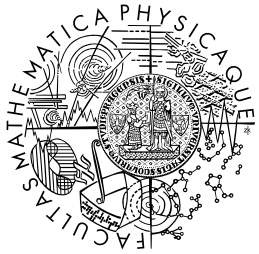


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Lukáš Kyjánek

Harmonisation of Language Resources for Word-Formation of Multiple Languages

Institute of Formal and Applied Linguistics

Supervisor of the master thesis: Mgr. Magda Ševčíková, Ph.D.
Study programme: Computer Science
Study branch: Computational Linguistics

Prague 2020

I declare that I carried out this master thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ..... Lukáš Kyjánek

I dedicate the thesis to Magda Ševčíková, Zdeněk Žabokrtský, Jonáš Vidra, and Anna Nedoluzhko. I thank them for their help and support.

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

Title:
Harmonisation of Language Resources for Word-Formation of Multiple Languages

Author:
Lukáš Kyjánek

Institute:
Institute of Formal and Applied Linguistics

Supervisor:
Mgr. Magda Ševčíková, Ph.D., Institute of Formal and Applied Linguistics

Abstract:
In the field of Natural Language Processing, word-formation is under-resourced comparing to inflectional morphology. Moreover, the existing resources capturing word-formation differ in many aspects. This thesis aims to review existing language resources for word-formation across languages and to unify them to a common data structure and file format. Basic notions of word-formation are followed by a review of existing language resources and their comparison in both quantitative and qualitative aspects. In the core part of the thesis, the harmonisation process is presented. Design decisions on the unification procedure are presented, and the selection of the resources to unify is described. The resources are unified to the rooted tree data structure and stored in a lexeme-based file format, which is already used in DeriNet 2.0. The procedure applies supervised machine learning model and the Maximum Spanning Tree algorithm. While the model scores word-formation relations, the MST algorithm uses the scores for identifying the rooted tree structure in each word-formation family. The resulting collection of harmonised resources covering 20 European languages was published under the title 'Universal Derivations' (UDer).

Keywords:
language resource, lexical resource, word-formation, derivation, harmonisation, natural languages, natural language processing

Contents

List of Figures

List of Tables

Introduction

Similarities in both form and meaning of some words can be easily noticed. For instance, words '*employer*', '*employee*', '*employable*', and '*employment*' relate formally and semantically to the verb '*employ*'. The form and meaning of '*employ*' is, however, slightly changed by *-er*, *-ee*, *-able*, and *-ment* to denote a person who employs other people ('*employer*'), a person who is employed ('*employee*'), a possession of enough abilities for being employed ('*employable*'), and a relation originated from employing someone ('*employment*'). Linguists address this phenomenon as *word-formation* or in a narrow sense as *derivational morphology*. Štekauer et al. (2012) attested it in many languages across the world.

In the recent two decades, electronic resources have been created to capture derivationally related words. These machine-trackable resources have been developed separately with minimal mutual influence (with a few exceptions) and different purposes. As a consequence, the situation around the resources seems fragmented, and the resources differ in many aspects. Even a list of the existing word-formation resources had not existed before the work on this thesis.

This thesis tries to change the situation. It reviews existing word-formation resources and describes their unification (harmonisation) in terms of data representation. A collection of harmonised resources is created as a result.

The idea of harmonising word-formation resources is inspired by the recent situation in syntactic treebanks. Collections of harmonised treebanks of many languages, e.g. HamleDT (Zeman et al., 2014), Universal Dependencies (Nivre et al., 2016), etc., have allowed subsequent development of multilingual syntactic analysers and knowledge-transfer methods for creating new treebanks. Harmonisation of word-formation resources might bring similar benefits to computational processing of word-formation.

The structure of the thesis is as follows. Chapter 1 describes basic notions of word-formation to provide the necessary linguistic background. The review of existing electronic language resources of word-formation available for different languages is presented in Chapter 2. Chapter 3 describes the harmonisation process, including the selection of the target data representation and resources for harmonisation. The resulting harmonised resources are quantitatively and qualitatively evaluated in Chapter 4, and they are assembled into a collection called *Universal Derivations*, which is freely available in the LINDAT/CLARIAH-CZ repository.

Chapter 1

Word-formation modelled in the resources

The opening chapter provides basic linguistic notions of word-formation of natural languages and especially the phenomena modelled in the existing word-formation resources. The structure of words is described, followed by a description of word-formation relations/processes.

If a word is taken, e.g. the verb '*play*', other words having a similar form and meaning (possible slightly shifted) can be observed, e.g. '*playing*', '*plays*', '*played*', '*player*', '*replay*', '*playable*', '*playtime*', '*playboy*'. The systematic combinations of form and meaning within words is studied by a linguistic discipline called *morphology*, which is subsequently subdivided into *inflectional morphology* and *derivational morphology* (Haspelmath & Sims, 2010, pp. 2, 18). The former one focuses on the relationship between word-forms belonging to the same word and expressing grammatical meanings (for instance, the third person singular present tense) so that the word can be used in a concrete sentence (Haspelmath & Sims, 2010, p. 16). For example, word-forms '*plays*', '*played*', '*playing*' belong to the verb '*play*'. The latter one studies the relationship between words that are not inflectionally related but still share form and meaning (Haspelmath & Sims, 2010, p. 17), such as words '*player*', '*replay*', '*playable*'. They together could create a set of derivationally related words, so-called *word family*. While the inflected word-forms '*plays*' or '*played*' stay for the same concept as the verb '*play*' and their main difference is only in the syntactic context whose formal requirements they satisfy, derivationally related words '*player*' or '*playable*' denote new concepts different from the concepts of the simple corresponding word '*play*'. Besides inflexion and derivation, some more complex relations also exist, e.g. in the case of compounding, some words (compounds), such as '*playtime*' and '*playboy*', could belong to more word families. Derivation, compounding and other more complex relationships are usually addressed as *word-formation* (Haspelmath and Sims, 2010, pp. 18–19; Štekauer et al., 2012, p. 15).

This thesis focuses on word-formation, especially on derivation. Although the borderline between inflexion vs. derivation is a wanted ideal only (Štekauer et al., 2012, p. 14), inflexion is not further described. Štekauer et al. (2012, pp. 19–35) and ten Hacken (2014, pp. 10–25) document corner-cases of delineating the borderline and claim that the phenomena should be treated as scales rather than as dichotomies (Štekauer et al., 2012, p. 19; ten Hacken, 2014, p. 11).

1.1 Word structure

For inflectional morphology and word-formation, the basic meaningful unit of a word is a *morpheme* (Matthews, 1991, p. 12). They are identified by similarities in forms and meanings of words, for example, *-s* means plural in the words '*dogs*', '*cats*', and '*birds*'. A morpheme is an abstract unit having a form and meaning, and its concrete surface form, so-called *morph*, does not have to be unique, e.g. '*dog-s*', '*potato-es*' (Matthews, 1991, p. 107). If one morpheme has more than one morph, then linguists use the term *allomorphs* to address individual morphs.

The process of decomposing a word into morphemes is usually called morphological segmentation. Lipka (1975, p. 179) proposed morpheme classification on the basis of two oppositions:

1. *lexical* vs. *grammatical morphemes*
   (a) lexical morphemes carry meaning,
   (b) grammatical morphemes convey grammatical functions of words;
2. *free* vs. *bound morphemes*
   (a) free morphemes can stand alone as words, or be combined with other morphemes as *roots*,
   (b) bound morphemes must be combined with other morphemes as *affixes*.

Every morpheme is assumed be classified into one of the four combinations: a lexical free morpheme (*content words*, e.g. '*play*', '*boy*', '*nice*'), a lexical bound morpheme (*derivational affixes*, e.g. *un-*, *dis-*, *-like*, *-ly*), a grammatical free morpheme (*function words*, e.g. '*the*', '*at*', '*and*'), a grammatical bound morpheme (*inflectional affixes*, e.g. *-s*, *-est*, *-ing*).

Based on the position in a word, the following morphemes are distinguished: (1) *root* as a nucleus of the word, (2) *prefix* preceding the root, (3) *suffix* following the root, (4) *circumfix* surrounding the root, (5) *infix* is inserted into another morpheme, (6) *interfix* connecting two (root) morphemes.

Even though the term *word* has been used so far, terms *lexeme* and *lemma* are used in linguistics to generalise and simplify the description of individual word-forms. The lexeme denotes a set of word-forms with the same root and related through inflexion (Hladká, 2017), whereas the lemma refers to one canonical representative form of a lexeme in a dictionary or language resource (Hladká & Cvrček, 2017). To give an example, '*plays*', '*played*', '*playing*' are word-forms of the same lexeme with the lemma '*play*'. The approaches to identification of lexemes and their lemmas (*lemmatisation*) can differ across languages.

1.2 Word-formation processes

Concurring with Lipka's (1975, p. 179) morpheme classification presented in the previous section, Kastovsky (1982, p. 73) claims that inflectional morphology focuses on grammatical free and bound morphemes through declination and conjugation, while word-formation deals with lexical free and bound morphemes. According to the used type of lexical morpheme, Štekauer et al. (2012, p. 15) distinguish three groups of word-formation processes: (a) with bound morphemes, (b) with free morphemes, (c) without additional derivational material.

1.2.1 Processes with bound morphemes

Derivation adds/changes/removes lexical bound morphemes to a lexical free morpheme or a lexeme (Štekauer et al., 2012, p. 135), e.g. verb '*to re-write*' derived from the verb '*to write*'. The entering lexeme is called a *base lexeme*, while the resulting lexeme is referred to as a *derivative* (also *derivational parent* and *child*). The process can change the part-of-speech category of the base lexeme, e.g. '*careful*' → '*careful-ly*', modify/add a non-grammatical meaning, e.g. '*to write*' → '*to re-write*', or do both, e.g. '*large*' → '*to en-large*'.

The meaning of derivatives can estimated by analogy in word structures. As an illustration, the meaning of the verb '*to rewrite*' derived from the verb '*to write*' can be deduced by analogy with other verbs using the same prefix *re-*, e.g. '*to restart*', '*to rebuild*', '*to remarry*', etc., which conveys the meaning '*do again*' (Cambridge Dictionary, 2020). Lexemes that are not derived are addressed as *unmotivated*, in contrast with *motivated lexemes* whose base lexeme exists (Dokulil, 1962, p. 103).

In general, Dokulil (1962, pp. 11–12) defines derivation as a relationship of both the form (*foundation*) and the meaning (*motivation*) between a derivative and its base lexeme. The form and meaning of the derivative are based on the form and the meaning of its base lexeme. Derivatives are expected to have more complex morphological structures, but their meanings are expected to be narrower. The relation between form and meaning expressed by morphemes is usually not one-to-one because the same meaning in a particular language can be conveyed by several different forms and vice versa. For instance, morphemes *-ka* in '*učitel-ka*' ('*female teacher*'), *-ová* in '*šéf-ová*' ('*female boss*'), *-yně* in '*ministr-yně*' ('*female minister*'), derive female counterparts of profession names in Czech. However, one morpheme can convey more than one meaning, e.g. *-ka* occurs not only in female nouns but also in instrument nouns as '*obál-ka*' ('*envelope*'), diminutives as '*skříň-ka*' ('*small cupboard*'), etc. (Ševčíková & Kyjánek, 2019, p. 420).

Several types of derivations (derivational processes) can be distinguished by the position of an attached lexical bound morpheme (illustrated in the Slovak language; Štekauer et al., 2012, pp. 143, 161, 199, 210):

- *prefixation* attaches a prefix so that it precedes the root of the base lexeme, e.g. '*písať*' ('*to write*') → '*pre-písať*' ('*to re-write*');
- *suffixation* attaches a suffix so that it follows the root of the base lexeme, e.g. '*ruka*' ('*a hand*') → '*rúč-ka*' ('*little hand*');
- *circumfixation* attaches a prefix and a suffix in one step whereas neither the prefixed root not the suffixed root are attested alone,
  e.g. '*mesto*' ('*town*') → '*pred-mest-ie*' ('*suburb*'), neither '*pred-mest(o)*', nor '*mest-ie*' exist;
- *infixation* inserts an infix into a free morpheme,
  e.g. '*dva*' ('*two*') → '*dv-aj-a*' ('*two male persons*').

Word-formation does not have to be reduced to binary derivational relations only. Dokulil (1962, pp. 12–14) views such pairs as a basis for modelling of more complex structures:

A derivational paradigm ('*slovotvorný svazek*' in Czech) is an ordered set of derivatives derived directly from the same base lexeme, e.g.

'*list*' ('*leaf*') → '*lísť-ek*' ('*small leaf*')
→ '*list-oví*' ('*leafage*')
→ '*list-natý*' ('*leafy*')

Furdík (2004, p. 74) postulates an idea of a system of *derivational cases* analogously to *inflectional cases* but less systematic.

A derivational series ('*slovotvorná řada*' in Czech) represents a subsequent derivation of lexeme from each other one by one, e.g.

'*list*' ('*leaf*') → '*lísť-ek*' ('*small leaf*') → '*lístk-ový*' ('*leafy by small leaves*')
→ '*lístkov-itý*' ('*being leafy by small leaves*')

A derivational nest ('*slovotvorná čeleď*' in Czech) comprises recursive combinations of above-described derivational paradigm and series so that all lexemes share the same root in one derivational nest (also *derivational cluster* or *family*), e.g.

'*list*' ('*leaf*') → '*lísť-ek*' ('*small leaf*')
→ '*lístk-ový*' ('*leafy by small leaves*')
→ '*lístkov-itý*' ('*being leafy by small leaves*')
→ '*lísteč-ek*' ('*really small leaf*')
→ '*lístečk-ový*' ('*leafy by really small leaves*')
→ '*lístečkov-itý*' ('*being leafy by r. s. leaves*')
→ '*list-oví*' ('*leafage*')
→ '*list-natý*' ('*leafy*')
→ '*listn-áč*' ('*leafy tree*')
→ '*listnat-ě*' ('*leafly*')

Dokulil's approach has been further elaborated on and is still being applied by Buzássyová (1974, pp. 24, 73–74), Horecký et al. (1989, pp. 38–47), Furdík (2004, pp. 73–77), and Štekauer (2005, p. 207).

Besides theory proposed by Dokulil, van Marle (1985) presents the paradigmatic approach discussing derivational paradigms in the broader context of word-formation and describing paradigms of derivationally-related lexemes in a similar way as it is done in the case of inflectionally-related lexemes. Bonami and Strnadová (2019, pp. 167–182) summarise a previous debate and provide definitions of individual used terms.[1] Figure 1.1 shows the key concepts in the paradigmatic approach (Bonami & Strnadová, 2019, pp. 169–173):

A morphological family is a tuple of morphologically related lexemes (having the same root) without any internal order in contrast with Dokulil's

[1]Definitions are formulated as relatively general using the word *morphological* to allow describing paradigms of derivationally-related and inflectionally-related lexemes at the same time.

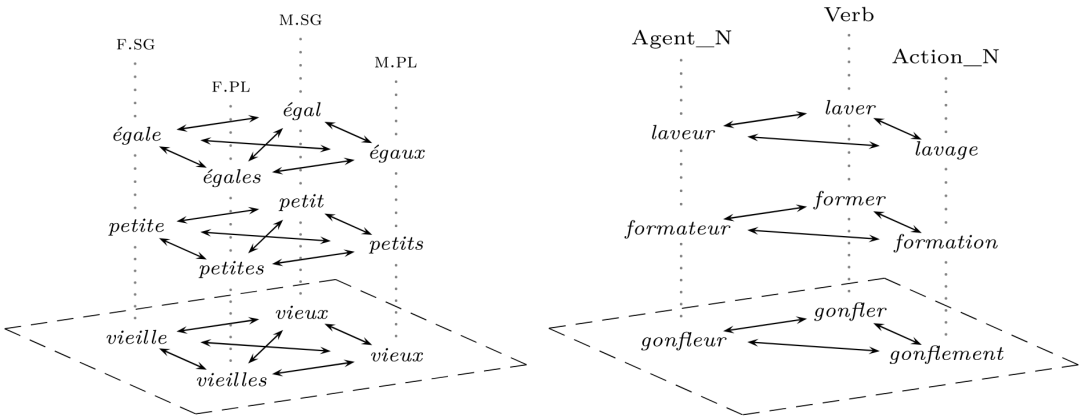


Figure 1.1: Paradigmatic systems of partial morphological families of inflectionally-related (left) and derivationally-related (right) lexemes in French (Bonami & Strnadová, 2019, p. 172).

derivational nests. An overlapping term (*derivational/inflectional*) *family* is also used for the tuple. A morphological family can be treated as *complete* or *partial*. While a partial family contains a subset of morphologically related lexemes only, a complete family includes all morphologically related lexemes.

An aligning relation represents a property of two pairs of morphologically related lexemes. If two pairs convey the same content, i.e. meaning, grammatical or non-grammatical category, then the pairs are aligned. The same form is not required. For example, in French, the pair '*laver*' ('*to wash*') ↔ '*lavage*' ('*to washing*') is aligned with '*former*' ('*to form*') ↔ '*formation*' ('*to forming*') because they are in the same relation (verb and its action noun).

A paradigmatic system is a set of morphological families of the same size of morphologically related lexemes such that the relations are aligned pairwise by the same aligning relations. Figure 1.1 shows paradigmatic systems of partial morphological families (horizontal levels) whose relations are aligned (vertical levels). The pairs in the vertical levels are usually called (*derivational/inflectional*) *series*. The paradigmatic system is also simply addressed as a (*derivational/inflectional*) *paradigm*. Although the terms overlap with Dokulil's ones, the individual concepts are different from Dokulil's.

1.2.2 Processes with free morphemes

Compounding combines two or more lexical free morphemes (Štekauer et al., 2012, p. 42). The prototypical *compound lexemes* (*compounds*) consists of two parts: free morphemes (roots) and possibly a linking element (an interfix), e.g. '*tmav-o-modrý*' ('*dark blue*') in Czech. Dokulil (1962, p. 22) considers compounds as an intermediate stage between derivation and syntax. In addition, Olsen (2014, pp. 26–49) and Štekauer et al. (2012, pp. 36–48) document that borderlines between compounding vs. derivation and compounding vs. syntax are fuzzy.

Reduplication repeats the same morpheme, e.g. '*neri neri*' ('*really black*') in Italian, '*čern-o-černý*' ('*really black*') in Czech (Štekauer et al., 2012, pp. 103–104). Despite the reduplication being attested in both derivation and inflexion, it seems to be more frequent in derivation (Bybee, 1985, p. 97), e.g. '*ma-li-…-li-nký*' ('*very . . . very small*') in Slovak (Štekauer et al., 2012, p. 94).

Blending reduces and joins two lexical free morphemes, e.g. '*photocopillage*' ('*illegal photocopying*') created from '*photocopy*' and '*pillage*' in French (Štekauer et al., 2012, pp. 131–132).

1.2.3 Processes without additional derivational material

Conversion forms a new lexeme having a different part-of-speech category without any formal changes, e.g. noun '*a pilot*' and verb '*to pilot*' (Štekauer et al., 2012, p. 213). However, the definition is not stable across individual linguistic traditions. Especially in languages with inflectional morphology, there are also other definitions of conversion because of vague notions of part-of-speech categories and lack of formal change. For instance, Dokulil (1962, pp. 24, 62–65) understood both vague conditions as the change of the set of inflectional features (inflectional paradigm) including phonetic alternations, so adjective '*zlý*' ('*evil*') and adverb '*zlo*' ('*an evil*') in Czech had been treated as conversion in the Czech tradition before Dokulil's (1982) reassessment of the process as so-called *transflection*. Besides that, Štekauer (1996, pp. 55–95) argues that stress shifting, e.g. noun '*ˈrecord*' and verb '*reˈcord*' (Štekauer et al., 2012, p. 225), and tone/pitch shifting, e.g. verb '*àô*' ('*to fly*') and noun '*àó*' ('*eagle*') in Cirecire (Štekauer et al., 2012, p. 227), should also be treated as a specific case of conversion.

Chapter 2

Language resources capturing word-formation

Existing language resources capturing word-formation across languages are presented in this chapter. The resources are described in terms of their origin and their technical and linguistic background. Basic statistic properties are also measured to allow a simple comparison of the reviewed resources.

Although the study of word-formation has been an established linguistic subdiscipline for a long time, in the field of Nature Language Processing, word-formation has not got much attention. Language resources focusing exclusively on word-formation have been developed only recently. Before that, word-formation had been captured marginally in language resources, or only incidentally in resources capturing other phenomena. The existing resources had not been listed, so a draft containing their list and description was published by Kyjánek (2018) before publishing this thesis. The draft is updated and extended here.

There exist several different types of the electronic word-formation resources:

- *morphological segmenters*, e.g. DériF for French (Namer, 2003), Frog for Dutch (Bosch et al., 2007), and *derivational analysers*, e.g. Derivancze for Czech (Pala & Šmerk, 2015);
- *digital datasets*, e.g. CatVar for English (Habash & Dorr, 2003), CroDeriV for Croatian (Šojat et al., 2014), CELEX for Dutch, English, and German (Baayen et al., 1995);
- various *supervised*, *semi-supervised*, and *unsupervised methods* to create digital datasets, e.g. Gaussier (1999), Baranes and Sagot (2014), Lango et al. (2020);
- *digitised monolingual dictionaries*, e.g. Algemeen Nederlands Woordenboek for Dutch (Tiberius & Niestadt, 2010), Wielki słownik języka polskiego for Polish (Żmigrodzki et al., 2007).

Since the thesis is not focused on the creation of new digital datasets using morphological segmenters, derivational analysers, or methods mentioned above, these types of resources are not described in more details here. Regarding the aim of the thesis, attention is paid to stable released digital datasets that can be harmonised. Hereafter, the term (*word-formation*) *resource* is used in a narrower sense for digital datasets capturing word-formation.

2.1 Resources specialised in word-formation

2.1.1 Morpheme-oriented resources

Resources capturing word-formation as a decomposition of an individual lexeme into morphemes are presented as *morpheme-oriented* here.

CELEX is a large manually created resource providing orthographic, phonetic, morphological, and syntactic annotations for Dutch, English, and German (Baayen et al., 1995). The three language parts of CELEX were developed separately for psycholinguistic research. Their sets of lexemes come from various dictionaries and corpora. The data, see slash-separated columns in Figure 2.1, provides three types of morphological segmentation: (a) immediate segmentation of lexemes into bases and affixes, (b) hierarchical segmentation of lexemes into morphemes organised into a tree structure, and (c) flat segmentation of lexemes into morphemes obtainable from the last tree level. Individual morphemes are also labelled in columns 13 (number or capital letter for the base, x for the affix) and 21 (A for the affix, S for the root). In the case of the German part of CELEX, the orthographic forms of lexemes do not comply with the current German orthographic standards.[1]

```
1 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
2 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
3 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
```

Figure 2.1: Slash-separated textual file format of CELEX. Some positions differ across the language versions of the resource. In the English part, each line contains: a lexeme (2nd position), an immediate morphological segmentation (12th), morpheme labels (13th, 21st), and bracketed hierarchical and flat morphological segmentation (22nd).

Morphological Treebank is created by Steiner (2016) who merged word-formation data from German part of CELEX and GermaNet (German WordNet). She named the resulting resource as Morphological Treebank because particular segmented morphemes are organised into trees, as in the original input resources. During the development of Morphological Treebank, the inaccurate orthographic standard in the German part of CELEX was fixed. Later, Steiner (2019) augmented and revised the Morphological Treebank.

DerIvaTario contains manually morphologically segmented Italian nouns, adjectives, verbs, and adverbs (see Figure 2.2) extracted from a large Italian corpus (Talamo et al., 2016). Each lexeme is linked to other Italian language resources using a unique ID which allows obtaining various information about the particular lexeme, e.g. morphological categories, phonetic transcription, etc. DerIvaTario can be queried online.[2]

[1] Steiner (2016) created an automatic orthographic correction for the German CELEX.

