M A T E M A T I C K O - F Y Z I K Á L N Í F A K U L T A

P R A H A

Towards Universal Segmentations:

Survey of Existing Morphosegmentation Resources

Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Madga Ševčíková,

Jonáš Vidra, Zdeněk Žabokrtský

ÚFAL Technical Report
TR-2021-69

ISSN 1214-5521

U N I V E R S I T A S C A R O L I N A P R A G E N S I S

Copies of ÚFAL Technical Reports can be ordered from:

Institute of Formal and Applied Linguistics (ÚFAL MFF UK)
Faculty of Mathematics and Physics, Charles University
Malostranské nám. 25, CZ-11800 Prague 1
Czechia

or can be obtained via the Web: http://ufal.mff.cuni.cz/techrep

Towards Universal Segmentations: Survey of Existing Morphosegmentation Resources

ÚFAL Technical Report

Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda,
Madga Ševčíková, Jonáš Vidra, Zdeněk Žabokrtský

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

December 2021

Abstract

This study presents a preliminary overview of 18 resources which contain morphematic segmentation of word forms or of lemmas in various languages, or from which such segmentation could be derived.

Contents

Chapter 1

Introduction

1.1 Motivation for Harmonisation Efforts

The availability of data that can easily be read automatically is growing increasingly important to facilitate data-driven research. However, linguistic resources may differ from each other in several factors. For example, they may be tailored to the particular language(s) they contain with respect to annotation (e.g. tagsets), they may differ in what phenomena they choose to handle or ignore, they may choose one or the other format, and so on.

This problem has been recognized and addressed in certain paradigms, such as for syntactic treebanks and semantic WordNets, and the harmonzation of these resources into Universal POS Tagging and Universal Dependencies has arguably encouraged large-scale monolingual as well as multilingual efforts in problems such as tagging and parsing, both directly as well as indirectly as preprocessing for another task. However, such an effort is still missing for morphological segmentation; despite the fact that such segmentation is required for understanding word formation and inflection in most of the languages in the world, we lack a unified data resource that we can look to in order to access segmentation information about a possibly unknown language. Currently, this information can only be found in disparate resources, within and across languages, focusing on slightly different phenomena, both structurally and principally different from each other. This is the problem that this work seeks to address, given the clear and historically validated benefits of having simple, harmonized resources for many languages for a given task.

The aim of this document is to describe the diversity of existing data resources in which segmentation information is stored directly in a formalized way, or from which such information could be derived with high reliability. Creating a collection of harmonized versions of the resources, possibly under the name “Universal Segmentations”, will be the natural next step.

1.2 Basic Notions

1.2.1 Morpheme, morph, and allomorphy

A morpheme is defined as a grapheme sequence associated with a particular meaning that cannot be further subdivided, i.e., it is considered the smallest linguistic sign.[1] Morphemes are smaller than words, or identical with them; cf. *chair* as an example of a one-morpheme word, *chair-s* or *play-er* as two-morpheme words, three morphemes in *play-er-s* or *en-rich-ment*, four morphemes in *dis-taste-ful-ly*, etc.

Two oppositions can be combined to classify morphemes. On the one hand, free morphemes, which can be used as separate words, are differentiated from bound morphemes, which can only be used in combination with another morpheme. On the other hand, lexical vs grammatical morphemes are distinguished based on the meaning they convey. While lexical morphemes have (more or less general) lexical meanings on their own, grammatical morphemes bear inflectional meanings (such as number or tense). Each type obtained by combining these values owes a short comment:

1. free lexical morphemes ("content words") are roots and stems; e.g. *book*, *book-s*, *play*, *play-er-s*;
2. bound lexical morphemes ("derivational morphemes") are used to form a new word; they change the meaning and/or the part-of-speech category of words (e.g. *book* → *book-ish*, *dark* → *dark-ness*, *like*.v → *dis-like*.v); they have specialized meanings, added in succession (*uč-i-t* → *uč-i-tel* → *uč-i-tel-ka*); they occur before inflectional morphemes (*play-er-s*);
3. free grammatical morphemes ("function words") are used to link word forms to syntactic phrases, clauses and sentences; the delimitation of these words is heavily depending on the theoretical framework applied cf. *in a book*, *but*, *that*, *them*);
4. bound grammatical morphemes ("inflectional morphemes") are used to create word forms of a given lexeme with the same lexical meaning but different inflections (e.g. *play-s*, *play-ed*, *play-ing*, *play-er-s*, *book-s*, *dis-lik-ed*); a single morpheme can express multiple inflectional meanings ("portmanteau morpheme"; cf. *-s* conveying the meanings *+3rd person +sg +present* in English verbs); inflectional morphemes occur outside derivational morphemes.

Based on the position with respect to the root, bound morphemes are distinguished into prefixes (in front of the root), suffixes (behind the root; final inflectional suffixes are called endings), circumfixes (around the root), and infixes (within the root).

Analogously to other (more complex) linguistic signs, assymetries between forms and meanings are also documented in morphemes. A particular form

[1] The term also applies to a sequence of phonemes when dealing with speech, which is, though, not the case of the present paper.

can be associated with more than one meaning – if this is interpreted as different form-meaning pairings where the form is identical (by chance), the terms "homonymy" or "polyfunctionality" apply, which are preferred to "polysemy" (and thus to the interpretation that it is a single form with multiple meanings) in recent accounts; cf. the derivational suffix *ka* used to form diminutives as in *skříň* 'cupboard' → *skříňka* 'small cupboard', while in *učitel* 'teacher' → *učitelka* 'female teacher' it coins the feminine agent noun, but it occurs also in the instrument noun *žehlička* 'iron' motivated by the verb *žehlit* 'to iron'. In contrast, a particular meaning can be expressed by multiple, formally different morphemes ("synonymy"), cf. the derivational suffixes *-ka, -yně, -ice, -ová* all used in Czech to coin feminine counterparts of masculine animate nouns.

Morphemes are assumed to repeat within sets of words, with so-called cranberry (unique) morphemes being the exception. In individual words, morphemes are represented by particular forms (morphs). The relationship between two or more different morphs of a single morpheme is called allomorphy (Haspelmath and Sims, 2010, pp. 22–26); cf. the allomorphs *br–bír–běr–bor* in words derivationally related to the Czech verb *br-á-t* 'to take' (in *vy-bír-a-t* 'to choose', *vý-běr* 'choice', and *vý-bor* 'committee', etc.). Allomorphs are assumed to occur in different contexts in complementary distribution (Aronoff, 2019).

1.2.2 Morphological segmentation

In general, words are expected to be fully decomposable into morphemes. Nevertheless, one can easily find words whose simple splitting yields strings that do not match any morph. This may happen when the words were made up of morphs that were hard to pronounce in succession, so that a simplification was necessary (cf. *český* 'Czech' ← *češ+ský*, *obléci* 'to dress up' ← *ob+vléci*). There can be also strings (e.g. *l* in the Czech verb *kres-l-i-t* 'draw') delimited that are not easily assigned a meaning in the synchronic perspective.

The task of decomposing a word into a sequence of minimal meaning-bearing units is called morphological segmentation in the present paper, but alternative names are also used.

The basic linguistic principle of delimiting morphemes on the basis of their recurrence in words is challenged by allomorphy, as exemplified above, but also by other issues, some of them related specifically to the morpheme position in the word structure.

For instance, in Czech there is a limited number of prefixes used in words of a particular part of speech (for instance, no more than 20 prefixes are attested in native verbs), showing relatively regular patterns when crossing the part-of-speech boundaries (e.g., vowel lengthening in verb-to-noun derivation *vy-br-a-t* 'to choose' > *vý-běr* 'choice'). The number of prefixes is limited to one or two in most words, concatenation of more prefixes being rare (*z-ne-pří-jemň-ova-t* 'to make unpleasant'). There are some, rather textbook examples of words that can be analyzed as containing a prefix or not; they are, though, disambiguated by different derivational parents in DeriNet (*proud-i-t* 'to stream' < *proud* 'stream' vs. *pro-ud-i-t* 'to smoke thoroughly' < *ud-i-t* 'to smoke (meat)').

Determining morpheme boundaries within the suffix part of Czech words is even more intricate, in particular, because it often consists of multiple segments, which can be delimited differently based on different analogies. For instance, the thematic suffix *ova* is delimited in *kup-ova-t* 'to buy.imperf' in contrast to *koup-i-t* 'to buy.perf', but if propagated to *kup-ová-va-t* 'to buy.imperf-iter', the lengthened variant obtained (*ová*) cannot be, though, found in other iterative verbs. An alternative, more subtle segmentation (*kup-ov-a-t* > *kup-ov-áv-a-t*) seems to be justified with regard to other iteratives (*plav-a-t* 'to swim.imperf' > *plav-áv-a-t* 'to swim.imperf-iter') but may be questioned by other formations.

Morphological segmentation as identification of all morphemes within the word structure (e.g., *lod'-k-a* 'small boat' is cut into the root morpheme *lod'*, the derivational suffix *k* and the inflectional suffix = ending *a*) can be distinguished from delimiting morphemes that distinguish a word from an immediately simpler word (e.g. *lod'-ka* 'small boat' from *lod'* 'boat').

1.3 Resource selection

The main criteria for including a particular resource into our study was availability of the data in an electronic form, existence of sufficient documentation of the data (e.g. in the form of a conference paper), and reasonable size of the data (toy sets containing e.g. only a few dozens segmented words are not included). Table 1.1 lists the resources surveyed in the next chapter.

We are aware of several other resources which are related to morphosegmentation task but which we have not included into our study, either because they do not contain segmented words or lemmas (but contain e.g. morpheme inventories), or because they are published under too restrictive licenses, or because they are simply not available in an electronic form.

Abbrev. name	Original name, version	Languages	License	Reference
CELEX	CELEX Lexical Database 2.0	Dutch, English, German	EULA (research use only)	Baayen et al. (1995)
CroDeriV	CroDeriV 1.0	Croatian	CC BY-SA-3.0	Šojat et al. (2014)
Démonette	Démonette-1.2	French	CC BY-NC-SA 3.0	Hathout and Namer (2014)
DerIvaTario	DerIvaTario	Italian	CC BY-SA 4.0	Talamo et al. (2016)
DerivBaseDE	DErivBase 2.0	German	CC BY-SA 3.0	Zeller et al. (2013)
DerivBaseRU	DerivBase.Ru 1.0	Russian	Apache-2.0	Vodolazsky (2020)
DeriNet	DeriNet 2.1	Czech	CC BY-NC-SA 3.0	Vidra et al. (2021)
MorphoDictKE	Dictionary of Morphemes of Russian	Russian	All rights reserved	Kuznetsova and Efremova (1986)
Échantinom	Échantinom	French	CC BY 4.0	Bonami and Tribout (2021)
KCIS	KCIS Resources	Marathi, Hindi, Malayalam, Kannada, Bangla	EULA (research use only)	(see Sec. 2.10)
MorphoChallenge	MorphoChallenge 2005, 2007-2010	English, Finnish, German, Turkish, (Arabic)	unspecified	Kurimo et al. (2010)
MorphoLex	MorphoLex, MorphoLex-FR	English and French	CC BY 4.0	Sánchez-Gutiérrez et al. (2018); Mailhot et al. (2020)
MorphyNet	MorphyNet, v1	15 languages	CC BY-SA 3.0	Batsuren et al. (2021)
PerSegLex	Persian Morphologically Segmented Lexicon 0.5	Persian	CC BY-NC-SA 4.0	Ansari et al. (2019)
Tikhonov's dictionary	Morphemic-spelling dictionary of the Russian language. Russian morphemics	Russian	All rights reserved	Tikhonov (1996)
UniMorph	UniMorph 3.0	141 languages	CC BY-SA 3.0 for most languages	McCarthy et al. (2020)
Uniparser	Uniparser morphological analyzer	7 languages	MIT	Arkhangelskiy et al. (2012)
WFL	Word Formation Latin 1.1	Latin	CC BY-NC-SA 4.0	Litta et al. (2016)

Table 1.1: Overview of morphological resources.

