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
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

1 Introduction ................................................................. 3
   1.1 Motivation for Harmonisation Efforts ............................ 3
   1.2 Basic Notions ...................................................... 4
      1.2.1 Morpheme, morph, and allomorphy ......................... 4
      1.2.2 Morphological segmentation ............................... 5
   1.3 Resource selection ............................................... 6

2 Overview of existing resources ........................................... 8
   2.1 CELEX ............................................................ 8
   2.2 CroDeriV ......................................................... 9
   2.3 Démonette ....................................................... 9
   2.4 DeriNet ........................................................ 10
   2.5 DerIvaTario ...................................................... 10
   2.6 D ErivBase ..................................................... 12
   2.7 DerivBase.RU ................................................... 12
   2.8 Dictionary of Morphemes of Russian ............................ 14
   2.9 Échantinom ....................................................... 14
   2.10 KCIS ............................................................. 15
      2.10.1 Marathi ..................................................... 16
      2.10.2 Kannada ................................................... 16
      2.10.3 Malayalam ................................................ 18
      2.10.4 Hindi ...................................................... 18
      2.10.5 Bangla .................................................... 18
   2.11 MorphoChallenge ................................................... 19
      2.11.1 Shared task in 2005 ....................................... 19
      2.11.2 Shared tasks in 2007-2009 ............................... 19
      2.11.3 Shared task in 2010 ...................................... 19
   2.12 MorphoLex ....................................................... 20
   2.13 Tikhonov’s dictionary .......................................... 22
   2.14 MorphyNet ....................................................... 22
   2.15 Persian Morphologically Segmented Lexicon ................... 22
   2.16 UniMorph ........................................................ 24
   2.17 Uniparser ........................................................ 24
   2.18 Word Formation Latin .......................................... 26

3 Similarities and Differences across Morphosegmentation Resources 28
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 harmonization 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. 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 (“portman-teau 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, asymmetries between forms and meanings are also documented in morphemes. A particular form