[2] http://derivatario.sns.it/derivatario.php

```
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;;;;
```

Figure 2.2: Semicolon-separated textual file format of DerIvaTario. Each line contains: an ID, a lexeme, a root, and affixes.

MorphoLex-like resources

MorphoLex-like resources are datasets created for research of word-formation in the field of psycholinguistics, cognitive psychology, and cognitive science. The datasets contain lexemes assigned several morphological categories, including morphological segmentation. The segmentation (see Figure 2.3) is arranged using the following characters: « for prefixes, » for suffixes, and {} for lexical bases.

MorphoLex-en is data created for research into English word-formation (Sánchez-Gutiérrez et al., 2018). It was developed based on English Lexicon Project (Balota et al., 2007) and English part of CELEX.

```
weightier    [...]  {(weigh)>t>}>y>>er>    [...]
weightiest   [...]  {(weigh)>t>}>y>>est>   [...]
weightily    [...]  {(weigh)>t>}>y>>ly>    [...]
weightiness  [...]  {(weigh)>t>}>y>>ness>  [...]
weightlessly [...]  {(weigh)>t>}>less>>ly> [...]
```

Figure 2.3: Microsoft Excel file format of MorphoLex-en. Each line contains a lexeme, its morphological segmentation, and many other variables (skipped).

MorphoLex-fr was developed and utilised for research in French word-formation (Mailhot et al., 2019). It is based on French Lexicon Project (Ferrand et al., 2010). Since one of the goals of creating the dataset was to provide a cross-linguistic comparison, the resource mirrors MorphoLex-en. MorphoLex-fr stores the data in the same file format as MorphoLex-en.

Unimorph also known as The Russian Morphological Database, is a lexicon of manually morphologically segmented Russian nouns, adjectives, verbs, and adverbs. It is based on large Russian grammar books, and it is available for queries.[3]

2.1.2 Lexeme-oriented resources

Resources capturing word-formation as relations between individual derivationally related lexemes are presented as *lexeme-oriented* here. By assembling all together connected lexemes, a word-formation family is obtained.

[3] http://courses.washington.edu/unimorph/

DeriNet-like resources

DeriNet-like resources are datasets capturing word-formation of different languages in a similar way as a monolingual word-formation resource for Czech, DeriNet. The resources model relations as directed edges between derivatives and their base lexemes, which concurs with Dokulil's (1962) description of the word-formation system. All DeriNet-like resources adhere to the principle that each lexeme (except for compound lexemes) can have at most one base lexeme. Thus, word-formation families are represented as rooted trees. The resources can be queried online.[4] The Polish and Spanish Word-Formation Networks described below were developed together using an unsupervised machine learning method proposed by Lango et al. (2018).

DeriNet is a semi-automatically created word-formation lexicon of derivationally related nouns, adjectives, verbs, and adverbs (Vidra, Žabokrtský, Kyjánek, et al., 2019). Its lexemes are taken from a large inflectional dictionary, and derivational relations between them originate from semi-automatic annotation procedures. The data structure and the file format of DeriNet have undergone significant changes (Vidra, Žabokrtský, Ševčíková, et al., 2019) in DeriNet version 2.0. The new data representation (see Figure 2.4) allows adding a lot of new features, such as morphological categories, morphological segmentation, semantic labels, etc. The data structure is prepared to capture compounds, which was not possible in the older file format (cf. Figure 2.5).

```
1 215108.0  šerif#NNM??-----A---?  šerif  N  Animacy=Anim&Gender=Masc  _  _  _
      _  {"techlemma": "šerif"}
2 215108.1  šerifka#NNF??-----A---?  šerifka  N  Gender=Fem  _  215108.0
    SemanticLabel=Female&Type=Derivation  _  {"techlemma": "šerifka_^(*2)"}
3 215108.2  šerifčin#AU????--------?  šerifčin  A  Poss=Yes  _  215108.1
    SemanticLabel=Possessive&Type=Derivation  _  {"techlemma": "šerifčin_^(*3
    ka)"}
4 215108.3  šerifský#AA???----??---?  šerifský  A  _  _  215108.0  Type=
    Derivation  _  {"techlemma": "šerifský"}
5 215108.4  šerifskost#NNF??-----?---?  šerifskost  N  Gender=Fem  _  215108.3
     Type=Derivation  _  {"techlemma": "šerifskost_^(*3ý)"}
6 215108.5  šerifsky#Dg-------??---?  šerifsky  D  _  _  215108.3  Type=
    Derivation  _  {"techlemma": "šerifsky_^(*1ý)"}
7 215108.6  šerifství#NNN??-----A---?  šerifství  N  Gender=Neut  _  215108.3
     Type=Derivation  _  {"techlemma": "šerifství"}
```

Figure 2.4: Tab-separated textual file format of DeriNet version 2.0. Each line consists of 10 columns containing: an ID, a unique lexeme ID, a written form of a lexeme, a part-of-speech category, morphological categories, a morphological segmentation, an ID referring to the base lexeme, an annotation of the relation, other relations, a JSON-encoded custom data. Empty columns are filled with underscores.

DeriNet.FA is an automatically developed word-formation lexicon of Persian (Haghdoost et al., 2019). Its construction is based on manually morphologically segmented lexemes. The lexemes have not yet been assigned part-of-speech categories. DeriNet.FA stores data in the same new file format as DeriNet 2.0.

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

DeriNet.ES is a word-formation resource of Spanish (Faryad, 2018). Its first version started as a revision of The Spanish Word-Formation Network. Faryad (2018) decided to revise the lexeme set and re-identify derivational relations between lexemes without considering the original relations. DeriNet.ES version 0.5 stores data in the older DeriNet file format.

The Polish Word-Formation Network is a semi-automatically created lexicon of the Polish word-formation (Lango et al., 2018). Its lexemes, without assigned part-of-speech categories, come from a large dictionary and Polish WordNet. After applying the machine learning model to create The Polish WFN, the relations extracted from Polish WordNet were included in the resulting data, too. The Polish WFN is stored in the older DeriNet format, see Figure 2.5.

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

Figure 2.5: Tab-separated textual file format of The Polish Word-Formation Network (the older file format of DeriNet-like resources that was used before the release of DeriNet version 2.0). Each line consists of 5 columns containing: an ID, a written form of a lexeme, a unique lexeme ID, a space for part-of-speech category, an ID referring to the base lexeme. If empty, then filled with underscores.

The Spanish Word-Formation Network was constructed together with The Polish WFN by Lango et al. (2018). Its lexemes came from a morphological and syntactic lexicon of Spanish. Since Faryad (2018) noticed that the lexeme set contains many French lexemes and proper nouns, he has revised the resource and published it as DeriNet.ES. The Spanish WFN is stored in the older DeriNet format.

DerivBase-like resources

German DErivBase has inspired the creation of other similar DerivBase-like word-formation resources. The resources have been constructed based on heuristic identification of derivational relations between lexemes using a rule-based approach. The approach has identified derivational relations between individual lexemes, and the *word-formation rules* are also included in the data. Word-formation families can be obtained by grouping all connected lexemes. DerivBase.Hr and DErivCELEX have also been inspired by DErivBase, but they are presented among *family-oriented* word-formation resources because they contain only word-formation families.

DErivBase is a word-formation resource for German that includes derivationally related nouns, adjectives, and verbs (Zeller et al., 2013). While its lexemes came from a large German web corpus, the rules used for identifying derivational relations were extracted from several German grammar books.

Steiner (2016) noticed that lexemes in DErivBase do not concur with the current German spelling standards. Zeller et al. (2014) split derivational families into semantically more consistent clusters in DErivBase version 2.0. The resource is distributed as a package of three files containing: (a) whole word-formation families without individual relations between lexemes, (b) individual derivational relations between lexemes see Figure 2.6, and (c) rules used to identify derivational relations.

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

Figure 2.6: Space-separated textual file format of DErivBase. Each line contains: a derivative, a derivationally related lexeme, a length of the shortest path between the lexemes, and the path separated by applied word-formation rules.

DerivBase.Ru is a word-formation resource for Russian capturing derivationally related nouns, adjectives, verbs, and adverbs (Vodolazsky, 2020). Its lexemes came from Russian Wikipedia, and the rules were extracted from several Russian grammar books. The file format of DerivBase.Ru slightly differs from DErivBase (see Figure 2.7).

```
детсад    noun  детсадик       noun  rule429(noun + ик/ок/ук -> noun)  SFX
детсад    noun  детсадовский   adj   rule630(noun + ск(ий) -> adj)     SFX
антиправо noun  антиправовой   adj   rule628(noun + ов(ый) -> adj)     SFX
```

Figure 2.7: Tab-separated textual file format of DErivBase.Ru. Each line contains: a lexeme and its part-of-speech category, its derivative and its part-of-speech category, applied word-formation rules and process.

Word Formation Latin also abbreviated as WFL, is a word-formation resource for Classical Latin (Litta et al., 2016). It is a semi-automatically created lexicon containing nouns, adjectives, verbs, adverbs, and few lexemes from other part-of-speech categories. WFL captures not only derivational relations but also compounding relations. In the first versions of WFL, at most one base lexeme has been preferred for a derivative (except for compound lexemes), so derivational families have been represented as rooted trees. However, Litta et al. (2019) presented a new version that organises the data in a morpheme-oriented approach. For each lexeme, WFL provides annotations of morphological categories, morphological segmentation, and the word-formation process used to derive (or compose) the lexeme. While the first versions of WFL have been integrated into SQL database of Latin morphological analyser LEMLAT3, the new version has been integrated to LiLa Knowledge Base infrastructure. The resource can be queried online.[5]

[5]http://wfl.marginalia.it/ and https://lila-erc.eu/sparql/

CroDeriV in full name Croatian Derivational Lexicon, is a manually created word-formation resource for Croatian (Šojat et al., 2014). In its first version, which can be queried online,[6] CorDeriV was morpheme-oriented, and it focused on the morphological structure of 14,500 Croatian verbs. Filko et al. (2019) presented significant changes and enrichment in the newest version, CroDeriV 2.0. It contains 21 thousand lexemes including nouns, adjectives, and verbs taken from a large Croatian web corpus. Besides manual morphological segmentation for each lexeme, the CorDeriV is enriched with links connecting derivationally related lexemes. Except for compound lexemes, at most one base lexeme is preferred for each derivative. CroDeriV also contains extensive manual annotations of morphological categories, morphological segmentation (including the normalisation of allomorphy), word-formation properties, and semantic labels. Moreover, each lexeme is assigned web links to other Croatian resources.

Resources of nominalisations

The following resources focus on nominalisations of verbs, i.e. verbs turned into nouns. For example, the English verb '*to combine*' can be turned into a noun '*combination*' by attaching derivational affix.

NOMLEX is a manually constructed lexicon of English nominalisations (Macleod et al., 1998). Its derivational relations (see Figure 2.8) were identified on the basis of a list of suffixes used to nominalise English verbs.

```
(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)))))
```

Figure 2.8: Textual file format of NOMLEX. The entry constains not only derivational relation but also other syntactic annotations.

VerbAction is a lexicon of French nominalisations (Hathout et al., 2002). Its lexemes came from several lexicons, and the relations (see Figure 2.9) were captured using a rule-based approach and manual annotations.

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

Figure 2.9: XML file format of VerbAction.

[6] http://croderiv.ffzg.hr/Croderiv

Nomage is a semi-automatically created lexicon of French nominalisations (Balvet et al., 2010). Its lexemes came from one of the French treebanks, and the relations were obtained based on a list of suffixes used to nominalise French verbs. It also includes 4 semantic labels for verbs (state, activity, achievement, perfective), and 3 semantic labels for nouns (habit, object, information object). Figure 2.10 illustrates the original file format of the resource.

```
<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 à 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>
```

Figure 2.10: XML file format of Noamage. A derivative is captured between *Lemma* tags and its base lexeme is in the *SenseRelation* tag.

NomLex-PT also known as NomLex-BR, consists of nominalisations in Brazilian Portuguese (De Paiva et al., 2014). Lexemes came from various language resources, and derivational relations were obtained based on a list of common suffixes. The relations can be extracted from links stored in the XML file format of the data, see Figure 2.11.

```
<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>
```

Figure 2.11: XML file format of NomLex-PT.

NomBank

NomBank collection of resources (Meyers et al., 2004) started as a revision of already existing English NOMLEX. However, several new language resources focusing on derivational relations among English lexemes were created and included in the collection. Their sets of lexemes came from various corpora and treebanks.

NOMLEXPlus represents a revised version of NOMLEX. Nominalisations of adjectives were added into NOMLEXPlus, see Figure 2.12.

```
(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)
```

Figure 2.12: Textual file format of NOMLEXPlus. The format resembles the NOMLEX format.

ADJADV captures derivationally related adjectives and adverbs (and also nine verbs). Figure 2.13 illustrates the original file format of the resource.

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

Figure 2.13: Textual file format of ADJADV. The format resembles the NOMLEX format.

NOMADV focuses on derivationally related English adverbs and nouns, see Figure 2.14.

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

Figure 2.14: Textual file format of NOMADV. The format resembles the NOMLEX format.

2.1.3 Paradigm-oriented resources

The *paradigm-oriented* resources capture word-formation using references between individual lexemes as lexeme-oriented word-formation resources do, but the goal of the paradigm-oriented resources is to model word-formation as paradigmatic systems consisting of aligned morphological relations as presented in Section 1.2.1. As a consequence, the paradigm-oriented resources often contain only lexemes involved in particular (sub)paradigms, but other potentially derivationally related lexemes are omitted.

Morphonette is an automatically created lexicon for French, which focuses on derivational series (using the terminology of the paradigmatic approach to word-formation) of derivationally related nouns, adjectives, verbs, and adverbs (Hathout, 2010; see Figure 2.15). In contrast with the current definition of a derivational series in the paradigmatic approach to word-formation presented in Section 1.2.1, lexemes in Morphonette are aligned in a derivational series only if their conveyed content is expressed by the same form.

```
<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>
```

Figure 2.15: XML file format of Morphonette. Besides derivational relation, each entry also contains derivational series.

Démonette merges the existing resources of French word-formation (morphological segmenters, VerbAction, and Morphonette) into one morpho-semantic network (Hathout & Namer, 2014). Démonette focuses on derivational families and derivational series (in the terminology of the paradigmatic approach to word-formation) of nouns, adjectives and verbs. It distinguishes direct and indirect relations within derivational families. While the direct relations connect lexemes with their base lexemes, indirect relations connect lexemes within the other more distant members of their derivational family. Démonette includes annotations of the morphological categories, morphological segmentation, and the semantics of derivational relations, see Figure 2.16. Namer and Hathout (2019) announced a new, significantly improved Démonette version 2.0.

```
<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>
```

Figure 2.16: XML file format of Démonette.

2.1.4 Family-oriented resources

Resources that group derivationally related lexemes into whole word-formation families without specifying individual relations between lexemes are presented as *family-oriented* resources here.

CatVar in full name the Categorial Variation Database, is an automatically constructed word-formation database of English derivationally related nouns, adjectives, verbs, and adverbs (Habash & Dorr, 2003). It was developed for improving Information Retrieval, Natural Language Generation, and Machine Translation systems. Word-formation families (see Figure 2.17) were based on the morphological segmentation obtained from several morphological segmenters and the English part of CELEX. Some relations were also included from ADJADV (NomBank). CatVar can be queried online.[7]

```
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
```

Figure 2.17: Hash-sign-separated textual file format of CatVar. Each line contains a word-formation family consisting of: lexemes, their part-of-speech categories (preceded by underscores), and IDs of the original language resources of the lexemes (preceded by per cent signs).

[7] https://clipdemos.umiacs.umd.edu/catvar/

Framorpho-FR is a semi-automatically developed word-formation resource for French (Hathout, 2005). It includes nouns, adjectives, verbs, and adverbs extracted from a dictionary containing words from the 19th and 20th century. Word-formation families (see Figure 2.18) originate from a manual revision of automatic morphological segmentation.

```
<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>
```

Figure 2.18: XML file format of Framorpho-FR.

DerivBase.Hr is an automatically created word-formation lexicon for Croatian (Šnajder, 2014) inspired by DErivBase and DErivCELEX for German. DerivBase.Hr includes nouns, adjectives, and verbs taken from a large Croatian web corpus. The resource is distributed in a data package that contains two variants of DerivBase.Hr created by: (a) an unsupervised clustering based on string distance, and (b) a knowledge-based approach using an inflectional lexicon and a set of word-formation rules. The authors recommend the knowledge-based version because of its higher quality, see Figure 2.19.

```
bojovnik_N bojić_N bojev_A bojo_N bojovan_A bojati_V bojište_N bojenje_N bojen_A
     bojanić_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
```

Figure 2.19: Space-separated textual file format of DerivBase.Hr. Each line contains a word-formation family consisting of: lexemes with their part-of-speech categories (preceded by underscores).

DErivCELEX automatically connects derivationally related German nouns, adjectives, verbs, and adverbs into word-formation families (Shafaei et al., 2017). The lexemes are taken from the German part of CELEX that contains manually morphologically segmented lexemes. Since the lexemes came from CELEX, their written forms do not concur with the current orthographic standards, as noticed by Steiner (2016). Based on the morphological structure of lexemes, Shafaei et al. (2017) automatically created whole word-formation families, see Figure 2.20.[8]

```
10 unabänderlich_A unveränderlich_A veränderbar_A abändern_V Veränderlichkeit_N
    Änderung_N umändern_V änderbar_A abänderlich_A ändern_V veränderlich_A Abä
    nderung_N verändern_V Unveränderlichkeit_N Umänderung_N Veränderung_N
```

Figure 2.20: Space-separated textual file format of DErivCELEX. Each line contains: a family ID, and a whole word-formation family, i.e. part-of-speech tagged lexemes.

[8]The proposed procedure could also be replicated for German and English parts of CELEX, but it has not been done so far.

2.2 Dictionaries containing word-formation

2.2.1 Wiktionary-originated resources

Wiktionary.org project[9] is a multilingual free content dictionary of many natural languages. Several language variants of Wiktionary exist. The entries in Wiktionary are created by humans and bots that automatically generate entries or import them from previously published dictionaries. Among annotations of etymology, pronunciation, inflictive forms, and semantic definitions of lexemes, the entries sometimes provide information on word-formation, too. Wiktionary, as well as Wikipedia, has served as a base for various language resources and Nature Language Processing systems. In this section, resources that are rooted in Wiktionary and contain word-formation relevant information.

WiktiWF is an ongoing project[10] of the author of the thesis. The goal of the project is to extract word-formation relations from as many language versions of Wiktionary as possible and provide them in a unified data structure and file format, see Figure 2.21. Although one language version of Wiktionary contains lexemes for more than one language, WiktiWF focuses on the main language of a given language version. Word-formation of five languages (English, French, Czech, Polish, German) has been processed and published. The WiktiWF framework is prepared to extract word-formation of another 20 languages.

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

Figure 2.21: Tab-separated textual file format of WiktiWF (example from English data). Each line contains two columns containing: a base lexeme and its derivative. Some lexemes are also part-of-speech tagged (if not, then marked _X).

Etymological WordNet was constructed using the data extracted from the English language version of Wiktionary (Gerard, 2014). Although it is named WordNet, its aim is different from WordNets (Miller, 1998). While WordNets focus on lexical-semantic relations between lexemes, the Etymological WordNet connects lexemes of multiple languages based on their etymology. Besides information about etymology, Etymological WordNet also provides other linguistic annotations, including word-formation, see Figure 2.22. It captures derivationally related lexemes for almost 180 languages (many languages have only a few relations between lexemes). The resource can be queried online.[11]

[9] https://www.wiktionary.org/
[10] https://github.com/lukyjanek/wiktionary-wf
[11] http://www.lexvo.com/

```
caramelise      rel:is_derived_from     caramel
caramelised     rel:is_derived_from     caramelise
caramelises     rel:is_derived_from     caramelise
caramelising    rel:is_derived_from     caramelise
caramelize      rel:is_derived_from     caramel
```

Figure 2.22: Tab-separated textual file format of Etymological WordNet (example from English data). Each line contains three columns containing: two lexemes and their relation (derivational relations here).

2.2.2 Morphological dictionaries

Sometimes word-formation relations are captured in various morphological dictionaries instead of separate specialised word-formation resources. These lexicons are presented here.

E-Lex also known as TST-lexicon (Department of Language and Speech at Radboud University Nijmegen and ELIS and University of Ghent and CGN Consortium, 2008), is a lexical database of Dutch. It was developed as an annotation part of large Dutch corpus. E-Lex provides linguistic information for each lexeme, e.g. word-forms, lemma, pronunciation, orthography, morphological categories, spelling variants, morphological segmentation, semantic taxonomy and definitions, etc. The morphological segmentation is bracketed in the same way as in CELEX, so particular morphemes are organised into trees, see Figure 2.23.

```
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>]\\
```

Figure 2.23: Slash-separated textual file format of E-Lex. It is similar as for CELEX: lexemes (2nd position), morphological segmentation and part-of-speech categories (3rd).

E-dictionary is a morphological lexicon of Serbian (Vitas & Krstev, 2005). Although its early versions did not contain any word-formation annotation, (regular) derivational relations among nouns, adjectives, verbs, and adverbs were added in later versions. It also puts semantic labels on possessives, diminutives, augmentatives, female counterparts of profession names, and relational adjectives. It is distributed in several different versions with and (more often) without word-formation annotation.

Sloleks is a large Slovene morphological lexicon (Dobrovoljc et al., 2019), which contains derivational relations among nouns, adjectives, verbs, adverbs and lexemes of some other part-of-speech categories, see Figure 2.24. Sloleks can be queried online.[12]

[12]http://eng.slovenscina.eu/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>
```

Figure 2.24: XML file format of Sloleks. An abbreviated record of one lexeme (between tags *Lemma*) and its derivatives (between tags *RelatedForm*) is presented.

2.2.3 WordNets

WordNets are lexical databases grouping lexemes into sets of cognitive synonyms, so-called *synsets*, containing definitions of meanings of the lexemes. The synsets are connected by various lexical-semantic relations, e.g. hypernymy, hyponymy, meronymy, etc., and the relations also include word-formation (usually called *morpho-semantic relations*) in some WordNet language versions. WordNet databases capturing word-formation are presented here.

The Morpho-Semantic Database is a database (Fellbaum et al., 2007) automatically extracted from English (Princeton) WordNet version 3.0 (Miller, 1998). The M-S Database focuses on derivationally related nouns and verbs (see Figure 2.25), and relations between them are assigned 14 semantic labels.

