formation Network (Lango et al., 2018) is a semi-automatically created lexicon of Spanish. It focuses on the construction of derivational networks between lemmas, especially using an unsupervised method. It was constructed together with *Polish WFN*.

Lemmas of Spanish WFN come from *Leffe lexicon*. Used method for creating Spanish WFN is based on a sequential pattern mining technique to construct useful morphological features in an unsupervised manner. For each lemma, the most probable base word is chosen from a resulting list of possible base words given by the model.

Listing 22 shows a format of Spanish WFN. It keeps the same data structure (the rooted tree of lemmas) and format as *DeriNet*, although the techlemma does not make sense (because it does not serve to connect other linguistic resources, nor does it bring new information about lemma), and the part-of-speech classification is unspecified. However, it could be considered as a first step to harmonize lexicons of derivational networks. Each line contains an id of the lemma, the lemma, its so-called techlemma (it is same as lemma), part-of-speech classification (empty), and id of the ascendant, i.e. derivational parent for the lemma, all separated by tabulators. The dataset is available for queries on the web (see Appendix A).

Listing 22: An example of format of Spanish WFN.

```
$ 36465   ilustración  ilustración      3170
$ 134192  ilustrador   ilustrador       3170
$ 3170    ilustrar     ilustrar
```

3.19 Unimorph

Unimorph (Augerot, 2002) is a manually created lexicon of Russian. It focuses on derivationally related nouns, adjectives, verbs and adverbs.

Lemmas (almost 100k) and derivational relations come from *Zaliznyak's Russian Grammar Dictionary*. Unimorph organizes families into the simplified rooted tree of lemmas. All derivationally related lemmas are connected to the root of the tree, i.e. an unmotivated word.

Unfortunately, the dataset of this resource is available only for queries (see Appendix A).

3.20 VerbAction

VerbAction (Hathout, Namer, & Dal, 2002) is a semi-automatically created lexicon of French. It focuses on the derivationally related nouns and verbs. This lexicon has become a part of *Démonette*.

Lemmas of VerbAction come from the *TLFnome* and *TLFindex lexica*. Used method for creating VerbAction is based on a rule-based approach. Then, resulted VerbAction was extended by data of *Webbaffix toolbox*, and it all was manually checked.

Listing 23 shows a format of VerbAction. It is an easy-to-read `xml` format including individual couples of verb and noun. VerbAction organizes its derivational families of nouns and verbs to the directed graph.

Listing 23: An example of format of French VerbAction.

```
<couple>
  <verb>
    <lemma>baguenauder</lemma>
    <tag>Vmn----</tag>
  </verb>
  <noun gender="feminine" number="singular">
    <lemma>baguenauderie</lemma>
    <tag>Ncfs</tag>
  </noun>
</couple>
```

3.21 WFL

Word-Formation Latin (Litta, Passarotti, & Culy, 2016) is a semi-automatically created lexicon of Classical Latin. It focuses on derivationally related nouns, adjectives, verbs and adverbs. It includes suffixation, prefixation, conversion and also composition.

Lemmas of WFL come from Latin *Lemlat*. Used method for creating WLF is based on manually and automatically obtained word-formation rules.

WFL is implemented to *Lemlat*, so it is distributed in the format of the SQL database. It organizes derivational families to the rooted trees of lemmas, and it is available for queries on the web (see Appendix A).

4 Lexicons of other linguistics phenomena

Lexicons containing another linguistics phenomena, however, also containing derivational word-formation relations to some extent are well-known but not for their derivational morphology. So, from this point of view, the situation about these resources is very messy. Among all existed linguistic resources, it is very difficult to find individual ones containing derivational morphology. Nevertheless, there exist bigger projects as WordNets or WiktiWF for which the desired relationships can be found for a larger set of languages.

There were discovered 20 datasets of this type covering 15 languages (usually Indo-European languages, except for two Uralic and one Altaic languages). The author of the report obtained 12 datasets to do more detailed statistics and observations, see Table 3 on page 21. The extraction of required relationships was needed before tabulating and calculating the statistic.

As was already said above, datasets differ in most properties, even for the datasets that which were created as part of the same project (cf. approaches to process so-called morphosemantic relations in WordNets described in Subsection 4.5). These resources are not much different in their formats and data structures, but they are very different in their distribution and licences (many of them are not downloadable), see Appendix B.

Following subsections describe each resource and/or project with its resources individually. Some of these resources can be queried online. The list of URL links for all discovered resources is listed in Appendix A.

4.1 E-Dictionary

E-dictionary (Vitas & Krstev, 2005) is a morphological lexicon of Serbian. In the beginning, E-dictionary did not contain any derivational relations. Derivational relations between and within nouns, adjectives, verbs and adverbs were added additionally in the form of so-called "derivational lexical transducers" (Vitas & Krstev, 2005).

E-dictionary covers prefixation and suffixation, but it focuses only on regular derivations. It would semantically label possessive adjective, diminutive, augmentative, gender motion, relational adjective (Vitas & Krstev, 2005).

E-dictionary was distributed in various versions (with and more often without derivational relations), so it is not easy to obtain the version with derivational relations.

4.2 E-Lex

E-Lex (Department of Language and Speech at Radboud University Nijmegen and ELIS and University of Ghent and CGN Consortium, 2008), also known as TST-lexicon, is a lexical database of Dutch. It originated as a part of CGN (Corpus Gesproken Nederlands; en. Spoken Dutch Corpus) with which it is connected (Hoekstra, Moortgat, Schuurman, & Van Der Wouden, 2001).

For each lemma, E-Lex annotates various morphological, syntactic and phonological information (e.g. word form, lemma, pronunciation, orthography, morphological categories and variants and segmentation, semantic taxonomy, definition, a frequency of occurrence in CGN, etc.). Derivational relations are covered as morphological segmentation inspired by CELEX format (hierarchical rooted tree of morphemes). Morphological variants are also promising.

E-Lex is distributed in two formats: in the `txt` similar to CELEX ones, and in the `xml`. Both are well documented. The authors of E-Lex provide two license agreements: commercial and

Table 3. Statistics for lexicons containing mainly other linguistic phenomena. The extraction has been done before calculating statistics (which explains zero singletons). Abbreviations for Structure coresponds to abbreviations in Section 2. Relations means edges connecting lemmas. Families means non-singleton families of lemmas. Singletons (singleton families) means derivational families consisting of only one lemma. Part-of-speech: N for noun, A for adjective, V for verb, D for adverb, O for others. Minuses in table means that information could not be obtained.

Language	Resource	Ver	Structure	Lemmas	Relations	Families	Singletons	Part-of-speech [%]				
								N	A	V	D	O
Bulgarian	BulNet[6]	5.0	DGL	-	-	-	-	-	-	-	-	-
Croatian	CroWN[6]	3.0	DGL	-	-	-	-	-	-	-	-	-
Czech	CS-WiktiWF[8]	1.0	DGL, CGL	50,526	57,902	8,387	0	27	9	6	1	57
Czech	Czech WordNet[6]	1.0	DGL	-	-	-	-	-	-	-	-	-
Dutch	E-Lex[7]	1.1.1	RTM	97,054	174,210	13,112	0	80	9	9	1	1
English	EN-WiktiWF[8]	1.0	DGL, CGL	23,044	20,319	2,908	0	54	32	5	3	6
English	Princeton WordNet	3.0	DGL	13,813	17,739	9,834	0	57	0	43	0	0
Estonian	EstWN	2.1	DGL	989	544	465	0	16	29	8	47	0
Finnish	FinnWordNet	2.0	DGL	20,035	42,136	6,349	0	55	29	16	0	0
French	FR-WiktiWF[8]	1.0	DGL, CGL	136,574	121,101	28,978	0	40	29	6	1	24
German	DE-WiktiWF[8]	1.0	DGL, CGL	140,896	132,637	14,605	0	33	4	4	1	57
German	GermaNet[6]	13.0	DGL	-	-	-	-	-	-	-	-	-
Polish	PL-WiktiWF[8]	1.0	DGL, CGL	106,699	249,584	18,089	0	36	11	5	2	46
Polish	PlWordNet	3.1	DGL	112,075	139,283	23,886	0	52	25	17	6	0
Portuguese	OpenWordNet-PT	2018	DGL	7,024	4,238	2,787	0	60	0	40	0	0
Romanian	RoWN[6]	2.0	DGL	-	-	-	-	-	-	-	-	-
Serbian	E-Dictionary[6]	1.0	DGL	-	-	-	-	-	-	-	-	-
Serbian	SrpWN[6]	3.0	DGL	-	-	-	-	-	-	-	-	-
Slovene	Sloleks	1.2	DGL	97,242	65,984	20,845	0	52	27	11	7	3
Turkish	Turkish WordNet[9]	06.14	DGL	0	0	0	0	0	0	0	0	0

[6] Author of this technical report did not obtain dataset of this resource because of the license.

[7] Numbers of Relations, Families and Singletons are for reference only. Used script for linking derivationally related lemmas was quite naive, which affected those numbers.

[8] Percentage of lemmas tagged as Other is increased by unspecified pos. Number of families is decreased by compounds.