\footnote{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 \( \text{skříň} \) ‘cupboard’ \( \rightarrow \text{skříňka} \) ‘small cupboard’, while in \( \text{učitel} \) ‘teacher’ \( \rightarrow \text{učitelka} \) ‘female teacher’ it coins the feminine agent noun, but it occurs also in the instrument noun \( \text{žehlička} \) ‘iron’ motivated by the verb \( \text{žehlit} \) ‘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 \( vý-bír-a-t \) ‘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’ \( \leftarrow \) \( češ+ský \), \( obléci \) ‘to dress up’ \( \leftarrow \) \( ob+vléci \)). 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 (\( z-ne-při-jemň-ova-t \) ‘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.
<table>
<thead>
<tr>
<th>Abbrev. name</th>
<th>Original name, version</th>
<th>Languages</th>
<th>License</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>CELEX Lexical Database 2.0</td>
<td>Dutch, English, German</td>
<td>EULA (research use only)</td>
<td>Baayen et al. (1995)</td>
</tr>
<tr>
<td>CroDerIV</td>
<td>CroDerIV 1.0</td>
<td>Croatian</td>
<td>CC BY-SA-3.0</td>
<td>Sojat et al. (2014)</td>
</tr>
<tr>
<td>Démonette</td>
<td>Démonette-1.2</td>
<td>French</td>
<td>CC BY-NC-SA 3.0</td>
<td>Hathout and Namer (2014)</td>
</tr>
<tr>
<td>DerIvaTario</td>
<td>DerIvaTario</td>
<td>Italian</td>
<td>CC BY-SA 4.0</td>
<td>Talamo et al. (2016)</td>
</tr>
<tr>
<td>DerivBaseDE</td>
<td>DErivBase 2.0</td>
<td>German</td>
<td>CC BY-SA 3.0</td>
<td>Zeller et al. (2013)</td>
</tr>
<tr>
<td>DerivBaseRU</td>
<td>DerivBase.Ru 1.0</td>
<td>Russian</td>
<td>Apache-2.0</td>
<td>Vodolazsky (2020)</td>
</tr>
<tr>
<td>DeriNet</td>
<td>DeriNet 2.1</td>
<td>Czech</td>
<td>CC BY-NC-SA 3.0</td>
<td>Vidra et al. (2021)</td>
</tr>
<tr>
<td>MorphoDictKE</td>
<td>Dictionary of Morphemes of Russian</td>
<td>Russian</td>
<td>All rights reserved</td>
<td>Kuznetsova and Efremova (1986)</td>
</tr>
<tr>
<td>Échantinom</td>
<td>Échantinom</td>
<td>French</td>
<td>CC BY 4.0</td>
<td>Bonami and Tribout (2021)</td>
</tr>
<tr>
<td>KCIS</td>
<td>KCIS Resources</td>
<td>Marathi, Hindi, Malayalam, Kannada, Bangla</td>
<td>EULA (research use only)</td>
<td>(see Sec. 2.10)</td>
</tr>
<tr>
<td>MorphoLex</td>
<td>MorphoLex, MorphoLex-FR</td>
<td>English and French</td>
<td>CC BY 4.0</td>
<td>Sánchez-Gutiérrez et al. (2018); Mailhot et al. (2020)</td>
</tr>
<tr>
<td>MorphyNet</td>
<td>MorphyNet, v1</td>
<td>15 languages</td>
<td>CC BY-SA 3.0</td>
<td>Batsuren et al. (2021)</td>
</tr>
<tr>
<td>PerSegLex</td>
<td>Morphologically Segmented Lexicon 0.5</td>
<td>Persian</td>
<td>CC BY-NC-SA 4.0</td>
<td>Ansari et al. (2019)</td>
</tr>
<tr>
<td>Tikhonov’s dictionary</td>
<td>Morphemic-spelling dictionary of the Russian language. Russian morphemics</td>
<td>Russian</td>
<td>All rights reserved</td>
<td>Tikhonov (1996)</td>
</tr>
<tr>
<td>UniMorph</td>
<td>UniMorph 3.0</td>
<td>141 languages</td>
<td>CC BY-SA 3.0 for most languages</td>
<td>McCarthy et al. (2020)</td>
</tr>
<tr>
<td>Uniparser</td>
<td>Uniparser morphological analyzer</td>
<td>7 languages</td>
<td>MIT</td>
<td>Arkhangelskiy et al. (2012)</td>
</tr>
<tr>
<td>WFL</td>
<td>Word Formation Latin 1.1</td>
<td>Latin</td>
<td>CC BY-NC-SA 4.0</td>
<td>Litta et al. (2016)</td>
</tr>
</tbody>
</table>

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 ([Sojat et al., 2014](http://croderiv.ffzg.hr)) 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, 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. 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.

![HTML code for a data sample from CroDeriv](image)

Figure 2.2: A data sample from CroDeriv

## 2.3 Démonette

Démonette ([Hathout and Namer, 2014](http://croderiv.ffzg.hr)) is a morphosemantic lexical database that is automatically built from the parsing system DériF ([Namer, 2009](http://croderiv.ffzg.hr())), the Morphonette network ([Hathout, 2011](http://croderiv.ffzg.hr())), and Verbaction ([Tanguy and Hathout, 2002](http://croderiv.ffzg.hr()); [Hathout et al., 2002](http://croderiv.ffzg.hr())). 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](http://croderiv.ffzg.hr/)), a (single) suffix (if any), a conversion process (if any),

[^1]: [http://croderiv.ffzg.hr/](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)](http://croderiv.ffzg.hr/).
[^3]: [www.cnrtl.fr/lexiques/morphalou/](http://croderiv.ffzg.hr/)
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","@" "abaissement","tlfnome","abaisser","tlfnome","Ncms","tlfnome","Vmn----","tlfnome",
→ "simple","demonette","suf","ment","demonette","@ACT","demonette","@",
→ "demonette","action de abaisser","demonette","action de @","demonette","descendant",
→ "demonette","abaiss","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","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 and 2,000 form-lemma pairs 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 unrecognizable (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.

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 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 inverse 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
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 inferable 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.

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: preposition, APRO and SPRO: pronoun, ANUM and NUM: numeral, ADVPRO: adverb, CONJ: conjunction, 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).

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 louchébém. 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. 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.