Chapter 2

Overview of existing resources

2.1 CELEX

CELEX 2 (Baayen et al., 1995) is a general phonological and morphological resource for German, Dutch and English, which, among other annotations, contains information about morphological segmentation of lemmas. Selected lemmas are divided into both their immediate constituent stems and affixes, and into individual morphemes, with indications of hierarchy that can be used to infer derivational series. In all, the German version lists 51,728 segmented lexemes, the Dutch version 125,611 and the English 52,447.

Although the resource also lists inflected word forms together with their morphological tags, segmentations are not given for these. For the lemmas, the segmentation given is generally complete, but some stems may be left unsegmented (e.g. prefixes such as in *bestellen* (“to order”) are usually not delimited).

As seen in Figure 2.1, neither the stems nor the morphemes need to correspond 1:1 to parts of the segmented word form, as the morphemes are listed as canonical allomorphs and stems often being in the form of lexemes related by word-formation.

Some morphemes listed in the segmentation may be completely elided from the word form due to changes through word formation – see again Figure 2.1, where the “-e” suffix of the first base does not correspond to any phoneme or grapheme of the word form.

The CELEX data may specify multiple alternative segmentation for a single lexeme by including the relevant columns multiple times on the same line in the data file.

```
  22845\Leuchtbombe\1\C\1\Y\Y\Y\Leuchte+Bombe\NN\N\N\N\
(((licht)[A],(e)[N|A.])[N],(Bombe)[N])[N]\Y\N\N\N\S3/P3\N
```

Figure 2.1: An example CELEX annotation of the German lexeme *Leuchtbombe* (“flash bomb”), broken into two lines. The bold parts are, in order: the lemma, the stem segmentation, and the hierarchical segmentation.

2.2 CroDeriV

CroDeriV (Šojat et al., 2014) is a lexical resource of derivational morphology for Croatian. In its first version, which is available in a searchable database on the web page of the project,[1] it includes manual morphological segmentation of more than 14,400 lemmas extracted from the Croatian morphological lexicon. All the lemmas are verbs, except for two nouns. The lemmas are segmented into morphs labelled as "Stem", "Prefix", "Suffix", or "Ending". A zero morpheme is used in two cases: the nouns *pis-ar-0* ("scribe") and *pis-ač-0* ("writer").

Allomorphy is handled in newer versions of the resource. However, these versions have not been released yet.[2] With the consent of the original authors, we present the first version of the data crawled from the web page. Due to the crawling procedure, the data is presented in the form of HTML code, see Figure 2.2.

```
<tr class="">
    <td class="text-left col-md-3 forma">
        <a href="/Entry/Details/116">barikadirati</a>
    </td>
    <td class="text-center col-md-5">
        <a class="Stem" [...]>barikad</a>
        <span class="Suffix">ir</span>
        <span class="Suffix">a</span>
        <span class="Ending">ti</span>
    </td>
    <td class="text-right col-md-4">
        <div class="btn-group">
            <a class="btn btn-info" href="/Croderiv/Details/116">Details</a>
        </div>
    </td>
</tr>
```

Figure 2.2: A data sample from CroDeriv

2.3 Démonette

Démonette (Hathout and Namer, 2014) is a morphosemantic lexical database that is automatically built from the parsing system DériF (Namer, 2009), the Morphonette network (Hathout, 2011), and Verbaction (Tanguy and Hathout, 2002; Hathout et al., 2002). It contains a total of 22,570 unique lemmas, taken from the TLFNome lexicon[3] and Verbaction. Each entry has a pair of morphologically related words (lemmas), and defines the first with respect to the second; see Figure 2.3. In addition, it also marks for each of them with a GRACE POS tag (Rajman et al., 1997), a (single) suffix (if any), a conversion process (if any),

[1] http://croderiv.ffzg.hr/

[2] They promise significant enrichment in terms of (i) lemmas of other part-of-speech categories including their morphological segmentation, (ii) derivational relations between lemmas, and (iii) labelling of semantics in the relations, cf. Filko et al. (2019).

[3] www.cnrtl.fr/lexiques/morphalou/

and, sometimes, a root. The dataset is built for words containing a select set of 32 suffixes as well as conversion; allomorphy is rare. A lemma may come from more than one resource and therefore have more than one segmentation.

```
"abaissement","tlfnome","abaisser","tlfnome","Ncms","tlfnome","Vmn----","tlfnome",
↪  "simple","derif","suf","ment","derif",,,,"@RES","demonette","@","demonette","résultat
↪  de abaisser","derif","résultat de @","demonette","descendant",
↪  "demonette","abaiss","derif",,,"derif"

"abaissement","tlfnome","abaisser","tlfnome","Ncms","tlfnome","Vmn----","tlfnome",
↪  "simple","demonette","suf","ment","demonette",,,"demonette","@ACT","demonette","@",
↪  "demonette", "action de abaisser","demonette", "action de @","demonette","descendant",
↪  "demonette","abaisse","demonette",,,"verbaction"

"abandon","tlfnome","abandonner","tlfnome","Ncms","tlfnome","Vmn----","tlfnome",
↪  "simple","demonette","conv",,"demonette","conv",,"demonette","@ACT","demonette","@",
↪  "demonette", "action de abandonner","demonette","action de @", "demonette",,
↪  "demonette",,"demonette",,,"verbaction"
```

Figure 2.3: A data sample from Démonette. The first two entries differ in the source of the derivation (Dérif vs. Verbaction); we see a conversion process marked in *abandon*.

2.4 DeriNet

DeriNet 2.1 (Vidra et al., 2021) is a Czech database of word-formation relations. Its lemmaset consists of 1 039 012 lemmas extracted from the MorfFlex (Hajič et al., 2020) dictionary together with part-of-speech categories conforming to the Universal POS tagset (Petrov et al., 2012). Apart from the word-formation relations, it also contains other additional annotations, such as automatically induced segmentation to morphs. Morphs are further tagged as Prefix/Root/Suffix.

The DeriNet project also published manually annotated morphological segmentation data in its source-control repository – 3,000 lemmas[4] and 2,000 form-lemma pairs[5] completely segmented to morphs. The data were sampled in multiple ways by dividing the 3,000 and 2,000 items into equally-sized parts and sampling each part in a different manner - uniformly, by corpus frequency and by corpus frequency classes (words were separated into groups by logarithm of corpus frequency and sampled uniformly from each group).

2.5 DerIvaTario

DerIvaTario (Talamo et al., 2016) is a morphemic segmentation dataset containing 11,000 manually annotated Italian derivatives, sampled from the CoLFIS corpus (Bertinetto et al., 2005).

[4] https://github.com/vidraj/derinet/tree/master/data/annotations/cs/2021_05_complete_morphseg_bandsampling

[5] https://github.com/vidraj/derinet/tree/master/data/annotations/cs/2021_11_complete_morphseg-forms_bandsampling

Each entry contains a lemma with its CoLFIS ID, its base, and a complete list of affixes in order of derivation, shown in Figure 2.4. The base is further marked with its type from a set of 9 possible labels, including suppletion, verbal theme (indicating a deverbal base), or if the base is unrecoverable (i.e. the word shows a certain affix, but the root is not interpretable synchronously). Affix fields contain four parts: the affix, the allomorph, morphotactic transparency, and morphosemantic transparency, respectively. There may also be a "-P" or "-G" flag; the former indicates that the current morphological process was simultaneous with another process (which would also be marked with the flag), and the latter indicates that the order of this morphological process was undecidable relative to another.

The allomorph does not manifest the phonological processes the morpheme may have gone through in the given lemma. (See Figure 2.4 for examples.) Fields marking a conversion process, which is a zero-morpheme lacking a span in the lemma, are marked with a label indicating its type, e.g. "N_V" indicates noun-to-verb conversion, in the appropriate position given the ordering of affixes. Intra-word hyphens are not treated specially; however, in words marked as compounds, they may be assumed to delineate root-boundaries. Homonymy between morphemes is explicitly marked by adding an index to each homonymous morpheme; the indexing is explained in documentation. (Talamo et al., 2016). See Figure 2.4 for examples of all the above.

```
6809,AMBIENTALISMO,AMBIENTE:root,ALE:ale:mt1:ms1,ISMO:ismo:mt1:ms1;;;
3940;ABBASSAMENTO;BASSO:root;ACons:ad:mt2:ms1-P;CONVERSION:A_V-P;MENTO:mento:mt1:ms1;;;
3951;ABBATTIMENTO;BATTERE:vrb_th;ACons:ad:mt2:ms2b;MENTO:mento:mt4:ms1;;;;
3958;ABBELLIMENTO;BELLO:root;ACons:ad:mt2:ms1-P;CONVERSION:A_V-P;MENTO:mento:mt1:ms1;;;
3969;ABBIGLIAMENTO;ABBIGLIARE:vrb_th;MENTO:mento:mt1:ms2b;;;;;
3972;ABBINAMENTO;ABBINARE:vrb_th;MENTO:mento:mt1:ms1;;;;;
4774;ADDOMESTICABILE;BASELESS:unrec;ICO:ico:mt8:ms3a;ACons:ad:mt1:ms2a-P;
CONVERSION:A_V-P;BILE:bile:mt1:ms1;;
7841;ANTI-COMUNISTA;COMUNE:root;ISMO:ismo:mt1:ms2a;ANTI:anti:mt1:ms1;ISTA:ista:mt6:ms1;;;
63412;POST-CUBISTA;POST-CUBISMO:root:neocl_cmp;
ISTA:ista:mt6:ms1;;;;;
26113;DEVIARE;VIA:root;1DE:de:mt1:ms2a-P;CONVERSION:N_V-P;;;;
41473;INDEFINITO;FINIRE:vrb_th;2DE:de:mt1:ms2b;CONVERSION:V_A;1IN:in:mt1:ms1;;;
```

Figure 2.4: A data sample from DerIvaTario. Note that marked allomorphs do not record phonological processes. For example, the allomorph "ad" undergoes a doubling process in the "ABBATTIMENTO"; similarly, "ale" is stripped of its final vowel in "AMBIENTALISMO". We see that the base is "unrecognized" in "ADDOMESTICABILE"; although the it's clear that the affixes "ad", "ico+bile" appear in this lemma, "domest" cannot be interpreted synchronously in Italian. "ANTI-COMUNISTA" is an instance of the highly frequent overlapping "ismo+ista" affixes. Also note that "post" in "POST-CUBISTA" is not marked as an affix; rather, lemma is marked as a neoclassical compound, and the hyphen may here be assumed to be a separator. Finally, "DEVIARE" and "INDEFINITO" both show the affix "de"; however, the first, marked "1DE", is a causative polyseme, while the second, marked "2DE", is the inversive polyseme.