[9] Turkish WordNet does not contain explicitly labeled derivational relations, but it includes them into other labels.

non-commercial. Both must be signed before downloading the datasets and its documentation. The whole package of E-Lex includes two datasets: single-word lexicon, and multi-word lexicon. For the extraction of derivational relations (and calculating basic statistics in Table 3), it was preferable to use the single-word lexicon.

Listing 24 shows a format of E-Lex. Each line consists of the lemma and various linguistic information separated by a slash. The most important are positions for the lemma (position 2), morphological segmentation and part-of-speech classification (position 3). However, more information would be useful to extract and process, e.g. the morphological variants (position 4). E-Lex is available for queries on the web (see Appendix A).

Listing 24: An example of format of Dutch E-Lex.

```
$ 500304\aanstippen\((aan)[P],(stip)[V])[V]\\\\\4317\aanstipten\WW(pv,verl,mv)\\C\
    anstIpt@\anstIpt@n\anstIpt@\'an-stIp-t@\V\0\[SU:NP][HD:<aanstipten >][OBJ1:CP<dat >]\\
$ 500308\aanstoppen\((aan)[P],(stop)[V])[V]\\\\\4355\aanstopt\WW(pv,tgw,met-t)\\C\anstOpt
    \anstOpt\anstOpt\'an-stOpt\V\0\\\
$ 8386\batig\((baat)[N],(ig)[A|N.])[A]\\\\\418662\batig\ADJ(nom,basis,zonder,zonder-n)\\C
    \bat@x\bat@x\bat@x\'ba-t@x\V\0\[HD:<batig >]\\
```

4.3 Sloleks

Sloleks (Dobrovoljc, Krek, Holozan, Erjavec, & Romih, n.d.) is a morphological lexicon of Slovene. It was base for various tools processing Slovene language, e.g. tagger (Grčar, Krek, & Dobrovoljc, 2012) and parser (Dobrovoljc, Krek, & Rupnik, 2012).

From the opportunity to extract derivational information, it is important that Sloleks contains lemmas and their related forms. It covers derivational relations between and within nouns, adjectives, verbs, adverbs and some other part-of-speech. It does not semantically label any relations.

The dataset is distributed in a large `xml` file which is freely downloadable for non-commercial use. Listing 25 shows an abbreviated record of one lemma and its related form. It seems that Sloleks organizes derivational families to the direct graphs. Sloleks is available for queries on the web (see Appendix A).

Listing 25: An example of format of Slovenian Sloleks.

```
  <LexicalEntry id="LE_984f1b971b3c5415cb3ff21dcb9823d7">
    <feat att="ključ" val="G_zasevati"/>
    <feat att="besedna_vrsta" val="glagol"/>
    <feat att="vrsta" val="glavni"/>
    <feat att="vid" val="dovršni"/>
    <Lemma>
      <feat att="zapis_oblike" val="zasevati"/>
    </Lemma>
    <WordForm>
      <feat att="msd" val="Ggdn"/>
      <feat att="oblika" val="nedoločnik"/>
      <FormRepresentation>
        <feat att="zapis_oblike" val="zasevati"/>
        <feat att="pogostnost" val="2"/>
      </FormRepresentation>
    </WordForm>
    [...]
    <RelatedForm>
      <feat att="idref" val="LE_bd7b6bb4b07406805f799b4a612cbdc7"/>
      <feat att="besedna_vrsta" val="samostalnik"/>
      <feat att="lema" val="zasevanje"/>
    </RelatedForm>
  </LexicalEntry>
```

4.4 WiktiWF

WiktiWF is a still ongoing project of the author of the report (`https://github.com/lukyjanek/wiktionary-wf`). Many language mutations of Wiktionary.org contain between syllable segmentation, definition of meaning, pronunciation also derivational relations. Therefore, the goal of WiktiWF project is to extract these relations and provide them in some common data structure.

All date in each language mutation of Wiktionary came from real speakers of those languages. It leads to a good precision of Wiktionary. On the other side, these speakers are usually not consistent in used format and data structure. It should be pointed out, that one language mutation contains more than one languages, but WiktiWF focuses on the main language of given language mutation. Five languages (English, French, Czech, Polish, German) has been processed and published so far. Original data from Wiktionary organizes derivational families into the directed graphs of lemmas (sometimes its really close to the rooted tree of lemmas), however, the data of WiktiWF are provided in the simple complete graph of lemmas, too.

Listing 26 shows a format of English En-WiktiWF in the structure of the complete graph of lemmas. Each line contains all derivationally related lemmas (with their part-of-speech classification after the underscore sign, if it was present in Wiktionary, *X* means it is not) separated by the tabulator. Listing 27 shows a format of English En-WiktiWF in the structure of the direct graph of lemmas. Each line contains a pair of derivational parent (first position) and child (second position).

Listing 26: An example of format of English EN-WiktiWF (complete graph of lemmas).

```
$ environmentalist_N  antienvironmental_A  geoenvironmental_A  microenvironmental_A
    environmentalism_N  environmental_A  environmentally_D  nonenvironmental_A
    paleoenvironmental_A  bioenvironmental_A  macroenvironmental_A  socioenvironmental_A
$ generalise_V  generalization_N  generally_D  general_A  generalize_V  generality_N
    generalisation_N
```

Listing 27: An example of format of English EN-WiktiWF (direct graph of lemmas).

```
$ environmental_A  bioenvironmental_A
$ environmental_A  environmentalism_N
$ general_A  generalisation_N
$ general_A  generalise_V
$ general_A  generality_N
```

4.5 WordNet

WordNets are lexical databases of nouns, adjectives, verbs and adverbs grouped into sets of cognitive synonyms, so-called synsets. They contain brief definitions of meaning and various semantic and lexical relations.

WordNets usually focus on synonyms, hypernyms, hyponyms, meronyms, holonyms, toponyms, entailments and coordinate terms. However, some language mutations of WordNets also deal with so-called morphosemantic relations, the derivational word-formation relations (often semantically labelled). These morphosemantic relations are identified semi-automatically or fully automatically (Leseva et al., 2015), but the approaches to process morphosemantic relations are inconsistent across all WordNets (cf. Fellbaum, Osherson, and Clark (2007); Koeva (2008); Maziarz, Piasecki, Szpakowicz, Rabiega-Wiśniewska, and Hojka (2011); Pala and Hlaváčková (2007)). Despite that, after various extraction methods, potential derivational families would be more or less organized to the direct graphs of lemmas.

In the Fellbaum et al. (2007) approach, the derivational relations are extracted through automatic identification of base-derived and semantically related pairs. It links and semantically labels them (using labels from WordNet).

In the Pala and Hlaváčková (2007) approach, so-called "derivational nests" (new derivative derived from stems by affixes with specific meanings) are automatically generated and incorporated into synsets. Two levels of the semantic network are built in this way: higher level (semantic relations between synsets, e.g. meronymy) and lower level (derivational relations between a single member of synset).

In the Koeva (2008) approach, morphosemantic and derivational relations are distinguished. The derivational relations in BulNet means automatically transferred "derivative", "derived" and "participle" relations from Princeton WordNet, whereas the morphosemantic relations in BulNet means relations of "words (synset members) similar in meaning, in which one word is derived from the other by means of a morphological affix." (Koeva, 2008, p. 365)

In the Maziarz et al. (2011) approach, relations between synsets and individual lexical units are distinguished and processed separately. In fact, it combines the best of Fellbaum et al. (2007) and Pala and Hlaváčková (2007) approaches.

In many WordNets, various derivational relations are included in the semantic relations between synsets, e.g. *role agent* in Czech WordNet, Estonian WordNet etc. often makes regular derivations. However, there would be needed more complex extraction of these derivational relations (as Fellbaum et al. (2007) did).

So far, many language mutations of WordNets has been made under patronizations of various projects (MultiWordNet, EuroWordNet, BalkaNet, etc.). Some of these projects had more ambitions than just creating monolingual data. They created new language mutations base on already existed WordNets, they linked synsets across languages etc.

Because of the existence of many language mutations in various versions with completely different licensing and distributing, it is difficult to find and obtain specific WordNet data or information about them. It would need deeper insight to clearly cover situation about morphosemantic relations in WordNets. To the best of our knowledge, the morphosemantic relations should be included in these WordNet language mutations: Bulgarian, Croatian, Czech, English, Estonian, Finnish, German, Polish, Romanian, Serbian and Turkish.

4.5.1 BulNet

BulNet (Koeva, Genov, & Totkov, 2004) is WordNet of Bulgarian. It is distributed under ELRA Agreement and it is not freely downloadable.

BulNet distinguishes between morphosemantic and derivational relations. The derivational relations in BulNet means automatically transferred "derivative", "derived" and "participle" relations from Princeton WordNet, whereas the morphosemantic relations in BulNet means relations of "words (synset members) similar in meaning, in which one word is derived from the other by means of a morphological affix." (Koeva, 2008, p. 365) It would cover prefixation, suffixation and also conversion between and within nouns, adjectives, verbs and adverbs using 4 semantic (Dimitrova, Tarpomanova, & Rizov, 2014; Koeva, Krstev, & Vitas, 2008).

BulNet is available for queries on the web (see Appendix A).