<table>
<thead>
<tr>
<th>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_phone, sfx_allomorph, der_stem_phone, edit_distance, pattern, pattern_tf, pattern_rel_tf, base_der_sim, offset_sim</th>
</tr>
</thead>
<tbody>
<tr>
<td>berlingue, m, bɛʁ.lɛ̃g, 0.34, 0, 34, nonconcat,</td>
</tr>
<tr>
<td>apocope, 0, 0, 0, 0, NA,</td>
</tr>
<tr>
<td>NA, NA, NA, NA, NA, NA,</td>
</tr>
<tr>
<td>corton, m, kɔʁ.tɔ̃, 0.27, 0.03, 398, suffix,</td>
</tr>
<tr>
<td>suffix, 0, 0, 0, on, on, cour,</td>
</tr>
<tr>
<td>N, TRUE, kwɔ, ɔ, kwɔr,</td>
</tr>
<tr>
<td>2, _u_ɔ_tɔ̃, 1, 0.015625, 0.222162783145905, 0.158108526129264</td>
</tr>
<tr>
<td>dabuche, f, da.byʃ, 0.54, 0.3, suffix,</td>
</tr>
<tr>
<td>suffix, 0, 0, 0, uche, Vche, dabe,</td>
</tr>
<tr>
<td>N, TRUE, UNKNOWN, yf, dab,</td>
</tr>
<tr>
<td>UNKNOWN, UNKNOWN, UNKNOWN, UNKNOWN, UNKNOWN, UNKNOWN</td>
</tr>
<tr>
<td>alpiniste, m, al.pi.nist, 1.49, 1.96, 5819, suffix,</td>
</tr>
<tr>
<td>suffix, 0, 0, 0, iste, iste, alpin,</td>
</tr>
<tr>
<td>A, TRUE, alpin, ist, alpin, 0, ~_ist, 53, 0.56982473, 0.4425928, 0.454843023</td>
</tr>
<tr>
<td>verlan, m, vɛʁ.lɑ̃, 0.34, 0.07, 1695, nonconcat,</td>
</tr>
<tr>
<td>verlan, 0, 0, 0, 0, NA,</td>
</tr>
<tr>
<td>NA, NA, NA, NA, NA, NA,</td>
</tr>
<tr>
<td>sueur, f, sɥœʁ, 60.34, 11.71, 35392, suffix,</td>
</tr>
<tr>
<td>suffix, 0, 0, 0, eurF, eurF, suer,</td>
</tr>
<tr>
<td>V, TRUE, su, su, su,</td>
</tr>
<tr>
<td>0, ~_m, 11, 0.8461532846153846, 0.474891513586894, 0.444119388776185</td>
</tr>
</tbody>
</table>

Figure 2.9: A data sample from Échantinom. The suffix in “alpiniste” is marked as “iste”, and its allomorph 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 contain treebanks for 5 languages: Hindi, Marathi, Kannada, Malayalam, and Bengali, from different domains such as tourism and agriculture.

---

6There 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.

7Can be downloaded here: [https://ltrc.iiit.ac.in/showfile.php?filename=downloads/kolhi/](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,lc	,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. 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.

Further, it has no meaning of its own and perhaps is better considered as a phonological intervention.
Figure 2.11: A data sample from KCIS (treebank structure from Marathi).
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 33,000 are marked with at least one suffix. Similarly as above, it marks morphemes, with similar characteristics regarding vowels and the “vīrama” 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.

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.

\[\text{with exactly two exceptions}\]
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:\[\text{vaccinates vaccine}\_N\_ate\_s +3SG.\]

2.11.3 Shared task in 2010

The 2010 dataset contains aligned segmentation to morphs and morphemes: \[\text{overbalanced over:over\_p balanc:balance}\_V\_ed:+PAST.\] 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.
Table 2.1: Number of words in the MorphoChallenge datasets for each year and language.