2.6 DErivBase

DErivBase v2 (Zeller et al., 2013) is a large-coverage lexicon of derivationally related lexemes for German. These derivational relations were identified on the basis of more than 190 (derivational) rules extracted from German reference grammar books. The rules are based on derivational changes (in the form of string substitutions) that happen when deriving a lexeme from its base lexeme. Lemmas of the lexicon (more than 280 thousand) were extracted from a large German web corpus SDeWAC. The homonymy of lemmas is partly handled by assigning part-of-speech categories (`N`: noun, `A`: adjective, `V`: verb) and gender for some nouns (`n`: neuter, `m`: masculine, `f`: feminine).

As for the morphological segmentation of individual lemmas into morphs, only a partial segmentation can be inferred from a reverse application of the derivational rules to lemmas. These rules also include labels for individual affixes, namely, `sfx` or `dsfx` for suffixes, and `pfx` or `dpfx` for prefixes. Allomorphy is not handled in the resource. The data is distributed in separate files containing documentation of all derivational rules (cf. Figure 2.5) and a list of subsequent derivations of two lemmas where a relation is always labelled by a derivational rule (cf. Figure 2.6).

```
 -- Bäcker -> Bäckerei, Rüpel -> Rüpelei, Träumer -> Träumerei, Türke -> Türkei
dNN01 = dPattern "dNN01"
  (sfx "ei" & try (dsfx "e")) mNouns fNouns

-- Bäcker -> Bäckerin, Idiot -> Idiotin, Türke -> Türkin, Vanille -> Vanillin
dNN02 = dPattern "dNN02"
  (sfx "in" & try (dsfx "e")) nouns nouns

-- Dieb -> Dieberei, Sklave -> Sklaverei, Abgott -> Abgötterei, Schwein -> Schweinerei
dNN03 = dPattern "dNN03"
  (sfx "erei" & opt uml & try (dsfx "e")) nouns fNouns

-- Anwalt -> Anwaltschaft, Freund -> Freundschaft, Friede -> Friedschaft
dNN04 = dPattern "dNN04"
  (sfx "schaft" & try (dsfx "e")) nouns fNouns
```

Figure 2.5: A data sample from DErivBase (derivational rules).

2.7 DerivBase.RU

DerivBase.RU v001 (Vodolazsky, 2020) is a data resource of derivationally related lexemes for Russian. The methodology of its construction has been inspired by the creation of DErivBase for German. Therefore, its derivational relations were also identified on the basis of (derivational) rules extracted from Russian reference grammar books. The rules are based on derivational changes (in the form of string substitutions) that happen when deriving a lexeme from its base lexeme. Lemmas of the lexicon (more than 270 thousand) were extracted from the Russian portion of Wikipedia and Wiktionary. As some lemmas are, for example, compounds, they include dashes. The homonymy of lemmas is partly

```
Abrechnen_Nn abrechnen_V 1 Abrechnen_Nn dNV09> abrechnen_V
Abrechnen_Nn abrechnend_A 2 Abrechnen_Nn dNV09> abrechnen_Ven dVA02> abrechnend_A
Abrechnen_Nn Abrechnung_Nf 2 Abrechnen_Nn dNV09> abrechnen_Ven dVN07> Abrechnung_Nf
abrechnen_V abrechnend_A 1 abrechnen_V dVA02> abrechnend_A
abrechnen_V Abrechnung_Nf 1 abrechnen_V dVN07> Abrechnung_Nf
abrechnend_A Abrechnung_Nf 1 abrechnend_A dNA26*> Abrechnung_Nf
Abrichten_Nn abrichten_V 1 Abrichten_Nn dNV09> abrichten_V
Abrichten_Nn Abrichtung_Nf 2 Abrichten_Nn dNV09> abrichten_Ven dVN07> Abrichtung_Nf
abrichten_V Abrichtung_Nf 1 abrichten_V dVN07> Abrichtung_Nf
anrechnen_V Anrechnung_Nf 1 anrechnen_V dVN07> Anrechnung_Nf
anreichern_V Anreicherung_Nf 1 anreichern_V dVN07> Anreicherung_Nf
anreichern_V reich_A 1 anreichern_V dAV08*> reich_A
anreichern_V Reiche_Nm 2 anreichern_V dAV08*> reich_A dAN01> Reiche_Nm
anreichern_V Reichere_Nf 1 anreichern_V dNV15*> Reichere_Nf
anreichern_V reichern_V 1 anreichern_V dVV13.3*> reichern_V
anreichern_V reichlich_A 2 anreichern_V dAV08*> reich_A dAA01> reichlich_A
```

Figure 2.6: A data sample from DErivBase (derivational relations).

handled by assigning part-of-speech categories to each lemma (`noun`, `adj`, `verb`, `adv`, `num`).

Just like in the case of German DErivBase, the morphological segmentation of individual lemmas into morphs is only inferrable from a reverse application of the derivational rules to lemmas. In this case, there are no explicit labels of morphs, but they can be inferred from the position of morphs, i.e., prefix for morphs preceding a base lemma, suffix for morphs following a base lemma, and ending for bracketed morphs. Moreover, the derivational step for coining a derivative is labeled in the last column. The data is distributed in separate six-column tab-separated files (separately for each part-of-speech category of base lexemes) containing derivational relations between a pair of lemmas. As can be seen in Figure 2.7, each relation is always labelled by a derivational rule and the morphological operation(s) used for deriving a lexeme.

```
вымор   noun  повыморить   verb  rule887(по + noun + и1(ть) -> verb)        PFX,SFX
вымор   noun  вымориться   verb  rule932(noun + и1(ть) + ся -> verb)        SFX,PTFX
баббит  noun  баббитный    adj   rule619(noun + н1(ый) -> adj)              SFX
баббит  noun  баббитовый   adj   rule628(noun + ов(ый) -> adj)              SFX
путин   noun  путинист     noun  rule343(noun + ист -> noun)               SFX
путин   noun  путинизм     noun  rule355(noun + изм -> noun)               SFX
путин   noun  путинизация  noun  rule366(noun + ацию(а)/изацию(а) -> noun)  SFX
путин   noun  путиноид     noun  rule400(noun + оид -> noun)               SFX
путин   noun  путинка      noun  rule410(noun + к(а) -> noun)              SFX
путин   noun  путинец      noun  rule414(noun + ец -> noun)               SFX
путин   noun  путинка      noun  rule434(noun + к(а) -> noun)              SFX
путин   noun  путинный     adj   rule619(noun + н1(ый) -> adj)              SFX
путин   noun  путинский    adj   rule630(noun + ск(ий) -> adj)              SFX
```

Figure 2.7: A data sample from DerivBase.RU.

2.8 Dictionary of Morphemes of Russian

Dictionary of Morphemes of the Russian Language (called MorphoDictKE in the tables) is a dictionary of around 52 thousand manually morphologically segmented lemmas (Kuznetsova and Efremova, 1986); it was digitised and enlarged to contain more than 74 thousand lemmas. The lexicon contains a complete morphological segmentation of lemmas into morphs; root morph(s) are always labelled; see Figure 2.8. The mapping of morphs to the graphemes of the lemma is straightforward.

While the homonymy of lemmas is partly handled by assigning part-of-speech categories to the lemmas, allomorphy is not handled in the resource. Except for the part-of-speech categories (A: adjective, S: noun, V: verb, ADV: adverb, PR: adposition, APRO and SPRO: pronoun, ANUM and NUM: numeral, ADVPRO: adverb, CONJ: conjuction, PART: particle, and some others), the resource does not include any other morphological categories. As some lemmas are, for example, compounds, they include dashes.

The data is distributed in a six-column comma-separated file format consisting of a lemma, morphological segmentation into morphs, a list of root morph(s), a part-of-speech category, initial indices of each morph in the lemma, and initial and final positions of root morph(s).

```
вязальщик,"['вяз', 'а', 'льщик']",['вяз'],S,"[0, 3, 4]","[[0, 2]]"
вязальщица,"['вяз', 'а', 'льщиц', 'а']",['вяз'],S,"[0, 3, 4, 9]","[[0, 2]]"
вязание,"['вяз', 'а', 'ни', 'е']",['вяз'],S,"[0, 3, 4, 6]","[[0, 2]]"
вязанка,"['вяз', 'а', 'н', 'к', 'а']",['вяз'],S,"[0, 3, 4, 5, 6]","[[0, 2]]"
вязаночка,"['вяз', 'а', 'н', 'оч', 'к', 'а']",['вяз'],S,"[0, 3, 4, 5, 7, 8]","[[0, 2]]"
вязаный,"['вяз', 'а', 'н', 'ый']",['вяз'],A,"[0, 3, 4, 5]","[[0, 2]]"
вязанье,"['вяз', 'а', 'н', 'ье']",['вяз'],S,"[0, 3, 4, 5]","[[0, 2]]"
вязаться,"['вяз', 'а', 'ть', 'ся']",['вяз'],V,"[0, 3, 4, 6]","[[0, 2]]"
вязать,"['вяз', 'а', 'ть']",['вяз'],V,"[0, 3, 4]","[[0, 2]]"
вязель,"['вяз', 'ел', 'ь']",['вяз'],S,"[0, 3, 5]","[[0, 2]]"
вязка,"['вяз', 'к', 'а']",['вяз'],S,"[0, 3, 4]","[[0, 2]]"
вязкий,"['вяз', 'к', 'ий']",['вяз'],A,"[0, 3, 4]","[[0, 2]]"
вязковатый,"['вяз', 'к', 'оват', 'ый']",['вяз'],A,"[0, 3, 4, 8]","[[0, 2]]"
```

Figure 2.8: A data sample from the Dictionary of Morphemes of Russian.