4.5.2 CroWN

CroWN (Raffaelli, Tadić, Bekavac, & Agić, 2008) is WordNet of Croatian. Three versions of Croatian WordNet should exist but only one (version 1.0) is downloadable. It is distributed in the `xml` and/or `json` file, but it does not contain many derivational relations. Versions 2.0 and 3.0 of CroWN were incorporated to the Open Multilingual WordNet project (in version 1.2 and 2.0), however, before that, both datasets of CroWN were edited to the format without derivations.

Since version 2.0 of CroWN, it should link and semantically label derivationally related nouns, adjectives, verbs and adverbs. It covers prefixation and suffixation following Pala and Hlaváčková (2007) approach. Some of the derivational relations of verbs would come from CroDeriV (Oliver, Šojat, & Srebačić, 2015; Šojat & Srebačić, 2014).

Listing 28 shows a format of CroWN in version 1.0 which contains derivational relations at least in promising labels: *causes*, *category domain*, *derived*, *eng derivative*, and some pairs of perfective/imperfective/iterative verbs are found in verbal synsets. In the Listing 28, the first entry defines lemma and its semantic relations, as well as second entry, however, the second entry refers to the first entry with label "type=derived".

Listing 28: An example of format of Croatian CroWN.

```
<SYNSET><ID>ENG30-07708798-n</ID><POS>n</POS><SYNONYM><LITERAL sense="2">mahuna</LITERAL
    ></SYNONYM><DEF>jestiva ljuštura povrća mahunarki u kojoj se nalaze sjemenke</DEF><
    ILR type="hypernym">ENG30-07707451-n</ILR><STAMP>Danduan 2011/11/09</STAMP><BCS>3</
    BCS><SUMO>FruitOrVegetable<TYPE>+</TYPE></SUMO><DOMAIN>gastronomy</DOMAIN></SYNSET>

<SYNSET><ID>ENG30-02917945-a</ID><POS>a</POS><SYNONYM><LITERAL sense="1">mahunast</
    LITERAL></SYNONYM><DEF>koji je nalik mahuni ili kojemu je plod mahuna</DEF><ILR type
    ="derived">ENG30-07708798-n</ILR><STAMP>igor@zzl-ht06.dhcp.ffzg.hr 2012-04-21
    17:18:59</STAMP><BCS>0</BCS><USAGE>Drvo rogača ima mahunast plod srednje veličine.</
    USAGE><SUMO>FloweringPlant<TYPE>+</TYPE></SUMO><DOMAIN>gastronomy</DOMAIN></SYNSET>
```

4.5.3 Czech WordNet

Czech WordNet (Pala & Smrž, 2004) is the mutation of WordNet of Czech. It is distributed under ELRA Agreement and it is not freely downloadable.

Morphosemantic relations in Czech Wordnet are based on the automatic generation of "derivational nests" (new derivative derived from stems by affixes with specific meanings). Thus Czech WordNet covers prefixation and suffixation between and within nouns, adjectives, verbs and adverbs. It uses 16 semantic labels and 10 word-formation rule labels (Pala & Hlaváčková, 2007).

4.5.4 EstWordNet

EstWordNet (Kerner, Orav, & Parm, 2010) is WordNet of Estonian. It is distributed in two ways, the `xml` (ver 2.1) and the `txt` (ver kb73). The dataset distributed in the `xml` is more machine-readablethan the `txt` format.

EstWordNet links and labels derivationally related nouns, adjectives, verbs and adverbs (Kahusk, Kerner, & Vider, 2010) as an individual lexical units. However, some derivational relations could be found in used semantic relations between synsets. Very promising appear to be the labels: *causes*, *is caused by*, *be in state*, *state of*, *involved agent*, *role*, *role agent*, *xpos fuzzynym*, and some pairs of perfective/imperfective/iterative verbs are found in verbal synsets. These promising derivational relations would need more complex extraction. Used approach to cover derivational relations combines Pala and Hlaváčková (2007) and Maziarz et al. (2011) approaches as described above.

Listing 29 shows a format of the direct derivational relations in the `xml` dataset. Listing 30, on the other hand, shows a format of the `txt` dataset. EstWordNet is available for queries on the web (see Appendix A).

Listing 29: An example of (`xml`) format of Estonian EstWN.

```
<LexicalEntry id="w526908">
   <Lemma partOfSpeech="r" writtenForm="aastaringselt" />
     <Sense id="s-aastaringselt-r1" status="unchecked" synset="estwn-et-47344-b">
       <SenseRelation confidenceScore="1.0" relType="derivation" status="unchecked"
           target="s-aastaringne-a1" />
       <Example language="et">Ka suusatamist treenitakse aastaringselt.</Example>
     </Sense>
</LexicalEntry>
```

Listing 30: An example of (`txt`) format of Estonian EstWN.

```
0 @2358@ WORD_MEANING
  1 PART_OF_SPEECH "n"
  1 VARIANTS
    2 LITERAL "populaarsus"
      3 SENSE 1
      3 EXAMPLES
        4 EXAMPLE "Populaarsusele tuleb alati midagi ohverdada."
      3 EXTERNAL_INFO
        4 SOURCE_ID 1
          5 TEXT_KEY "2358"
        [...]
    2 LITERAL "üldtuntus"
      3 SENSE 1
    2 LITERAL "menukus"
      3 SENSE 1
  1 INTERNAL_LINKS
    2 RELATION "has_hyperonym"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "n"
        4 LITERAL "omadus"
          5 SENSE 1
      3 SOURCE_ID 1001
    [...]
    2 RELATION "derivation"
      3 TARGET_CONCEPT
        4 PART_OF_SPEECH "a"
        4 LITERAL "populaarne"
          5 SENSE 1
      3 SOURCE_ID 1016
  1 EQ_LINKS
    2 EQ_RELATION "eq_synonym"
      3 TARGET_ILI
        4 PART_OF_SPEECH "n"
        4 WORDNET_OFFSET 3383002
```

4.5.5 FinnWordNet

FinnWordNet (Lindén & Carlson, 2010) is WordNet of Finnish. It is distributed in easy-to-read machine-readable columns style (`tsv`) file.

FinnWordNet does not semantically label derivational relations but it labels them only as "derivationally related", so it similar to Fellbaum et al. (2007) approach as described above. It covers derivations between and within nouns, adjectives and verbs (Lindén, Niemi, & Hyvärinen, 2012).

Listing 31 shows a format of FinnWordNet. Derivationally related lemmas are found in the columns 2 and 3, their ids and part-of-speech classifications (a letter after the colon) are found in the columns before them. Column 5 contains the same information as column 6 in form of an alphanumeric sign. FinnWordNet is available for queries on the web (see Appendix A).

Listing 31: An example of format of Finnish FinnWordNet.

```
$ fi:a00001740  kykenevä     fi:n05200169  kyky          +  derivationally related
$ fi:a00006336  absorboiva   fi:n04940964  absorboivuus  +  derivationally related
$ fi:a00006336  absorboiva   fi:v01539633  absorboitua   +  derivationally related
$ fi:n00043195  löytäminen   fi:v02285629  löytää        +  derivationally related
```

4.5.6 GermaNet

GermaNet (Hamp & Feldweg, 1997) is WordNet of German. This WordNet is distributed under several types of licenses (academic and commercial). Before downloading, it is necessary to pick and sign the license and meet the license requirements (in the academic license, there is an obligation to send reports to the distributor every year).

In addition to the licensed version of GermaNet, there is a list of 82 thousand processed compounds (Henrich & Hinrichs, 2011). This list containing segmentation of compounds is freely downloadable for academic research (without the necessity of signing and sending any license) from the official GermaNet webpage.

4.5.7 PlWordNet

PlWordNet (Piasecki, Szpakowicz, & Broda, 2009) is WordNet of Polish. The dataset is distributed in the `xml` file under a non-commercial license. Before downloading, it is necessary to register and agree to the license agreement. After that, a link for downloading PlWordNet is sent to a registered e-mail address.

PlWordNet links and semantically labels derivationally related nouns and verbs. It covers prefixation and suffixation using 11 semantic labels (Maziarz et al., 2011).

Listing 32 shows a format of PlWordNet. The dataset consists of three parts defining: lexical units, syntagmas, relations between lexical units. The first and second entries in the Listing shows the definition of lexical units, and the third entry contains the definition of the relation. PlWordNet is available for queries on the web (see Appendix A).

Listing 32: An example of format of Polish PlWordNet.

```
<lexical-unit id="40116" name="robić" pos="czasownik" tagcount="0" domain="cwyt" desc="co
    ć konkretnego, wytwarzać to, np. robić rzećbę. Jest to czasownik teliczny &lt;##VLC:
    DZn>"workstate="Nieprzetworzony" source="użytkownika" variant="2"/>

<lexical-unit id="77915" name="odrobić" pos="czasownik" tagcount="0" domain="sp" desc="##
    K: og. ##D: wykonać jakąć czynnoćć, którą miało się wykonać w przeszłoćci lub którą
    ma się wykonać w przyszłoćci. [##P: Nie odrobię już w tym semestrze zajęć z wuefu, na
     których mnie nie było.] &lt;##VLC: DZd>" workstate="Nowy" source="użytkownika"
    variant="1"/>

<lexicalrelations parent="40116" child="77915" relation="111" valid="true" owner="
    Agnieszka.Dziob"/>
```

4.5.8 OpenWordNet-PT

OpenWordNet-PT (Paiva, Rademaker, & Melo, 2012) is WordNet of (Brazilian) Portuguese. It is automatically created by machine learning methods, and it is distributed on GitHub under Creative Common license.