```
survive%2:42:00::  202616713  state        survival%1:26:00::   113962166  [...]
rule%2:36:00::     201690020  instrument   ruler%1:06:00::      104118776  [...]
infer%2:32:00::    200944924  event        inference%1:09:00::  105774614  [...]
refer%2:32:12::    200877083  undergoer    reference%1:10:04::  106417598  [...]
```

Figure 2.25: Microsoft Excel file format of The Morpho-Sem. Database. Each line contains: base lexemes and their WordNet IDs, semantic labels, derivatives and their WordNet IDs, and definitions of both lexemes (not displayed). Part-of-speech categories are encoded in the first number preceded by the per cent sign (1 for nouns, 2 for verbs).

BulNet is the Bulgarian WordNet (Koeva et al., 2004), and it distinguishes morpho-semantic and derivational relations. While the derivational relations represent relations extracted from English WordNet, the morpho-semantic relations capture word-formation (Koeva, 2008, p. 365). BulNet can be queried online.[13]

[13] http://dcl.bas.bg/bulnet/

CroWordNet is the Croatian WordNet (Raffaelli et al., 2008). Its word-formation annotation came from the first versions of CroDeriV (Oliver et al., 2015; Šojat & Srebačić, 2014). Several versions of CroWordNet have been already published, however, without derivational relations.

Czech WordNet is a WordNet database for Czech (Pala & Smrž, 2004). It includes derivationally related nouns, adjectives, verbs, and adverbs obtained on the basis of ten word-formation rules and automatic generation of derivatives by attaching affixes with specific meanings (Pala & Hlaváčková, 2007). The resulting relations are assigned 16 semantic labels.

EstWordNet is the Estonian WordNet (Kahusk et al., 2010; Kerner et al., 2010). It connects derivationally related nouns, adjectives, verbs, and adverbs, see Figure 2.26. EstWordNet can be queried online.[14]

```
<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>
```

Figure 2.26: XML file format of EstWordNet.

FinnWordNet is the Finnish WordNet (Lindén & Carlson, 2010; Lindén et al., 2012). It includes derivationally related nouns, adjectives, and verbs, see Figure 2.27. FinnWordNet can be queried online.[15]

```
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
```

Figure 2.27: Tab-separated file format of FinnWordNet. Each line contains: unique IDs of derivatives, the derivatives, unique IDs of base lexemes, the base lexemes, marks specifying relations (plus for the derivational ones).

GermaNet is the German WordNet (Hamp & Feldweg, 1997). It captures not only derivational relations but also many compound lexemes. Lexemes are morphologically segmented into hierarchical segmentation (Henrich & Hinrichs, 2011), as it is done in CELEX.

OpenWordNet-PT is a WordNet for Brazilian Portuguese, and it contains word-formation annotation extracted from NomLex-PT (Paiva et al., 2012; Rademaker et al., 2014).

[14] https://teksaurus.keeleressursid.ee/

[15] https://sanat.csc.fi/wiki/Toiminnot:WordNet

PlWordNet is the Polish WordNet (Piasecki et al., 2009). It captures word-formation of Polish nouns, adjectives, and verbs, see Figure 2.28. The relations are assigned 11 semantic labels (Maziarz et al., 2011). PlWordNet can be queried online.[16]

```
1 <lexical-unit id="40116" name="robić" pos="czasownik" tagcount="0" domain="cwyt"
      desc="coć konkretnego, wytwarzać to, np. robić rzećbę. Jest to czasownik
     teliczny &lt;##VLC: DZn>"workstate="Nieprzetworzony" source="użytkownika"
     variant="2"/>
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"/>
3 [...]
4 <lexicalrelations parent="40116" child="77915" relation="111" valid="true" owner
     ="Agnieszka.Dziob"/>
```

Figure 2.28: XML file format of PlWordNet.

RoWordNet is the Romanian WordNet (Mititelu, 2012; Tufis et al., 2006). It contains word-formation relations between nouns, adjectives, verbs, and adverbs.

SrpWordNet is the Serbian WordNet (Krstev et al., 2004). It includes semantically labelled word-formation relations among nouns, adjectives, and verbs.

2.3 Corpora containing word-formation

Prague Dependency Treebank is a large morphologically and syntactically annotated treebank of Czech (also simply abbreviated as PDT; Hajič et al., 2018). Its annotation style is rooted in Functional Generative Description (cf. Sgall, 1967; Sgall et al., 1986). In the data, sentences are linguistically annotated on morphological, surface-syntactic (analytical), and tectogrammatical layers. While the first one contains lemmatised and morphologically annotated lexemes, the analytical layer analyses surface-syntactic structure, and the tectogramatical layer reflects the underlying (deep) structure of a given sentence. The morphological and tectogrammatical layers also include word-formation annotations capturing derivation of pronominal adjectives, pronouns, numerals, adverbs, and deadjectival adverbs and possessive lexemes (Razímová Ševčíková & Žabokrtský, 2006). The file format of PDT uses the Prague Markup Language, which is an XML-based format for linguistic annotations.

Russian National Corpus is a collection of diachronic Russian texts (Zakharov, 2013). It covers the period primarily from the middle of the 18th to the early 21st century. Neither morphological segmentation nor word-formation relations between lexemes are included in the corpus. However, some lexemes in the corpus are assigned 35 semantic labels, e.g. diminutive, augmentative, nominal agent, verbal nouns, etc.

[16] http://plwordnet.pwr.wroc.pl/wordnet/

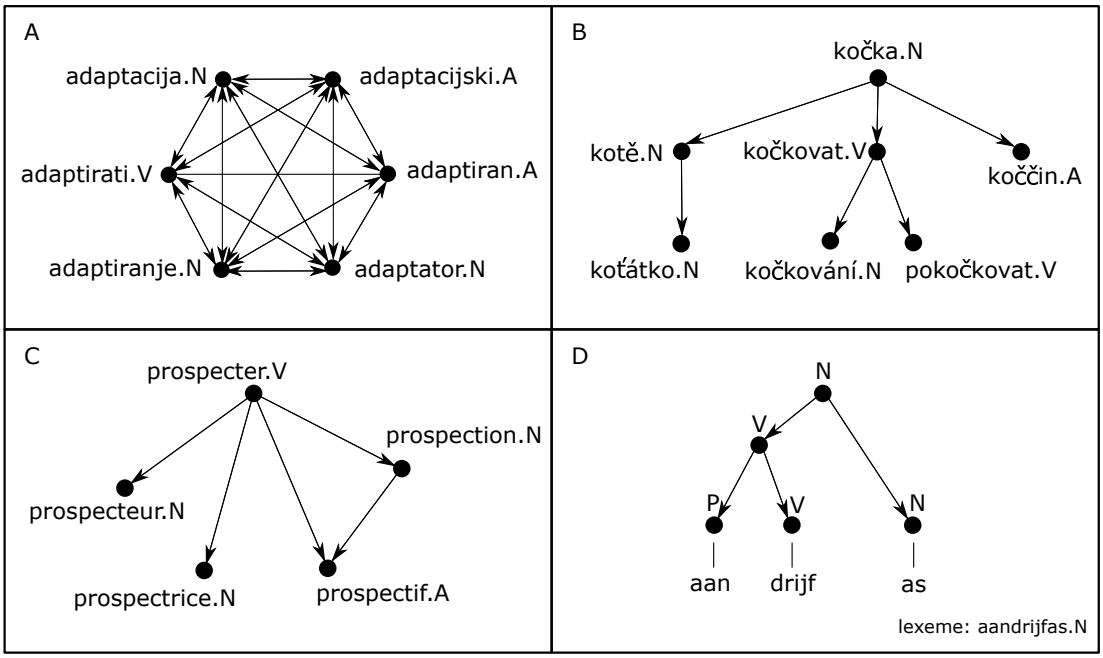


Figure 2.29: Observed data structures in reviewed language resources.

2.4 Observations and summarisations

The word-formation resources differ in many aspects regarding not only theoretical backgrounds and practical realisations but also technical details. As was already presented in this chapter, the resources differ in their purpose, scope, process of creation, distribution, accessibility and availability, etc. Table 2.1 provides basic statistics to illustrate the difference in sizes between individual resources.

From the harmonisation point of view, the data structure used for storing the data is the crucial aspect. Hereafter, in this thesis, the Graph theory terminology is used in order to describe data structures of the reviewed word-formation resources in a unified manner. Graph theory, cf. Matoušek and Nešetřil (2009), is the study of *graphs*, which are mathematical structures used for modelling relations between objects. A graph consists of *nodes* (also *vertices*) connected by *directed* or *undirect edges*. Processing word-formation families as (sub)graphs allows using already existing graph algorithms during the harmonisation process. From the graph theory perspective, four data structures can be observed in the data, see Figure 2.29.[17] Based on the following description, Table 2.2 specifies the data structure used in each resource presented in this chapter.

A. Some resources list only derivationally related lexemes (nodes) from derivational families. Individual derivational relations (edges) between lexemes are unspecified. Complete subgraphs could represent such families; however, because of the modelling of linguistic derivation, it would be rather *complete directed subgraphs* (cf. DerivBase.hr for Croatian; A in Figure 2.29). Although approaching edges as directed might seem redundant, it allows applying graph algorithms during the harmonisation procedure.

[17]The data structures have already been presented by Kyjánek (2018, pp. 4–5) and Kyjánek et al. (2019a, p. 102). The descriptions are summarised and specified here.

B. Resources allowing at most one base lexeme for each derivative represent derivational families as *rooted trees* (cf. DeriNet for Czech; B in Figure 2.29). The tree root represents the simplest (unmotivated) lexeme in terms of morphological complexity (and it has the broadest meaning), while leaf nodes contain the most morphologically complex lexemes (with the narrowest meaning) in a particular derivational family. The rooted tree data structure cannot capture relations of compounding because of the one-base-lexeme constraint.

C. If the derivative can have more than one base lexeme, then the data structure capturing derivational relations within lexemes in derivational family corresponds to a *weakly connected subgraph* (cf. Démonette for French; C in Figure 2.29). Since the base lexeme for the derivative is not always clear, capturing more than one base lexeme for the derivative is acceptable from the linguistic point of view, especially when compounding is captured.

D. Some resources focus on morphological segmentation of lexemes rather than on grouping lexemes into derivational families. On the one hand, a basic listing individual morphemes of a given lexeme is a way to represent morphological segmentation (cf. DerIvaTario for Italian; data in Figure 2.2). On the other hand, a hierarchical arrangement of morphemes also occurred in the reviewed resources (cf. Dutch part of CELEX; D in Figure 2.29). The hierarchical segmentation resembles *derivation tree* data structure (in the terminology of Context-Free Grammars, cf. Hopcroft et al., 2000, pp. 169–216) in which particular morphemes are placed in leaf nodes of a tree, and non-terminal nodes represent a combination of individual morphemes. Capturing compound lexemes is not a problem when using the derivation tree data structure. In addition, if the root morphemes are labelled, then word-formation relations between composed lexemes can also be considered.

Table 2.1: Basic quantitative properties of the original word-formation resources. The column *Lang* represents language of the particular resource, *Resource* specifies name and version, *Lex* for the number of lexemes, *Rel* counts edges between lexemes, *NFam* sums up families having more than one lexeme, *SFam* includes the number of families consisting of only one lexeme, *Part-of-speech* presents percent distribution of nouns (N), adjectives (A), verbs (V), adverbs (D), and other (O) part-of-speech categories. The last column is filled by zeroes or the number of O category is high, if the resource is only partly tagged or not tagged at all. Only lexemes relevant for word-formation are extracted from resources that are not specialised in word-formation. Relations in resources capturing word-formation in form of morpheme segmentation are not counted. Only the languages with at least one thousand derivational relations captured in Etymological WordNet are extracted from the data and presented.

Lang	Resource	Lex	Rel	NFam	SFam	Part-of-speech N/A/V/D/O
Armenian	EtymWordNet-xcl 2013	27,526	32,519	406	0	0/0/0/0/0
Asturian	EtymWordNet-ast 2013	3,132	2,547	585	0	0/0/0/0/0
Bulgarian	EtymWordNet-bul 2013	1,856	1,045	843	0	0/0/0/0/0
Catalan	EtymWordNet-cat 2013	7,496	4,613	2,918	1	0/0/0/0/0
Croatian	DerivBase.Hr 1.0	99,606	3,056,962	14,818	40,733	59/30/12/0/0
Czech	Cs-WiktiWF 1.0	50,526	57,902	8,387	0	27/9/5/1/57
Czech	DeriNet 2.0	1,027,665	809,882	122,175	96,208	44/35/5/16/0
Czech	EtymWordNet-ces 2013	7,633	5,331	2,354	0	0/0/0/0/0
Danish	EtymWordNet-dan 2013	22,957	20,368	2,987	3	0/0/0/0/0
Dutch	D-CELEX 2.0	121,787	0	5,672	35,429	64/8/8/1/19
Dutch	E-Lex 1.1.1	97,054	0	13,112	0	80/10/10/0/0
Dutch	EtymWordNet-nld 2013	40,446	37,485	3,508	0	0/0/0/0/0
English	ADJADV 1.0	5,005	2,581	2,424	0	0/51/0/48/0
English	CatVar 2.1	82,675	155,064	13,368	38,604	60/24/11/5/0
English	E-CELEX 2.0	43,649	0	10,535	3,164	56/18/16/9/1
English	En-WiktiWF 1.0	23,044	20,319	2,908	0	54/32/5/3/6
English	EtymWordNet-eng 2013	263,239	170,927	93,184	22	0/0/0/0/0
English	MorphoLex-en 1.0	40,899	0	234,765	150,093	52/12/35/1/0
English	NOMADV 1.0	318	161	158	0	50/0/0/50/0
English	NOMLEX 2001	1,964	1,025	941	0	52/0/48/0/0
English	NOMLEXPlus 1.0	7,756	4,450	3,298	5	57/6/37/0/0
English	The M-S Database 1.0	13,813	17,739	5,818	0	57/0/43/0/0
English (old)	EtymWordNet-ang 2013	2,291	1,830	479	0	0/0/0/0/0
Esperanto	EtymWordNet-epo 2013	103,970	95,002	9,124	0	0/0/0/0/0
Estonian	EstWordNet 2.1	989	544	457	0	16/29/8/47/0
Finnish	EtymWordNet-fin 2013	73,052	58,311	16,260	30	0/0/0/0/0
Finnish	FinnWordNet 2.0	20,035	42,136	6,347	2	55/29/15/0/0
French	Démonette 1.2	22,620	96,027	7,542	0	64/2/33/0/0
French	EtymWordNet-fra 2013	257,196	231,137	26,923	128	0/0/0/0/0
French	Famorpho-FR 1.0	635	4,456	119	54	63/24/10/3/0
French	Fr-WiktiWF 1.0	136,574	121,101	28,978	0	41/28/6/1/24
French	MorphoLex-fr 1.0	15,954	0	48,415	71,088	0/0/0/0/0
French	Morphonette 0.1	29,310	96,107	8,607	0	58/25/14/4/0
French	Nomage 1.0	1,298	667	656	11	51/0/49/0/0
French	VerbAction 1.0	15,885	9,393	6,513	0	58/0/42/0/0
Gaelic	EtymWordNet-gla 2013	7,524	5,091	2,469	0	0/0/0/0/0
Galician	EtymWordNet-glg 2013	17,119	16,552	1,537	8	0/0/0/0/0
Georgian	EtymWordNet-kat 2013	3,866	3,515	359	0	0/0/0/0/0
German	DErivBase 2.0	281,387	57,689	19,796	214,916	85/10/5/0/0
German	DErivCELEX 2.0	46,644	378,530	5,422	20,774	58/19/19/0/3

Table 2.1 – continued from the previous page

Lang	Resource	Lex	Rel	NFam	SFam	Part-of-speech N/A/V/D/O
German	De-WiktiWF 1.0	140,896	132,637	14,605	0	33/5/5/0/58
German	EtymWordNet-deu 2013	71,190	57,571	13,763	2	0/0/0/0/0
German	G-CELEX 2.0	51,338	0	6,138	4,263	53/18/18/2/9
Greek (anc.)	EtymWordNet-grc 2013	3,151	2,154	1,091	0	0/0/0/0/0
Greek (mod.)	EtymWordNet-ell 2013	1,872	1,352	522	0	0/0/0/0/0
Hungarian	EtymWordNet-hun 2013	26,010	21,873	4,339	0	0/0/0/0/0
Icelandic	EtymWordNet-isl 2013	8,245	7,202	1,114	0	0/0/0/0/0
Ido	EtymWordNet-ido 2013	3,611	2,171	1,451	0	0/0/0/0/0
Irish	EtymWordNet-gle 2013	6,053	4,372	1,780	1	0/0/0/0/0
Italian	DerIvaTario 1.0	11,147	0	4,872	1,348	51/26/13/10/0
Italian	EtymWordNet-ita 2013	422,322	383,800	45,760	1	0/0/0/0/0
Japanese	EtymWordNet-jpn 2013	7,999	7,391	1,055	5	0/0/0/0/0
Korean	EtymWordNet-kor 2013	385	270	121	0	0/0/0/0/0
Latin	EtymWordNet-lat 2013	629,181	605,763	24,504	4	0/0/0/0/0
Latin	WFL 2019	36,097	34,737	2,811	0	46/29/22/0/3
Latvian	EtymWordNet-lav 2013	1,561	1,263	358	0	0/0/0/0/0
Lithuanian	EtymWordNet-lit 2013	2,063	1,737	354	0	0/0/0/0/0
Mandarin	EtymWordNet-cmn 2013	3,371	2,357	1,125	0	0/0/0/0/0
Manx	EtymWordNet-glv 2013	2,060	1,343	751	0	0/0/0/0/0
Norwegian	EtymWordNet-nob 2013	1,748	1,440	314	1	0/0/0/0/0
Persian	DeriNet.FA 0.5	43,357	35,745	7,612	0	0/0/0/0/0
Polish	EtymWordNet-pol 2013	27,797	24,985	2,881	0	0/0/0/0/0
Polish	Pl-WiktiWF 1.0	106,699	249,584	18,089	0	36/11/5/1/46
Polish	PlWordNet 4.0	112,898	140,686	23,745	0	52/24/17/6/0
Polish	The Polish WFN 0.5	262,887	189,217	32,337	41,333	0/0/0/0/0
Portuguese	EtymWordNet-por 2013	2,797	1,627	1,175	6	0/0/0/0/0
Portuguese	NomLex-PT 2016	7,024	4,238	2,787	0	60/0/40/0/0
Romanian	EtymWordNet-ron 2013	4,056	2,703	1,396	2	0/0/0/0/0
Russian	DerivBase.Ru 1.0	265,358	289,893	17,946	114,762	62/18/17/3/0
Russian	EtymWordNet-rus 2013	4,005	3,400	750	1	0/0/0/0/0
Serbo-Croat.	EtymWordNet-hbs 2013	8,033	6,349	1,714	0	0/0/0/0/0
Slovene	Sloleks 1.2	97,242	65,984	19,889	956	52/27/10/7/3
Spanish	DeriNet.ES 0.5	151,173	36,935	15,912	98,326	0/0/0/0/0
Spanish	EtymWordNet-spa 2013	232,041	219,161	13,925	8	0/0/0/0/0
Spanish	The Spanish WFN 0.5	162,751	18,441	11,322	132,988	0/0/0/0/0
Swedish	EtymWordNet-swe 2013	7,333	4,451	2,885	0	0/0/0/0/0
Telugu	EtymWordNet-tel 2013	1,512	1,038	474	0	0/0/0/0/0
Turkish	EtymWordNet-tur 2013	7,774	5,956	1,921	0	0/0/0/0/0
Venetian	EtymWordNet-vec 2013	3,268	1,936	1,334	0	0/0/0/0/0
Volapük	EtymWordNet-vol 2013	6,585	6,666	337	1	0/0/0/0/0

Table 2.2: Licenses and data structures of all presented word-formation resources. The column *Resource* specifies the name and version, *Structure* represent the data structure, and *License* specifies the original license.

Resource	Structure	License
ADJADV 1.0	weakly connected subgraphs	LDC User Agreement
BulNet 3.0	weakly connected subgraphs	ELRA License Agreement
CatVar 2.1	complete directed subgraphs	OSL-1.1
CELEX 2.0	derivation trees	CELEX Agreement
CroDeriV 2.0	rooted trees	unspecified
CroWordNet 1.0	weakly connected subgraphs	ELRA License Agreement
Czech WordNet 1.0	weakly connected subgraphs	ELRA License Agreement
DeriNet 2.0	rooted trees	CC BY-NC-SA 3.0
DeriNet.ES 0.5	rooted trees	CC BY-NC-SA 3.0
DeriNet.FA 0.5	rooted trees	CC BY-NC-SA 4.0
DErivBase 2.0	weakly connected subgraphs	CC BY-SA 3.0
DerivBase.Hr 1.0	complete directed subgraphs	CC BY-SA 3.0
DerivBase.Ru 1.0	weakly connected subgraphs	Apache 2.0
DerIvaTario 1.0	listed segmentation	CC BY
DErivCELEX 2.0	complete directed subgraphs	CC BY-SA 3.1
Démonette 1.2	weakly connected subgraphs	CC BY-SA-NC 3.0
E-Dictionary 1.1.1	derivation trees	unspecified
E-Lex 1.1.1	derivation trees	E-Lex Agreement
EstWordNet 2.1	weakly connected subgraphs	CC BY-SA
Etymological WordNet 2013	weakly connected subgraphs	CC BY-SA 3.0
Famorpho-FR 1.0	complete directed subgraphs	CC BY-SA-NC 2.0
FinnWordNet 2.0	weakly connected subgraphs	CC BY 3.0
GermaNet 13.0	derivation trees	GermaNet Agreement
MorphoLex-en 1.0	listed segmentation	CC BY 4.0
MorphoLex-fr 1.0	listed segmentation	CC By 4.0
Morphological Treebank 2019	derivation trees	CELEX+GermaNet Agr.
Morphonette 0.1	weakly connected subgraphs	CC BY-NC-SA 2.0
NOMADV 1.0	weakly connected subgraphs	LDC User Agreement
Nomage 1.0	weakly connected subgraphs	CC BY-SA 4.0
NOMLEX 2001	weakly connected subgraphs	unspecified
NOMLEXPlus 1.0	weakly connected subgraphs	LDC User Agreement
NomLex-PT 2016	weakly connected subgraphs	CC BY 4.0
OpenWordNet-PT 2019	weakly connected subgraphs	CC BY 4.0
PlWordNet 4.0	weakly connected subgraphs	plWordNet 3.0 License
Prague Dependency Treebank 3.5	rooted trees	CC BY-NC-SA 4.0
RoWordNet 3.6	weakly connected subgraphs	Meta-Share License
Russian National Corpus	annotated meaning	RNC Agreement
Sloleks 1.2	complete directed subgraphs	CC BY-NC-SA 4.0
SrpWordNet 3.0	weakly connected subgraphs	Meta-Share License
The Morpho-Semantic Database	weakly connected subgraphs	WordNet 3.0 license
The Polish WFN 0.5	rooted trees	plWordNet 3.0 License
The Spanish WFN 0.5	rooted trees	CC BY-ND
Unimorph	listed segmentation	restricted
VerbAction 1.0	weakly connected subgraphs	CC BY-NC-SA 2.0
WiktiWF 1.0	weakly connected subgraphs	CC BY-NC-SA 4.0
Word Formation Latin 2019	weakly connected subgraphs	CC BY-NC-SA 4.0

Chapter 3

Harmonisation of word-formation resources

This chapter describes the harmonisation process of language resources capturing word-formation of multiple languages. The proposed procedure, its parameters, and evaluation are the core of the effort, but a selection of a target data structure and a file format are equally important.[1]

As presented in the previous chapter, dozens of word-formation resources of multiple languages exist. They differ significantly in many aspects, which complicates processing the data in multilingual systems. The situation resembles the story of the development of syntactic treebanks (Kyjánek et al., 2019a). In the area of syntactic treebanks, efforts have been made to convert (harmonise) the existing treebanks to the same annotation styles, cf. CoNLL Shared Task 2006 (Buchholz & Marsi, 2006), the HamleDT treebank collection (Zeman et al., 2014), Google Universal Treebanks (McDonald et al., 2013), and Universal Dependencies project (Nivre et al., 2016; Zeman et al., 2019). Thanks to the availability of the treebanks in the same annotation styles, the multilingual systems for tokenisation, lemmatisation, morphological tagging, and dependency parsing have been developed or, at least, have been improved (cf. Manning et al., 2014; Straka and Straková, 2017). Notable progress has also been made in the field of creating new treebanks using knowledge transfer from well-resourced to under-resourced languages (cf. Agić et al., 2015; Hwa et al., 2005; Rosa, 2018; Rosa et al., 2017; Yarowsky et al., 2001; Zeman and Resnik, 2008).

Being inspired by the harmonisation of syntactic treebanks, harmonisation of several word-formation resources is presented here. As a result, a collection of harmonised word-formation resources is created. Similarly to the evolution of syntactic treebanks, the collection could open a discussion on annotating word-formation resources for different languages, and it could facilitate knowledge transfer experiments, research in word-formation, etc.

[1]The description of the target data structure, the file format, and the harmonisation procedure involved in this chapter has already been published (Kyjánek et al., 2019a; Vidra, Žabokrtský, Ševčíková, et al., 2019). In this chapter, they are described in more details, and the procedure is improved.

3.1 Resources selected for harmonisation

The following four selection criteria are considered while deciding which resources should be harmonised in this thesis.

- **Input data structure.** One of the goals is to show that all data structures observed in the existing word-formation resources can be harmonised into the target representation. It allows to apply proposed harmonisation procedure to other existing resources. It could also accelerate a further discussion of the suitability of the existing data structures and the target representations for the word-formation data.

- **Processed language.** The collection should cover as many different languages as possible to be utilisable for multilingual projects and cross-linguistic research, eventually. Harmonising a resource covering language not yet included in the collection is preferred rather than harmonisation of many resources for one language.

- **Purpose of the creation.** The previous chapter presents three types of existing word-formation resources in terms of their scope: resources specialised in word-formation, dictionaries containing word-formation as one of their parts, and corpora. Specialised resources are preferred over dictionaries and corpora.

- **Availability and licensing.** The last criterion focuses on replicability and evaluation of the harmonisation procedure, and on the utilisation of the collection. If a resource is easily available, the harmonisation can be replicated and evaluated by anyone. Moreover, a resulting harmonised resource can be compared with the original resource. It closely relates to the licensing of the original resources. Open licenses of the original resources are preferred for publishing the final collection of the harmonised resources.

For the harmonisation, 17 original resources covering word-formation of 20 languages were selected. In alphabetical order, namely: **CatVar** for English; **CELEX** for Dutch, English, and German; **DeriNet** for Czech; **DeriNet.ES** for Spanish; **DeriNet.FA** for Persian; **DerIvaTario** for Italian; **DErivBase** for German; **DerivBase.Hr** for Croatian; **DerivBase.Ru** for Russian; **Démonette** for French; **EstWordNet** for Estonian; **Etymological WordNet** for Czech, Catalan, Gaelic, Polish, Portuguese, Russian, Serbo-Croatian, Swedish, Turkish; **FinnWordNet** for Finnish; **NomLex-PT** for Portuguese; **The Morpho-Semantic Database** for English; **The Polish Word-Formation Network**; **Word Formation Latin**.

The set covers resources organising data in all data structures presented in the previous chapter. Some languages, e.g. English, are in the collection more than once. Their data is harmonised and stored separately; the harmonised resources are not merged even if they cover the same language. All resources mentioned above specialise in capturing word-formation, except for three WordNets. They all are distributed under the open licenses, except for CELEX. However, CELEX organises data in derivation trees, unlike other selected resources.

As to the content of the selected resources, only CELEXes, Word Formation Latin and partly DeriNet distinguish derivation and compounding explicitly. DeriNet, DErivBase, Démonette, and Word Formation Latin include relatively rich annotation of various features: part-of-speech and other morphological categories, labels for derivational processes, semantic labels, and morphological segmentation. DErivBase and DerivBase.Ru include labels for derivational processes and morphological segmentation in word-formation rules. Besides direct (derivational) relations, there are indirect (subparadigmatic) relations captured in Démonette. The Morpho-Semantic Database contains only nouns and verbs, and it annotates semantics. NomLex-PT captures only nominalisations. CELEX and DerIvaTario contain a detailed morphological segmentation. Except for DeriNet.ES, DeriNet.FA, Etymological WordNet, and The Polish Word-Formation Network, lexemes are assigned with part-of-speech categories. For a detailed overview of the resources, see the previous Chapter 2.

For clarification, CELEX is a collection of three separate datasets for Dutch (referred to as D-CELEX), English (E-CELEX), and German (G-CELEX), so they are presented as three harmonised resources in the final collection. The opposite situation concerns Etymological WordNet, which merges data for more than a hundred of languages in one dataset. The dataset is split and harmonised according to individual languages. Only the languages having at least one thousand relations are selected.[2] Harmonised resources resulting from Etymological WordNet are presented as EtymWordNet-x where x is a language abbreviation taken from Etymological WordNet, i.e. ISO 639-2 Code.[3]

3.2 Target data structure and file format

As presented in the previous chapter, individual word-formation resources are anchored in different approaches to data storage (hereafter also called *annotation schema*). The harmonisation of the annotation schemata has to start with the selection of a target data structure and a file format for the final harmonised resources. The selection balances two opposing aspects – expressiveness and uniformity (Kyjánek et al., 2019a, p. 104). Heavy pressure on expressiveness, flexibility and completeness leads to a preservation of all linguistic and technical features from the original resources. Forcing uniformity and generalisation too much can cause negligence of important features that are characteristic of the original resources, eventually of the particular languages. The harmonisation is a trade-off between the two aspects.

The resulting **target data structure** combines rooted tree and weakly connected subgraph data structures (see Figure 3.1).[4] Tree-shaped skeletons are identified for all derivational families in each harmonised resource, and non-tree

[2]The list of the language data selected from Etymological WordNet is not limited only by the chosen size threshold of derivational relations but also by the ability of the author to annotate word-formation of a particular language.

[3]Hereafter: *cat* for Catalan, *ces* for Czech, *gla* for Gaelic, *pol* for Polish, *por* for Portuguese, *rus* for Russian, *hbs* for Serbo-Croatian, *swe* for Swedish, *tur* for Turkish.

[4]This decision resembles decision made in Universal Dependencies collection which used trees in the beginning, although trees are not sufficient for modelling all syntactic relations. In the recent versions, a set of secondary non-tree edges was added; however, the tree-shaped skeletons remain.

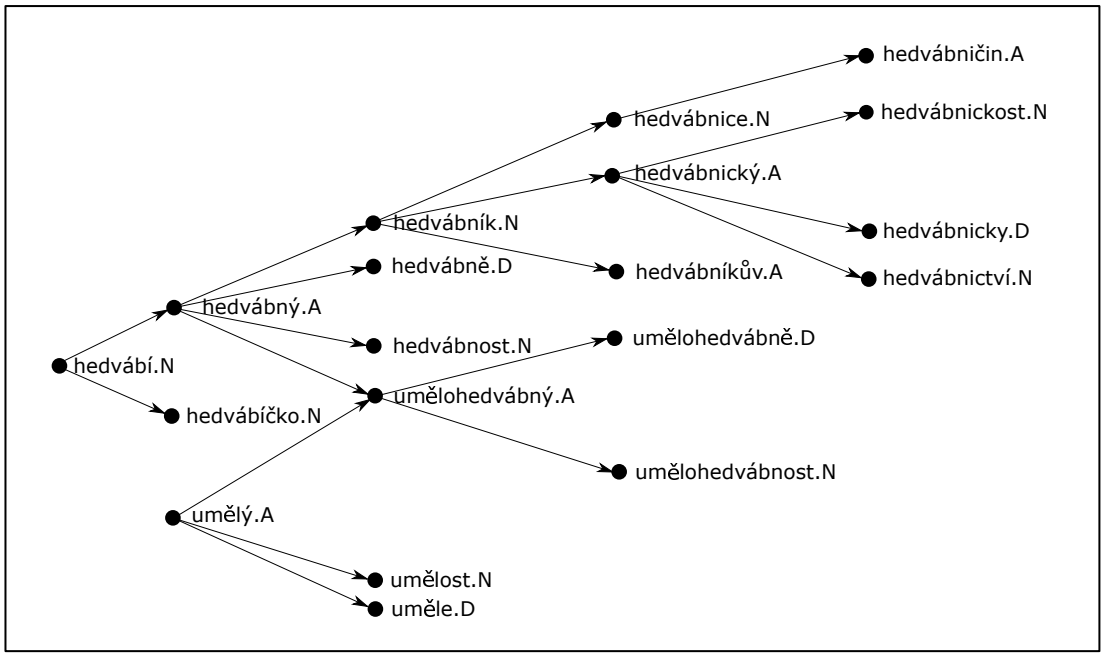


Figure 3.1: Target data structure represented by the word-formation family of the lexeme *hedvábí* (*silk*) and a part of the family of the lexeme *umělý* (*arficial*) from DeriNet 2.0.

edges represent many-to-one relations as compounding, or they store non-tree edges from the original resources (in a less prominent place). This data structure proposed by Vidra, Žabokrtský, Ševčíková, et al. (2019) is already used in DeriNet 2.0. The content of other existing language resources was considered during the creation of the data structure.

If compared to other less constrained graphs, the selected target data structure might seem limited by the tree-constraint. However, it is an advantage in terms of technical aspects, because it simplifies data traversing and visualisation. From the linguistic point of view, the data structure concurs with the description of derivation as a process of adding an affix to a base to form a new lexeme (Dokulil, 1962, pp. 11–14).

Regarding the **target file format**, a textual lexeme-based format consisting of tab-separated columns was developed together with the target data structure by Vidra, Žabokrtský, Ševčíková, et al. (2019) for DeriNet 2.0. The format is inspired by the CoNLL-U format (Nivre et al., 2016) used to organise Universal Dependencies treebanks and other syntactic annotations. Each line of the simple target format contains a lexeme annotated by key-value pairs specifying various features. The format aims at containing all relevant word-formation pieces of information/annotations.

In the target file format, lexemes are kept together with the other related lexemes belonging to the same derivational family; an empty line separates individual families. The format allows to save both annotations of lexemes and relations. Each annotated feature is represented as a key-value pair. Ampersands or vertical bars are used for concatenations of the pairs. While ampersands (`key1=value1&key2=value2`) concatenate pairs describing a single entity, vertical bars (`key1=value1&key2=value2|keyA=valueA`) concatenate pairs of multiple