2.9 Échantinom

Échantinom (Bonami and Tribout, 2021) is a morphological resource for French nouns, documenting 5,000 nominal lemmas sampled from the Lexique and flexique databases, based on frequency, and manually annotated. It records the last morphological process applied to the lemma, labelling prefixation, suffixation, conversion, compounding, or a non-concatenative process. Each entry is also marked with a finer-grained label for this process from a set of 29 labels, e.g. back-formation, reduplication, or type of conversion, as well as rarer French-specific processes such as verlan or louchébem. In cases of affixation, the prefix or suffix morphs are recorded; suffixes are also marked with their corresponding

morphemes. The derivational base of the lemma is also provided, along with its part-of-speech.[6] The database also has several other fields, including gender, type of compound if applicable, phonetic transcription of the stem, and the allomorphs of the suffix if any; see Figure 2.9. Homonymy with respect to gender is marked by adding a final "M" or "F" to the suffix annotation.

```
lemma,gen,phon,freq_lex_books,freq_lex_subtitles,freq_frcow,last_process_broad,
  last_process_narrow,prefix,compound,conversion,suffix,suffix_broad,sfx_base,
  sfx_base_pos,autonomous_base,base_stem_phon,sfx_allomorph,der_stem_phon,
  edit_distance,pattern,pattern_tf,pattern_rel_tf,base_der_sim,offset_sim

berlingue,m,bɛʁ.lɛ̃g,0.34,0,34,nonconcat,
  apocope,0,0,0,0,0,NA,
  NA,NA,NA,NA,NA,
  NA,NA,NA,NA,NA,NA

corton,m,kɔʁ.tɔ̃,0.27,0.03,398,suffix,
  suffix,0,0,0,on,on,cour,
  N,TRUE,kuʁ,ɔ̃,kɔʁt,
  2,_u_~_ɔ_tɔ̃,1,0.015625,0.222162783145905,0.158108526129264

dabuche,f,da.byʃ,0.54,0,3,suffix,
  suffix,0,0,0,uche,Vche,dabe,
  N,TRUE,UNKNOWN,yʃ,dab,
  UNKNOWN,UNKNOWN,UNKNOWN,UNKNOWN,UNKNOWN,UNKNOWN

alpiniste,m,al.pi.nist,1.49        1.96,5819,suffix,
suffix,0,0,0,iste,iste,alpin,
A,TRUE,alpin,ist,alpin,0,_~_ist,53,0.569892473,0.4425928,0.454843023

verlan,m,vɛʁ.lɑ̃,0.34,0.07,1695,nonconcat,
verlan,0,0,0,0,0,NA,
NA,NA,NA,NA,NA,
NA,NA,NA,NA,NA,NA

sueur,f,sɥœʁ,60.34,11.71,35392,suffix,
suffix,0,0,0,eurF,eurF,suer,
V,TRUE,sɥ,œʁ,sɥ,
0,_~_œʁ,11,0.846153846153846,0.474891513586044,0.444119388776185
```

Figure 2.9: A data sample from Échantinom. The suffix in "alpiniste" is marked as "iste", and its allomoph is marked "ist". We see "verlan" itself marked as an example of the morphological process of verlan, coming from "l'envers", meaning "backwards". The gender homonymous "eur" suffix is marked as "eurF" in "sueur".

2.10 KCIS

The KCIS datasets[7] contain treebanks for 5 languages: Hindi, Marathi, Kannada, Malayalam, and Bengali, from different domains such as tourism and agriculture.

[6] There are 7 tags in total - N:noun, NP:proper noun, A:adjective, V:verb, ADV:adverb, NUM:numeral, and NA, e.g. for an unknown base.

[7] Can be downloaded here: `https://ltrc.iiit.ac.in/showfile.php?filename=downloads/kolhi/`. The annotation was funded by KCIS, DeiTY, Govt. of India.

Each word in a sentence is marked with a POS tag according to the AnnCorra scheme (Bharati et al., 2006), and with a feature structure of the form from Figure 2.10.

In particular, the "suffixes" field contains a manually recorded complete list of all suffixes of the word form (as it occurs in the sentence), such as case-markers, postpositional suffixes, or verbal inflections; cf. Figure 2.11.

2.10.1 Marathi

The Marathi treebanks contain roughly 41,000 unique tokens, or word forms, in total. The suffix field in the feature structure shows morphs separated by a connector, such that it is trivial to induce exact morphemic boundaries in the word. However, there is a margin of annotation error, with some feature structures containing corresponding morphemes instead of morphs. Further, the same word may be annotated in different ways (any field may be missing, the root stem may contain a base instead, or a morph may be missing), sometimes leading to several dozens of feature structures for a single word. For example, there are 27 different features structures for the wordform गेली , the irregular past tense (both simple and participle) of the verb "to go", for a feminine subject. In total, roughly 23,000 wordforms are marked with at least one suffix.

```
<fs af='root,lcat,gender,number,person,case,case/tam marker,suffixes' ...>
```

Figure 2.10: A general data structure of KCIS. The seventh field contains a case marker (if any) for nouns and TAM (tense, aspect, modality) marker for verbs.

2.10.2 Kannada

The Kannada treebanks contain about 30,000 unique tokens in total. Of these, about 25,000 are marked with at least one suffix. Feature structures record morphemes rather than morphs in the suffixes field; allomorphy is very frequent and is not handled. Vowels represented as matras (diacritics) in the word form may be annotated in their swara (standalone letter) form in the morpheme if they are "word"-initial. Marked suffixes may overlap with each other and with the root as given. Further, the schwa character, or the inherent vowel in Kannada consonants, may be represented as a separate "morpheme", even though it occurs as part of any consonant and therefore has no representation or span in the wordform.[8] This usually happens when the previous morpheme ends with a "virama", the vowel-suppressing character. Thus, a consonantal character can often be a part of two overlapping "morphemes" as marked, the first including its purely consonantal part (sans vowel), and the second one the schwa character. See Figure 2.12 for an example. Forms of allomorphy include stripping of any intial vowels of the vowel, or interchanging of long and short vowels.

[8]Further, it has no meaning of its own and perhaps is better considered as a phonological intervention.

```
<Sentence id='25'>
  1     ((      NP      <fs name='NP' drel='k5:VGNF'>
  1.1     तेथून      PR_PRP      <fs af='तेथून,pn,,,,,,' name='तेथून'>
        ))
  2     ((      NP      <fs name='NP2' drel='k7t:VGNF'>
  2.1     रात्री      N_NN      <fs af='रात्र,n,f,sg,,o,ी,ी' name='रात्री'>
        ))
  3     ((      VGNF      <fs af=',,,,,,,' name='VGNF' drel='vmod:VGINF'>
  3.1     परतत      V__VM      <fs af='परत,v,,,,,त,त' type='kr' name='परतत'>
  3.2     असताना      V__VAUX      <fs af='अस,v,,,,,ताना,ताना' type='tn' name='असताना'>
        ))
  4     ((      NP      <fs name='NP3' drel='k7p:VGINF'>
  4.1     गाडीत      N_NN      <fs af='गाडी,n,f,sg,,o,त,त' name='गाडीत'>
        ))
  5     ((      NP      <fs name='NP4' drel='k2:VGINF'>
  5.1     गॅस      N_NN      <fs af='गॅस,n,m,sg,,d,,' name='गॅस'>
        ))
  6     ((      VGINF      <fs af=',,,,,,,' name='VGINF' drel='rt:VGNF2'>
  6.1     भरण्यासाठी      V_VM      <fs af='भर,v,,,,,साठी,ण्यासाठी' kr='n_kr' name='भरण्यासाठी'>
        ))
  7     ((      NP      <fs name='NP5' drel='k1:VGNF2'>
  7.1     ते      PR_PRP      <fs af='तो,pn,m,pl,,d,,े' name='ते'>
        ))
  8     ((      NP      <fs name='NP6' drel='k7:VGNF2'>
  8.1     तिरुपती      N__NNP      <fs af='तिरुपती,n,,,,,,' name='तिरुपती'>
  8.2     पंपावर      N_NN      <fs af='पंप,n,m,sg,,o,वर,ा_वर' name='पंपावर'>
        ))
  9     ((      VGNF      <fs af=',,,,,,,' name='VGNF2' drel='vmod:VGF'>
  9.1     आले      V_VM      <fs af='ये,v,m,pl,3,,ले,ले' t='pas' a='p' type='ak' name='आले'>
  9.2     असता      V__VAUX      <fs af='अस,v,,,,,ता,ता' type='kr' name='असता'>
  9.3     ,      RD_PUNC      <fs af='&sbquo;,punc,,,,,,' name=','>
        ))
  10     ((      NP      <fs name='NP7' drel='k7p:VGF'>
  10.1     तेथे      PR__PRP      <fs af='तेथे,pn,,,,,,' name='तेथे'>
        ))
  11     ((      NP      <fs name='NP8' drel='k1:VGF'>
  11.1     अज्ञात      JJ      <fs af='अज्ञात,adj,,,,,,' name='अज्ञात'>
  11.2     तरुण      N_NN      <fs af='तरुण,n,m,pl,,d,,' name='तरुण'>
        ))
  12     ((      VGF      <fs name='VGF' drel='ccof:CCP'>
  12.1     आले      V_VM      <fs af='ये,v,m,pl,3,,ले,ले' t='pas' a='p' type='ak' name='आले2'>
        ))
  13     ((      CCP      <fs name='CCP' drel='ccof:VGF2'>
  13.1     व      CC_CCD      <fs af='व,avy,,,,,,' name='व'>
        ))
  14     ((      NP      <fs name='NP9' drel='k5:VGF2'>
  14.1     कुठून      PR_PRQ      <fs af='कुठ,pn,,,,,ून,ून' name='कुठून'>
        ))
  15     ((      VGF      <fs name='VGF2' drel='rs:NP10'>
  15.1     आला      V_VM      <fs af='ये,v,m,sg,3,,ला,ला' t='pas' a='p' type='ak' name='आला'>
  15.2     ,      RD_PUNC      <fs af='&sbquo;,punc,,,,,,' name=',2'>
        ))
  16     ((      NP      <fs name='NP10' drel='k2:VGF3'>
  16.1     असे      DM__DMD      <fs af='असे,pn,n,sg,,d,,' name='असे'>
        ))
  17     ((      VGF      <fs name='VGF3'>
  17.1     विचारले      V_VM      <fs af='विचार,v,n,sg,3,,ले,ले' t='pas' a='p' type='ak' name='विचारले'>
  17.2     .      RD_PUNC      <fs af='&sdot;,punc,,,,,,' name='.'>
        ))
</Sentence>
```

Figure 2.11: A data sample from KCIS (treebank structure from Marathi).

```
12.1	ಆಕರ್ಷಿಸುತ್ತದೆ	V__VM__VF	<fs
↪  af='ಆಕರ್ಷಿಸು,v,,sg,3,,ಇ+ಇಸ್+ಉ+ಉತ್ತ್+ಅ+ಅದ್+ಎ,i+is+u+uww+a+ax+eV' name='ಆಕರ್ಷಿಸುತ್ತದೆ'>
```

Figure 2.12: The first marked suffix “i” is written in its letter form in the suffix list; however, it occurs as a matra or diacritic in the wordform. The first three suffixes also overlap with the root as marked; this is common throughout the dataset. Finally, the fifth marked “morpheme” is “a”, the inherent vowel of the previous consonant “t”. It has no span of its own in the wordform. The suffixes field can therefore be interpreted rather as a segmentation rather than a string of suffixes, with any morphemic parts marked as morphemes rather than as (allo)morphs as they occur in the word form.

2.10.3 Malayalam

The Malayalam treebanks contain a total of roughly 46,000 unique tokens. Of these, about 33000 are marked with at least one suffix. Similarly as above, it marks morphemes, with similar characteristics regarding vowels and the “virama” character. Allomorphy is also not handled here; in addition to vowel stripping, allomorphs may show certain consonants changed (such as “ka” to “nga”, or “ta” to “ra/rra”) as well as doubling of consonants. Morphemes may be marked with the vowel-suppressing virama that may not occur in the surface wordform string.

2.10.4 Hindi

The Hindi treebanks have a total of 10,513 unique tokens, with only about 900 word forms marked with at least one suffix. These treebanks have morphemes rather than morphs (e.g. they may use a default gender while marking the suffix, rather than the observed gender inflection) marked in the suffix field of the feature structure of words, so it is not trivial to induce morphemic boundaries in the word form. Further, only certain inflectional and derivational morphemes are marked; there are several missing segmentations of multi-morphemic wordforms, many for highly productive morphemes. (See Figure 2.13 for examples). It also seems that at most one suffix per word form has been marked.[9]

2.10.5 Bangla

The Bangla dataset has about 24,000 unique tokens. Similarly to Hindi, the suffixes are under-annotated; only about 900 wordforms are marked as containing at least one suffix. In total, only 29 unique suffixes are marked; no word form is marked as containing more than a single suffix.