From the derivational morphology point of view, there exists a project which enlarges OpenWordNet-PT with *Nomlex-PT* which is realized by links to Nomlex-PT (Rademaker et al., 2014).

4.5.9 Princeton WordNet

Princeton WordNet (Miller, 1998) was the first WordNet ever, and so it was called simply WordNet without any attribute. When another language mutation of WordNet was created, Princeton WordNet got its attribute (the project began in the Princeton University). It is WordNet for English.

The newest version of Princeton WordNet is the version 3.1, but it does not contain the morphosemantic relations. However, there exists a separated standoff (`.xls`) file containing morphosemantic relations from WordNet 3.0 by Fellbaum et al. (2007). This dataset of derivational relations was extracted through automatic identification of base-derived and semantically related noun-verb pairs (Fellbaum et al., 2007). It links and semantically labels (using 14 semantic labels) derivationally related nouns and verbs.

Listing 33 shows a format of standoff file extracted from the Princeton WordNet. Each line consists of 7 cells (in text separated by slashes) containing derivational parent, its id, semantic label, derivational child, its id, the definition of derivational parent, and the definition of derivational child. The part-of-speech classification is codded after percentage sign (1 means nouns, 2 means verb).

Listing 33: An example of format of English Princeton WordNet.

```
$ cannibalise%2:34:00:: / 201162291 / agent / cannibal%1:18:00:: / 109891079 / eat human
    flesh / a person who eats human flesh
$ survive%2:42:00:: / 202616713 / state / survival%1:26:00:: / 113962166 / support
    oneself; "he could barely exist... / a state of surviving; remaining alive
$ rule%2:36:00:: / 201690020 / instrument / ruler%1:06:00:: / 104118776 / mark or draw
    with a ruler; "rule the mar... / measuring stick consisting of a strip of...
$ center%2:38:00:: / 201852701 / location / center%1:15:00:: / 108521816 / move into the
    center; "That vase in the ... / a point equidistant from the ends of a l...
```

4.5.10 RoWN

RoWN (Tufis, Mititelu, Bozianu, & Mihaila, 2006) is WordNet of Romanian. This WordNet is distributed under Meta-Share Creative Commons license, so before downloading, it is necessary to sign the license.

Morphosemantic relations were added to RoWN additionally. They should cover prefixation and suffixation between and within nouns, adjectives, verbs and adverbs following Pala and Hlaváčková (2007) approach (Mititelu, 2012).

4.5.11 SrpWN

SrpWN (Krstev, Pavlovic-Lazetic, Vitas, & Obradovic, 2004) is WordNet of Serbian. This WordNet is distributed under Meta-Share Creative Commons license, so before downloading, it is necessary to sign the license.

Morphosemantic relations should be semantically labelled between and within nouns, adjectives and verbs following Koeva (2008) approach (Koeva et al., 2008).

4.5.12 Turkish WordNet

Turkish WordNet (Bilgin, Çetinoğlu, & Oflazer, 2004) is the mutation of WordNet for Turkish. The distribution of the dataset is unclear because of the broken official link.

Turkish WordNet uses the same approach as Bulgarian BulNet, so it distinguishes morphosemantic and derivational relations and follows Koeva (2008) approach. It covers prefixation and suffixation between and within nouns, adjectives, verbs and adverbs using 15 morphosemantic labels (e.g. *become*, *be in state*, *causes*, etc.) (Bilgin, Cetinoglu, & Oflazer, 2004). However, Turkish WordNet does not contain any direct derivational relation between or within lemmas (only between synsets which making the extraction more difficult), thus Table 3 shows zeroes in all its columns.

Listing 34 shows a format of Turkish WordNet. First entry defines lemma and its semantic relations, as well as second entry, however, the second entry refers to the first entry with label "<TYPE>causes</TYPE>".

Listing 34: An example of format of Turkish WordNet.

```
<SYNSET><ID>ENG20-00014429-v</ID><POS>v</POS><SYNONYM><LITERAL>uyumak<SENSE>1</SENSE></
    LITERAL></SYNONYM><ILR>ENG20-00014092-v<TYPE>hypernym</TYPE></ILR><ILR>ENG20
    -00019757-v<TYPE>near_antonym</TYPE></ILR><ILR>ENG20-00015493-v<TYPE>also_see</TYPE
    ></ILR><ILR>ENG20-01141196-v<TYPE>also_see</TYPE></ILR><ILR>ENG20-01141387-v<TYPE>
    also_see</TYPE></ILR><DEF>Uyku durumunda olmak</DEF><BCS>2</BCS></SYNSET>

<SYNSET><ID>ENG20-00018968-v</ID><SYNONYM><LITERAL>uyutmak<SENSE>1</SENSE></LITERAL></
    SYNONYM><POS>v</POS><ILR>ENG20-00121430-v<TYPE>hypernym</TYPE></ILR><ILR>ENG20
    -00018508-v<TYPE>near_antonym</TYPE></ILR><ILR>ENG20-00014429-v<TYPE>causes</TYPE></
    ILR><STAMP>orhanb 2004/05/21</STAMP></SYNSET>
```

5 Corpora

Corpora which contain derivational word-formation relations to some extent are quite rare and they contain regular derivational relations. There were discovered 5 datasets of this type covering 4 languages. The author of the report does more detailed statistics and observations for 3 obtained datasets, see Table 4 on page 31. The extraction of required relationships was needed before tabulating and calculating the statistic.

Compared to resources described in Section 3 and 4, corpora does not link two lemmas. They tag one lemma (semantically, or formally). Because of 3 discovered resources come from Universal Dependencies (UD) project, they are very similar. Prague Dependency Treebank is also Treebank as well as UD, but it is different from UD. Russian National Corpus is also different from UD, but it is quite interesting because it tags lemmas exclusively semantically. Unfortunately, the distribution of this corpus is complicated because of the license, see Appendix B.

Following subsections describe each resource and/or project with its resources individually. Some of these resources can be queried online. The list of URL links for all discovered resources is listed in Appendix A.

5.1 Russian National Corpus

Russian National Corpus (Zakharov, 2013) is the national corpus of Russian. It is based on a collection of Russian texts in electronic form covering primarily the period from the middle of the 18th to the early 21st centuries. It respects principles of building corpora as well as other national corpora (*British National Corpus*, *Czech National Corpus*). After signing the license (for academic research and language teaching only), the dataset is downloadable.

Russian National Corpus consists of more than 52 thousand texts including totally over 149 million tokens (January 2008). Corpus is morphologically and semantically annotated (Apresjan et al., 2006). The morphological information (part of speech, gender, case, aspect, etc.) is mainly based on the morphological model suggested by Zalizniak in the *Grammatical dictionary of Russian*. The semantical information includes some derivations (using 35 labels, e.g. diminutive, augmentative, nominal agent, verbal nouns etc.). A complete list of used semantic tags can be found at `http://www.ruscorpora.ru/en/corpora-sem.html`.

Unfortunately, lemmas/tokens in the Corpus are only tagged with these (derivational) labels; they are not linked to their parents/children. The dataset is available for queries on the web (see Appendix A).

5.2 Prague Depencency Treebank

Prague Dependency Treebank (PDT) (Hajič et al., 2018) is a large project providing a framework for annotation of grammatical annotation of Czech. Dataset of PDT is distributed in xml based PML (Prague Markup Language) format.

PDT annotates each Czech sentence on 4 layers (wordform layer, morphological layer, analytical layer, tectogrammatical layer) with respect to the Functional Generative Description proposed by Sgall, Hajicová, and Panevová (1986).

From the derivational morphology point of view, PDT annotates so-called pro-forms on the morphological and tectogrammatical layers (Razímová & Žabokrtskỳ, 2006). Pro-forms are pronouns, numerals and adverbs which can be involved to derivational relations.

Table 4. Statistics for corporas containing also derivational word-formation relations. Abbreviations for Structure coresponds to abbreviations in Section 2. Relations means edges connecting lemmas. Families means non-singleton families of lemmas. Singletons (singleton families) means derivational families consisting of only one lemma. Part-of-speech: N for noun, A for adjective, V for verb, D for adverb, O for others. Minuses in table means that information could not be obtained.

Relations, Families and Singletons are filled by minuses because used TL format does not connect two lemmas, so a complex decesions would have to be done to find based lemma and after that it could be possible to calsulate those numbers.

Language	Resource	Ver	Structure	Lemmas	Relations	Families	Singletons	Part-of-speech [%]				
								N	A	V	D	O
Czech	Prague Dependency Treebank[10]	3.5	TL	-	-	-	-	-	-	-	-	-
Finnish	PUD Treebank	2.0	TL	758	-	-	-	45	45	0	10	1
Finnish	TDT Treebank	5.0	TL	4,498	-	-	-	53	36	0	11	0
Komi-Zyrian	Lattice Treebank	1.0	TL	11	-	-	-	0	0	100	0	0
Russian	Rus. National Corpus[11]	-	TL	-	-	-	-	-	-	-	-	-

[10] The extraction would be quite uncomfortable because of using more layers.