```
1.0  hedvábí#NNN??-----A---?  hedvábí  NOUN  Gender=Neut  _  _  _  _  {"
    techlemma": "hedvábí"}
1.1  hedvábný#AA???----??---?  hedvábný  ADJ  _  _  1.0  Type=Derivation  _  {"
    techlemma": "hedvábný"}
1.2  hedvábně#Dg-------??---?  hedvábně  ADV  _  _  1.1  Type=Derivation  _  {"
    techlemma": "hedvábně_^(*1ý)"}
1.3  hedvábník#NNM??-----A---?  hedvábník  NOUN  Animacy=Anim&Gender=Masc  _
    1.1  Type=Derivation  _  {"techlemma": "hedvábník"}
1.4  hedvábnice#NNF??-----A---?  hedvábnice  NOUN  Gender=Fem  _  1.3
    SemanticLabel=Female&Type=Derivation  _  {"techlemma": "hedvábnice_^(*3ík)"}
1.5  hedvábničin#AU????--------?  hedvábničin  ADJ  Poss=Yes  _  1.4
    SemanticLabel=Possessive&Type=Derivation  _  {"techlemma": "hedvábničin_^(*3
    ce)"}
1.6  hedvábnický#AA???----??---?  hedvábnický  ADJ  _  _  1.3  Type=Derivation
    _  {"techlemma": "hedvábnický"}
1.7  hedvábnickost#NNF??-----?---?  hedvábnickost  NOUN  Gender=Fem  _  1.6
    Type=Derivation  _  {"techlemma": "hedvábnickost_^(*3ý)"}
1.8  hedvábnicky#Dg-------??---?  hedvábnicky  ADV  _  _  1.6  Type=Derivation
    _  {"techlemma": "hedvábnicky_^(*1ý)"}
1.9  hedvábnictví#NNN??-----A---?  hedvábnictví  NOUN  Gender=Neut  _  1.6  Type
    =Derivation  _  {"techlemma": "hedvábnictví"}
1.10  hedvábníkův#AU???M--------?  hedvábníkův  ADJ  Poss=Yes  _  1.3
    SemanticLabel=Possessive&Type=Derivation  _  {"techlemma": "hedvábníkův_
    ^(*2)"}
1.11  hedvábnost#NNF??-----?---?  hedvábnost  NOUN  Gender=Fem  _  1.1  Type=
    Derivation  _  {"techlemma": "hedvábnost_^(*3ý)"}
1.12  umělohedvábný#AA???----??---?  umělohedvábný  ADJ  _  _  1.1  Sources
    =3.258,1.1&Type=Compounding  _  {"is_compound": true, "techlemma": "umě
    lohedvábný"}
1.13  umělohedvábnost#NNF??-----?---?  umělohedvábnost NOUN  Gender=Fem  _  1.12
      Type=Derivation  _  {"techlemma": "umělohedvábnost_^(*3ý)"}
1.14  umělohedvábně#Dg-------??---?  umělohedvábně  ADV  _  _  1.12  Type=
    Derivation  _  {"techlemma": "umělohedvábně_^(*1ý)"}
1.15  hedvábíčko#NNN??-----A---?  hedvábíčko  NOUN  Gender=Neut  _  _  1.0
    SemanticLabel=Diminutive&Type=Derivation  _  {"techlemma": "hedvábíčko"}
...
3.258  umělý#AA???----??---?  umělý  ADJ  _  End=2&Morph=um&Start=0&Type=Root
    3.4  Type=Derivation  _  {"segmentation": "(um)ělý", "techlemma": "umělý"}
3.259  uměle#Dg-------??---?  uměle  ADV  _  End=2&Morph=um&Start=0&Type=Root
    3.258  Type=Derivation  _  {"segmentation": "(um)ěle", "techlemma": "uměle_
    ^(*1ý)"}
3.340  umělost#NNF??-----?---?  umělost  NOUN  Gender=Fem  End=2&Morph=um&Start
    =0&Type=Root  3.258  Type=Derivation  _  {"segmentation": "(um)ělost", "
    techlemma": "umělost_^(*3ý)"}
```

Figure 3.2: Target file format which illustrates the word-formation family of the lexeme *hedvábí* (*silk*) and a part of the family of the lexeme *umělý* (*arficial*) from DeriNet 2.0. If empty, columns are filled with underscores for illustrative purposes.

different entities (Vidra, Žabokrtský, Ševčíková, et al., 2019, p. 87). During the harmonisation process, one of the essential tasks is to find uniformity of key-value pairs across the harmonised resources (without affecting the original meaning of the key-value pairs from the original resources; cf. Zeman, 2010), e.g. applying the same part-of-speech tags. The target file format comprises ten columns separated by tabulators as presented in Figure 3.2. An application programming interface (API) for developing and managing the data in the target format is available on GitHub.[5]

1. An internal ID consisting of the word-formation family number and the lexeme number separated by a dot. The ID changes across released versions of datasets as it depends on relations captured in the datasets.

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

2. A language-dependent unique identifier for each lexeme (LEMID) involved in the data.[6]

3. The written form of the lexeme.

4. A tag representing the part-of-speech category.

5. Morphological features describing the lexeme using relevant linguistic categories (e.g. gender, animation, verbal aspect, etc.) The set of included morphological features can be customised.

6. Outcome of (surface) morphological segmentation which splits the written form of the lexeme into morphemes. Each morpheme is described by the first and the last position (counted from zero) and the type (e.g. root, prefix, suffix, etc.), see lines 18, 19, 20 in the Figure 3.2.

7. Internal IDs referring to the base lexeme. If the relation type is compounding this column contains the relation to the "main base lexeme" and the following column (8) lists all relations.

8. Annotation of the relation referenced to by the internal ID (column 7). The relations can be annotated by various features (e.g. the type of the word-formation process, semantic labels, etc.). In the case of compounding, this column lists all base lexemes of the resulting compound lexeme.

9. A column reserved for other potential relations.

10. A JSON-encoded data (Bray, 2017) providing potentially unlimited space for various custom annotations and extensions in the form of key-value pairs.

3.3 Fundamental decisions

Harmonisation of individual resources aims at unifying annotation schemata, i.e. data structure, file format, and feature-value pairs. After the harmonisation process presented in this thesis, data of all harmonised resources should be organised in the same data structure and stored in the same file format. However, the data, i.e. lexeme sets and word-formation relations, can be affected during the harmonisation, too. Before the harmonisation, the fundamental decisions have to be made to specify the extent to which the original data will be affected by the harmonisation process proposed in the thesis.

3.3.1 Lexeme sets

The individual lexeme sets vary greatly from resource to resource. While some resources as DeriNet or DerivBase contain more lexemes than a common native speaker vocabulary is, NomLex-PT is limited to nominalisations only. The small lexeme sets limits usefulness in the case of further use in multilingual systems and data-based oriented word-formation research. Enlarging the sets would be

[6] In DeriNet 2.0, it consists of the written form of the lexeme and its morphological categories.

a solution; however, it would have to involve the identification of new word-formation relations. This full-fledged development of the original resources is not possible to manage for all resources individually during the harmonisation.

The insight into the individual lexeme sets reveals that different approaches to tokenisation and lemmatisation are used across the resources. It is evident, especially in the following phenomena:

Inflexion & Derivation While most of the resources try to separate derivation from inflexion (inflected forms of lexemes are not captured in the data), for example, DeriNet.FA and Etymological WordNet do not distinguish derivation and inflexion at all. Even if resources distinguish inflexion and derivation, the boundary between them is not explicitly specified, and it varies across the resources. For instance, DeriNet does not contain negation and reflexives, but DerivBase.Ru does. As Štekauer et al. (2012, pp. 14, 19–35) documented, the boundary is not clear-cut even from the linguistic perspective.

Spelling variants Many resources contain spelling variants, but none of the resources explicitly marks them. For example, in NomLex-PT and DErivBase, spelling variants are treated as any other lexemes, e.g. noun '*comunhão*' ('*communion*') is derived from verbs '*comunhar*' and '*comungar*', which are both spelling variants of the same lexeme '*to commune*'. In DeriNet, on the other hand, spelling variants are processed inconsistently. For example, the spelling variants '*čistění*' and '*čištění*' (both '*cleaning*') are derived from the same verb '*čistit*' ('*to clean*'), and they both motivates different lexemes; however, '*brambora*' is derived from '*brambor*' in DeriNet, despite they both are spelling variants ('*potato*'), too.

Multi-word lexemes In most of the resources selected for harmonisation, multi-word lexemes do not occur, except for FinnWordNet, The Morpho-Semantic Database, and DerivBase.Ru. For instance, while The Morpho-Semantic Database uses multi-word lexemes for phrasal verbs, FinnWordNet suffers from incorrect tokenisation because it uses multi-word lexemes for whole expressions as '*alkion rakkulavaiheen keskusontelon aukkoon liittyvä*' ('*associated with an opening in the central cavity of the embryonic vesicle*').

Named entities Named entities occur in most of the resources. DerivBase.Ru contains multi-word lexemes to capture named entities, in contrast with DeriNet and Word Formation Latin, which contain only those name entities that are expressed in one-word lexemes.

Although reducing lexeme sets would help to unify the phenomena mentioned above, none of those issues is explicitly labelled in the original data, and their identification would be complicated in one resource, let alone all harmonised resources. Moreover, forcing some arbitrary boundaries, e.g. for inflexion and derivation, could damage the data.

The main decision concerning the lexeme sets and arising from the above-presented information is not to affect the original lexeme sets.

3.3.2 Word-formation relations

Word-formation relations captured in the individual resource are affected by a lexeme set, the technical features of the resource, and the linguistic tradition in a particular language. If a lexeme set contains compounds, then the lexemes are very often connected to at least one of their base lexemes, however, except for CELEX, DeriNet and Word Formation Latin, none of the selected resources explicitly labels those relations as compounding.

> *Since compound lexemes cannot be identified easily, they remain intact (except for CELEX, DeriNet and Word Formation Latin).*

Relatively regular word-formation relations with similar meaning, e.g. negation, reflexivity, and gradation, are captured differently in the selected resources. For instance, it is often possible to capture affirmatives and negatives in two separate parallel subgraphs, e.g. '*impolitely*' would be derived from '*impolite*', and '*politely*' from '*polite*'. However, some resources prefer to derive negatives directly from the corresponding affirmatives, e.g. '*impolitely*' would be derived from '*politely*', and '*impolite*' from '*polite*'. Figure 3.4 (the third example) in Section 3.4.2 illustrates both approaches.

> *To avoid damaging the original data, no new relations are added but, if it is possible, the unification of regular word-formation relations is done, e.g. in the case of capturing negation, and described during the harmonisation procedure.*

Finally, the target rooted tree data structure cannot capture all relations included in most of the original resources, especially those that store data in the complete directed or weakly connected subgraphs, as presented in Section 2.4.

> *Although the tree-shaped skeletons will be based on just a part of the proposed relations, the rest of non-tree relations (hereafter also called secondary relations) will be stored in the harmonised data, too.*

3.3.3 Additional features

Resources selected for harmonisation do not provide the same set of features. While some resources assign many different features, e.g. morphological categories, morphological segmentation, and semantic labels in DeriNet, some resources do not even contain part-of-speech tags, e.g. The Polish Word-Formation Network.

Adding features is a challenging task because it could cause new problems in the data and its processing. For example, in the case of additional part-of-speech tagging, homonyms would be one of the issues. Considering Polish lexeme '*przepaść*' ('*a gap*'/'*to get lost*'), either one tag would have to be chosen, i.e. noun/verb, or a new lexeme would have to be created to cover both cases. However, both solutions would affect word-formation relations captured in the original data.

> *The final decision is, therefore, not to add new features during the harmonisation process.*

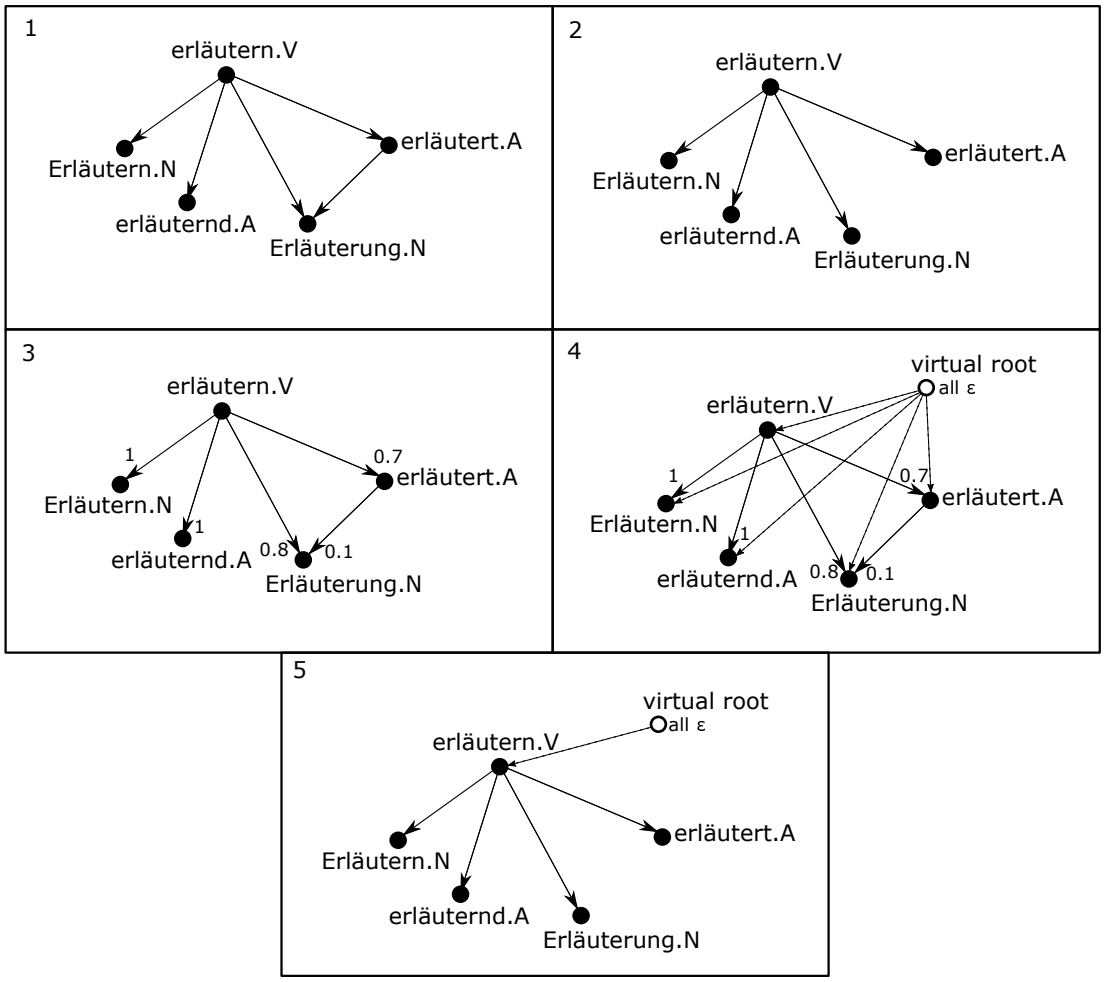


Figure 3.3: Harmonisation procedure illustrated on data from DErivBase.

3.4 Harmonisation procedure

The proposed procedure harmonises annotation schemata of selected resources into the same target data structure and file format. As a result, a collection of several harmonised word-formation resources is created. The procedure consists of five parts, each briefly introduced here, and described in details separately in the following sections. Figure 3.3 illustrates the individual steps of the procedure.

1. **Importing original data.** The procedure starts with importing data from the original resources and identifying its data structure (or representing the original data as a data structure described in Section 2.4, respectively). Step 1 in Figure 3.3 shows a word-formation family from German DErivBase represented as a weakly connected subgraph. Since the family is not a rooted tree which is the selected target data structure, a tree-shaped skeleton has to be identified in the family.

2. **Annotating word-formation families.** The rooted trees are identified on the basis of manual annotations, cf. step 2 in Figure 3.3. If the original resource has many families that are not organised in rooted trees, only a random sample of those families is annotated. The sample is used for the development of a machine learning model.

3. **Scoring word-formation relations.** Based on the manually annotated random sample, a machine learning model for scoring relations in the families is developed and applied to the original data, see step 3 in Figure 3.3.

4. **Identifying rooted trees.** Before identifying rooted trees, a temporary virtual root is added and connected to all lexemes in the family, see step 4 in Figure 3.3. More details on the virtual root are described in Section 3.4.4. The tree-shaped skeleton is obtained using the Maximum Spanning Tree algorithm for finding maximum spanning arborescence of maximum scores.

5. **Converting data into the target representation.** The roots of resulting rooted trees are attached below the virtual root, see step 5 in Figure 3.3. The virtual root is removed from the family, and the resulted rooted tree(s) is/are converted to the target file format using the DeriNet API.[7] The non-tree relations (cf. step 1 in Figure 3.3) are also stored, but in a less prominent place than the tree-shaped relations.

3.4.1 Importing data from the input resources

The input resources differ in file formats (see Chapter 2). While Word Formation Latin stores the data in a SQL database, Démonette, NomLex-PT, and Sloleks use XML format, and other resources distribute the data in various types of textual file formats with different separators. For that reason, the data from the input resources needs to be converted into the same common file format at the beginning of the harmonisation process.

As many relevant pieces of information as possible were imported from all resources. Table 3.1 lists features imported from the input resources, which often include lexemes, derivational relations (DER), relations of compounding (COM), part-of-speech tags (POS), morphological categories (MCG), morphological segmentation (SEG), and semantic labels (SEM). Some resources also include the individual custom features, such as bracketed hierarchical morphological segmentation in CELEX (see Figure 2.1), subparadigmatic relations in Démonette (see the paradigm system described in Section 1.2.1), technical lemma identifiers in DeriNet, unique IDs connecting lexemes to other Italian resources in DerIvaTario, types of derivational process (e.g. suffixation, prefixation, etc.) in DerivBase.Ru, and word-formation rules serving as a basis of morphological segmentation and identification of derivational relations during the creation of the resources, e.g.

in DErivBase:
– '*Bäcker*' → '*Bäckerei*', '*Rüpel*' → '*Rüpelei*', '*Türke*' → '*Türkei*'
dNN01 = dPattern 'dNN01'
(sfx 'ei' & try (dsfx 'e')) mNouns fNouns

in DerivBase.Ru:
– 'детсад' → 'детсадик'