<table>
<thead>
<tr>
<th>Language</th>
<th>2005</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>N</td>
<td>500</td>
<td>690</td>
<td>N</td>
</tr>
<tr>
<td>English</td>
<td>532</td>
<td>484</td>
<td>466</td>
<td>1,686</td>
</tr>
<tr>
<td>Finnish</td>
<td>660</td>
<td>506</td>
<td>634</td>
<td>1,835</td>
</tr>
<tr>
<td>German</td>
<td>N</td>
<td>557</td>
<td>525</td>
<td>1,779</td>
</tr>
<tr>
<td>Turkish</td>
<td>774</td>
<td>541</td>
<td>581</td>
<td>1,760</td>
</tr>
</tbody>
</table>

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

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 (Fer-rand 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.

<table>
<thead>
<tr>
<th>Цанга</th>
<th>Ца'нг/а</th>
</tr>
</thead>
<tbody>
<tr>
<td>Цанговый</td>
<td>Ца'нг/ов/ый</td>
</tr>
<tr>
<td>Цап-Царган</td>
<td>Цап/-Цара'п, глаг. междом. (от цара'п/ну/ть)</td>
</tr>
<tr>
<td>Цапать</td>
<td>Ца'п/а/ть</td>
</tr>
<tr>
<td>Цапаться</td>
<td>Ца'п/а/ть/ся</td>
</tr>
<tr>
<td>Цапка</td>
<td>Ца'п/к/а</td>
</tr>
<tr>
<td>Цапля</td>
<td>Ца'пл/я</td>
</tr>
</tbody>
</table>

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: 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 categories 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
<table>
<thead>
<tr>
<th>German Word Forms</th>
<th>Word Forms</th>
<th>Ambiguity</th>
<th>Segment 1</th>
<th>Segment 2</th>
<th>Segment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>anfangen</td>
<td>anfangen</td>
<td>V;NFIN</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anfangen</td>
<td>anfangend</td>
<td>V;V;PTCP;PRS;PRS;1;SG ADP</td>
<td>an-</td>
<td>fang</td>
<td>-end</td>
</tr>
<tr>
<td>anfangen</td>
<td>fange an</td>
<td>V;IND;PRS;1;SG ADP</td>
<td>fang</td>
<td>-e an</td>
<td></td>
</tr>
<tr>
<td>anfangen</td>
<td>fange an</td>
<td>V;IND;PRS;1;PL ADP</td>
<td>fang</td>
<td>-en an</td>
<td></td>
</tr>
<tr>
<td>anfangen</td>
<td>fange an</td>
<td>V;SBJ;PRS;1;SG ADP</td>
<td>fang</td>
<td>-e an</td>
<td></td>
</tr>
<tr>
<td>anfangen</td>
<td>fange an</td>
<td>V;SBJ;PRS;1;PL ADP</td>
<td>fang</td>
<td>-en an</td>
<td></td>
</tr>
<tr>
<td>anfangen</td>
<td>fängst an</td>
<td>V;IND;PRS;2;SG ADP</td>
<td>fang</td>
<td>-st an</td>
<td></td>
</tr>
<tr>
<td>anfangen</td>
<td>fangt an</td>
<td>V;IND;PRS;2;PL ADP</td>
<td>fang</td>
<td>-t an</td>
<td></td>
</tr>
<tr>
<td>anfangen</td>
<td>fangt an</td>
<td>V;SBJ;PRS;2;SG ADP</td>
<td>fang</td>
<td>-est an</td>
<td></td>
</tr>
<tr>
<td>anfangen</td>
<td>fängt an</td>
<td>V;SBJ;PRS;2;PL ADP</td>
<td>fang</td>
<td>-et an</td>
<td></td>
</tr>
<tr>
<td>Zahl</td>
<td>zählen</td>
<td>N;V</td>
<td>en</td>
<td>suffix</td>
<td></td>
</tr>
<tr>
<td>zählen</td>
<td>zählbar</td>
<td>V;J</td>
<td>bar</td>
<td>suffix</td>
<td></td>
</tr>
<tr>
<td>Arzt</td>
<td>Ärztin</td>
<td>N;N</td>
<td>in</td>
<td>suffix</td>
<td></td>
</tr>
<tr>
<td>schlagen</td>
<td>beschlagen</td>
<td>V;V</td>
<td>V</td>
<td>be</td>
<td>prefix</td>
</tr>
<tr>
<td>suchen</td>
<td>Besuch</td>
<td>V;N</td>
<td>be</td>
<td>prefix</td>
<td></td>
</tr>
<tr>
<td>Bote</td>
<td>Botschaft</td>
<td>N;N</td>
<td>schaft</td>
<td>suffix</td>
<td></td>
</tr>
<tr>
<td>stören</td>
<td>zerstören</td>
<td>V;V</td>
<td>zer</td>
<td>prefix</td>
<td></td>
</tr>
<tr>
<td>Störung</td>
<td>Zerstörung</td>
<td>N;N</td>
<td>zer</td>
<td>prefix</td>
<td></td>
</tr>
<tr>
<td>Zier</td>
<td>zierlich</td>
<td>N;J</td>
<td>lich</td>
<td>suffix</td>
<td></td>
</tr>
<tr>
<td>Rücken</td>
<td>zurück</td>
<td>N;R</td>
<td>zu</td>
<td>prefix</td>
<td></td>
</tr>
<tr>
<td>bringen</td>
<td>zurückbringen</td>
<td>V;V</td>
<td>zurück</td>
<td>prefix</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Word</th>
<th>Lemma</th>
<th>Form</th>
<th>Ambiguity</th>
<th>Segment 1</th>
<th>Segment 2</th>
<th>Segment 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>zählen</td>
<td>zählen</td>
<td>V</td>
<td>en</td>
<td>suffix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>zählen</td>
<td>zählbar</td>
<td>V</td>
<td>bar</td>
<td>suffix</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ärztin</td>
<td>Arzt</td>
<td>N</td>
<td>in</td>
<td>suffix</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.18: A general data structure of 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 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.