[11] Author of this technical report did not obtain dataset of this resource because of the license.

5.3 Universal Dependencies

Universal Dependencies (UD) is a large project providing a framework for cross-linguistically consistent grammatical annotation. More than 200 contributors produced more than 100 treebanks in over 60 languages (all distributed in Creative Common license, and freely downloadable from the project webpage).

The treebanks consist of morphologically and mainly syntactically (dependency syntax) annotated sentences following CONLL-U format (tab separated columns in plain-text). The strength of UD is that each treebank tries to follow the same annotation principles as much as possible. Thanks to that, UD Treebanks are used for cross-lingual learning and parsing research from a language typology perspective.

From the derivational relations point of view, the morphological annotation is the most important part of UD Treebanks. Some Treebanks also includes information about (regular) derivations in their morphological tagset. However, as in the Russian National Corpus, lemmas/tokens in UD Treebanks are in all cases only tagged, not linked to their parents/children.

The main field of interest of UD Treebank is the syntactic annotation, so the information about derivations can be found only in 3 treebanks (2 for Finnish, 1 for Komi-Zyrian).

5.3.1 Lattice Treebank

Lattice Treebank (Partanen, Lim, & Poibeau, 2018) is a dependency treebank of Komi-Zyrian (Uralic language with about 160 thousand speakers).

The whole Treebank consists of written standard openly available fiction, especially of Lev Uspenskiy's book (Нёль боевӧй случай) and Ivan Belyx's short story.

From the derivational morphology point of view, Lattice Treebank tags 4 simple different regular derivations for verbs. However, a number of processed verbs is small.

Listing 35: An example of format of Komi-Zyrian Lattice Treebank.

```
# sent_id = belyx-011.003
# text =Шондіыс нем жалиттӧг пӧжис.
# text_rus = Солнце пекло, ничего не жалея.
1 Шондіыс шонді NOUN N Case=Nom|Number=Sing|Number[psor]=Sing|Person[psor]=3 4 nsubj _ _
2 нем нем PRON Pron Case=Nom|Number=Sing|Polarity=Neg 3 obj _ _
3 жалиттӧг жалитны VERB V Derivation=Tog|Polarity=Neg|VerbForm=Conv 4 advcl _ _
4 пӧжис пӧжны VERB V Mood=Ind|Person=3|Tense=Past|VerbForm=Fin 0 root _ SpaceAfter=No
5 . . PUNCT CLB _ 4 punct _ _
```

5.3.2 PUD Treebank

PUD Treebank (Kanerva, Ginter, Ojala, & Missilä, 2017), Parallel Universal Dependency Treebank, is a dependency treebank of Finnish. It was created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies.

The whole Treebank consists of 1,000 sentences which are taken from the news and from Wikipedia. First 750 sentences were taken from English, and the rest 250 sentences from other European languages (German, French, Italian, Spanish) and they were translated and annotated.

From the derivational morphology point of view, PUD Treebank tags 11 simple and 8 combined different regular derivations for nouns, adjectives, adverbs and (one) pronouns. PUD Treebank only tags lemmas/tokens (derivational children), labels derivational suffix, but it does not link or explicitly point to the derivational base (parent).

Listing 36: An example of format of Finnish PUD Treebank.

```
# sent_id = n01139027
# text = Tällä hetkellä analyytikot epäröivät julistaa palvelua kuolleeksi.
# english_text = For now, analysts are hesitant to write off the service for dead.
1 Tällä tämä PRON _ Case=Ade|Number=Sing|PronType=Dem 2 det _ _
2 hetkellä hetki NOUN _ Case=Ade|Number=Sing 4 obl _ _
3 analyytikot analyytikko NOUN _ Case=Nom|Number=Plur 4 nsubj _ _
4 epäröivät epäröidä VERB _ Mood=Ind|Number=Plur|Person=3|Tense=Past|VerbForm=Fin|Voice=
    Act 0 root _ _
5 julistaa julistaa VERB _ InfForm=1|Number=Sing|VerbForm=Inf|Voice=Act 4 xcomp _ _
6 palvelua palvelu NOUN _ Case=Par|Derivation=U|Number=Sing 5 obj _ _
7 kuolleeksi kuollut NOUN _ Case=Tra|Number=Sing 5 xcomp:ds _ SpaceAfter=No
8 . . PUNCT _ _ 4 punct _ _
```

5.3.3 TDT Treebank

TDT Treebank (Haverinen et al., 2014), Turku Dependency Treebank, is a dependency treebank of Finnish. This treebank was created before the UD project, but it conversed to UD, which required extensive manual checks and corrections to keep the UD format.

The whole Treebank covers numerous genres extracted from Wikipedia articles, Wikinews articles, University online news, Blog entries, Student magazine articles, Grammar examples, Europarl speeches, JRC-Acquis legislation, Financial news, and Fiction sourced from 674 individual documents.

From the derivational morphology point of view, it tags 11 simple and 5 combined different regular derivations for nouns, adjectives, (three) verbs, adverbs and (two) pronouns. TDT Treebank only tags lemmas/tokens (derivational children), labels derivational suffix, but it does not link or explicitly point to the derivational base (parent).

Listing 37: An example of format of Finnish TDT Treebank.

```
# text = Viikonlopun pyöritys alkoi H&M:n järjestämällä bloggaajabrunssilla Helsingissä.
# sent_id = b204.2
1 Viikonlopun viikon#loppu NOUN N Case=Gen|Derivation=U|Number=Sing 2 nmod:poss _ _
2 pyöritys pyöritys NOUN N Case=Nom|Number=Sing 3 nsubj _ _
3 alkoi alkaa VERB V Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act 0
    root _ _
4 H&M:n H&M PROPN N Abbr=Yes|Case=Gen|Number=Sing 5 nsubj _ _
5 järjestämällä järjestää VERB V Case=Ade|Degree=Pos|Number=Sing|PartForm=Agt|VerbForm=
    Part|Voice=Act 6 acl _ _
6 bloggaajabrunssilla bloggaaja#brunssi NOUN N Case=Ade|Number=Sing 3 obl _ _
7 Helsingissä Helsinki PROPN N Case=Ine|Number=Sing 3 obl _ SpaceAfter=No
8 .     . PUNCT Punct _ 3 punct _ _
```

6 Digital explanatory dictionaries

Digital explanatory dictionaries which contain derivational word-formation relations to some extent are valuable resources for linguists. Unfortunately, it is hardly possible to process this type of resources by machine because their datasets are generally not downloadable. They are available in printed publications (which are later converted to the online dictionaries) or online for querying, but not as a raw dataset. There were discovered 12 online explanatory dictionaries covering 10 languages (usually Indo-European languages).

Although this type or resources is often made manually, so the precision is very high, the obtaining of the data would have to be done by web crawling which is unethical and against the licenses. (Licenses of these dictionaries are usually not specified which in Czech law means "fully restricted".) Moreover, the data in these dictionaries is often unstructured or semi-structured. They all can be queried online. The list of URL links is listed in Appendix A.

Because of the inability to work with this type of resources, each discovered online explanatory dictionary is described individually and briefly in the following paragraphs.

Algemeen Nederlands Woordenboek (ANW) is a corpus-based, digital dictionary of contemporary (since 1970 to present) Dutch (Tiberius & Niestadt, 2010). It contains much linguistic information for each lemma. From the derivational morphology point of view, ANW provides morphological segmentation, base word, used affixes, derivational children and so on for each word meaning of given lemma.

Das Digitale Wörterbuch der deutschen Sprache (DWDS) is a corpus-based, digital dictionary of historical and contemporary German (Klein & Geyken, 2010). It contains much linguistic information for each lemma (e.g. pronunciation, syllable segmentation, collocation, etymology etc.). From the derivational morphology point of view, DWDS provides morphological segmentation, base word, used affixes, and a list of derivational children for each lemma. DWDS is connected with various German linguistic resources. One of them, `www.canoo.net`, shows derivational children from DWDS in a rooted tree of lemmas, and it also shows given lemma from DWDS in a rooted of morphemes.

Elexiko is a digital dictionary of contemporary (since 2013 to present) German (Klosa, Schnörch, & Storjohann, 2006). It is included in OWID project (Müller-Spitzer & Möhrs, 2008) and contains much linguistic information for each lemma. From the derivational morphology point of view, Elexiko provides base word and used affixes for some lemmas.

Hrvatski jezični portal (HJP) is an online distribution of the Croatian language dictionary (Liber, 2008). It was made on the basis of respected Croatian dictionaries and lexicons. Various derivational word-formation relations can be found in a part which focuses on etymology. This part very often extends to the Proto-Slavic language.

Lèxic Obert Flexionat de Català (LOFC) is a morphological dictionary of Catalan. It contains nominal verbs, a formation of feminines etc. LOFC is included in Open Source Lexical Information Network project (Janssen, 2005).

Léxico Abierto Flexionado del Español (OSLIN-ES) is a morphological dictionary of Spanish. It is still under construction, but it should be similar to Calatan LOFC, or OSLIN-AST, or VOP. OSLIN-ES is included in Open Source Lexical Information Network project (Janssen, 2005).

Lexicón abiertu de la llingua asturiana (OSLIN-AST) is a morphological dictionary of Asturian. It contains diminutives, augmentatives, nominal verbs etc. OSLIN-AST is included in Open Source Lexical Information Network project (Janssen, 2005).