rule429(noun + 'ик/ок/ук' → noun)

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

Table 3.1: Imported features from the individual resources: *DER* for derivational relations, *COM* for compounding relations, *POS* for part-of-speech categories, *MCG* for morphological categories, *SEG* for morphological segmentation, *SEM* for semantic labels, *CST* for additional custom features. Tick marks (✓) denote imported features, while dashes occur if the resource does not contain the particular feature.

Input resource	Imported features						
	DER	**COM**	**POS**	**MCG**	**SEG**	**SEM**	**CST**
CatVat	✓	–	✓	–	–	–	–
D-CELEX	✓	✓	✓	–	✓	–	✓
Démonette	✓	–	✓	✓	✓	✓	✓
DeriNet	✓	✓	✓	✓	✓	✓	✓
DeriNet.ES	✓	–	–	–	–	–	–
DeriNet.FA	✓	–	–	–	–	–	–
DerIvaTario	✓	–	✓	–	✓	–	✓
DErivBase	✓	–	✓	✓	–	–	✓
DerivBase.Hr	✓	–	✓	–	–	–	–
DerivBase.Ru	✓	–	✓	–	–	–	✓
E-CELEX	✓	✓	✓	–	✓	–	✓
EstWordNet	✓	–	✓	–	–	–	–
EtymWordNet (9x)	✓	–	–	–	–	–	–
FinnWordNet	✓	–	✓	–	–	–	–
G-CELEX	✓	✓	✓	–	✓	–	✓
NomLex-PT	✓	–	✓	–	–	–	–
The M-S Database	✓	–	✓	–	–	✓	–
The Polish WFN	✓	–	–	–	–	–	–
WFL	✓	✓	✓	✓	✓	–	–

Not all pieces of information were imported from the original resources, for instance, labels referring to the origin of each feature involved in Démonette were left out. In the case of EstWordNet, FinnWordNet and Etymological WordNet, only derivationally related lexemes were imported, disregarding the wordnet architecture.

In most of the resources, it is not sufficient to represent lexemes by using only their written forms. Lemmatisation of lexemes in each original resource is crucial because of lexeme homonymy. The representations of lexemes vary across the resources. Word Formation Latin assigns a unique numerical ID to each lexeme. DErivBase (and many other resources) distinguishes lexemes based on combinations of morphological categories, e.g. part-of-speech class and gender. DeriNet uses the written form of a particular lexeme and a *tag masks* consisting of stable morphological categories in the paradigm of the particular lexeme.

In the case of resources that do not contain any word-formation relations among lexemes, i.e. CELEXes and DerIvaTario, the relations were generated using the morphological segmentation, which is included in the data. Having the segmentation, especially in the hierarchical form, potential base lexemes can be automatically proposed for individual lexemes.

After the imports, input data structures of the imported resources were identified; respectively, the input data was represented as a data structure which was the most suitable for the data according to the description in Section 2.4. Since the rooted trees were selected as the target data structure for representing word-formation families, resources organising the families in rooted trees, i.e. DeriNet,

DeriNet.ES, DeriNet.FA, and The Polish Word-Formation Network, did not need any harmonisation of the data structure. Harmonisation of these resources laid in the transformation of their file formats to the target file format, and possibly in unifying different key-value pairs, see Section 3.4.5. In the remaining resources, tree-shaped skeletons were identified.

3.4.2 Annotating word-formation families

Word-formation families in most of the resources are represented using less constrained graphs than the rooted tree is, which can be caused by not only technical but also linguistic reasons. The target rooted tree data structure focuses directly on a subsequent derivation of lexemes from each other one by one using derivational processes. The other data structures allow additional non-tree relations to capture other phenomena, such as compounding or *double motivation*, i.e. the situations when the lexeme can be derived from two or more base lexemes, see example 7 in Figure 3.4. However, the additional relations can also be only a by-product resulting from a method which has been used to connect lexemes within word-formation families in a particular resource. For instance, the rule-based approach in DErivBase and DerivBase.Ru over-generates (additional non-tree) relations to ensure that all lexemes belonging to the same word-formation family are connected, even if any (base) lexeme is missing from the lexeme set. Table 3.2 shows the amount of tree-shaped and non-tree-shaped word-formation families in each resource selected for harmonisation.[8] To obtain tree-shaped word-formation families for the following development of supervised machine learning models, manual annotations of word-formation families that are not represented as rooted trees, i.e. contain additional non-tree relations, is needed.

As shown in Table 3.2, CELEXes, CatVar, DErivBase, DerivBase.Hr, DerivBase.Ru, and FinnWordNet contain so many non-tree word-formation families that only (random) samples of those families were annotated from the mentioned resources. The sample sizes vary between 400-600 word-formation families depending on several factors, such as repetitions of annotated phenomena,[9] sizes of the families in terms of lexemes and relations, and time consumption. The samples serve for the development of machine learning models to score relations automatically in the next phase of the harmonisation process (Section 3.4.3). Nevertheless, Démonette, EstWordNet, EtymWordNet-{cat, ces, gla, pol, por, rus, hbs, swe, tur}, NomLex-PT, The Morpho-Semantic Database, and Word Formation Latin were annotated completely manually because they contain less than 300 families that are not organised in rooted trees.

The annotation task should be specified precisely, and adequate conditions to accomplish the task should be provided to the annotator(s). In the case of harmonisation of word-formation data presented in this thesis, both the annotation task and the technical conditions are designed from scratch. Therefore, both aspects are described separately in the following subsections.

[8]The non-tree-shaped families were identified using the Breadth-First Search graph algorithm. Families consisting of only one lexeme (so-called *singletons*) were excluded.

[9]Since most of the resources have been created (semi)automatically, the additional non-tree relations are often systematically repeated across word-formation families in particular resources. In those cases, the annotation of large samples (e.g. 600 families) would be not be sufficient in terms of time management, so the smaller samples were annotated.

Table 3.2: The numbers of tree-shaped and non-tree-shaped word-formation families (and relations within them) in the input resources selected for harmonisation. Families consisting of only one lexeme (so-called *singletons*), and relations explicitly labelled as compounding are not considered.

Input resource	Tree-shaped		Non-tree-shaped	
	families	relations	families	relations
CatVat	0	0	13,367	155,064
D-CELEX	0	0	5,449	1,733,364
Démonette	7,050	12,849	286	1,303
DerIvaTario	0	0	1,992	28,088
DErivBase	15,831	21,795	3,962	33,215
DerivBase.Hr	0	0	14,818	3,056,962
DerivBase.Ru	7,653	10,076	10,293	279,817
E-CELEX	0	0	6,725	109,002
EstWordNet	428	470	28	65
EtymWordNet-cat	2,879	4,422	40	191
EtymWordNet-ces	2,284	4,788	70	543
EtymWordNet-gla	2,412	4,688	57	403
EtymWordNet-pol	2,822	24,106	59	879
EtymWordNet-por	1,166	1,586	15	41
EtymWordNet-rus	715	2,926	36	474
EtymWordNet-hbs	1,694	6,111	20	238
EtymWordNet-swe	2,865	4,075	20	376
EtymWordNet-tur	1,837	5,188	84	769
FinnWordNet	2	2	6,345	29,781
G-CELEX	0	0	5,615	145,936
NomLex-PT	2,751	4,124	34	111
The M-S Database	5,690	7,580	128	420
WFL	5,230	21,946	43	741

The annotation task

For all non-tree-shaped word-formation families, the annotator's task was to identify derivational relations that would form a tree-shaped word-formation family and that would concur with the linguistic view of derivation described in Section 1.2.1. Moreover, the resulting families had to be organised as rooted tree(s). Splitting the family was allowed, but all new families had to be still tree-shaped. Annotators were not allowed to add any new relations or lexemes because of the conservative approach to the harmonisation, which is discussed in Section 3.3.

As for the annotators, the annotation sample of word-formation families from DerivBase.Ru was annotated by Anna Nedoluzhko, who is a Russian native speaker with a linguistic background. The samples from the rest of the resources were annotated by the author of the thesis, who is a Czech native speaker with a linguistic background and knowledge of English, German, Polish, and Slovak. Besides the language experience, annotators used several electronic translation dictionaries[10], monolingual and specialised lexicons[11], and other resources[12] while

[10]https://slovniky.lingea.cz/ and https://translate.google.cz/

[11]http://anw.inl.nl/ and https://wsjp.pl/ and http://drevoslov.ru/ and http://slovnikafixu.cz/ and https://dwds.de/ and https://www.owid.de/ and http://hjp.znanje.hr/ and https://cnrtl.fr/ and http://etymologiebank.nl/

[12]https://wiktionary.org/

the annotating of the data. Wiktionary was a very useful resource during the annotating. The language portion of Wiktionary suitable for a particular annotated language was used; however, the English language portion of Wiktionary contains lexemes and many pieces of information for not only English but also for other languages annotated here. Morphological segmentation included in CELEXes, DerIvaTario, and partly in DErivBase, DerivBase.Ru, Démonette, and WFL was also helpful.

During the manual annotations, several phenomena with fuzzy solutions (and also identification) were observed, see Figure 3.4. Some of them were specific for a particular resource, but most of them repeated across the resources.

Lemmatisation The approach to lemmatisation differs in individual word-formation resources. Especially resources of morphologically rich languages, e.g. EtymWordNet-ces, also contain inflected forms of lexemes, e.g. plural forms of lexemes, cf. example 1 in Figure 3.4. The inflected forms were kept as close as possible to their representative lexemes.

Spelling variants i.e. several different realisations denoting the same meaning, occurring, for example, in DErivBase and The Morpho-Semantic Database, are a problem similar to inflected forms of lexemes, cf. example 2 in Figure 3.4. If it was possible, one of the spelling variants was chosen to become a base lexeme for the other ones.

Negation, reflexivity, and grammatical aspect represent corner-cases of the problem with lemmatisation. Two approaches to capturing the phenomena were observed in the resources, see examples 3 and 4 in Figure 3.4: (1) negative/reflexive lexemes were connected directly to their affirmative/irreflexive lexemes (solid lines in the examples), (2) negative/reflexive created parallel sub-trees of negative/reflexive lexemes and affirmative/irreflexive lexemes separately (dotted lines in the examples). The former solution was selected, because it simplifies dealing with situations in which some (negative/reflexive) lexeme is missing. The resources of Slavic languages, e.g. DeriNet, The Polish Word-Formation Network, and DerivBase.Ru, contain verbal aspectual counterparts because the grammatical aspect is conveyed by derivational affixes in Slavic languages. In the case of grammatical aspect, perfective verbs were mostly annotated as derived from imperfective verbs, except in the case of secondary imperfectivisation, cf. example 5 in Figure 3.4.

Loanwords Most of the resources contain loanwords, see example 6 in Figure 3.4. If possible, they were captured as derivation.

Compounding and double motivation Other problematic phenomena were compound and double motivated lexemes; both are defined by having more than one base, see examples 7 and 8 in Figure 3.4. If a compound lexeme was explicitly labelled in the input resource (e.g. in DeriNet, Word Formation Latin), no additional annotation was needed. Otherwise, the compounds were disconnected from their base lexemes, except for subsequent

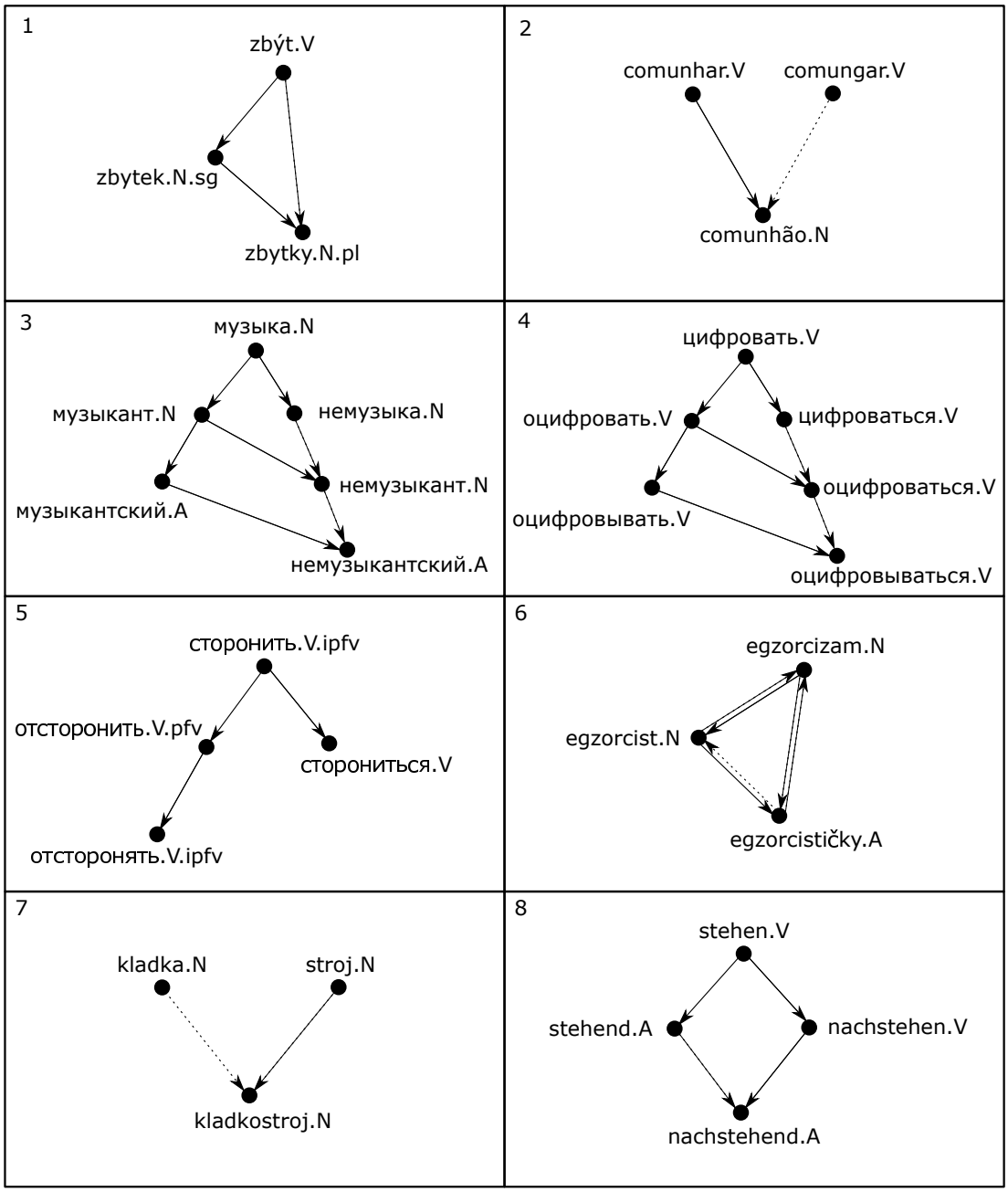


Figure 3.4: Several annotated phenomena illustrating manual annotation of inflexion (1; EtymWordNet-ces), spelling variants (2; NomLex-PT), negation (3; DerivBase.Ru), reflexivity (4; DerivBase.Ru), gramatical aspect (5; DerivBase.Ru), loanwords (6; DerivBase.Hr), compound lexemes (7; EtymWordNet-ces), and double motivation (7; DErivBase). Solid lines represent resulting tree-shaped skeletons, dotted lines represent other possible relations provided in particular resources.

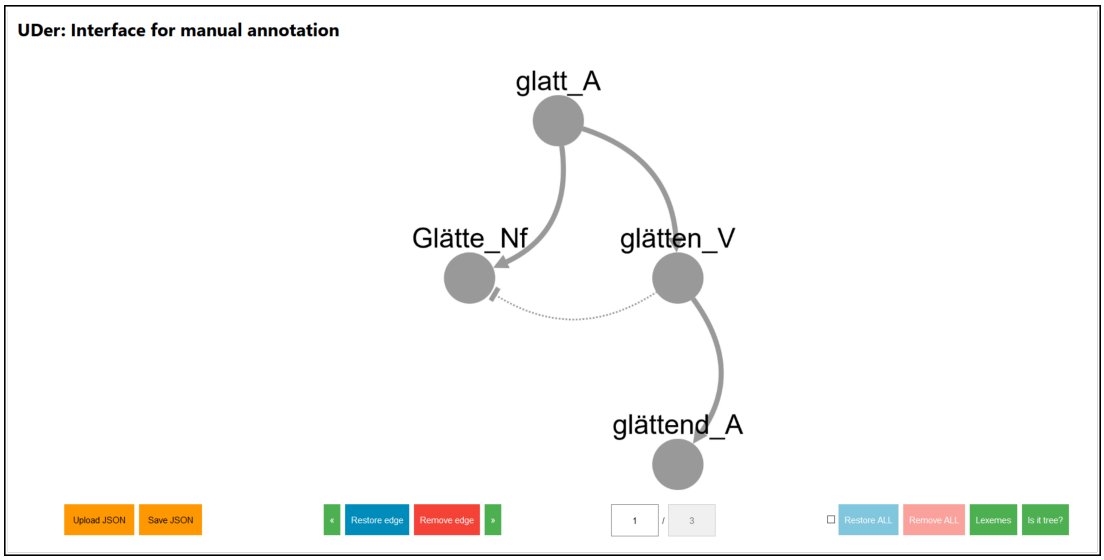
derivations of compounds – they are still annotated as derivational relations. Any future annotation could focus on identifying compound lexemes and connecting them with all their base lexemes. As for double motivation, one derivational process is always chosen (e.g. prefixation is understood as 'more stable' than suffixation, see example 8 in Figure 3.4), and the rest of similar situations is annotated consistently with the option. The other possibilities are still preserved but in a less prominent place in the target file format.

Interface for manual annotations

The annotator usually gets a text file containing the individual word-formation relations whose presence/absence is decided by the annotator. In the case of harmonisation presented in this thesis, the annotator has to annotate all relations included in the individual word-formation families. The resulting families have to be organised into the rooted tree data structures. Resolving the task, especially the accomplishment of treeness, is difficult without visual control. Therefore, a visual interface for manual annotations has been developed by the author of the thesis to facilitate the annotation, see Figure 3.5. Word-formation families can be displayed, edited and saved using the interface. Some additional features, such as the automatic check whether the annotated family is already represented as rooted tree(s), were also added to facilitate the process of annotation.

When the annotator uploads data using the `Upload_JSON` button, the interface displays the first family. After the annotator finishes the annotation or he/she wants to stop working, the data can be saved by pressing `Save_JSON`. Lexemes are represented as nodes and relations are represented as directed edges (arrows) pointing from the base lexeme to the derivatives. Although nodes are placed randomly at the initial screen, the interface saves the positions of nodes for comfortable repetitive annotations. The screen can be zoomed using the mouse wheel; nodes can be moved by holding the left mouse button, and edges can be selected by clicking (more of them can be selected by holding the `Ctrl` key).

The annotator has to select non-tree-shaped edges and 'remove' them using `Remove_edge` button (or pressing `Delete` key). After that, the edge line is dotted, and its head is a small rectangle instead of a triangle (tree-shaped edges are represented as solid lines). Setting the solid lines back is possible using `Restore_edge` button (or `Shift` key). For annotation of word-formation families organised in complete directed subgraphs, `Restore_ALL` and `Remove_ALL` buttons are useful. They can be enabled by ticking the checkbox. The button `Lexemes` (or pressing key `l`) lists all lexemes displayed on the screen. Thanks to that, the annotator can copy the lexemes, and they do not have to write them. It is helpful if the annotator wants to search a lexeme on the internet or in the other language resources. By clicking on the button `Is_it_tree?` (or by pressing the key `t`), the Breadth-First Search graph algorithm checks and notifies whether the displayed family is already organised in the rooted tree(s). After the annotation of the displayed family, the next one can be displayed by pressing the green button with the right arrow (or pressing the right arrow on the keyboard). The green button with the left arrow (or pressing the left arrow on the keyboard) serves for displaying the previous family. The number of the currently displayed family occurs in the textbox. Annotator can also write the number of the particular family to the textbox, and after they press Enter key, the required family is displayed. It