https://unimorph.github.io/
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
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\(^\text{12}\), 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](http://udmurt.web-corpora.net/index_en.html#about_tagset)
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 sub-string 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).

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1865.1 malandria</td>
<td>NOUN</td>
<td>Declension=c &amp; Gender=Neut</td>
<td>1865.0 Type=Conversion</td>
<td></td>
</tr>
<tr>
<td>1865.2 malandriosus</td>
<td>ADJ</td>
<td>AdjClass=f</td>
<td>1865.1 Suffix=os &amp; Type=Derivation</td>
<td></td>
</tr>
<tr>
<td>1866.0 malaxo</td>
<td>VERB</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1866.1 commalaxo</td>
<td>VERB</td>
<td>1866.0 Prefix=con &amp; Type=Derivation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1866.2 malaxatio</td>
<td>NOUN</td>
<td>Declension=c &amp; Gender=Fem</td>
<td>1866.0 Suffix=(t)io(n) &amp; Type=Derivation</td>
<td></td>
</tr>
<tr>
<td>1867.0 malleus</td>
<td>NOUN</td>
<td>Declension=c &amp; Gender=Masc</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1867.1 commalleo</td>
<td>VERB</td>
<td>1867.0 Prefix=con &amp; Type=Derivation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1867.2 malleator</td>
<td>NOUN</td>
<td>Declension=c &amp; Gender=Masc</td>
<td>1867.0 Suffix=(t)or &amp; Type=Derivation</td>
<td></td>
</tr>
<tr>
<td>1867.3 malleatus</td>
<td>ADJ</td>
<td>AdjClass=f</td>
<td>1867.0 Suffix=at &amp; Type=Derivation</td>
<td></td>
</tr>
<tr>
<td>1867.4 malleolus</td>
<td>NOUN</td>
<td>Declension=c &amp; Gender=Masc</td>
<td>1867.0 Suffix=ol &amp; Type=Derivation</td>
<td></td>
</tr>
<tr>
<td>1867.5 malleolum</td>
<td>NOUN</td>
<td>Declension=c &amp; Gender=Neut</td>
<td>1867.4 Type=Conversion</td>
<td></td>
</tr>
<tr>
<td>1867.6 malleolaris</td>
<td>ADJ</td>
<td>AdjClass=f</td>
<td>1867.4 Suffix=ar &amp; Type=Derivation</td>
<td></td>
</tr>
<tr>
<td>1867.7 commalliolo</td>
<td>VERB</td>
<td>1867.4 Prefix=con &amp; Type=Derivation</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 2.22: A data sample from Word Formation Latin (an 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.
<table>
<thead>
<tr>
<th>resource</th>
<th>type of morphological information</th>
<th>segmentation origin</th>
<th>segmented units</th>
<th>completeness of segmentation units</th>
</tr>
</thead>
<tbody>
<tr>
<td>CELEX</td>
<td>hierarchical segmentation</td>
<td>manual</td>
<td>lemmas</td>
<td>mostly complete</td>
</tr>
<tr>
<td>CroDeriV</td>
<td>lemmas</td>
<td>manual</td>
<td>lemmas</td>
<td>completed with labels</td>
</tr>
<tr>
<td>Démonette</td>
<td>lemmas + suffix, derivational relations</td>