[9] with exactly two exceptions

```
18.2    सहायक    N_NN    <fs af='सहायक,n,m,sg,3,d,0,0' name='सहायक' posn='270'>
16.1    उग्रता    N_NN    <fs af='उग्रता,n,f,sg,3,d,0,0' name='उग्रता' posn='230'>
11.1    दशाएँ    N_NN    <fs af='दशा,n,f,pl,3,d,0,0' name='दशाएँ' posn='170'>
3.1    गंदगी    N_NN    <fs af='गंदगी,n,f,sg,3,d,0,0' name='गंदगी' posn='50'>
```

Figure 2.13: Samples from the KCIS Hindi treebanks, where the feature structures falsely mark "0", or no suffixes. These wordforms contain a noun-to-actor suffix, an adjective-to-noun suffix, a nominal plural suffix, and a different adjective-to-noun suffix respectively, none of which are marked. These allomorphs are in fact never marked in entire dataset.

2.11 MorphoChallenge

This dataset comes from the MorphoChallenge shared tasks for morphological segmentation (Kurimo et al., 2010). It contains complete morphological segmentation of lemmas but the precise format of data contained depends on the year. The contained languages are Arabic, English, Finnish, German, Turkish (see Figure 2.14 for a sample of the data and Table 2.1 for data sizes in individual years). The data encoding depends on the language. English uses simple text, with all words lower-cased (including the proper names). Finnish is in ISO LATIN 1 with all characters encoded as single bytes. German is lower-cased and transliterated (ö => oe, ß => ss), all the remaining characters are in ISO LATIN 1. Turkish is lower-cased and the letters specific to Turkish are replaced by letters from the Latin alphabet. The Arabic is transliterated via the Buckwalter transliteration. The year 2007 also contains vowelized Arabic.

2.11.1 Shared task in 2005

The 2005 dataset contains words segmented to morphs, without additional annotation. The dataset occasionally contains multiple variants of segmentation.

2.11.2 Shared tasks in 2007-2009

The 2007 and the 2008 datasets are exactly the same, except for the addition of Arabic. The 2007-2009 datasets contain words segmented to highly abstract morphemes: *vaccinates vaccine_N ate_s +3SG*.

2.11.3 Shared task in 2010

The 2010 dataset contains aligned segmentation to morphs and morphemes: *overbalanced over:over_p balanc:balance_V ed:+PAST*. The only exception from this is the German dataset, which only contains the 2007-2009 data format. The 2010 dataset uses null morphs, as well as null morphemes. Hyphens are handled either as separate morphs (representing null morpheme) or as part of a morph.

Language	2005	2008	2009	2010
Arabic	N	500	690	N
English	532	484	466	1,686
Finnish	660	506	634	1,835
German	N	557	525	1,779
Turkish	774	541	581	1,760

Table 2.1: Number of words in the MorphoChallenge datasets for each year and language.

```
CantasIz            Canta sIz
CarpIlmanIz         Carp Il man Iz, Carp Il ma nIz
CatalcayI           Catalca yI
CeSnilerin          CeSni ler in
Cekilirdik          Cek il ir di k
Cekmemesi           Cek me me si
Cenede              Cened e, Cene de
Cevirmektir         Cevir mek tir
CiCeklerinden       CiCek ler in den, CiCek ler i nden, CiCek leri nden
Cindeydi            Cin de ydi
Cizgisine           Ciz gi si ne
CobanlIGa           Coban lIG a

aravalainoitettua   arava_N lainoittaa_V +PSSPCP2 +PTV
armahdusoikeus      armahtaa_V +DV-US oikea_A +DA-US
arveli              arvella_V +PAST
arveluttava         arvella_V +DV-UTTA +PCP1
arviointimenetelmän arvioida_V +DV-NTI menetelmä_N +GEN
asiaansa            asia_N +ILL +3SGPL, asia_N +PTV +3SGPL
asiallisen          asia_N +DN-LLINEN +GEN
asiavirheet         asia_N virhe_N +PL
asuntolainoitus     asunto_N lainoittaa_V +DV-US
autoradiossa        auto_N radio_N +INE
avoimella           avoin_A +ADE
avoimille           avoin_A +PL +ALL
avoimin             avoin_A +PL +INS, avoin_A +SUP
avuton              apu_N +DN-TON

accompanied         ac:ac_p compani:co\mpany_N ed:+PAST
accompaniment       ac:ac_p compani:company_N ment:ment_s
accorded            accord:accord_V ed:+PAST
acknowledging       ac:ac_p knowledg:knowledge_N ing:+PCP1
acquisition         acquis:acquire_V ition:ition_s
acquisitions'       acquis:acquire_V ition:ition_s s:+PL ':+GEN
actions             act:act_V ion:ion_s s:+PL
actress'            act:act_V r:or_s ess:ess_s ':+GEN
acupuncture         acupuncture:acupuncture_N
```

Figure 2.14: Data samples from MorphoChallenge 2005, 2008 and 2010, respectively.

2.12 MorphoLex

MorphoLex is a manually-segmented lexicon for English (Sánchez-Gutiérrez et al., 2018) and French (Mailhot et al., 2020) annotated with morphological variables, such as morphological family sizes and corpus frequencies of individual mor-

```
pendentif                   (pendre)>ant/ent>>if>
plaidoirie                  (plaider)>oy[VB]>>erie>
rafraîchissant              <re<<a<(frais)>[VB]>>sant>

wrongheadedness  NN         {(wrong)}{(head)}>ness>
artistically     RB         {(art)}>ist>>ic>>ly>
americanizing    VB|NN|JJ   {(america)}>n>>ize>
```

Figure 2.15: Three example records from the French and three from the English MorphoLex variant.

phemes. The main aim of the project is to facilitate research into morphological processing of language, by providing a canonical resource annotated with corpus frequencies and other data. See Figure 2.15 for examples from both datasets.

The lexicon of the English resource consists of 68,624 English words taken from the English Lexicon Project (Balota et al., 2007) and contains part-of-speech categories of lexemes from the Penn Treebank tag inventory (Santorini, 1990), with possibly multiple categories listed for lexemes which undergo conversion (e.g. *publicized* is listed as *VB|JJ*, as it can function both as a verb and an adjective). Some lexemes are spelled in all-uppercase, some in all-lowercase even when they are proper nouns, some in true case. We were unable to find whether these distinctions have any meaning.

The morpheme segmentation is based on the segmentation given in the source data (Balota et al., 2007), but with amendments made to regularize the annotation: Affixes inside the stems were originally not typed as prefixes, suffixes or roots, but simply delimited, and parts of neoclassical compounds were sometimes marked as affixes and other times as roots. These were normalized based on clear rules (Sánchez-Gutiérrez et al., 2018).

The French version of the data is based on the French Lexicon Project (Ferrand et al., 2010) with a vocabulary of 38,840 words. No part-of-speech categories or other morphological features of lexemes are listed. The segmentation was made manually from scratch, as the French source does not contain segmentational information (Mailhot et al., 2020).

Both datasets contains a single possible segmentation for each word form, meaning that homonyms and other words with ambiguous segmentation are only listed once. Allomorphy is disambiguated and the segmentation only lists canonical forms. In addition to canonicalizing the morph forms, the French data unifies all possible verbal suffixes and lists them as an *[VB]* morpheme.

The segmentation in both resources does not contain all morphemes, because some inflectional morphemes are not listed at all, even when occurring inside the word stem (e.g., *accordingly* is segmented as *accord* + *ly* and listed as a single-root, single-suffix lexeme, omitting the *ing* morpheme completely).

2.13 Tikhonov's dictionary

Morphological dictionary of Aleksandr Nikolaevich Tikhonov (Tikhonov, 1996) contains 100,000 Russian lemmas segmented to morphs. The segmentation is complete and morph types are not annotated. Part of speech tags or any other additional information is not present. The dataset is written in Cyrillic and some words contain hyphens or apostrophes as accent marks.

```
цанга | ца'нг/а
цанговый | ца'нг/ов/ый
цап-царап | цап/-цара'п, глаг. междом. (от цара'п/ну/ть)
цапать | ца'п/а/ть
цапаться | ца'п/а/ть/ся
цапка | ца'п/к/а
цапля | ца'пл/я
```

Figure 2.16: A data sample from Tikhonov's dictionary.

2.14 MorphyNet

MorphyNet (Batsuren et al., 2021) is a multilingual database of derivational and inflectional morphology. Currently, MorphyNet contains 13.5 million inflectional and 696 thousand derivational instances of 15 languages:[10] Catalan, Czech, English, Finnish, French, German, Hungarian, Italian, Mongolian, Polish, Portuguese, Russian, Serbo-Croatian, Spanish, and Swedish.

MorphyNet was extracted from Wiktionary using both hand-crafted and automated methods. Morphological information explicitly contained in Wiktionary was enriched by inferring more general patterns from data, both for inflection and derivation.

For each language, there are two files in the MorphyNet resource, one for inflection and one for derivation; see two fragments for German in Figure 2.17. In the inflectional file, for each lemma there is a set of lines corresponding to inflected forms; for each, its inflectional catogories are specified and inflectional boundary in front of the inflectional ending is given.

In the derivational file, for each derived lemma, its derivational antecedent is specified and the last derivational affix (prefix or suffix) is separated.

2.15 Persian Morphologically Segmented Lexicon

Persian Morphologically Segmented Lexicon (Ansari et al., 2019) is a specialised resource of morphological segmentation. It includes complete morphological segmentation of word forms that originate from Persian Wikipedia, popular Persian corpus BijanKhan, and Persian Named Entity corpus. The homonymy of word forms is handled by classifying them into four categories (V: verb, E: named

[10] https://github.com/kbatsuren/MorphyNet

```
anfangen        anfangen        V;NFIN  -
anfangen        anfangend       V;V.PTCP;PRS    an-|fang|-end
anfangen        fange an        V|IND;PRS;1;SG ADP      fang|-e an
anfangen        fangen an       V|IND;PRS;1;PL ADP      fang|-en an
anfangen        fange an        V|SBJV;PRS;1;SG ADP     fang|-e an
anfangen        fangen an       V|SBJV;PRS;1;PL ADP     fang|-en an
anfangen        fängst an       V|IND;PRS;2;SG ADP      fang|-st an
anfangen        fangt an        V|IND;PRS;2;PL ADP      fang|-t an
anfangen        fangest an      V|SBJV;PRS;2;SG ADP     fang|-est an
anfangen        fanget an       V|SBJV;PRS;2;PL ADP     fang|-et an
anfangen        fängt an        V|IND;PRS;3;SG ADP      fang|-t an

Zahl    zahlen  N       V       en      suffix
zählen  zählbar V       J       bar     suffix
Arzt    Ärztin  N       N       in      suffix
schlagen        beschlagen      V       V       be      prefix
suchen  Besuch  V       N       be      prefix
Bote    Botschaft       N       N       schaft  suffix
stören  zerstören       V       V       zer     prefix
Störung Zerstörung      N       N       zer     prefix
Zier    zierlich        N       J       lich    suffix
Rücken  zurück  N       R       zu      prefix
bringen zurückbringen   V       V       zurück  prefix
```

Figure 2.17: An inflectional and a derivational fragment from the German section of MorphyNet.

entity, `I`: irregular plural, `X`: none of the above) and labelling them by `0` or `1` if a word form is ambiguous. Except for these, the lexicon does not include any additional annotation.