```
[
  {
    "nodes":
      [
        {"data":{"name":"glättend_A","id":"glättend_A"}},
        {"data":{"name":"Glätte_Nf","id":"Glätte_Nf"}},
        {"data":{"name":"glatt_A","id":"glatt_A"}},
        {"data":{"name":"glätten_V","id":"glätten_V"}}
      ],
    "edges":
      [
        {"data":{"target":"glatt_A","intoTree":"solid","source":"Glätte_Nf"}},
        {"data":{"target":"glatt_A","intoTree":"solid","source":"glätten_V"}},
        {"data":{"target":"glätten_V","intoTree":"dotted","source":"Glätte_Nf
    "}},
        {"data":{"target":"glätten_V","intoTree":"solid","source":"glättend_A"}}
      ]
  }
]
```

Figure 3.5: Interface for manual annotations and an example of one word-formation family captured in the input JSON file format, which is loaded by the interface.

is also possible to write a particular lexeme, and the interface displays the family containing that lexeme.

Technically, the interface is designed for running in common web browsers. It is optimised for Microsoft Edge, Microsoft Explorer, Google Chrome, and Mozilla Firefox. The interface is developed using HTML5, CSS3 (including W3.CSS), and JavaScript (jQuery, CytoScape.js and Notify.js libraries were used). Input and output data are expected to be encoded in JSON, cf. Figure 3.5.

3.4.3 Scoring word-formation relations

Based on manually annotated samples, supervised machine learning classification models were developed to annotate data from CELEXes, CatVar, DerIvaTario, DErivBase, DerivBase.Hr, DerivBase.Ru, and FinnWordNet. The models predicted scores estimating a chance of presence/absence of derivational relations proposed by the resources.

The relations were equipped with several features to create a feature vector. Most of the features were converted to binary (Boolean data type) using one-hot

Table 3.3: The numbers of families (and relations within them) included in train, validation, and holdout datasets.

Input resource	TRAIN		VALIDATION		HOLDOUT	
	fams	relats	fams	relats	fams	relats
CatVat	390	5,068	90	1,070	120	1,480
D-CELEX	274	4,246	62	1,082	83	1,268
DerIvaTario	286	3,520	66	856	88	1,078
DErivBase	281	3,416	64	753	86	1,057
DerivBase.Hr	397	5,042	91	1,084	122	1,458
DerivBase.Ru	361	6,914	83	1,688	111	2,152
E-CELEX	268	4,382	61	990	82	1282
FinnWordNet	246	1,564	56	382	75	486
G-CELEX	293	3,670	67	820	89	1,230

encoding. The following features were acquired:

- part-of-speech categories and other morphological categories, e.g. gender, aspect, etc., of the proposed base lexeme and derivative, if present in the particular resource; (Boolean);
- Levenshtein distance/similarity (Levenshtein, 1966) counting the minimum number of single-character edits between two lexemes; (Number);
- Jaro-Winkler distance/similarity (Jaro, 1989; Winkler, 1990) measuring an edit distance biased by the idea that initial lexeme differences (prefixes) are more significant than differences near the end of the lexemes; (Number);
- Jaccard distance/similarity (Jaccard, 1912) calculating (dis)similarity of character n-gram sets in two lexemes; (Number);
- length of the longest common substring; (Number);
- boolean values manifesting whether the base lexeme and derivative have the same one/two initial or final characters; (Boolean);
- initial and final character n-grams of the base and derivative; (Boolean);
- other custom features from the original resource, e.g. derivational rules documented in DErivBase and DerivBase.Ru; (Boolean).

Features included in the final models vary resource by resource. The conditional entropy calculated between each feature and the output variable, i.e. decision on the presence or absence of a particular relation, helped to select a suitable set of features for developing supervised machine learning models.

Manually annotated samples containing both the positive and negative examples (relations) were always divided into the training, validation, and holdout datasets, see Table 3.3. The training dataset (65% of families from the sample) was used for learning classifiers. The validation dataset (15%) served for testing the model during a development phase, and the holdout dataset (20%) provided a final estimate of machine learning model performance.

Several machine learning classification methods implemented in the Python `scikit-learn` library (Pedregosa et al., 2011) were tested, namely: *Naive Bayes*,

K-Nearest Neighbour, *Logistic Regression*, *Decision Tree*, *Random Forest*, *AdaBoost*, *Perceptron*, and *Multi-Layer Perceptron*. For each predicted relation, the probability[13] of being tree-shaped were always estimated by the model. The returned probabilities were used as scores of the individual relations (edges), regardless their scaling, normalisation, or transformation made by the models. The scores have, therefore, different nature across the methods in terms of absolute values but still estimate the presence/absence of the particular relations in the tree-shaped word-formation families.

The performance of the models was evaluated using the established F-measure (also known as F-score; Chinchor, 1992; Van Rijsbergen, 1979) which is the harmonic mean of precision and recall:

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

Precision is the fraction of relations predicted as tree-shaped (*true positives*) divided by all predicted relations as tree-shaped (*true positives* plus *false positives*):

$$precision = \frac{true\ positives}{true\ positives + false\ positives}$$

Recall is the fraction of relations predicted as tree-shaped (*true positives*) divided by relations that should have been predicted as tree-shaped (*true positives* plus *true negatives*):

$$recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

Models having the best results of performance were chosen as the final ones; see Table 3.4 for the results and parameters of the models. However, these results of performance should be considered only as a proxy measure. The models predict only a probability of being tree-shaped, but the final performance can be evaluated only after the identification of rooted trees, which is described in the next section. Decision Tree models were the best for predicting data from CatVar, DerIvaTario, DerivBase.Hr, D-CELEX, E-CELEX, and G-CELEX. Random Forest models were used for predicting relations in FinnWordNet. Logistic Regression model reached the highest F-measure while predicting relations from DErivBase and DerivBase.Ru.

[13]Here should be mentioned, that K-Nearest Neighbour method has only a limited concept of probability which estimates probabilities as a fraction of votes among nearest neighbours.

Table 3.4: Evaluation of the machine learning models on validation (V) and holdout (H) datasets. Values of F-measure are in percents. A quality of a split in Decision Tree and Random Forest Classifiers is set to criterion='entropy'. The bold value indicates the chosen model for final harmonisation of the particular resource.

	CatVar		D-CELEX		DerIvaTario		DErivBase		DerivBase.Hr		DerivBase.Ru		E-CELEX		FinnWordNet		G-CELEX	
Machine Learning model (Python sklearn)	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**
GaussianNB()	50.2	46.2	56.4	50.6	43.8	45.2	72.7	71.6	52.6	54.1	63.7	61.9	48.3	39.0	52.1	55.5	43.2	55.3
BernoulliNB()	68.2	67.9	59.1	55.2	61.5	63.0	84.1	83.7	65.8	59.8	76.9	76.8	59.6	60.4	69.2	69.4	62.6	63.0
ComplementNB()	67.7	64.6	68.7	61.4	63.4	63.7	82.7	82.8	61.7	68.1	76.8	77.6	59.6	60.2	68.4	69.0	65.1	67.9
MultinomialNB()	68.2	67.5	63.7	59.3	64.8	70.0	83.9	83.5	65.5	69.7	77.0	77.6	60.5	57.6	69.8	69.7	66.7	67.1
LogisticRegression(solver='liblinear', penalty='l2')	76.2	75.5	79.2	76.3	70.6	75.3	87.8	85.9	71.6	79.8	82.3	82.8	65.0	65.3	70.2	68.0	73.6	72.2
LogisticRegression(solver='saga', l1_ratio=0.3)	76.0	76.4	77.9	74.7	70.9	75.8	88.5	86.7	74.2	80.0	**83.0**	**83.1**	65.7	64.8	70.4	68.5	73.3	70.8
LogisticRegression(solver='saga', l1_ratio=0.5)	76.4	76.5	76.6	74.4	69.9	76.2	88.2	86.6	74.9	79.8	82.6	83.5	63.9	64.5	71.5	68.6	74.2	70.6
LogisticRegression(solver='saga', l1_ratio=0.8)	76.5	76.0	76.3	74.9	69.7	77.2	**88.6**	**85.8**	75.1	79.2	82.7	83.8	63.3	63.0	71.4	68.1	72.2	70.0
DecisionTreeClassifier(min_samples_split=5)	**82.4**	**80.7**	78.5	78.1	76.5	75.4	86.5	84.3	77.1	80.1	79.2	81.4	73.2	73.1	73.9	67.3	74.6	77.8
DecisionTreeClassifier(max_depth=10)	81.6	80.2	**81.1**	**77.1**	**77.5**	**76.0**	86.4	85.3	72.7	73.9	74.9	75.9	64.9	67.6	72.1	66.6	73.2	74.8
DecisionTreeClassifier(min_samples_split=20)	82.0	81.5	80.7	75.2	75.5	75.5	87.8	84.6	**77.2**	**80.7**	79.1	81.5	**74.0**	**74.0**	73.2	66.7	**75.6**	**76.8**
DecisionTreeClassifier(min_samples_split=50)	80.1	78.7	78.8	77.8	74.6	75.2	88.3	85.5	74.7	80.4	77.7	80.7	67.1	72.7	72.4	66.9	71.4	72.8
RandomForestClassifier(min_samples_split=5)	82.3	77.4	70.5	71.3	72.6	78.5	88.1	87.1	77.1	79.4	82.2	83.7	61.0	57.1	**74.0**	**70.1**	69.9	71.0
RandomForestClassifier(max_depth=20)	52.9	57.2	55.2	56.7	57.2	71.5	86.1	84.5	62.1	62.9	80.2	80.0	58.8	54.9	71.2	68.3	49.0	51.2
RandomForestClassifier(min_samples_split=20)	80.9	76.0	69.3	70.5	66.9	74.4	87.7	85.8	73.9	77.8	81.0	83.6	56.2	53.3	73.6	68.3	66.4	68.0
RandomForestClassifier(min_samples_split=50)	76.5	75.1	65.9	67.5	60.5	66.8	87.7	85.9	72.3	76.4	80.9	83.5	53.1	52.8	73.9	68.0	61.4	66.8
AdaBoostClassifier(n_estimators=100)	71.9	70.9	70.7	71.7	72.4	75.8	87.2	84.7	76.0	76.2	80.1	80.6	61.4	62.4	72.1	64.5	70.2	70.5
Perceptron(max_iter=50, tol=-np.infty)	53.4	58.5	67.2	63.1	64.2	70.8	86.6	85.1	53.1	54.7	81.3	82.1	49.5	50.0	66.2	67.1	59.6	57.0
Perceptron(max_iter=100, tol=-np.infty)	54.8	59.8	61.4	62.2	62.4	66.8	86.7	86.0	53.5	56.8	80.6	81.5	46.3	47.1	60.6	61.5	56.4	57.2
KNeighborsClassifier(n_neighbors=5)	76.2	70.7	75.0	70.8	71.5	77.6	85.1	84.0	65.3	63.3	79.2	79.2	58.3	64.7	69.5	67.5	71.1	70.4
KNeighborsClassifier(n_neighbors=2)	71.8	68.2	68.1	68.6	69.5	76.7	84.5	82.0	69.6	71.5	76.7	76.3	64.3	63.4	71.9	63.7	72.1	72.9
MLPClassifier()	81.0	80.5	77.3	77.2	73.7	77.3	86.8	85.3	72.0	80.3	79.7	81.6	71.6	68.2	73.6	64.3	74.1	77.5

3.4.4 Identifying rooted trees

Having relations assigned with scores using a machine learning model, the tree-shaped skeleton can be identified by maximising the sum of scores for each word-formation family, see A in Figure 3.6. Maximum Spanning Tree algorithm (Chu & Liu, 1965; Edmonds, 1967) was used for finding the skeleton.[14]

However, some word-formation families cannot be covered by a single tree-shaped spanning tree because of various phenomena presented in Section 3.4.2. Therefore, some families needed to be divided to obtain tree-shaped skeleton(s). Due to these families, a temporary virtual root was added to each family, and it was connected with all lexemes in the family, see B in Figure 3.6. Yet adding the virtual root may seem only a technical step to avoid failing Maximum Spanning Tree algorithm, it also brings an important parameter ε that provides scoring the edges between the virtual roots and other lexemes. While $\varepsilon = \infty$ would lead to disconnection of all relations between lexemes, $\varepsilon = -\infty$ allows successful completion of the algorithm even in families that do not have one tree-shaped skeleton. The scores assigned by the machine learning model are in the range from 0 to 1 (zero for relations preferred as absent, one for the opposite). Setting ε in the same range can serve as a parameter for smoothing the resulting families.

As for the evaluation of the final tree-skeleton(s) identification, the F-score was used. Table 3.5 shows a dependency of F-score and parameter ε evaluated on validation and holdout datasets of each resource harmonised using the selected machine learning model.

3.4.5 Converting data into the target representation

If tree-shaped skeletons are identified in all word-formation families, i.e they fit the target data structure, they can be converted into the target file format. At the same time, other additional annotations are harmonised and converted.

Converting correctly distinguished lexemes is one of the key steps. A unique identifier for each lexeme (LEMID) has to be used; however, the harmonised resources distinguish lexemes in different ways, as was already mentioned in Section 3.4.1. These ways were more or less respected. In most cases, the written form of the lexeme and its part-of-speech tag, if present, (separated by hash sign) were enough. In DErivBase, the gender of nouns was also added to the LEMID. The written form of lexeme and tag mask are still used in DeriNet. For distinguishing lexemes in Word Formation Latin and in CELEXes, original IDs were taken, so the LEMID always consists of the written form of the lexeme, part-of-speech tag, and an original ID. For instance, in Word Formation Latin, it was necessary to use original IDs because homonymy/polysemy of lexemes, e.g. lexeme '*gallus*' has three meanings with different derivatives: '*a farmyard cock*', '*an inhabitant of Gaul*', and '*an emasculated priest of Cybele*' (Glare, 1968).

Relations between lexemes were converted as expected. The identified rooted trees represent skeletons of harmonised word-formation families from the original resources. Non-tree-shaped relations are stored in a less prominent place (JSON-encoded column 10) in the target file format. However, they are not preserved

[14]It was used Maximum Spanning Tree algorithm implemented in the Python library NetworkX (Hagberg et al., 2008). In the thesis, all graphs were processed by this library.

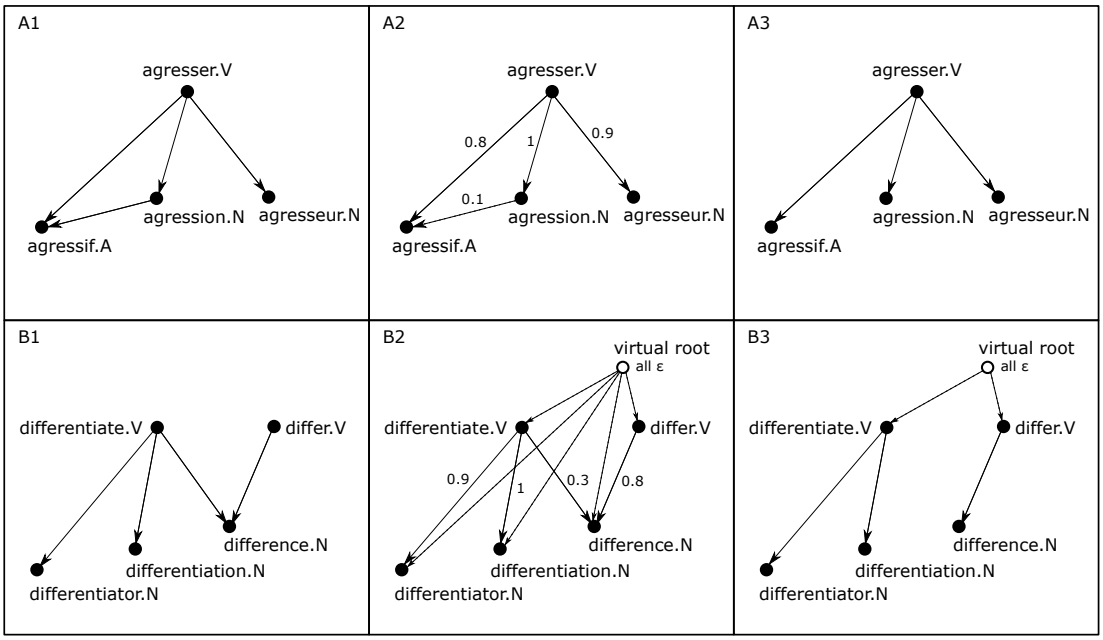


Figure 3.6: Illustration of identifying rooted trees by maximising a sum of scores. While just one tree is obtainable from family A (The Morpho-Semantic Database), family B (Démonette) has to be divided. The virtual root prevents failing Maximum Spanning Tree algorithm, and provides smoothing based on the value of ε.

Table 3.5: A dependency of F-score and parameter ε evaluated on validation (V) and holdout (H) datasets. The bold value indicates the chosen ε for final harmonisation of particular resource.

	CatVar		**D-CELEX**		**DerIvaTario**		**DErivBase**		**DerivBase.Hr**		**DerivBase.Ru**		**E-CELEX**		**FinnWordNet**		**G-CELEX**	
ε	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**	**V**	**H**
−1M	64.8	63.7	59.1	63.2	62.2	59.7	87.8	87.8	58.9	61.4	82.9	82.9	62.4	64.9	66.3	62.9	64.4	63.5
0.0	80.6	80.8	68.3	68.4	66.6	66.3	90.0	88.9	80.5	82.7	**84.4**	**85.5**	69.2	72.7	78.2	79.3	75.0	74.4
0.1	82.1	82.8	78.9	78.6	75.7	74.4	**93.4**	**92.1**	80.6	81.3	83.7	85.0	74.5	75.7	78.7	79.9	78.7	76.8
0.2	82.7	82.5	80.7	77.8	77.3	75.8	92.8	91.7	**81.1**	**81.0**	83.6	85.1	74.4	76.6	77.9	78.0	77.8	78.2
0.3	82.5	80.5	**81.1**	**79.5**	76.7	75.9	93.0	91.5	79.6	81.2	83.2	84.3	74.3	75.2	**80.2**	**76.9**	77.8	75.8
0.4	82.8	81.1	80.2	76.0	78.0	76.8	92.0	90.7	78.5	80.6	82.9	83.9	74.0	74.3	79.8	74.5	77.9	77.5
0.5	**83.1**	**81.0**	80.9	77.7	77.1	75.0	90.6	89.5	77.9	81.2	81.9	82.7	**74.9**	**73.8**	77.6	72.9	**79.5**	**77.4**
0.6	82.1	80.5	78.7	75.3	**78.1**	**75.1**	89.0	88.6	76.6	79.8	36.5	37.2	73.1	72.7	75.2	68.6	76.4	77.4
0.7	80.6	81.4	78.9	75.1	78.1	74.5	87.6	87.1	78.2	79.1	78.9	78.7	68.5	68.5	73.5	65.2	75.2	78.0
0.8	81.3	81.6	63.9	63.3	77.3	75.3	85.0	84.8	74.8	75.5	75.3	76.1	65.9	67.9	64.8	58.3	72.2	76.2
0.9	82.3	81.1	63.9	62.6	77.3	75.9	80.6	80.4	73.7	74.0	67.6	69.7	63.5	66.1	50.6	49.4	71.1	73.1
1.0	57.3	59.4	61.2	58.7	49.0	51.4	24.9	25.4	66.0	66.2	36.5	37.0	47.9	56.1	38.2	37.8	49.6	57.2
+1M	44.6	44.9	47.9	47.8	47.7	47.5	24.9	25.4	45.2	45.4	35.1	34.1	47.0	47.0	38.2	37.8	45.4	46.7

for the harmonised versions of CatVar and DerivBase.Hr because their word-formation families are represented as complete directed graphs. If the original word-formation family was divided into more rooted trees, links connecting the root lexemes of the trees were always saved (column 10). It allows for the original graphs to be reconstructed, see Section 3.6.

As for the harmonisation of feature-value pairs, traditional categories, such as part-of-speech category, gender, number, etc., were harmonised, if present. Although semantic labels occur in several resources, namely DeriNet, Démonette, and The Morpho-Semantic Database (cf. Section 2.1.2 and 2.1.3), they have not been harmonised so far because their meaning can significantly differ resource by resource. Their values were only converted as features of particular relations. Partial or full morphological segmentation was converted to CELEXes, DeriNet, DerIvaTario, Démonette, and Word Formation Latin; however, since each resource processes the segmentation in different ways (cf. Section 2.1.1, 2.1.2, and 2.1.3), the original segmentation is only stored in a JSON-encoded column of the target file format in most of the harmonised resources. Word-formation rules annotated in DErivBase and DerivBase.Ru were converted as features of the particular relations in form `Rule=`x where x is the original identifier of the rule. The descriptions of the rules are, however, stored in a separate file. So-called sub-paradigmatic relations from Démonette were also converted to the JSON-encoded column in the target file format. The resulting collection of harmonised resources is presented in more details in the next chapter.

3.5 Remarks on evaluation

Both the prediction made by particular machine learning models and the identification of rooted trees are evaluated and presented in the description of the harmonisation procedure (see Section 3.4). In this section, a simple baseline for scoring word-formation relations in the harmonised resources is presented to illustrate the task difficulty.

For each resource harmonised by machine learning, the baseline was developed as a simple probabilistic model. Using the training dataset, the model trains probabilities of a word-formation relation in terms of part-of-speech categories in base_lexeme-derivative pairs, e.g. probabilities of V-N, V-A, N-V, N-A, etc., relations. The probabilities are used for scoring the rest of (unknown) relations in the validation and holdout datasets. The baseline model assigned scores to all word-formation relations, and the rooted trees were identified using MST-approach. Table 3.6 shows the resulting F-scores of identifying rooted trees (the complete harmonisation) using the best machine learning model vs. the baseline model (parameter ε with the highest F-score was chosen) for the resources.

No baseline model reached better results of F-score than the best machine learning models. The differences between the F-scores of the baseline and machine learning models illustrate how much better the machine learning models are in the harmonisation task than the simple baseline. Although the use of a machine learning model needs time-consuming manual annotations of at least a sample of the original data, the differences of F-score prove that the approach is useful when harmonising word-formation resources.