<td>automatically collated from other resources</td>
<td>lemmas</td>
<td>single suffix</td>
</tr>
<tr>
<td>DeriNet</td>
<td>segmented lemmas</td>
<td>manual + automatically derived</td>
<td>forms, lemmas</td>
<td>complete</td>
</tr>
<tr>
<td>DerivaTario</td>
<td>base + affixes, derivational, ordered derivational relations</td>
<td>automatic (grammar based)</td>
<td>lemmas</td>
<td>only single suffix segmentation</td>
</tr>
<tr>
<td>DerivBaseDE</td>
<td>derivational relations</td>
<td>automatic (grammar based)</td>
<td>lemmas</td>
<td>only single suffix segmentation</td>
</tr>
<tr>
<td>DerivBaseRU</td>
<td>derivational relations</td>
<td>automatic (grammar based)</td>
<td>lemmas</td>
<td>with labels</td>
</tr>
<tr>
<td>MorphoDictKE</td>
<td>lemmas</td>
<td>manual</td>
<td>lemmas</td>
<td>complete with labelled roots</td>
</tr>
<tr>
<td>Échantinom</td>
<td>lemmas + affix</td>
<td>manual</td>
<td>lemmas</td>
<td>last derivational process</td>
</tr>
<tr>
<td>KCIS</td>
<td>root + suffixes</td>
<td>manual</td>
<td>wordforms</td>
<td>complete</td>
</tr>
<tr>
<td>MorphoChallenge 2005</td>
<td>segmented lemmas</td>
<td>manual, automatic (grammar based)</td>
<td>lemmas</td>
<td>complete</td>
</tr>
<tr>
<td>MorphoChallenge 2007-2009</td>
<td>segmented lemmas</td>
<td>manual, automatic (grammar based)</td>
<td>lemmas</td>
<td>abstract morphemes</td>
</tr>
<tr>
<td>MorphoChallenge 2010</td>
<td>segmented lemmas</td>
<td>manual, automatic (grammar based)</td>
<td>lemmas</td>
<td>complete</td>
</tr>
<tr>
<td>MorphoLex</td>
<td>segmentation of roots and derivational morphemes</td>
<td>manual</td>
<td>forms</td>
<td>complete</td>
</tr>
<tr>
<td>Tikhonov’s dictionary</td>
<td>segmented lemmas</td>
<td>manual</td>
<td>lemmas</td>
<td>complete</td>
</tr>
<tr>
<td>MorphyNet</td>
<td>inflectional clusters, derivational pairs wordform segmentations</td>
<td>mostly automatic</td>
<td>lemmas + forms</td>
<td>partial</td>
</tr>
<tr>
<td>PerSegLex</td>
<td>inflectional clusters</td>
<td>manual</td>
<td>word forms</td>
<td>complete</td>
</tr>
<tr>
<td>UniMorph</td>
<td>derived</td>
<td>derived</td>
<td>word forms</td>
<td>only single suffix separation</td>
</tr>
<tr>
<td>Uniparser</td>
<td>lemmatization + segmentation</td>
<td>automatic (grammar based)</td>
<td>forms</td>
<td>inflectional + incomplete derivational suffixation delimited</td>
</tr>
<tr>
<td>WFL</td>
<td>lemmas</td>
<td>manual + semiautomatically derived</td>
<td>lemmas</td>
<td>only single suffix separation</td>
</tr>
</tbody>
</table>

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


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-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 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