The segmentation of word forms was made by the Hazm toolkit (Persian preprocessing and tokenisation tools); however, words with more than 10 occurrences in the corpus collection (around 80 thousand word forms) were morphologically segmented manually. The original file format adheres to the Arabic ordering (from right to left) and is kept in space-separated columns format which has no fixed number of columns; cf. Figure 2.19. The first column contains a word while the following columns includes its lemma, form, ambiguity annotation, specification of class into which the word belongs, and a list of morphs; see Figure 2.18.

```
word lemma form ambiguity segment\_1 segment\_2 ... segment\_n
```

Figure 2.18: A general data structure of Persian Morphologically Segmented Lexicon.

آرمیک	آرمیک	E	0	آرمیک		
آرنا	آرنا	E	0	آرنا		
آرناس	آرناس	E	0	آرناس		
آرنالدو	آرنالدو	E	0	آرنالدو		
آرنت	آرنت	E	0	آرنت		
آرنج	آرنج	X	0	آرنج		
آرنجش	آرنج	X	0	آرنج	ش	
آرنج‌ها	آرنج	X	0	آرنج	ها	
آرندت	آرندت	E	0	آرندت		
آرنلد	آرنلد	E	0	آرنلد		
آرنو	آرنو	E	0	آرنو		
آرنور	آرنور	E	0	آرنور		
آرنولف	آرنولف	E	0	آرنولف		
آرنی	آرنی	E	0	آرنی		
آرنیوس	آرنیوس	E	0	آرنیوس		
آرنیکه‌ها	آرنیکه	X	0	آرنیکه	ها	
آره	آره	X	0	آره		
آرواره	آرواره	X	0	آرواره		
آرواره‌ای	آرواره	X	0	آرواره	ای	
آرواره‌ها	آرواره	X	0	آرواره	ها	
آرواره‌های	آرواره	X	0	آرواره	ها	ی
آروزی	آروز	X	0	آروز	ی	
آروس	آروس	E	0	آروس		
آروشا	آروشا	E	0	آروشا		
آروغ	آروغ	X	0	آروغ		
آروماتیک	آروماتیک	X	0	آرومات	یک	
آروماتیکی	آروماتیک	X	0	آرومات	یک	ی
آروماتیک‌ها	آروماتیک	X	0	آرومات	یک	ها
آرونا	آرونا	E	0	آرونا		
آرونسون	آرونسون	E	0	آرونسون		

Figure 2.19: A data sample from Persian Morphologically Segmented Lexicon.

2.16 UniMorph

The UniMorph project (McCarthy et al., 2020) deals with inflectional morphology for large number of languages. The data collection[11] currently contains around 9 million lemma – features – inflected form triples, for 141 languages.

The triples have been extracted from Wictionary and other inflectional resources. UniMorph does not contain any morphematic segmentation, however, one could assume that at least some morpheme boundaries (especially those in front of inflectional endings) could be heuristically derived by string comparisons of inflected word forms within a lemma's inflectional cluster.

2.17 Uniparser

Uniparser is a finite-state-transducer-like morphological analyzer, optionally combined with constraint grammars (Arkhangelskiy et al., 2012). Its initial goal was to process under-resourced languages of the Uralic region of Russia, but over time, grammars were written for many other small languages.

[11] https://unimorph.github.io/

```
Absage  Absage  N;ACC;SG
Absage  Absage  N;DAT;SG
Absage  Absage  N;GEN;SG
Absage  Absagen N;ACC;PL
Absage  Absagen N;DAT;PL
Absage  Absagen N;GEN;PL
Absage  Absagen N;NOM;PL
Absage  Absage  N;NOM;SG
absagen abgesagt        V.PTCP;PST
absagen absagend        V.PTCP;PRS
absagen absagen V;NFIN
absagen sag ab  V;IMP;2;SG
absagen sage ab V;IND;PRS;1;SG
absagen sage ab V;SBJV;PRS;1;SG
absagen sage ab V;SBJV;PRS;3;SG
```

Figure 2.20: An sample fragment from the German section of UniMorph 3.0.

Currently, a loose collection of parsers is publicly available for 11 languages, namely Adyghe (Arkhangelskiy and Lander, 2015), Albanian , Eastern Armenian (Khurshudian and Daniel, 2009), Erzya, Komi-Zyrian, Meadow Mari, Moksha (all described in Arkhangelskiy (2019)), Tajik (Iskandarova, 2021), Turoyo , Udmurt (Arkhangelskiy and Medvedeva, 2016) and Urmi . Parsers for several other languages are reported in the literature, but not publicly available: Buryat, Greek, Kalmyk, Lezgian and Ossetic (Arkhangelskiy et al., 2012). The authors of the Uniparser project also publish lexicons of annotated words extracted from various corpora of the languages.

In addition to lemmatizing and tagging texts, the grammar description can be used to delimit boundaries between the inflectional morphemes of word forms and produce linguistically-appropriate glosses conforming to the Leipzig glossing rules (Bickel et al., 2008; Lehmann, 1982), see Figure 2.21 for an example from the analyzed lexicon of Eastern Armenian. Such glosses are present in 7 of the 11 grammar descriptions (Eastern Armenian, Erzya, Komi-Zyrian, Meadow Mari, Moksha, Tajik and Udmurt). The data for Turoyo don't contain glosses for affixes, but it lists the consonants forming the root. While most grammars only list morphs with an overt form, the data for Tajik contain explicit zero affixes.

The annotation is XML-like, with each lexeme "<w>" record containing the word form in plain text and one or several analyses in "<ana>" tags. The presence of multiple analyses indicates ambiguity between multiple functions of the inflectional ending or homonymy. The structure is, however, not a well-formed XML, because the annotation file consists of a list of lexeme "<w>" records without an enclosing element, and the glosses and parts (lists of segments) may contain unencoded "<" and ">" signs. Minor errors are also present: Some lexemes are preceded or followed by underscores, which serve no function and appear to be a programming artifact, and sometimes the gloss contains doubled STEM markings (STEMSTEM instead of STEM).

Another problem for processing is the usage of the dash to mark segment and gloss boundaries, while also leaving them unescaped in word forms such as *Санкт-Петербург* ("St. Petersburg"). This causes the morph list to have more

```
<w>
  <ana lex="առաջ"
       gr="POST,sg,obl,def"
       parts="առաջ-ի-ս"
       gloss="before-OBL-DEF"
       trans_en="before, front side, ahead">
  </ana>
  <ana lex="առաջի"
       gr="A,sg,nom,def"
       parts="առաջի-ս"
       gloss="STEM-DEF"
       trans_en="">
  </ana>
  <ana lex="առաջիս"
       gr="NUM,A,sg,nom,nonposs"
       parts="առաջիս"
       gloss="first"
       trans_en="first">
  </ana>
  առաջիս
</w>
```

Figure 2.21: A data sample from Uniparser for Eastern Armenian.

apparent elements than the gloss list and the literal dash has to be detected by comparing the morph list to the word form.

Uniparser uses a very similar part-of-speech tagset for all languages[12], using a format inspired by the tagset of the Russian National Corpus.

2.18 Word Formation Latin

Word Formation Latin database (Litta et al., 2016) encompasses Latin derivation, compounding and conversion (but not inflection). It covers lemmas of all major parts-of-speech. The lemma list was compiled from three Classical and Late Latin dictionaries and contains 36,258 lemmas. Most of the derivational relationships were either created automatically using a set of different rules or semi-automatically. Only the last derivational step is segmented, but every lexeme contains a reference to all derivational/compounding parents; in order to construct the full segmentation, one has to recursively run through all references to reconstruct the segmentation; therefore, the segmentation is incomplete, but can be back-tracked.

Inflection is not handled at all, which presents difficulty in Latin, because the conceptual boundary between a stem allomorph and an ending may be ambiguous in the nominative case.

Both POS and inflectional categories are available. The POS tagset goes as follows: {'ADP': Adposition, 'PRON': Pronoun, 'PART': Participle, 'SCONJ': Subordinative conjunction, 'VERB': Verb, 'ADV': Adverb, 'ADJ': Adjective, 'INTJ': Interjection, 'NUM': Numeral, 'NOUN': Noun, 'CCONJ': Coordinative conjunction,

[12]For example, the variant for Udmurt is described at `http://udmurt.web-corpora.net/index_en.html#about_tagset`

'PROPN': Proper noun}

Allomorphy is handled using regular-expression-like syntax, but this excludes stems/roots. What this means that for all morphemes apart from stems/roots (the status of the latter two may in Latin be somewhat ambiguous), the substring shared by all allomorphs is represented as a simple string and variants are represented using an ad-hoc system that can be re-interpreted as a regular expression. For example, "Suffix=(i/e)ll" represents both the suffix "ill" and the suffix "ell".

Vowel length is captured if and only if it serves a phonological function, which is not particularly often (dozens of cases).

```
1865.1 malandria    NOUN Declension=c&Gender=Neut 1865.0 Type=Conversion
1865.2 malandriosus ADJ  AdjClass=f               1865.1 Suffix=os&Type=Derivation

1866.0 malaxo       VERB
1866.1 commalaxo    VERB                          1866.0 Prefix=con&Type=Derivation
1866.2 malaxatio    NOUN Declension=c&Gender=Fem  1866.0 Suffix=(t)io(n)&Type=Derivation

1867.0 malleus      NOUN Declension=c&Gender=Masc
1867.1 commalleo    VERB                          1867.0 Prefix=con&Type=Derivation
1867.2 malleator    NOUN Declension=c&Gender=Masc 1867.0 Suffix=(t)or&Type=Derivation
1867.3 malleatus    ADJ  AdjClass=f               1867.0 Suffix=at&Type=Derivation
1867.4 malleolus    NOUN Declension=c&Gender=Masc 1867.0 Suffix=ol&Type=Derivation
1867.5 malleolum    NOUN Declension=c&Gender=Neut 1867.4 Type=Conversion
1867.6 malleolaris  ADJ  AdjClass=f               1867.4 Suffix=ar&Type=Derivation
1867.7 commalliolo  VERB                          1867.4 Prefix=con&Type=Derivation
```

Figure 2.22: A data sample from Word Formation Latin (en excerpt from the harmonised version of WFL in the Universal Derivations collection).

Chapter 3

Similarities and Differences across Morphosegmentation Resources

We see in Table 3.1 that most of the considered resources differ from each other in a structural manner (clusters vs. hierarchical segmentation vs. string of affixes), as well as in the principled decisions made about which type of morphology is handled (inflectional vs. derivational) and how it is represented (marking morphs vs. morphemes, or ignoring vs. preserving ordering). Note that this information about the dataset is not always explicitly given, and must sometimes be inferred from the data.

While this diversity is obviously valuable in providing several perspectives on morphological information about possibly the same languages, it presents the drawback that researchers interested in accessing this information must first identify the above characteristics of datasets, and parse them appropriately in order to use them separately or together. This may function as a deterrent to large-scale studies on morphological information from different languages and resources. Especially in the age of neural data-crunching, different formats for similar information can be a limitation to further work and insights in morphology. This is our motivation for trying to harmonize at least some of the surveyed datasets under a common scheme in the near future.