Lexique Ouvert Flexionnel du Français (OSLIN-FR) is a morphological dictionary of French. It is still under construction, but it should be similar to Calatan LOFC, or OSLIN-AST, or VOP. OSLIN-FR is included in Open Source Lexical Information Network project (Janssen, 2005).

Open Source Lexical Information Network for Russian (OSLIN-RU) is a morphological dictionary of Russian. It is still under construction, but it should be similar to Calatan LOFC, or VOP or OSLIN-AST. OSLIN-RU is included in Open Source Lexical Information Network project (Janssen, 2005).

Slovník spisovné češtiny (SSČ) is a printed explanatory dictionary of Czech (Filipec, 2005) which was not published online, however, it is accessible through the Internetová jazyková příručka (Pravdová & Svobodová, 2014). It is widely respected, although, it is much smaller then SSJČ. Various derivational word-formation relations can be found inside the description in semi-structured form.

Slovník spisovného jazyka českého (SSJČ) is a printed explanatory dictionary of Czech (Havránek et al., 1960) which was published online. It is included and accessible through the Internetová jazyková příručka (Pravdová & Svobodová, 2014). SSJČ is widely respected, although, some parts are outdated because this dictionary was created between 1960 and 1971. Various derivational word-formation relations can be found inside the description in semi-structured form.

Vocabulário Ortográfico Comum da Língua Portuguesa (VOP) is a morphological dictionary of Portuguese. It contains many various derivational relations (e.g. adverbs from adjectives, nominal verbs, a formation of feminines, augmentatives, diminutives, etc.). VOP is included in Open Source Lexical Information Network project (Janssen, 2005).

Wielki słownik języka polskiego (WSPJ) is an online dictionary of Polish (Żmigrodzki, 2007) following modern approaches of 20th and 21st centuries to lexicography. Dictionary creation is still ongoing. Until August 2018, tens of thousands of words and phrases were published. It contains various linguistic information for each lemma (e.g. valency, inflectional morphology, semantically related words etc.). From the derivational morphology point of view, WSPJ provides base word, and sometimes also used affixes in part dedicated to etymology. This part very often extends to the Proto-Slavic language.

7 Conclusion

The report reviewed 63 resources containing derivational word-formation relations which cover 22 languages (usually from Indo-European language family). The situation about these resources was quite disorganized. Until this report, there was not any list of existed morphological resources of derivational word-formation relations.

Given that the derivational morphology is gaining more attention, it is valuable and helpful to know all existed resources. These resources differ a lot not even in their sizes, scopes or part-of-speech distributions, but also in used formats and data structures which makes work with them more complex. Therefore, a harmonization of some resources is going to be the next part of the author's Master Theses. Before that, it would be appropriate to add a list of resources that the author has not found[12].

Acknowledgement

I would like to thank Magda Ševčíková and Zdeněk Žabokrtský for all their help and advice. Last but not least, I would also like to thank Krešimir Šojat and Matea Filko for their time and consultations about CroDeriV, and Nizar Habash, and Piere Marco Bertinetto for specifying information about and helping with CatVar and DerIvaTario.

This work was supported by the Grant No. GA16-18177S of the Grant Agency of the Czech Republic.

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

[12] If you know any of them, please contact the author of the report.

References

Apresjan, J., Boguslavsky, I., Iomdin, B., Iomdin, L., Sannikov, A., & Sizov, V. (2006). A Syntactically and Semantically Tagged Corpus of Russian: State of the Art and Prospects. In *Proceedings of lrec* (pp. 1378–1381).

Augerot, J. (2002). *Russian Morphological Database.* Retrieved from `http://courses.washington.edu/unimorph/` (Russian Derivational Morphology Resource)

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1996). Celex2 LDC96L14. *Linguistic Data Consortium, Philadelphia.*

Balvet, A., Barque, L., & Marín, R. (2010). Building a lexicon of french deverbal nouns from a semantically annotated corpus. In *Lrec 2010.*

Bilgin, O., Çetinoğlu, Ö., & Oflazer, K. (2004). Building a wordnet for turkish. *Romanian Journal of Information Science and Technology*, *7*(1-2), 163–172.

Bilgin, O., Cetinoglu, O., & Oflazer, K. (2004). Morphosemantic relations in and across wordnets. In *Proceedings of the global wordnet conference* (pp. 60–66).

Chomsky, N., et al. (1968). *Remarks on nominalization.* Linguistics Club, Indiana University.

De Paiva, V., Real, L., Rademaker, A., & De Melo, G. (2014). Nomlex-pt: A lexicon of portuguese nominalizations. In *Lrec* (pp. 2851–2858).

Department of Language and Speech at Radboud University Nijmegen and ELIS and University of Ghent and CGN Consortium. (2008). *eLex.*

Dimitrova, T., Tarpomanova, E., & Rizov, B. (2014). Coping with derivation in the bulgarian wordnet. In *Proceedings of the seventh global wordnet conference* (pp. 109–117).

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., & Romih, M. (n.d.). *Morphological lexicon Sloleks 1.2. Slovenian language resource repository CLARIN. SI (2015).*

Dobrovoljc, K., Krek, S., & Rupnik, J. (2012). Skladenjski razčlenjevalnik za slovenščino. In *Zbornik osme konference jezikovne tehnologije* (pp. 42–47).

Dokulil, M. (1962). Tvoření slov v češtině. *Teorie odvozování slov.*

Dokulil, M. (1967). *Tvoření slov v češtině* (Vol. 2). Nakl. Československé akademie věd.

Ševčíková, M., & Žabokrtský, Z. (2014). Word-Formation Network for Czech. In *Proceedings of the ninth international conference on language resources and evaluation (lrec-2014).*

Fellbaum, C., Osherson, A., & Clark, P. E. (2007). Putting semantics into WordNet's" morphosemantic" links. In *Language and technology conference* (pp. 350–358).

Filipec, J. (2005). *Slovník spisovné češtiny pro školu a veřejnost: s dodatkem ministerstva školství, mládeže a tělovỳchovy české republiky.* Academia.

Gala, N., Rey, V., & Zock, M. (2010). A tool for linking stems and conceptual fragments to enhance word access. In *Lrec.*

Grčar, M., Krek, S., & Dobrovoljc, K. (2012). Obeliks: statistični oblikoskladenjski označevalnik in lematizator za slovenski jezik. In *Zbornik osme konference jezikovne tehnologije.*

Habash, N., & Dorr, B. (2003). 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* (pp. 17–23).

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., ... Žabokrtský, Z. (2018). *Prague dependency treebank 3.5.* Retrieved from `http://hdl.handle.net/11234/1-2621` (LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University)

Hamp, B., & Feldweg, H. (1997). Germanet – a Lexical-Semantic Net for German. *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications.*

Haspelmath, M. (2001). The european linguistic area: standard average european. In *Language typology and language universals.(handbücher zur sprach-und kommunikationswissenschaft)* (pp. 1492–1510). de Gruyter.

Hathout, N. (2005). Exploiter la structure analogique du lexique construit: une approche computationnelle. *Cahiers de lexicologie: Revue internationale de lexicologie et lexicographie*(87), 5–28.

Hathout, N. (2010). Morphonette: a morphological network of French. *arXiv preprint arXiv:1005.3902*.

Hathout, N., & Namer, F. (2014). Démonette, a French derivational morpho-semantic network. *LiLT (Linguistic Issues in Language Technology)*, *11*.

Hathout, N., Namer, F., & Dal, G. (2002). An experimental constructional database: the MorTAL project. *Many morphologies*, 178–209.

Haverinen, K., Nyblom, J., Viljanen, T., Laippala, V., Kohonen, S., Missilä, A., ... Ginter, F. (2014). Building the essential resources for Finnish: the Turku Dependency Treebank. *Language Resources and Evaluation*, *48*(3), 493–531.

Havránek, B., et al. (1960). *Slovník spisovného jazyka českého.* Nakl. Československé akademie věd.

Henrich, V., & Hinrichs, E. (2011). Determining immediate constituents of compounds in GermaNet. In *Proceedings of the international conference recent advances in natural language processing 2011* (pp. 420–426).

Hoekstra, H., Moortgat, M., Schuurman, I., & Van Der Wouden, T. (2001). Syntactic annotation for the spoken dutch corpus project (cgn). *Language and Computers*, *37*, 73–87.

Janssen, M. (2005). Open source lexical information network. In *Third international workshop on generative approaches to the lexicon* (pp. 400–401).

Kahusk, N., Kerner, K., & Vider, K. (2010). Enriching Estonian WordNet with Derivations and Semantic Relations. In *Baltic hlt* (pp. 195–200).

Kanerva, J., Ginter, F., Ojala, S., & Missilä, A. (2017). *Finnish Parallel Universal Dependencies (PUD) Universal Dependency Treebank.* (Created for the CoNLL 2017 shared task on Multilingual Parsing from Raw Text to Universal Dependencies)

Kerner, K., Orav, H., & Parm, S. (2010). Growth and revision of Estonian wordnet. *Principles, Construction and Application of Multilingual Wordnets*, 198–202.

