Towards Universal Segmentations: UniSegments 1.0

Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek,
Emil Svoboda, Magda Ševčíková, Jonáš Vidra
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
{zabokrtsky, bafna, bodnar, kyjanek, svoboda, sevcikova, vidra}@ufal.mff.cuni.cz

Abstract

Our work aims at developing a multilingual data resource for morphological segmentation. We present a survey of 17 existing data resources relevant for segmentation in 32 languages, and analyze diversity of how individual linguistic phenomena are captured across them. Inspired by the success of Universal Dependencies, we propose a harmonized scheme for segmentation representation, and convert the data from the studied resources into this common scheme. Harmonized versions of resources available under free licenses are published as a collection called UniSegments 1.0.

Keywords: morphology, morpheme, morph, segmentation, multilingual language resources

1. Introduction

In natural languages, especially in those with rich morphology, a huge number of word forms exist (both potential and corpus attested), with some sub-parts of word forms being clearly "recycled" many times. This motivates the linguistic notion of the morpheme, the smallest meaningful unit of language (Aronoff and Fudeman, 2011). In NLP, various reincarnations of a loosely related notion of "subword" are used too, but defined, rather, by technical means (such as segmentation resulting from the Byte-Pair Encoding algorithm used in various contemporary deep learning NLP approaches, see e.g. (Sennrich et al., 2016)).[1]

For numerous languages, we have annotated datasets with varying sizes, underlying models, and annotation quality. The task of morphological segmentation seems relatively straightforward and less sensitive to local linguistic traditions than, e.g., syntactic analysis. However, to the best of our knowledge, there is no reasonably standardized and widely accepted approach to morphological segmentation that would be comparable to Universal Dependencies (de Marneffe et al., 2021), a widely multilingual collection of dependency treebanks, which was the source of inspiration for our work.

This paper has two goals. First, we analyze diversity of existing segmentation data resources. Second, we present a novel harmonized scheme for representing morphological segmentation and convert data from 17 data resources into the common scheme. All our converters are fully automatic, and we do not insert any new manual annotations. However, we also infer some missing parts of information by heuristic approximations (for instance, when inducing morph boundaries given a sequence of morphemes).

The remainder of the paper is structured as follows. Section 2 summarizes basic linguistic notions relevant for morphological segmentation. Section 3 presents a survey of data resources included in our study, and Section 4 compares annotation schemes used in them. In Section 5, the harmonization scheme is proposed, and the resulting multilingual data collection is described. Section 6 concludes and outlines near future goals.

2. Basic Linguistic Notions

A morpheme is defined to be the smallest unit of language that has a meaning. Morphemes are smaller than words (cf. three morphemes in *play+er+s*), or identical with them (e.g. *chair* consisting solely of a root morpheme). A root morpheme conveys lexical meaning. Other morphemes, if present in the word's structure, are classified with respect to the root: the root is preceded by one or more prefixes (*re-* in *re+play*) and followed by one or more suffixes (*-er* in *play+er*); a final suffix that expresses inflectional categories (*-s* in *play+er+s*) can be distinguished by the term ending. In words with multiple roots (compounds), interfixes are often used to link the roots (*-s-* in *Arbeit+s+amt* 'employment office').

Morphemes repeat across sets of words, with certain (so-called, cranberry) morphemes forming the exception (Aronoff, 1976). As morphemes are the basic building blocks in inflection and word-formation processes, many of them are expressed by multiple different morphs in different contexts (allomorphy); cf. the root allomorphs *sheep* and *shep* in the nouns *sheep* and *shep+herd*). *Vice versa*, a single form can link to different morphemes; cf. homonymy of both the root and the inflectional marker in the noun *bear+s* and the verb *bear+s*.

In general, words are expected to be fully decomposable into morphs. In the present paper, this task is called morphological segmentation, but alternative names are also used (morphemic segmentation, morphemic analysis, etc.). Contrary to this expectation, one can easily

[1] It is open to conjecture whether the linguistic and the NLP views on word form segmentation will eventually converge or not.