Acknowledgements

The present work has been supported by grants no. GX20-16819X (LUSyD), no. 19-14534S of the Czech Science Foundation, and no. START/HUM/010 of Grant schemes at Charles University (r. no. CZ.02.2.69/0.0/0.0/19_073/0016935); LM2018101 (LINDAT/CLARIAH-CZ) of the Ministry of Education, Youth, and Sports of the Czech Republic.

resource	type of morphological information	segmentation origin	segmented units	completeness of segmentation	units
CELEX	hierarchical segmentation	manual	lemmas	mostly complete	morphemes
CroDeriV	lemmas	manual	lemmas	completed with labels	morphs
Démonette	lemmas + suffix, derivational relations	automatically collated from other resources	lemmas	single suffix	morphs
DeriNet	segmented lemmas	manual + automatically derived	forms, lemmas	complete	morphs
DerIvaTario	base + affixes, derivational, ordered	manual	lemmas	complete	morphemes
DerivBaseDE	derivational relations	automatic (grammar based)	lemmas	only single suffix segmentation	morphs
DerivBaseRU	derivational relations	automatic (grammar based)	lemmas	only single suffix segmentation with labels	morphs
MorphoDictKE	lemmas	manual	lemmas	complete with labelled roots	morphs
Échantinom	lemmas + affix	manual	lemmas	last derivational process	morphs
KCIS	root + suffixes	manual	wordforms	complete	morphs (Marathi) / morphemes (all others)
MorphoChallenge 2005	segmented lemmas	manual, automatic (grammar based)	lemmas	complete	morphs
MorphoChallenge 2007-2009	segmented lemmas	manual, automatic (grammar based)	lemmas	abstract morphemes	
MorphoChallenge 2010	segmented lemmas	manual, automatic (grammar based)	lemmas	complete	morphs and abstract morphemes
MorphoLex	segmentation of roots and derivational morphemes	manual	forms	complete	morphemes
Tikhonov's dictionary	segmented lemmas	manual	lemmas	complete	morphs
MorphyNet	inflectional clusters, derivational pairs	mostly automatic	lemmas + forms	partial	morphs
PerSegLex	wordform segmentations	manual	word forms	complete	morphs
UniMorph	inflectional clusters	derived	word forms	only single suffix separation	morphs
Uniparser	lemmatization + segmentation	automatic (grammar based)	forms	inflectional + incomplete derivational suffixation delimited	morphs
WFL	lemmas	manual + semiautomatically derived	lemmas	only single single suffix separation	morphemes

Table 3.1: Diversity of morphological information relevant for morphemic segmentation.

Bibliography

Ansari, E., Žabokrtský, Z., Haghdoost, H., and Nikravesh, M. (2019). Persian Morphologically Segmented Lexicon 0.5. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Arkhangelskiy, T. (2019). Corpora of social media in minority Uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 125–140, Tartu, Estonia. Association for Computational Linguistics.

Arkhangelskiy, T., Belyaev, O., and Vydrin, A. (2012). The creation of large-scale annotated corpora of minority languages using UniParser and the EANC platform. In *Proceedings of COLING 2012: Posters*, pages 83–92, Mumbai, India. The COLING 2012 Organizing Committee.

Arkhangelskiy, T. and Lander, Y. (2015). Some challenges of the West Circassian polysynthetic corpus. Research Paper WP BRP 37/LNG/2015, Higher School of Economics.

Arkhangelskiy, T. and Medvedeva, M. (2016). Developing morphologically annotated corpora for minority languages of Russia. In Kübler, S. and Dickinson, M., editors, *Proceedings of Corpus Linguistics Fest 2016 (CLiF 2016)*, volume 1607 of *CEUR Workshop Proceedings*, pages 1–6, Bloomington, Indiana, USA. CEUR-WS.org.

Aronoff, M. (2019). Competitors and alternants in linguistic morphology. In Rainer, F., Gardani, F., and Dressler, W.and Luschützky, H., editors, *Competition in Inflection and Word-Formation*, pages 39–66. Spinger, Cham.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database (CD-ROM). Catalogue No. LDC96L14.

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., and Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3):445–459.

Batsuren, K., Bella, G., and Giunchiglia, F. (2021). MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics,*

Phonology, and Morphology, pages 39–48, Online. Association for Computational Linguistics.

Bertinetto, P. M., Burani, C., Laudanna, A., Marconi, L., Ratti, D., Rolando, C., and Thornton, A. M. (2005). Colfis (corpus e lessico di frequenza dell'italiano scritto). *Available on http://www. istc. cnr. it/material/database*, pages 67–73.

Bharati, A., Sangal, R., Sharma, D. M., and Bai, L. (2006). Anncorra: Annotating corpora guidelines for POS and chunk annotation for Indian languages. *LTRC-TR31*, pages 1–38.

Bickel, B., Comrie, B., and Haspelmath, M. (2008). The leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses.

Bonami, O. and Tribout, D. (2021). Echantinom.

Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A., Augustinova, M., and 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., and Štefanec, V. (2019). Redesign of the Croatian derivational lexicon. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 71–80, Prague.

Hajič, J., Hlaváčová, J., Mikulová, M., Straka, M., and Štěpánková, B. (2020). MorfFlex CZ 2.0.

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

Hathout, N. (2011). Morphonette: a paradigm-based morphological network. *Lingue e linguaggio*, 10(2):245–264.

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

Hathout, N., Namer, F., and Dal, G. (2002). An Experimental Constructional Database: The MorTAL Project. In Boucher, P., editor, *Many Morphologies*, pages 178–209. Cascadilla, Somerville, Mass.

Iskandarova, D. M. (2021). Национальный корпус таджикского языка как инструмент лингвистических исследований (National corpus of the Tajik language as a tool for linguistic research). *Kazan Science*, 1:94–97.

Khurshudian, V. and Daniel, M. (2009). Eastern Armenian national corpus. *Dialog'2009*, pages 509–518.

Kurimo, M., Virpioja, S., Turunen, V., and Lagus, K. (2010). Morpho challenge competition 2005–2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, SIGMORPHON '10, page 87–95, USA. Association for Computational Linguistics.

Kuznetsova, A. I. and Efremova, T. F. (1986). *Slovar' morfem russkogo jazyka [Dictionary of morphemes of the Russian language]*. Russkij jazyk, Moscow.

Lehmann, C. (1982). Directions for interlinear morphemic translations. *Folia Linguistica*, 16(1-4):199–224.

Litta, E., Passarotti, M., and Culy, C. (2016). Formatio formosa est. Building a word formation lexicon for Latin. In *Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it)*.

Mailhot, H., Wilson, M. A., Macoir, J., Deacon, S. H., and Sánchez-Gutiérrez, C. H. (2020). MorphoLex-FR: A derivational morphological database for 38,840 French words. *Behavior Research Methods*, 52(3):1008–1025.

McCarthy, A. D., Kirov, C., Grella, M., Nidhi, A., Xia, P., Gorman, K., Vylomova, E., Mielke, S. J., Nicolai, G., Silfverberg, M., et al. (2020). Unimorph 3.0: Universal morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Namer, F. (2009). *Morphologie, lexique et traitement automatique des langues*. Hermès-Lavoisier.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Rajman, M., Lecomte, J., and Paroubek, P. (1997). Format de description lexicale pour le français, partie 2: Description morpho-syntaxique. Technical Report GRACE GTR-3-2.1, EPFL & INaLF.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Šojat, K., Srebačić, M., Pavelić, T., and Tadić, M. (2014). CroDeriV: A New Resource for Processing Croatian Morphology. In *Proceedings of the Language Resources and Evaluation (LREC-2014)*, volume 14, pages 3366–3370, Reykjavik. Citeseer.

Sánchez-Gutiérrez, C. H., Mailhot, H., Deacon, S. H., and Wilson, M. A. (2018). MorphoLex: A derivational morphological database for 70,000 English words. *Behavior Research Methods*, 50(4):1568–1580.

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

Tanguy, L. and Hathout, N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In Pierrel, J.-M., editor, *Actes de la 9ème Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*, pages 245–254, Nancy. ATALA.

Tikhonov, A. N. (1996). *Морфемно-орфографический словарь русского языка. Русская морфемика (Morphemic-spelling dictionary of the Russian language. Russian morphemics)*. Shkola-Press, Moscow, Russia.

Vidra, J., Žabokrtský, Z., Kyjánek, L., Ševčíková, M., Dohnalová, Š., Svoboda, E., and Bodnár, J. (2021). DeriNet 2.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Vodolazsky, D. (2020). DerivBase.Ru: A derivational morphology resource for Russian. In *Proceedings of the Language Resources and Evaluation (LREC-2020)*, volume 20, pages 3930–3936, Marseille, France.

Zeller, B., Šnajder, J., and Padó, S. (2013). DErivBase: Inducing and evaluating a derivational morphology resource for German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1201–1211. ACL.

The ÚFAL/CKL Technical Report Series

ÚFAL

ÚFAL (Ústav formální a aplikované lingvistiky; http://ufal.mff.cuni.cz) is the Institute of Formal and Applied linguistics, at the Faculty of Mathematics and Physics of Charles University, Prague, Czech Republic. The Institute was established in 1990 after the political changes as a continuation of the research work and teaching carried out by the former Laboratory of Algebraic Linguistics since the early 60s at the Faculty of Philosophy and later the Faculty of Mathematics and Physics. Together with the "sister" Institute of Theoretical and Computational Linguistics (Faculty of Arts) we aim at the development of teaching programs and research in the domain of theoretical and computational linguistics at the respective Faculties, collaborating closely with other departments such as the Institute of the Czech National Corpus at the Faculty of Philosophy and the Department of Computer Science at the Faculty of Mathematics and Physics.

CKL

As of 1 June 2000 the Center for Computational Linguistics (Centrum komputační lingvistiky; http://ckl.mff.cuni.cz) was established as one of the centers of excellence within the governmental program for support of research in the Czech Republic. The center is attached to the Faculty of Mathematics and Physics of Charles University in Prague.

Technical Reports

The ÚFAL/CKL technical report series has been established with the aim of disseminate topical results of research currently pursued by members, cooperators, or visitors of the Institute. The technical reports published in this Series are results of the research carried out in the research projects supported by the Grant Agency of the Czech Republic, GAČR 405/96/K214 ("Komplexní program"), GAČR 405/96/0198 (Treebank project), grant of the Ministry of Education of the Czech Republic VS 96151, and project of the Ministry of Education of the Czech Republic LN00A063 (Center for Computational Linguistics). Since November 1996, the following reports have been published.