Table 3.6: F-scores calculated for harmonisation procedure that uses the best machine learning model vs. simple baseline on validation and holdout datasets of each harmonised resource. Results are represented in form `simple_baseline / ml_model`.

Resource	Scoring relations		Identifying trees	
	VALIDATION	HOLDOUT	VALIDATION	HOLDOUT
CatVar	44.6 / 82.4	44.9 / 80.7	51.6 / 83.1	53.3 / 81.0
D-CELEX	47.2 / 81.1	47.7 / 77.1	54.2 / 81.1	53.0 / 79.5
DerIvaTario	47.7 / 77.5	47.5 / 76.0	48.7 / 78.1	50.0 / 75.1
DErivBase	24.9 / 88.6	25.4 / 85.8	75.1 / 93.4	78.9 / 92.1
DerivBase.Hr	45.2 / 77.2	45.4 / 80.7	56.4 / 81.1	58.3 / 81.0
DerivBase.Ru	35.1 / 83.0	34.1 / 83.1	49.3 / 84.4	45.0 / 85.5
E-CELEX	47.1 / 74.0	47.1 / 74.0	59.7 / 74.9	59.4 / 73.8
FinnWordNet	38.2 / 74.0	37.8 / 70.1	62.0 / 80.2	62.9 / 76.9
G-CELEX	45.8 / 75.6	46.1 / 76.8	57.5 / 79.5	57.5 / 77.4

3.6 Rebuilding the original data

The additional non-tree relations are still stored in the harmonised data as secondary edges. The main reasons for preserving them is the opportunity to provide the same expressiveness of the harmonised version of the original data, as discussed in Section 3.2. The original data can be reconstructed from the harmonised version, too. To verify the expressiveness, this section describes rebuilding original data from the harmonised versions of DerivBase.Hr (complete directed subgraphs), DerivBase.Ru (weakly connected subgraphs), and DerIvaTario (derivation trees / listed segmentation).

Since the original lexeme sets of the harmonised resources have been taken, and the original relations are stored either as primary tree-shaped or secondary relations in the harmonised data, the basic conditions for the data rebuilding are maintained. As was mentioned in Section 3.4.4, during the identification of rooted trees, some original families were split. However, links between the resulting trees belonging to the same original word-formation family are stored (in the tenth column under the key `was_in_family_with`). At the beginning of rebuilding the original data, the rooted trees containing the link to other rooted trees need to be connected, for example, the roots of the trees are connected to the same virtual root. Then all rooted trees are traversed from the root (or the virtual root) to the leaf nodes to obtain the original relations:

- Derivation trees/listed segmentation (original structure of DerIvaTario) are obtained easily from each visited node because the original forms of derivation trees or morphological segmentation are stored in the tenth JSON column.
- Weakly connected subgraphs (DerivBase.Ru) are extracted from the harmonised data as the primary and secondary relations that point to each visited node (except relations pointing from the virtual root).
- In the case of complete directed subgraphs (DerivBase.Hr), each visited node (except for the virtual root) are appended to the list, which represents a particular word-formation family.

Chapter 4

Universal Derivations collection

The resulting collection of the harmonised word-formation resources is presented in this chapter. The name of this collection, *Universal Derivations* (*UDer*), is admittedly inspired by Universal Dependencies in the field of syntactic treebanks. The following sections summarise both the basic quantitative and qualitative characteristics of the resulting UDer collection, and the information about the availability of the collection and software/tools that are used for harmonising, querying, and visualising the harmonised data.

UDer assembles word-formation resources unified into the DeriNet-like annotation schema proposed by Vidra, Žabokrtský, Ševčíková, et al. (2019). Based on the discussion on the needs of various existing word-formation resources, the schema was developed to be general, extensible, and language-agnostic. To deal with that, the perspective of graph theory was used for representing word-formation; specifically, lexemes were represented as nodes, and relations were represented as edges between the nodes. In the target data structure, a rooted tree is the backbone of each word-formation family; however, in general, word-formation families are represented as weakly connected subgraphs because of the phenomena that cannot be modelled as tree-shaped, e.g. compounding and/or double motivation. These secondary edges are used for the original derivational relations that were not identified as tree-shaped during the harmonisation process. Thus, all derivational relations from the original resources (except for resources whose word-formation families are represented as complete directed subgraphs) are still stored in the harmonised data because of the effort to design the structural transformation after the harmonisation as reversible as possible, cf. Section 3.6.

The UDer collection version 0.5 (Kyjánek et al., 2019b) was already released, and the harmonisation procedure used to create the version 0.5 described by Kyjánek et al. (2019a). However, the procedure has been improved, and applied to more word-formation resources as described in Chapter 3 in this thesis. As a result, the new Universal Derivations collection version 1.0 (Kyjánek et al., 2020), which is released and presented here,[1] consists of 27 word-formation resources covering 19 or 20 languages depending on whether Croatian and Serbo-Croatian are considered as the same language. Figure 4.1 and 4.2 illustrate word-formation families in all harmonised resources.

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

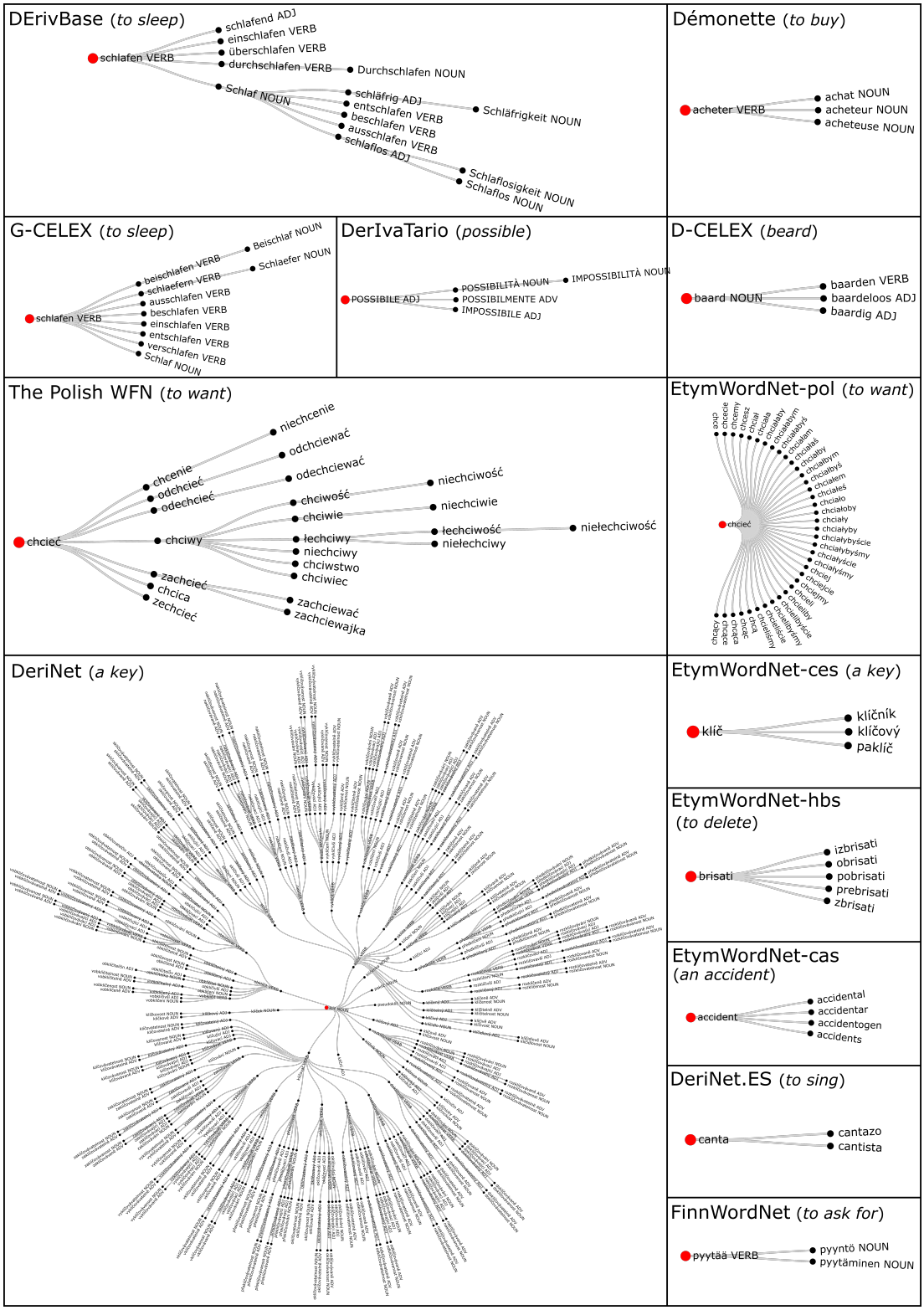


Figure 4.1: Harmonised word-formation families (part 1) from all resources included in UDer version 1.0.

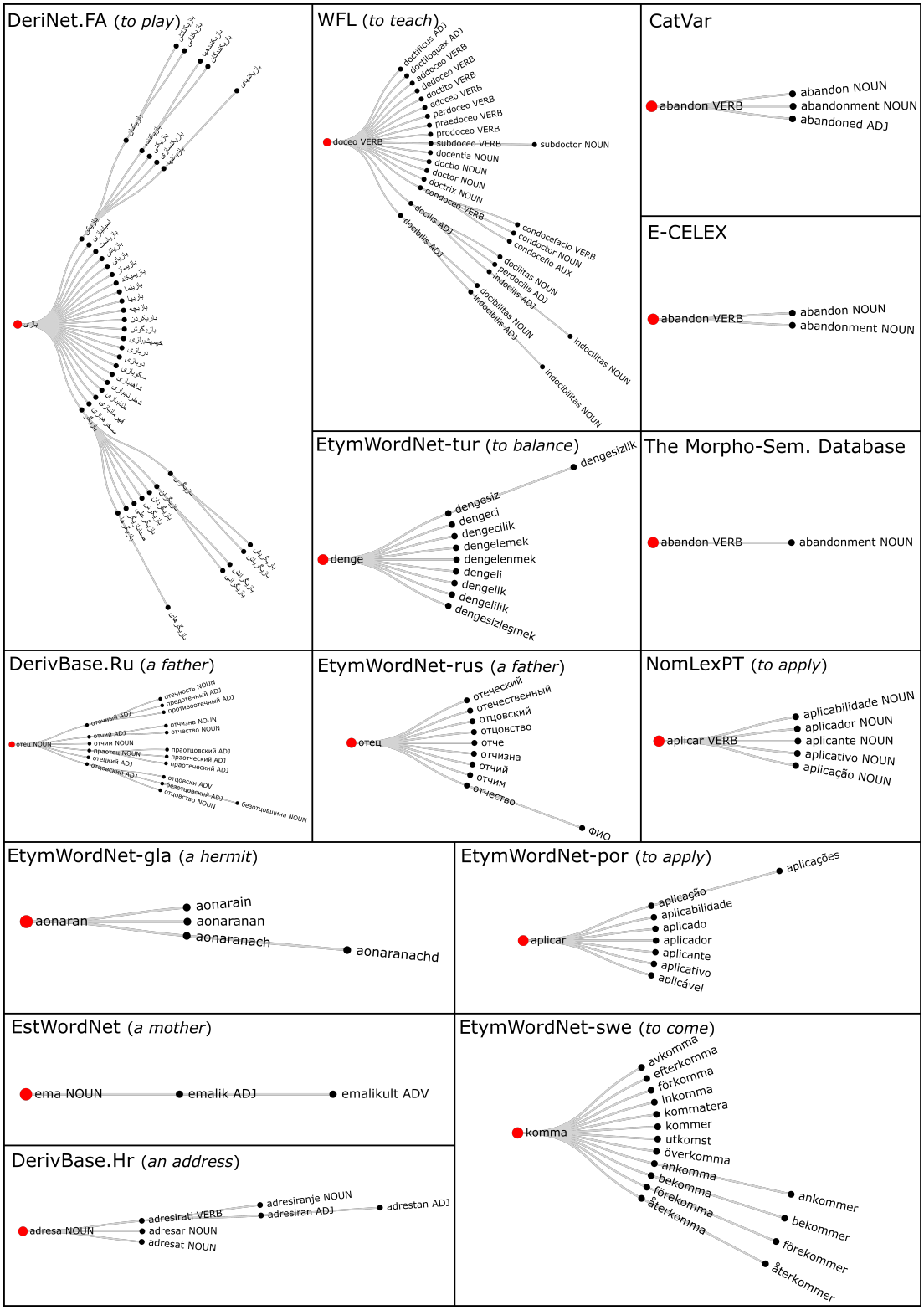


Figure 4.2: Harmonised word-formation families (part 2) from all resources included in UDer version 1.0.

4.1 Quantitative and qualitative description

To review the resulting harmonised resources included in UDer version 1.0, several quantitative characteristics are selected, cf. Table 4.1. They are described with some qualitative properties in the following paragraphs.

Resources. The presented collection consists of 27 word-formation resources. There were 16 input resources, but CELEX and Etymological WordNet contain more than one language, so their data was divided according to the individual languages. Only those language parts of Etymological WordNet that have the most word-formation relations were extracted and harmonised (others are planned to be harmonised in the future). The set of input resources consists of many resources specialised in word-formation, and it also represents all data structures observed in the existing word-formation resources, cf. Chapter 2. In addition, the original resources (except for CELEX) are published under the open licenses allowing direct redistribution of their harmonised versions in the UDer collection.

Languages. Languages captured by the UDer collection version 1.0 are mostly Indo-European languages. They are listed in Table 4.1. Czech, English, German, Polish, Portuguese, and Russian are each represented by two resources. If the Croatian (in DerivBase.Hr) and Serbo-Croatian (in Etymological WordNet) are considered as the same language as is proposed in WALS[2] by Dryer and Haspelmath (2013), then the Croatian is represented by two resources, too; however, for instance, they are distinguished as separate languages in Ethnologue[3] (Simons et al., 2020).

Lexemes. The lexeme sets were fully adopted from the input resources. The only exceptions are EstWordNet, FinnWordNet and Etymological WordNet from which only derivationally related lexemes were imported because word-formation in these resources is only a by-product while the main focus is laid on lexical relations. The amounts of lexemes in each resource differ significantly. DeriNet, DerivBase, DerivBase.Ru, The Polish Word-Formation Network, DeriNet.ES, D-CELEX, and DerivBase.Hr are the largest resources, which correlates to the way they were developed (except for D-CELEX). First, their lexeme sets were created, and second, word-formation relations between included lexemes were sought. This approach led to an increase in the number of so-called *singletons* (lexemes that have neither a base lexeme nor are further derived).

Tokenisation/lemmatisation also differ across the resources. Multi-word lexemes (their numbers are given in brackets) appear in the following resources: E-CELEX (6,600), FinnWordNet (1,297), The Morpho-Semantic Database (105), DerivBase.Ru (60), EstWordNet (14), DerIvaTario (6), Démonette (2), Word Formation Latin (1). During the manual annotations and browsing the data, many spelling variants of the same lexeme were observed in DeriNet, The Morpho-Semantic Database, and NomLex-PT. The harmonised resources also take different lemmatisation approaches to negation and reflexives. For example, while DeriNet and The Polish Word-Formation Network do not add special lemmas

[2]https://wals.info/

[3]https://www.ethnologue.com/

Resource	Language	Lexemes	Relations	Families	Singleton nodes	#Nodes	Tree depth	Tree out-degree	Part-of-speech distr. [%] Noun	Adj	Verb	Adv	Other
CatVar	English	82,675	24,873	57,802	45,954	1.4 / 18	0.3 / 7	0.3 / 10	60	24	11	5	0
D-CELEX	Dutch	125,611	13,435	112,176	107,112	1.1 / 301	0.1 / 11	0.1 / 73	63	8	8	1	21
Démonette	French	21,290	13,808	7,482	69	2.8 / 12	1.1 / 4	1.8 / 8	63	2	34	0	0
DeriNet	Czech	1,027,665	809,282	218,383	96,208	4.7 / 1638	0.8 / 10	1.1 / 40	44	35	5	16	0
DeriNet.ES	Spanish	151,173	36,935	114,238	98,325	1.3 / 35	0.2 / 5	0.3 / 14	0	0	0	0	0
DeriNet.FA	Persian	43,357	35,745	7,612	0	5.7 / 180	1.5 / 6	3.3 / 114	0	0	0	0	0
DerIvaTario	Italian	8,267	1,787	6,480	5,255	1.3 / 13	0.2 / 5	0.2 / 6	51	26	14	9	0
DErivBase	German	280,775	43,368	237,407	216,982	1.2 / 46	0.1 / 5	0.1 / 13	86	10	5	0	0
DerivBase.Hr	Croatian	99,606	35,289	64,317	50,100	1.5 / 945	0.3 / 21	0.4 / 863	59	30	12	0	0
DerivBase.Ru	Russian	270,473	133,759	136,714	116,037	2.0 / 1142	0.3 / 13	0.4 / 36	62	18	17	3	0
E-CELEX	English	53,103	9,826	43,277	37,951	1.2 / 51	0.2 / 8	0.2 / 33	47	15	13	7	18
EstWordNet	Estonian	988	507	481	22	2.1 / 3	1.0 / 2	1.0 / 3	16	29	8	47	0
EtymWordNet-cat	Catalanian	7,496	4,568	2,928	8	2.6 / 13	1.1 / 4	1.5 / 13	0	0	0	0	0
EtymWordNet-ces	Czech	7,633	5,237	2,396	14	3.2 / 48	1.1 / 4	2.0 / 42	0	0	0	0	0
EtymWordNet-gla	Gaelic	7,524	5,013	2,511	15	3.0 / 15	1.1 / 3	1.8 / 13	0	0	0	0	0
EtymWordNet-pol	Polish	27,797	24,876	2,921	19	9.5 / 75	1.1 / 3	8.3 / 66	0	0	0	0	0
EtymWordNet-por	Portuguese	2,797	1,610	1,187	9	2.4 / 57	1.0 / 3	1.3 / 57	0	0	0	0	0
EtymWordNet-rus	Russian	4,005	3,227	778	15	5.1 / 44	1.0 / 3	4.0 / 44	0	0	0	0	0
EtymWordNet-hbs	Serbo-Croat.	8,033	6,303	1,730	6	4.6 / 108	1.0 / 3	3.6 / 107	0	0	0	0	0
EtymWordNet-swe	Swedish	7,333	4,423	2,910	3	2.5 / 116	1.0 / 3	1.5 / 116	0	0	0	0	0
EtymWordNet-tur	Turkish	7,774	5,837	1,937	11	4.0 / 42	1.1 / 4	2.8 / 22	0	0	0	0	0
FinnWordNet	Finnish	20,035	11,922	8,113	1,461	2.5 / 20	1.0 / 5	1.3 / 14	55	29	15	0	0
G-CELEX	German	53,282	13,553	39,729	34,156	1.3 / 39	0.2 / 11	0.3 / 35	52	17	17	2	12
Nomlex-PT	Portuguese	7,020	4,201	2,819	17	2.5 / 7	1.0 / 1	1.5 / 7	60	0	40	0	0
The M-S Database	English	13,813	7,855	5,958	65	2.3 / 6	1.0 / 1	1.3 / 6	57	0	43	0	0
The Polish WFN	Polish	262,887	189,217	73,670	41,332	3.6 / 214	1.0 / 8	1.1 / 38	0	0	0	0	0
Word Formation Latin	Latin	36,417	32,414	4,003	121	9.1 / 524	1.7 / 6	4.3 / 236	46	29	21	0	4

Table 4.1: Some basic quantitative features of the UDer collection. Column *Relations* counts only tree-shaped derivational relations. Columns *#Nodes*, *Tree depth*, and *Tree outdegree* are presented in average/maximum value format.

for the phenomena, DerivBase.Hr contains special lemmas for negatives but not for reflexives, and DerivBase.Ru includes special lemmas for both into the lexeme set. From the word-formation perspective, the lemmatisation is notable in Word Formation Latin. It lemmatises lexemes based on their meaning and further derivational potential as is shown on the example of lexeme '*gallus*' (in Section 3.4.5).

Relations. The numbers of relations given in Table 4.1 count derivational tree-shaped relations after the harmonisation of each particular resource. It seems that the number decreased, compared to the total number of relations captured in the original resources (see Table 2.1); however the rest of original relations are stored as secondary relations in a less prominent place in the harmonised data. Word Formation Latin is the only resource that explicitly labels 3,882 relations as conversion. Compound lexemes are explicitly labelled and connected with their base lexemes in D-CELEX (3,949), G-CELEX (2,563), Word Formation Latin (1,747), E-CELEX (621), and DeriNet (600). DeriNet also labels 32,479 compound lexemes but does not connect them to their base lexemes.

Families and singletons. After the harmonisation process, the number of derivational families remained the same for resources organising the families in rooted trees. The number increased in other resources because of dividing the original derivational families represented as complete directed subgraphs, weakly connected subgraphs, or derivation trees, cf. Figure 3.6 and Section 3.4.4. Nevertheless, all families resulting from splitting the original family are inter-linked in the harmonised data. These links are stored under the key `was_in_family_with` in the tenth JSON-encoded column, and they connect the roots of the new rooted trees identified in the original family. As for the number of singleton nodes, most of the input resources include singletons in their original versions. The high number of singletons corresponds to the way the resource was built, as already mentioned above. Moreover, their number could increase due to splitting the original family during the harmonisation process.

Tree size. Tree size represents the number of nodes included in the rooted tree (derivational family). Average and maximum tree size of derivational families in the particular harmonised resources are in column *#Nodes* in Table 4.1. The biggest derivational families can be found in resources of Persian, Latin, and Slavic languages not only on average but also in absolute numbers. The biggest tree with 1,638 lexemes is in DeriNet, and it has the root '*dát*' ('*to give*'). The second biggest tree is in DerivBase.Ru with root 'лить' ('*to pour*').

Tree depth and out-degree. Tree depth represents the distance of the furthest node from the tree root. Tree out-degree is the highest number of direct children of a single node. As for the average and maximum tree depths and out-degree, they illustrate a general condition of each harmonised resource. Since NomLex-PT and The Morpho-Semantic Database are lexicons of nominalisations, their tree depth is expected to be just one. However, in the case of Etymological WordNet, small absolute maximum numbers of tree depth but high absolute

maximum numbers of tree out-degrees point to the fact that the families in Etymological WordNet are spread, but most of their lexemes are connected to one 'central' lexeme. These spread families were also observed during manual annotations.

Distribution of part-of-speech categories. Lexemes are assigned part-of-speech tags only in less than a half of the harmonised resources. Word-formation of nouns, adjectives, verbs, and adverbs is captured in CatVar, DeriNet, DerivBase.Ru, D-CELEX, E-CELEX, EstWordNet, and G-CELEX. Démonette, DerIvaTario, DerivBase, DerivBase.HR, FinnWordNet, Word Formation Latin lack adverbs. However, Word Formation Latin includes a few pronouns, auxiliaries, and unspecified lexemes. As already mentioned, NomLex-PT and The Morpho-Semantic Database consist of nominalisations, so they are limited to verbs and nouns only. In all harmonised resources, the part-of-speech tags were unified to the tags that are suggested by the Universal Features annotation scheme (Nivre et al., 2016).

Semantic labels. The meaning of derivational relations is labelled in Démonette, DeriNet, and The Morpho-Semantic Database. The Morpho-Semantic Database assigns labels that come from WordNet semantic types, i.e. Agent, Body, By, Destination, Event, Instrument, Location, Material, Property, Result, State, Undergoer, Uses, and Vehicle. Démonette uses labels obtained based on morpho-syntactic analysis, i.e. ACT, RES, AGF, AGM, and PRP. DeriNet version 2.0 has begun to label derivational relations by labels rooted in comparative semantic concepts proposed by (Bagasheva, 2017), i.e DIMINUTIVE, POSSESSIVE, FEMALE, ITERATIVE, and ASPECT (Ševčíková & Kyjánek, 2019). Since the resources use different labels, and their semantic labelling is anchored in different approaches, the labels have not been harmonised so far. Their harmonisation will require more detailed research into the semantics of derivational relations.

Morphological segmentation. Morphological segmentation appears in CELEXes, Démonette, DeriNet, DerIvaTario, DerivBase, DerivBase.Ru, and Word Formation Latin. The approaches to segmentation vary across the resources, and the morphological segmentation is only partial in all the resources except for CELEXes and DerIvaTario. Démonette and Word Formation Latin segment only those morphemes involved in a particular derivational relation. Since Démonette focuses on suffixation, the segmented morphemes are always suffixes. Word Formation Latin segments suffixes, prefixes, and also interfixes (in compound lexemes). Moreover, allomorphy of prefixes and suffixes is normalised in Démonete and Word Formation Latin. Due to rich allomorphy of Czech morphemes, DeriNet version 2.0 has started the morphological segmentation by root morphemes only. It includes 243,793 lexemes with identified boundaries of their root morphemes. Morphological segmentation in DerivBase and DerivBase.Ru is only potential/theoretical. The segmentation of individual derivational relations is described in the form of derivational rules with normalised allomorphy. It would have to be extracted from the rules. Since the annotation schema for morphological segmentation is designed for direct segmentation of particular string forms of lexemes, so it does not support normalisation of morphemes yet, the

harmonisation of morphological segmentation is intended to be realised in the future version. The segmentation from the original resources is only imported to the tenth JSON-encoded column in the harmonised data.

4.2 Publishing and licensing

4.2.1 Data

The UDer collection version 1.0 is freely available in a single data package in the LINDAT/CLARIAH-CZ repository[4] under the open licenses listed in Table 4.2. The file structure of the package is illustrated in Figure 4.3.