Klein, W., & Geyken, A. (2010). Das digitale wörterbuch der deutschen sprache (dwds). In *Lexicographica: International annual for lexicography* (pp. 79–96). De Gruyter.

Klosa, A., Schnörch, U., & Storjohann, P. (2006). Elexiko-a lexical and lexicological, corpus-based hypertext information system at the institut für deutsche sprache, mannheim. In *Atti del xii congresso internazionale di lessicografia: Torino, 6-9 settembre 2006* (pp. 425–429).

Koeva, S. (2008). Derivational and Morphosemantic Relations in Bulgarian WordNet. *Intelligent Information Systems*, *16*, 359–369.

Koeva, S., Genov, A., & Totkov, G. (2004). Towards bulgarian wordnet. *Romanian Journal of Information Science and Technology*, *7*(1-2), 45–60.

Koeva, S., Krstev, C., & Vitas, D. (2008). Morpho-semantic relations in Wordnet – a case study for two Slavic languages. In *Global wordnet conference* (pp. 239–253).

Krstev, C., Pavlovic-Lazetic, G., Vitas, D., & Obradovic, I. (2004). Using Textual and Lexical Resources in Developing Serbian WordNet. *Romanian Journal of Information Science and Technology*, *7*(1-2), 147–161.

Lango, M., Ševčíková, M., & Žabokrtský, Z. (2018). Semi-Automatic Construction of Word-Formation Networks (for Polish and Spanish). In *Lrec.*

Leseva, S., Stoyanova, I., Todorova, M., Dimitrova, T., Rizov, B., & Koeva, S. (2015). Automatic classification of wordnet morphosemantic relations. In *The 5th workshop on balto-slavic natural language processing* (pp. 59–64).

Liber, N. (2008). *Srce: Hrvatski jezični portal.* Retrieved from `http://hjp.znanje.hr/`

Lindén, K., & Carlson, L. (2010). FinnWordNet–Finnish WordNet by Translation. *LexicoNordica–Nordic Journal of Lexicography*, *17*, 119–140.

Lindén, K., Niemi, J., & Hyvärinen, M. (2012). Extending and updating the Finnish Wordnet. In *Shall we play the festschrift game?* (pp. 67–98). Springer.

Litta, E., Passarotti, M., & Culy, C. (2016). Formatio formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the third italian conference on computational linguistics (clic–it 2016)* (pp. 185–189).

Macleod, C., Grishman, R., Meyers, A., Barrett, L., & Reeves, R. (1998). Nomlex: A lexicon of nominalizations. In *Proceedings of euralex* (Vol. 98, pp. 187–193).

Maziarz, M., Piasecki, M., Szpakowicz, S., Rabiega-Wiśniewska, J., & Hojka, B. (2011). Semantic Relations between Verbs in Polish WordNet 2.0. *Cognitive Studies| Études cognitives*(11).

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., & Grishman, R. (2004). The nombank project: An interim report. In *Proceedings of the workshop frontiers in corpus annotation at hlt-naacl 2004*.

Miller, G. (1998). *WordNet: An electronic lexical database.* MIT press.

Mititelu, V. B. (2012). Adding Morpho-semantic Relations to the Romanian WordNet. In *Lrec* (pp. 2596–2601).

Müller-Spitzer, C., & Möhrs, C. (2008). First ideas of user-adapted views of lexicographic data exemplified on owid and elexiko. In *Proceedings of the workshop on cognitive aspects of the lexicon* (pp. 39–46). Stroudsburg, PA, USA: Association for Computational Linguistics. Retrieved from `http://dl.acm.org/citation.cfm?id=1598848.1598856`

Šnajder, J. (2014). DerivBase.hr: A High-Coverage Derivational Morphology Resource for Croatian. In *Proceedings of the ninth international conference on language resources and evaluation (lrec 2014). elra, reykjavik.*

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

Oliver, A., Šojat, K., & Srebačić, M. (2015). Enlarging the croatian wordnet with wn-toolkit and cro-deriv. In *Proceedings of the international conference recent advances in natural language processing* (pp. 480–487).

Paiva, V. d., Rademaker, A., & Melo, G. d. (2012). *Openwordnet-pt: An open brazilian wordnet for reasoning* (Tech. Rep.). COLING 2012.

Pala, K., & Hlaváčková, D. (2007). Derivational Relations in Czech WordNet. In *Proceedings of the workshop on balto-slavonic natural language processing: Information extraction and enabling technologies* (pp. 75–81).

Pala, K., & Šmerk, P. (2015). Derivancze—Derivational Analyzer of Czech. In *International conference on text, speech, and dialogue* (pp. 515–523).

Pala, K., & Smrž, P. (2004). Building Czech WordNet. *Romanian Journal of Information Science and Technology*, *7*(1-2), 79–88.

Partanen, N., Lim, K., & Poibeau, T. (2018). *Komi-Zyrian Lattice Universal Dependency Treebank.*

Piasecki, M., Szpakowicz, S., & Broda, B. (2009). *A Wordnet from the Ground Up.* Wroclaw: Oficyna Wydawnicza Politechniki Wroclawskiej. Retrieved from `http://www.dbc.wroc.pl/Content/4220/Piasecki_Wordnet.pdf`

Pravdová, M., & Svobodová, I. (2014). *Akademická příručka českého jazyka.* Academia.

Rademaker, A., De Paiva, V., de Melo, G., & Coelho, L. M. R. (2014). Embedding nomlex-br nominalizations into openwordnet-pt. In *Proceedings of the seventh global wordnet conference* (pp. 378–382).

Raffaelli, I., Tadić, M., Bekavac, B., & Agić, Ž. (2008). Building Croatian WordNet. In *Fourth global wordnet conference (gwc 2008).*

Razímová, M. Š., & Žabokrtskỳ, Z. (2006). Systematic parameterized description of pro-forms in the prague dependency treebank 2.0.

Sgall, P., Hajicová, E., & Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects.* Springer Science & Business Media.

Shafaei, E., Frassinelli, D., Lapesa, G., & Padó, S. (2017). DErivCELEX: Development and

Evaluation of a German Derivational Morphology Lexicon based on CELEX. In *Proceedings of the derimo workshop.* Milan, Italy.

Šojat, K., & Srebačić, M. (2014). Morphosemantic relations between verbs in Croatian WordNet. In *Proceedings of the seventh global wordnet conference* (pp. 262–267).

Šojat, K., Srebačić, M., Pavelić, T., & Tadić, M. (2014). CroDeriV: a new resource for processing Croatian morphology. *Proceedings of the Language Resources and Evaluation (LREC-2014)*, *14*, 3366–3370.

Steiner, P. (2016). Refurbishing a morphological database for german. In *Lrec.*

Steiner, P. (2017). Merging the Trees-Building a Morphological Treebank for German from Two Resources. In *Proceedings of the 16th international workshop on treebanks and linguistic theories* (pp. 146–160).

Talamo, L., Celata, C., & Bertinetto, P. M. (2016). DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure*, *9*(1), 72–102.

Tiberius, C., & Niestadt, J. (2010). The anw: an online dutch dictionary. In *Proceedings of the xiv euralex international congress. ljouwert, fryske akademy/afûk.*

Tufis, D., Mititelu, V. B., Bozianu, L., & Mihaila, C. (2006). Romanian wordnet: New developments and applications. In *Proceedings of the 3rd conference of the global wordnet association* (pp. 337–344).

Vitas, D., & Krstev, C. (2005). Derivational Morphology in an E-Dictionary of Serbian. In *Proceedings of 2nd language & technology conference* (pp. 139–143).

Zakharov, V. (2013). Corpora of the Russian language. In *International conference on text, speech and dialogue* (pp. 1–13).

Zeller, B., Padó, S., & Šnajder, J. (2014, August). Towards Semantic Validation of a Derivational Lexicon. In *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 1728–1739). Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Zeller, B., Šnajder, J., & Padó, S. (2013). DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the 51st annual meeting of the association for computational linguistics (volume 1: Long papers)* (Vol. 1, pp. 1201–1211).