ÚFAL TR-1996-01 Eva Hajičová, *The Past and Present of Computational Linguistics at Charles University*
Jan Hajič and Barbora Hladká, *Probabilistic and Rule-Based Tagging of an Inflective Language – A Comparison*

ÚFAL TR-1997-02 Vladislav Kuboň, Tomáš Holan and Martin Plátek, *A Grammar-Checker for Czech*

ÚFAL TR-1997-03 Alla Bémová at al., Anotace na analytické rovině, *Návod pro anotátory (in Czech)*

ÚFAL TR-1997-04 Jan Hajič and Barbora Hladká, *Tagging Inflective Languages: Prediction of Morphological Categories for a Rich, Structural Tagset*

ÚFAL TR-1998-05 Geert-Jan M. Kruijff, *Basic Dependency-Based Logical Grammar*

ÚFAL TR-1999-06 Vladislav Kuboň, *A Robust Parser for Czech*

ÚFAL TR-1999-07 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (in Czech)*

ÚFAL TR-2000-08 Tomáš Holan, Vladislav Kuboň, Karel Oliva, Martin Plátek, *On Complexity of Word Order*

ÚFAL/CKL TR-2000-09 Eva Hajičová, Jarmila Panevová and Petr Sgall, *A Manual for Tectogrammatical Tagging of the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-10 Zdeněk Žabokrtský, *Automatic Functor Assignment in the Prague Dependency Treebank*

ÚFAL/CKL TR-2001-11 Markéta Straňáková, *Homonymie předložkových skupin v češtině a možnost jejich automatického zpracování*

ÚFAL/CKL TR-2001-12 Eva Hajičová, Jarmila Panevová and Petr Sgall, *Manuál pro tektogramatické značkování (III. verze)*

ÚFAL/CKL TR-2002-13 Pavel Pecina and Martin Holub, *Sémanticky signifikantní kolokace*

ÚFAL/CKL TR-2002-14 Jiří Hana, Hana Hanová, *Manual for Morphological Annotation*

ÚFAL/CKL TR-2002-15 Markéta Lopatková, Zdeněk Žabokrtský, Karolína Skwarská and Vendula Benešová, *Tektogramaticky anotovaný valenční slovník českých sloves*

ÚFAL/CKL TR-2002-16 Radu Gramatovici and Martin Plátek, *D-trivial Dependency Grammars with Global Word-Order Restrictions*

ÚFAL/CKL TR-2003-17 Pavel Květoň, *Language for Grammatical Rules*

ÚFAL/CKL TR-2003-18 Markéta Lopatková, Zdeněk Žabokrtský, Karolina Skwarska, Václava Benešová, *Valency Lexicon of Czech Verbs VALLEX 1.0*

ÚFAL/CKL TR-2003-19 Lucie Kučová, Veronika Kolářová, Zdeněk Žabokrtský, Petr Pajas, Oliver Čulo, *Anotování koreference v Pražském závislostním korpusu*

ÚFAL/CKL TR-2003-20 Kateřina Veselá, Jiří Havelka, *Anotování aktuálního členění věty v Pražském závislostním korpusu*

ÚFAL/CKL TR-2004-21 Silvie Cinková, *Manuál pro tektogramatickou anotaci angličtiny*

ÚFAL/CKL TR-2004-22 Daniel Zeman, *Neprojektivity v Pražském závislostním korpusu (PDT)*

ÚFAL/CKL TR-2004-23 Jan Hajič a kol., *Anotace na analytické rovině, návod pro anotátory*

ÚFAL/CKL TR-2004-24 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2004-25 Jan Hajič, Zdeňka Urešová, Alevtina Bémová, Marie Kaplanová, *The Prague Dependency Treebank, Annotation on tectogrammatical level*

ÚFAL/CKL TR-2005-27 Jiří Hana, Daniel Zeman, *Manual for Morphological Annotation (Revision for PDT 2.0)*

ÚFAL/CKL TR-2005-28 Marie Mikulová a kol., *Pražský závislostní korpus (The Prague Dependency Treebank) Anotace na tektogramatické rovině (úroveň 3)*

ÚFAL/CKL TR-2005-29 Petr Pajas, Jan Štěpánek, *A Generic XML-Based Format for Structured Linguistic Annotation and Its application to the Prague Dependency Treebank 2.0*

ÚFAL/CKL TR-2006-30 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Magda Razímová, Petr Sgall, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Annotation manual)*

ÚFAL/CKL TR-2006-31 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall, Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Anotace na tektogramatické rovině Pražského závislostního korpusu (Referenční příručka)*

ÚFAL/CKL TR-2006-32 Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolařová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, Petr Sgall,Magda Ševčíková, Jan Štěpánek, Zdeňka Urešová, Kateřina Veselá, Zdeněk Žabokrtský, *Annotation on the tectogrammatical level in the Prague Dependency Treebank (Reference book)*

ÚFAL/CKL TR-2006-33 Jan Hajič, Marie Mikulová, Martina Otradovcová, Petr Pajas, Petr Podveský, Zdeňka Urešová, *Pražský závislostní korpus mluvené češtiny. Rekonstrukce standardizovaného textu z mluvené řeči*

ÚFAL/CKL TR-2006-34 Markéta Lopatková, Zdeněk Žabokrtský, Václava Benešová (in cooperation with Karolína Skwarska, Klára Hrstková, Michaela Nová, Eduard Bejček, Miroslav Tichý) *Valency Lexicon of Czech Verbs. VALLEX 2.0*

ÚFAL/CKL TR-2006-35 Silvie Cinková, Jan Hajič, Marie Mikulová, Lucie Mladová, Anja Nedolužko, Petr Pajas, Jarmila Panevová, Jiří Semecký, Jana Šindlerová, Josef Toman, Zdeňka Urešová, Zdeněk Žabokrtský, *Annotation of English on the tectogrammatical level*

ÚFAL/CKL TR-2007-36 Magda Ševčíková, Zdeněk Žabokrtský, Oldřich Krůza, *Zpracování pojmenovaných entit v českých textech*

ÚFAL/CKL TR-2008-37 Silvie Cinková, Marie Mikulová, *Spontaneous speech reconstruction for the syntactic and semantic analysis of the NAP corpus*

ÚFAL/CKL TR-2008-38 Marie Mikulová, *Rekonstrukce standardizovaného textu z mluvené řeči v Pražském závislostním korpusu mluvené češtiny. Manuál pro anotátory*

ÚFAL/CKL TR-2008-39 Zdeněk Žabokrtský, Ondřej Bojar, *TectoMT, Developer's Guide*

ÚFAL/CKL TR-2008-40 Lucie Mladová, *Diskurzní vztahy v češtině a jejich zachycení v Pražském závislostním korpusu 2.0*

ÚFAL/CKL TR-2009-41 Marie Mikulová, *Pokyny k překladu určené překladatelům, revizorům a korektorům textů z Wall Street Journal pro projekt PCEDT*

ÚFAL/CKL TR-2011-42 Loganathan Ramasamy, Zdeněk Žabokrtský, *Tamil Dependency Treebank (TamilTB) – 0.1 Annotation Manual*

ÚFAL/CKL TR-2011-43 Nguy Giang Linh, Michal Novák, Anna Nedoluzhko, *Coreference Resolution in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-44 Anna Nedoluzhko, Jiří Mírovský, *Annotating Extended Textual Coreference and Bridging Relations in the Prague Dependency Treebank*

ÚFAL/CKL TR-2011-45 David Mareček, Zdeněk Žabokrtský, *Unsupervised Dependency Parsing*

ÚFAL/CKL TR-2011-46 Martin Majliš, Zdeněk Žabokrtský, *W2C – Large Multilingual Corpus*

ÚFAL TR-2012-47 Lucie Poláková, Pavlína Jínová, Šárka Zikánová, Zuzanna Bedřichová, Jiří Mírovský, Magdaléna Rysová, Jana Zdeňková, Veronika Pavlíková, Eva Hajičová, *Manual for annotation of discourse relations in the Prague Dependency Treebank*

ÚFAL TR-2012-48 Nathan Green, Zdeněk Žabokrtský, *Ensemble Parsing and its Effect on Machine Translation*

ÚFAL TR-2013-49 David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Daniel Zemana, Zdeněk Žabokrtský, Jan Hajič *Cross-language Study on Influence of Coordination Style on Dependency Parsing Performance*

ÚFAL TR-2013-50 Jan Berka, Ondřej Bojar, Mark Fishel, Maja Popović, Daniel Zeman, *Tools for Machine Translation Quality Inspection*

ÚFAL TR-2013-51 Marie Mikulová, *Anotace na tektogramatické rovině. Dodatky k anotátorské příručce (s ohledem na anotování PDTSC a PCEDT)*

ÚFAL TR-2013-52 Marie Mikulová, *Annotation on the tectogrammatical level. Additions to annotation manual (with respect to PDTSC and PCEDT)*

ÚFAL TR-2013-53 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *Úpravy a doplňky Pražského závislostního korpusu (Od PDT 2.0 k PDT 3.0)*

ÚFAL TR-2013-54 Marie Mikulová, Eduard Bejček, Jiří Mírovský, Anna Nedoluzhko, Jarmila Panevová, Lucie Poláková, Pavel Straňák, Magda Ševčíková, Zdeněk Žabokrtský, *From PDT 2.0 to PDT 3.0 (Modifications and Complements)*

ÚFAL TR-2014-55 Rudolf Rosa, *Depfix Manual*

ÚFAL TR-2014-56 Veronika Kolářová, *Valence vybraných typů deverbativních substantiv ve valenčním slovníku PDT-Vallex*

ÚFAL TR-2014-57 Anna Nedoluzhko, Eva Fučíková, Jiří Mírovský, Jiří Pergler, Lenka Šíková, *Annotation of coreference in Prague Czech-English Dependency Treebank*

ÚFAL TR-2015-58 Zdeňka Urešová, Eva Fučíková, Jana Šindlerová, *CzEngVallex: Mapping Valency between Languages*

ÚFAL TR-2015-59 Kateřina Rysová, Magdaléna Rysová, Eva Hajičová, *Topic–Focus Articulation in English Texts on the Basis of Functional Generative Description*

ÚFAL TR-2016-60 Kira Droganova, Daniel Zeman, *Conversion of SynTagRus (the Russian dependency treebank) to Universal Dependencies*

ÚFAL TR-2018-61 Lukáš Kyjánek, *Morphological Resources of Derivational Word-Formation Relations*

ÚFAL TR-2019-62 Zdeňka Urešová, Eva Fučíková, Eva Hajičová, *CzEngClass: Contextually-based Synonymy and Valency of Verbs in a Bilingual Setting (CzEngClass: Kontextová synonymie a valence sloves v bilingvním prostředí)*

ÚFAL TR-2019-63 Ján Faryad,
Identifikace derivačních vztahů ve španělštině

ÚFAL TR-2020-64 Marie Mikulová, Jan Hajič, Jiří Hana, Hana Hanová, Jaroslava Hlaváčová, Emil Jeřábek, Barbora Štěpánková, Barbora Vidová Hladká, Daniel Zeman,
Manual for Morphological Annotation. Revision for Prague Dependency Treebank – Consolidated 2020 release

ÚFAL TR-2021-65 Rudolf Rosa, *Technická zpráva o vývoji projektu THEaiTRE v roce 2020*

ÚFAL TR-2021-66 Anna Nedoluzhko, Michal Novák, Martin Popel, Zdeněk Žabokrtský, Daniel Zeman
Coreference meets Universal Dependencies – a pilot experiment on harmonizing coreference datasets for 11 languages

ÚFAL TR-2021-67 THEaiTRobot 1.0, David Košťák, Daniel Hrbek, Rudolf Rosa, Ondřej Dušek,
AI: When a Robot Writes a Play

ÚFAL TR-2021-68 Markéta Lopatková, Václava Kettnerová, Anna Vernerová, Eduard Bejček, Zdeněk Žabokrtský,
Valenční slovník českých sloves VALLEX

ÚFAL TR-2021-69 Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Madga Ševčíková, Jonáš Vidra, Zdeněk Žabokrtský
Towards Universal Segmentations: Survey of Existing Morphosegmentation Resources