```
UDer-1.0
├── ca-EtymWordNetCA
│   ├── LICENSE
│   ├── README.md
│   └── UDer-1.0-ca-EtymWordNetCA.tsv.gz
├── cs-DeriNet
│   ├── LICENSE
│   ├── README.md
│   └── UDer-1.0-cs-DeriNet.tsv.gz
├── de-DerivBase
│   ├── LICENSE
│   ├── README.md
│   ├── UDer-1.0-de-DErivBase.tsv.gz
│   └── UDer-1.0-de-DErivBase-rules.txt
└── ...
```

Figure 4.3: The UDer collection version 1.0 package structure.

Each harmonised resource is stored in a folder labelled by the language code (ISO 639) and the slightly modified original name (see Table 4.2) of the resource. `README.md` and `LICENSE` files specify more details about the particular resource. They briefly introduce the resource and provide a list of the original authors, recommended citation for referencing the resource, and machine-readable metadata of the harmonised version of the resource. In the case of DerivBase and DerivBase.Ru, the folders also contain descriptions of derivational rules that are labelled in the resources. Since the license terms do not allow the CELEX resources to be redistributed directly, software that harmonises them is provided in their folder. However, the user needs to obtain CELEX from its original provider.

4.2.2 Software

The software developed in this thesis for harmonising all above-described resources and building the UDer collection is available in the GitHub repository[5]. The software architecture was designed as modular so harmonisation of any new

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

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

Resource	Language	UDer name	License in UDer
CatVar	English	CatVar	OSL-1.1
D-CELEX	Dutch	DCelex	–
Démonette	French	Demonette	CC BY-NC-SA 3.0
DeriNet	Czech	DeriNet	CC BY-NC-SA 3.0
DeriNet.ES	Spanish	DeriNetES	CC BY-NC-SA 3.0
DeriNet.FA	Persian	DeriNetFA	CC BY-NC-SA 4.0
DerIvaTario	Italian	DerIvaTario	CC BY-SA 4.0
DErivBase	German	DerivBase	CC BY-SA 3.0
DerivBase.Hr	Croatian	DerivBaseHR	CC BY-SA 3.0
DerivBase.Ru	Russian	DerivBaseRU	Apache 2.0
E-CELEX	English	ECelex	–
EstWordNet	Estonian	EstWordNet	CC BY-SA 3.0
EtymWordNet-cat	Catalanian	EtymWordNetCA	CC BY-SA 3.0
EtymWordNet-ces	Czech	EtymWordNetCS	CC BY-SA 3.0
EtymWordNet-gla	Gaelic	EtymWordNetGD	CC BY-SA 3.0
EtymWordNet-pol	Polish	EtymWordNetPL	CC BY-SA 3.0
EtymWordNet-por	Portuguese	EtymWordNetPT	CC BY-SA 3.0
EtymWordNet-rus	Russian	EtymWordNetRU	CC BY-SA 3.0
EtymWordNet-hbs	Serbo-Croat.	EtymWordNetSH	CC BY-SA 3.0
EtymWordNet-swe	Swedish	EtymWordNetSV	CC BY-SA 3.0
EtymWordNet-tur	Turkish	EtymWordNetTR	CC BY-SA 3.0
FinnWordNet	Finnish	FinnWordNet	CC BY-SA 4.0
G-CELEX	German	GCelex	–
Nomlex-PT	Portuguese	NomLexPT	CC BY-SA 4.0
The M-S Database	English	WordNet	CC BY-NC-SA 3.0
The Polish WFN	Polish	PolishWFN	CC BY-NC-SA 3.0
Word Formation Latin	Latin	WFL	CC BY-NC-SA 4.0

Table 4.2: Technical details about resources included in UDer version 1.0.

resource can be added without affecting the rest of the collection and harmonisation procedure can be easily replaced or improved.

The collection is created by a set of Makefiles and Python scripts that run individual parts of the harmonisation procedure. The whole collection is built by typing `make UDer-collection` to Shell Terminal and possibly specifying a required version of the collection, e.g. `make UDer-collection version=1.0`. An individual harmonised resource can be constructed by specifying the language, the UDer name (see Table 4.2), and the UDer version of the required resource, e.g. `make UDer-resource language=en resource=CatVar version=1.0`. If it is possible, the software automatically downloads the original resource and harmonises it. During the harmonisation, the following packages are used: Virtualenv,[6] NetworkX,[7] SciPy,[8] scikit-learn,[9] NumPy,[10] pandas,[11] matplotlib,[12] textdistance,[13] and xlrd.[14]

[6]https://virtualenv.pypa.io/
[7]https://networkx.github.io/
[8]https://www.scipy.org/
[9]https://scikit-learn.org/stable/
[10]https://numpy.org/
[11]https://pandas.pydata.org/
[12]https://matplotlib.org/
[13]https://pypi.org/project/textdistance/
[14]https://pypi.org/project/xlrd/

4.2.3 Tools

The repository with software for building the UDer collection also contains a web interface for manual annotations developed during the harmonisation project. Technical details were described in Section 3.4.2.

Harmonised resources from the UDer collection can also be processed by other software and tools developed within the DeriNet project, especially Python application interface[15] for data management, and DeriSearch tool[16] for querying and data visualisation (Vidra & Žabokrtský, 2017). Resources from the UDer collection version 1.0 (and older version 0.5) are already available in DeriSearch.

[15] https://github.com/vidraj/derinet/tree/master/tools/data-api/derinet2
[16] http://ufal.mff.cuni.cz/universal-derivations/derisearch

Conclusion

The attention to capturing word-formation of multiple languages in machine-readable resources rose in the last decade. Word-formation has been added to various already existing resources of other phenomena, but many new resources focusing exclusively on word-formation have been developed, too.

Before working on the Universal Derivations project, the individual existing resources had been relatively isolated from each other. Moreover, neither their list nor their description had existed together in one document, which had also been the reason for publishing at least a draft (see Kyjánek, 2018) of the current Chapter 2. The chapter listed the existing resources and documented their similarities and differences.

The resources differed in many technical and linguistic aspects. To allow using the resources in multilingual systems, the harmonisation procedure was proposed and applied to several selected existing resources, described in Chapter 3. DeriNet-like data structure (rooted trees) and file format (textual lexeme-based format consisting of tab-separated columns) were selected as target representation of the harmonised data. Although the procedure involves manual annotations, development of supervised machine learning classifiers, and identifications of the rooted trees based on scores assigned by the classifier, the procedure was developed as modular as possible, so it is easily reusable for other potential resources.

This thesis described the harmonisation of 27 resources that covers 20 mostly European languages. Being inspired by Universal Dependencies that resulted from similar harmonisation task in the field of syntactic treebanks, the final collection of harmonised word-formation resources was named *Universal Derivations* (UDer). The harmonised resources were included in the UDer collection v1.0. Chapter 4 presented the harmonised data included in the collection.

In future work, not only quantitative improvement in the form of new harmonised resources but also qualitative enhancements are planned. There is still space in unifying morphological segmentation and semantic labelling in already harmonised resources, in need of deeper insight into the issues. In addition, further development of individual harmonised resources, e.g. part-of-speech tagging, assigning other new features, enlarging or merging sets of lemmas, etc., would be valuable, too.

References

Agić, Ž., Hovy, D., & Søgaard, A. (2015). If All you Have is a Bit of the Bible: Learning POS Taggers for Truly Low-resource Languages. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Bagasheva, A. (2017). Comparative Semantic Concepts in Affixation. In J. Santana-Lario & S. Valera (Eds.), *Competing Patterns in English Affixation* (pp. 33–65). Peter Lang.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior research methods*, *39*(3), 445–459.

Baranes, M., & Sagot, B. (2014). A Language-independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. In *Proceedings of the Language Resources and Evaluation (LREC-2014)*.

Bonami, O., & Strnadová, J. (2019). Paradigm Structure and Predictability in Derivational Morphology. *Morphology*, *29*, 167–197.

Bray, T. (2017). The JavaScript Object Notation (JSON) Data Interchange Format.

Buzássyová, K. (1974). *Sémantická struktúra slovenských deverbatív*. Veda.

Bybee, J. L. (1985). *Morphology. A Study of the Relation between Meaning and Form* (Vol. 7). John Benjamins Publishing Company.

Chinchor, N. (1992). The Statistical Significance of the MUC-4 Results. In *Proceedings of the 4th conference on Message understanding*. Association for Computational Linguistics.

Chu, Y. J., & Liu, T. H. (1965). On the Shortest Arborescence of a Directed Graph. *Scientia Sinica*, *14*, 1396–1400.

Dokulil, M. (1962). *Tvoření slov v češtině 1: Teorie odvozování slov*. Academia.

Dokulil, M. (1982). K otázce slovnědruhových převodů a přechodů, zvl. transpozice. *Slovo a slovesnost*, *43*(4), 257–271.

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *WALS Online*. Max Planck Institute for Evolutionary Anthropology. https://wals.info/

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

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., & Pallier, C. (2010). The French Lexicon Project: Lexical Decision Data for 38,840 French Words and 38,840 Pseudowords. *Behavior research methods*, *42*(2), 488–496.

Filko, M., Šojat, K., & Štefanec, V. (2019). Redesign of the Croatian derivational lexicon. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*.

Furdík, J. (2004). *Slovenská slovotvorba*. NÁUKA.

Gaussier, É. (1999). Unsupervised learning of derivational morphology from inflectional lexicons. In *Unsupervised Learning in Natural Language Processing*.

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

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX. In *Proceedings of the 7th Python in Science Conference*.

Haspelmath, M., & Sims, A. D. (2010). *Understanding Morphology*. Hodder Education.

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*.

Hladká, Z. (2017). Lexém. In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *CzechEncy – Nový encyklopedický slovník češtiny*. NLN.

Hladká, Z., & Cvrček, V. (2017). Lemma. In P. Karlík, M. Nekula, & J. Pleskalová (Eds.), *CzechEncy – Nový encyklopedický slovník češtiny*. NLN.

Hopcroft, J. E., Motwani, R., & Ullman, J. D. (2000). *Introduction to Automata Theory, Languages and Computability*. Addison-Wesley Longman Publishing.

Horecký, J., Buzássyová, K., Bosák, J., et al. (1989). *Dynamika slovnej zásoby súčasnej slovenčiny*. Veda.

Hwa, R., Resnik, P., Weinberg, A., Cabezas, C., & Kolak, O. (2005). Bootstrapping Parsers via Syntactic Projection across Parallel Texts. *Natural language engineering*, *11*(3), 311–325.

Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone. *New Phytologist*, *11*(2), 37–50.

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

Kastovsky, D. (1982). *Wortbildung und Semantik*. Schwann-Bagel.

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

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

Kyjánek, L. (2018). *Morphological Resources of Derivational Word-Formation Relations* (tech. rep. TR-2018-61). Institute of Formal and Applied Linguistic, Faculty of Mathematics and Physics, Charles University. Prague.

Kyjánek, L., Žabokrtský, Z., Ševčíková, M., & Vidra, J. (2019a). Universal Derivations Kickoff: A Collection of Harmonized Derivational Resources for Eleven Languages. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*.

Lango, M., Žabokrtský, Z., & Ševčíková, M. (2020). Semi-automatic Construction of Word-formation Networks. *Language Resources and Evaluation*, 1–30. https://doi.org/https://doi.org/10.1007/s10579-019-09484-2

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

Lindén, K., Niemi, J., & Hyvärinen, M. (2012). Extending and Updating the Finnish WordNet. In D. Santos, K. Lindén, & W. Ng'ang'a (Eds.), *Shall We Play the Festschrift Game?* (pp. 67–98). Springer.

Lipka, L. (1975). Prolegomena to 'Prolegomena to a Theory of Word-Formation'. In E. F. K. Koerner (Ed.), *The Transformational-Generative Paradigm and Modern Linguistic Theory* (pp. 175–184). John Benjamins Publishing.

Litta, E., Passarotti, M., & Mambrini, F. (2019). The Treatment of Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*.

Matoušek, J., & Nešetřil, J. (2009). *Invitation to Discrete Mathematics*. Oxford University Press.

Matthews, P. H. (1991). *Morphology*. Cambridge University Press.

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).

Mititelu, V. B. (2012). Adding Morpho-semantic Relations to the Romanian WordNet. In *Proceedings of the Language Resources and Evaluation (LREC-2012)*.

Namer, F., & Hathout, N. (2019). ParaDis and Démonette: From Theory to Resources for Derivational Paradigms. In *Proceedings of the 2nd Workshop on Resources and Tools for Derivational Morphology*.

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*.

Olsen, S. (2014). Delineating derivation and compounding. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Derivational Morphology* (pp. 26–49). Oxford University Press.

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*. Association for Computational Linguistics.

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*.

Razímová Ševčíková, M., & Žabokrtský, Z. (2006). Systematic Parameterized Description of Pro-forms in the Prague Dependency Treebank 2.0. In *Fifth Workshop on Treebanks and Linguistic Theories*.

Re. (2020). In *Cambridge Dictionary*. Cambridge University Press. Retrieved March 20, 2020, from https://dictionary.cambridge.org/dictionary/english/re?q=re-

Rosa, R. (2018). *Discovering the Structure of Natural Language Sentences by Semi-supervised Methods* (Doctoral dissertation). Charles University, Faculty of Mathematics and Physics.

Rosa, R., Zeman, D., Mareček, D., & Žabokrtský, Z. (2017). Slavic forest, Norwegian wood. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*.

Ševčíková, M., & Kyjánek, L. (2019). Introducing semantic labels into the derinet network. *Journal of Linguistics/Jazykovedný casopis*, *70*(2), 412–423.

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

Simons, G. F., Eberhard, D. M., & Fennig, C. D. (Eds.). (2020). *Ethnologue: Languages of the World* (Twenty-third). SIL international. https://www.ethnologue.com/

Šojat, K., & Srebačić, M. (2014). Morphosemantic Relations between Verbs in Croatian WordNet. In *Proceedings of the Seventh Global WordNet Conference*.

Steiner, P. (2019). Augmenting a German Morphological Database by Data-Intense Methods. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Štekauer, P. (1996). *A Theory of Conversion in English*. Peter Lang Verlag.

Štekauer, P. (2005). Onomasiological Approach to Word-Formation. In P. Štekauer & R. Lieber (Eds.), *Handbook of Word-Formation* (pp. 207–232). Springer.

Štekauer, P., Valera, S., & Körtvélyessy, L. (2012). *Word-Formation in the World's Languages: A Typological Survey*. Cambridge University Press.

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

ten Hacken, P. (2014). Delineating derivation and inflection. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Derivational Morphology* (pp. 10–25). Oxford University Press.

Tiberius, C., & Niestadt, J. (2010). The ANW: An Online Dutch Dictionary. In *Proceedings of the XIV Euralex International Congress*.

van Marle, J. (1985). *On the Paradigmatic Dimension of Morphological Creativity*. Walter de Gruyter GmbH & Co KG.

Van Rijsbergen, C. J. (1979). *Information retrieval*. Butterworths, London.

Vidra, J., Žabokrtský, Z., Ševčíková, M., & Kyjánek, L. (2019). DeriNet 2.0: Towards an All-in-One Word-Formation Resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research*.

Yarowsky, D., Ngai, G., & Wicentowski, R. (2001). Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora. In *Proceedings of the First International Conference on Human Language Technology Research*. Association for Computational Linguistics.

Zeller, B., Padó, S., & Šnajder, J. (2014). Towards Semantic Validation of a Derivational Lexicon. In *Proceedings of COLING 2014*.

Zeman, D., & Resnik, P. (2008). Cross-language Parser Adaptation between Related Languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Żmigrodzki, P. et al. (2007). Wielki słownik języka polskiego PAN. *Instytut Języka Polskiego PAN, Kraków*. https://wsjp.pl/

Language resources and tools

Baayen, H. R., Piepenbrock, R., & Gulikers, L. (1995). CELEX2 [Linguistic Data Consortium, Catalogue No. LDC96L14].

Balvet, A., Barque, L., & Marín, R. (2010). Building a Lexicon of French Deverbal Nouns from a Semantically Annotated Corpus. In *Proceedings of the Language Resources and Evaluation (LREC-2010)*.

Bosch, A. v. d., Busser, B., Canisius, S., & Daelemans, W. (2007). An Efficient Memory-based Morphosyntactic Tagger and Parser for Dutch. *LOT Occasional Series*, *7*, 191–206.

Buchholz, S., & Marsi, E. (2006). CoNLL-X Shared Task on Multilingual Dependency Parsing. In *Proceedings of the 10th Conference on Computational Natural Language Learning*.

De Paiva, V., Real, L., Rademaker, A., & De Melo, G. (2014). NomLex-PT: A Lexicon of Portuguese Nominalizations. In *Proceedings of the Language Resources and Evaluation (LREC-2014)*.

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

Dobrovoljc, K., Krek, S., Holozan, P., Erjavec, T., Romih, M., Arhar Holdt, Š., Čibej, J., Krsnik, L., & Robnik-Šikonja, M. (2019). Morphological Lexicon Sloleks 2.0 [Slovenian language resource repository CLARIN.SI]. http://hdl.handle.net/11356/1230

Faryad, J. (2018). *Identifikace derivačních vztahů ve španělštině* (tech. rep. TR-2018-63). Institute of Formal and Applied Linguistic, Faculty of Mathematics and Physics, Charles University. Prague.

Fellbaum, C., Osherson, A., & Clark, P. E. (2007). Putting Semantics into WordNet's "morphosemantic"links. In *Language and Technology Conference*. Springer.

Gerard, d. M. (2014). Etymological Wordnet: Tracing The History of Words. In *Proceedings of the Language Resources and Evaluation (LREC-2014)*.

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*. Association for Computational Linguistics.

Haghdoost, H., Ansari, E., Žabokrtský, Z., & Nikravesh, M. (2019). Building a Morphological Network for Persian on Top of a Morpheme-Segmented Lexicon. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*.

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas,

P., Panevová, J., Poláková, L., ... Žabokrtský, Z. (2018). Prague Dependency Treebank 3.5 [Digital library LINDAT/CLARIN at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University]. http://hdl.handle.net/11234/1-2621

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

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* (tech. rep. arXiv: 1005.3902).

Hathout, N., & Namer, F. (2014). Démonette, a French Derivational Morpho-Semantic Network. *Linguistic Issues in Language Technology*, *11*, 125–162.

Hathout, N., Namer, F., & Dal, G. (2002). An Experimental Constructional Database: The MorTAL Project. *Many Morphologies*, 178–209.

Kahusk, N., Kerner, K., & Vider, K. (2010). Enriching Estonian WordNet with Derivations and Semantic Relations. In *Baltic HLT*.

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

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.

Kyjánek, L., Žabokrtský, Z., Vidra, J., & Ševčíková, M. (2019b). Universal Derivations v0.5 [LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University]. http://hdl.handle.net/11234/1-3041

Kyjánek, L., Žabokrtský, Z., Vidra, J., & Ševčíková, M. (2020). Universal Derivations v1.0 [LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University]. http://hdl.handle.net/11234/1-3236

Lango, M., Ševčíková, M., & Žabokrtský, Z. (2018). Semi-Automatic Construction of Word-Formation Networks (for Polish and Spanish). In *Proceedings of the Language Resources and Evaluation (LREC-2018)*.

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

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)*.

Macleod, C., Grishman, R., Meyers, A., Barrett, L., & Reeves, R. (1998). Nomlex: A Lexicon of Nominalizations. In *Proceedings of EURALEX*.

Mailhot, H., Wilson, M. A., Macoir, J., Deacon, H. S., & Sánchez-Gutiérrez, C. H. (2019). MorphoLex-FR: A Derivational Morphological Database for 38,840 French Words. *Behavior research methods*, 1–18.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In

Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations.

McDonald, R., Nivre, J., Quirmbach-Brundage, Y., Goldberg, Y., Das, D., Ganchev, K., Hall, K., Petrov, S., Zhang, H., Oscar, T., Claudia, B., Núria, C. B., & Jungmee, L. (2013). Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*.

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.

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

Nivre, J., De Marneffe, M. C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Reut, T., & Daniel, Z. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Language Resources and Evaluation (LREC-2016)*.

Paiva, V. d., Rademaker, A., & Melo, G. d. (2012). OpenWordNet-PT: An Open Brazilian WordNet for Reasoning. In *COLING 2012*.

Pala, K., & Šmerk, P. (2015). Derivancze—Derivational Analyzer of Czech. In *International Conference on Text, Speech, and Dialogue*. Springer.

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

Piasecki, M., Szpakowicz, S., & Broda, B. (2009). *A Wordnet from the Ground Up*. Oficyna Wydawnicza Politechniki Wroclawskiej.

Raffaelli, I., Tadić, M., Bekavac, B., & Agić, Ž. (2008). Building Croatian WordNet. In *Fourth Global WordNet Conference (GWC 2008)*.

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, H. S., & Wilson, M. A. (2018). MorphoLex: A Derivational Morphological Database for 70,000 English Words. *Behavior research methods*, *50*(4), 1568–1580.

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 Workshop on Resources and Tools for Derivational Morphology*.

Šnajder, J. (2014). DerivBase.hr: A High-Coverage Derivational Morphology Resource for Croatian. In *Proceedings of the Language Resources and Evaluation (LREC-2014)*.

Šojat, K., Srebačić, M., Pavelić, T., & Tadić, M. (2014). CroDeriV: A New Resource for Processing Croatian Morphology. In *Proceedings of the Language Resources and Evaluation (LREC-2014)*.

Steiner, P. (2016). Refurbishing a Morphological Database for German. In *Proceedings of the Language Resources and Evaluation (LREC-2016)*.

Straka, M., & Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*.

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*.

Vidra, J., & Žabokrtský, Z. (2017). Online Software Components for Accessing Derivational Networks. In *Proceedings of the Workshop on Resources and Tools for Derivational Morphology*.

Vidra, J., Žabokrtský, Z., Kyjánek, L., Ševčíková, M., & Dohnalová, Š. (2019). DeriNet 2.0 [LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University]. http://hdl.handle.net/11234/1-2995

Vitas, D., & Krstev, C. (2005). Derivational Morphology in an E-Dictionary of Serbian. In *Proceedings of 2nd Language & Technology Conference*.

Vodolazsky, D. (2020). DerivBase.Ru: A Derivational Morphology Resource for Russian. In *Proceedings of the Language Resources and Evaluation (LREC-2020)*.

Zakharov, V. (2013). Corpora of the Russian Language. In *International conference on text, speech and dialogue*. Springer.

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*.

Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z., & Hajič, J. (2014). HamleDT: Harmonized Multi-Language Dependency Treebank. *Language Resources and Evaluation*, *48*(4), 601–637.

Zeman, D., Nivre, J., Abrams, M., Aepli, N., Agić, Ž., Ahrenberg, L., Aleksandravičiūtė, G., Antonsen, L., Aplonova, K., Aranzabe, M. J., Arutie, G., Asahara, M., Ateyah, L., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Ballesteros, M., Banerjee, E., ... Zhu, H. (2019). Universal Dependencies 2.5 [LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University]. http://hdl.handle.net/11234/1-3105