Żmigrodzki, P. (2007). Wielki słownik języka polskiego pan. *Instytut Języka Polskiego PAN, Kraków.*

Appendix A Where to view derivational data online

Language	Resource	URL
Asturian	OSLIN-AST	http://ast.oslin.org/
Bulgarian	BulNet	http://dcl.bas.bg/bulnet/
Catalan	LOFC	http://ca.oslin.org/
Croatian	CroDeriV	http://croderiv.ffzg.hr/Croderiv
Croatian	HJP	http://hjp.znanje.hr/
Czech	DeriNet	http://ufal.mff.cuni.cz/derinet/derinet-search
Czech	SSČ	http://prirucka.ujc.cas.cz/
Czech	SSJČ	http://ssjc.ujc.cas.cz/
Dutch	ANW	http://anw.inl.nl/search
Dutch	E-Lex	http://tst.inl.nl/producten/eLex/?db_select=eLex_001
English	CatVar	https://clipdemos.umiacs.umd.edu/cgi-bin/catvar/webCVsearch.pl
Estonian	EstWN	https://teksaurus.keeleressursid.ee/
Finnish	FinnWordNet	http://kielipankki-tools.dy.fi/cgi-bin/fiwn/fiwn.cgi
French	OSLIN-FR	http://fr.oslin.org/
French	POLYMOTS	http://polymots.lif.univ-mrs.fr/v2/
German	DWDS	https://www.dwds.de/, http://www.canoo.net/
German	Elexiko	http://www.owid.de/suche/elex/erweitert
Italian	DerIvaTario	http://derivatario.sns.it/derivatario.php
Latin	WFL	http://wfl.marginalia.it/
Polish	PlWordNet	http://plwordnet.pwr.wroc.pl/wordnet/
Polish	Polish WFN	http://ufal.mff.cuni.cz/derinet/derinet-search
Polish	WSJP	http://wsjp.pl/
Portuguese	OpenWordNet-PT	http://wnpt.brlcloud.com/wn/
Portuguese	VOP	http://www.portaldalinguaportuguesa.org/index.php
Russian	OSLIN-RU	http://ru.oslin.org/
Russian	Rus. National Corpus	http://www.ruscorpora.ru/en/search-main.html
Russian	Unimorph	http://courses.washington.edu/unimorph/userInterface/rvnkur.php
Slovene	Sloleks	http://www.slovenscina.eu/sloleks
Spanish	OSLIN-ES	http://es.oslin.org/
Spanish	Spanish WFN	http://ufal.mff.cuni.cz/derinet/derinet-search

Appendix B Formats and Licenses of discovered resources

Language	Resource	License	Format
Bulgarian	BulNet	ELRA Agreement	`xml`
Croatian	CroDeriV	CC-BY-SA 3.0	`web`
Croatian	CroWN (WordNet)	CC BY-SA-NC 3.0	`xml`
Croatian	DerivBase.hr	CC BY-SA 3.0	`txt`
Czech	CS-WiktiWF	CC BY-SA-NC 4.0	`tsv`
Czech	Czech WordNet	ELRA Agreement	`xml`
Czech	DeriNet	CC BY-SA-NC 3.0	`tsv`
Czech	Prague Dependency Teebank	CC BY-SA-NC 4.0	`pml`
Dutch	D-CELEX	CELEX Agreement	`txt`
Dutch	E-Lex	License Agreement	`txt`
English	ADJAVD	LDC	`txt`
English	CatVar	OSL-1.1	`txt`
English	E-CELEX	CELEX Agreement	`txt`
English	EN-WiktiWF	CC BY-SA-NC 4.0	`tsv`
English	NOMADV	LDC	`txt`
English	NOMLEX	LDC	`txt`
English	NOMLEXPlus	LDC	`txt`
English	Princeton WordNet	WordNet 3.0	`xls`
Estonian	EstWN	CC BY-SA	`xml`
Finnish	FinnWordNet	CC BY 3.0	`tsv`
Finnish	PUD Treebank	CC BY-SA 4.0	`conllu`
Finnish	TDT Treebank	CC BY-SA 4.0	`conllu`
French	Démonette	CC BY-SA-NC 3.0	`xml`
French	Famorpho-FR	CC BY-SA-NC 2.0	`xml`
French	FR-WiktiWF	CC BY-SA-NC 4.0	`tsv`
French	Morphonette	CC BY-SA-NC 2.0	`xml`
French	Nomage	CC BY-SA 4.0	`xml`
French	POLYMOTS	CC BY-NC-ND 2.0	`web`
French	VerbAction	CC BY-SA-NC 2.0	`xml`
German	DE-WiktiWF	CC BY-SA-NC 4.0	`tsv`
German	DErivBase	CC BY-SA 3.0	`xml`
German	DErivCelex	CC BY-SA 3.1	`txt`
German	G-CELEX	CELEX Agreement	`txt`
German	GermaNet	License Agreement	`xml`
German	Morph. Treebank	CC BY-SA-NC	`script`
Italian	DerIvaTario	CC BY	`csv`
Komi-Zyrian	Lattice Treebank	CC BY-SA 4.0	`conllu`
Latin	WFL	CC BY-SA-NC 4.0	`sql`
Polish	PL-WiktiWF	CC BY-SA-NC 4.0	`tsv`
Polish	PlWordNet	plWordNet 3.0	`xml`
Polish	Polish WFN	CC BY-ND 3.0	`tsv`
Portuguese	Nomlex-PT	CC BY 4.0	`rdf`
Portuguese	OpenWordNet-PT	CC BY 4.0	`xml`
Romanian	RoWN	CC BY-SA-NC	`xml`
Russian	Rus. National Corpus	License Agreement	`-`
Russian	Unimorph	fully restricted	`web`
Serbian	E-Dictionary	MetaShare NC-NoReD	`xml`
Serbian	SrpWN	CC BY-NC 3.0	`xml`
Slovene	Sloleks	CC BY-SA-NC 4.0	`xml`
Spanish	Spanish WFN	LGPL LR	`tsv`
Turkish	Turkish WordNet	WordNet 3.0	`xml`

Appendix C Where to find reviewed resources

The list below contains useful URL links to a webpage or to some repository with data of given reviewed resource. For some resources, the author of the report did not find a webpage or repository with the current version of data, e.g. E-Dictionary, BulNet, etc. These resources are not included in list as well as digital explanatory dictionaries because their URLs are listed in Appendix A.

- CatVar [English]
  `https://clipdemos.umiacs.umd.edu/catvar/`
- CELEX [English, German, Dutch]
  `https://catalog.ldc.upenn.edu/LDC96L14`
- CroDeriV [Croatian]
  `http://croderiv.ffzg.hr/`
- DeriNet [Czech]
  `http://ufal.mff.cuni.cz/derinet`
- DerIvaTario (and CoLFIS corpus) [Italian]
  `http://derivatario.sns.it/`
  `http://linguistica.sns.it/CoLFIS/Descrizione.htm`
- DErivBase [German]
  `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/DErivBase.html`
- DerivBase.hr [Croatian]
  `http://takelab.fer.hr/data/derivbasehr/`
- DErivCelex [German]
  `http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/DErivBase.html`
- Démonette [French]
  `http://redac.univ-tlse2.fr/lexiques/demonette_en.html`
- E-Lex [Dutch]
  `https://ivdnt.org/downloads/taalmaterialen/tstc-e-lex`
- EstWordNet (and in `xml` format) [Estonian]
  `https://www.cl.ut.ee/ressursid/teksaurus/`
  `https://www.doi.org/10.15155/1-00-0000-0000-0000-0013AL`
- Famorpho-FR [French]
  `http://redac.univ-tlse2.fr/lexiques/famorpho-fr.html`
- FinnWordNet [Finnish]
  `http://www.ling.helsinki.fi/en/lt/research/finnwordnet/index.shtml`
- GermaNet [German]
  `http://www.sfs.uni-tuebingen.de/GermaNet/`
- Lattice Treebank [Kom-Zyrian]
  `http://universaldependencies.org/`
  `https://github.com/UniversalDependencies/UD_Komi_Zyrian-Lattice`
- Morphological Treebank [German]
  `https://github.com/petrasteiner/morphology`
- Morphonette [French]
  `http://redac.univ-tlse2.fr/lexiques/morphonette.html`

- Nomage [French]
  `https://sites.google.com/site/nomagesite/`
  `https://github.com/abalvet/nomage`
- NomBank [English]
  `https://nlp.cs.nyu.edu/meyers/NomBank.html`
  `https://nlp.cs.nyu.edu/nomlex/`
- Nomlex-PT [Portuguese]
  `https://github.com/own-pt/nomlex-pt`
- OpenWordNet-PT [Portuguese]
  `https://github.com/own-pt/openWordnet-PT`
- PlWordNet [Polish]
  `http://plwordnet.pwr.wroc.pl/wordnet/`
- Polish and Spanish WFN [Polish, Spanish]
  `http://ufal.mff.cuni.cz/derinet` (at the bottom of the webpage)
- POLYMOTS [French]
  `http://polymots.lif.univ-mrs.fr/v2/`
- Prague Dependency Treebank [Czech]
  `http://ufal.mff.cuni.cz/pdt3.5`
  `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2621`
- Princeton WordNet [English]
  `https://wordnet.princeton.edu/`
- PUD Treebank [Finnish]
  `http://universaldependencies.org/`
  `https://github.com/UniversalDependencies/UD_Finnish-PUD`
- Russian National Corpus [Russian]
  `http://www.ruscorpora.ru/en/`
- Sloleks [Slovene]
  `http://eng.slovenscina.eu/sloleks/opis`
- SrpWN [Serbian]
  `http://korpus.matf.bg.ac.rs/SrpWN/`
- TDT Treebank [Finnish]
  `http://universaldependencies.org/`
  `https://github.com/UniversalDependencies/UD_Finnish-TDT`
- Turkish WordNet [Turkish]
  `https://bitbucket.org/ozlemc/twn/downloads/`
- Unimorph [Russian]
  `http://courses.washington.edu/unimorph/`
- VerbAction [French]
  `http://redac.univ-tlse2.fr/lexiques/verbaction_en.html`
- WFL (and its blog and data) [Latin]
  `https://progetti.unicatt.it/progetti-milan-wfl-home`
  `https://wflblog.wordpress.com/`
  `https://github.com/CIRCSE/LEMLAT3`
- WiktiWF [Czech, English, French, German, Polish]
  `https://github.com/lukyjanek/wiktionary-wf`