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. *obléci* 'to dress up' ← *ob+vléci*). From the perspective of segmentation, taken in the present paper, such overlaps make it impossible to assign characters to morphs unambiguously.

When words are cut into segments that can no further be divided into smaller meaningful units, we speak of complete morphological segmentation henceforth. Nevertheless, some of the resources analyzed here record an incomplete (partial) segmentation; for example, resources that focus on derivation may delimit only the derivational affix that distinguishes a word from the word which it is immediately based on (cf. the suffix -*ung* in the noun *Abbezahlung*, which is derived from the verb *abbezahlen* in Example 13.[2] If the string that remains after affix separation is more complex than a simple root (usually, root and a derivational suffix; cf. *alpin* after removing *iste* in Example 8), it is called a stem (Haspelmath and Sims, 2010).

3. Existing Language Resources Relevant for Segmentation

Although there is a large number of published resources that are directly or indirectly concerned with morphological segmentation, we considered only a subset of these in our analysis, for a variety of reasons. Most importantly, we preferred resources that are available under free licenses; other factors included the number of handled languages, association with past shared tasks, and our ability to read at least some of the languages from the given resource. The 17 resources selected for harmonization will be briefly described in the rest of this section; see also Table 1.[3]

3.1. Data Resources with Free Licenses

CroDeriV. CroDeriV (Šojat et al., 2014) is a lexical resource of derivational morphology for Croatian, with verbal lemmas extracted from the Croatian morphological lexicon (in its first version[4]).

Démonette. Démonette (Hathout and Namer, 2014) is a morphosemantic lexical database automatically built from the parsing system DériF (Namer, 2009), the Morphonette network (Hathout, 2011), and Verbaction (Tanguy and Hathout, 2002; Hathout et al., 2002). Each entry has a pair of morphologically related lemmas, and defines the first with respect to the second, marking each with a GRACE POS tag (Rajman et al., 1997), affixation, and conversion (if any). The dataset marks a select set of 32 suffixes; allomorphy is rare.

DeriNet. DeriNet 2.1 (Vidra et al., 2021) is a database of word-formation relations in Czech. Its lemmaset is extracted from the MorfFlex dictionary (Hajič et al., 2020) together with Universal POS tags (Petrov et al., 2012). It contains automatic segmentation to morphs induced from derivational trees using an algorithm that traverses the trees recursively and compares base and derived lemmas (Bodnár et al., 2020). The DeriNet project also published manually annotated (complete) morphological segmentation data in its source-control repository – lemmas[5] and form-lemma pairs[6]. The data were sampled with multiple frequency-based strategies.

DerIvaTario. DerIvaTario (Talamo et al., 2016) is a morphological segmentation dataset containing manually annotated Italian lemmas, sampled from the CoLFIS corpus (Bertinetto et al., 2005), marking each lemma with its base and affixes (in order and disambiguated for homonymy). The base is further marked with its type from a set of 9 possible labels, including suppletion, verbal theme, or if the base is unrecoverable.

DerivBaseDE. DErivBase v2 (Zeller et al., 2013) is a wide-coverage lexicon of derivationally related lexemes for German. Derivational relations were identified on the basis of more than 190 rules extracted from German reference grammar books; rules are based on derivational changes (given as string substitutions). The lexemes were extracted from a German web corpus SDeWAC. Homonymy is partly handled by assigning POS categories and gender for some nouns; allomorphy is not handled.

DerivBaseRU. DerivBase.RU 1.0 (Vodolazsky, 2020) is a data resource of derivationally related lexemes for Russian. The methodology of its construction and its format was inspired by and is very similar to that of DErivBase for German, e.g. its creation on the basis of rules extracted from grammar books, and its handling of homonymy. Lemmas of the lexicon were extracted from the Russian portion of Wikipedia and Wiktionary.

Échantinom. Échantinom (Bonami and Tribout, 2021) is a manually annotated morphological resource

[2] All numbered examples are presented in Table 3.

[3] We primarily focused on morphologically segmented data and so did not include e.g. word-formation data such as derivational tree datasets (such as multilingual Universal Derivations (Kyjánek et al., 2021)) or derivational nests (such as POLYMOTS (Gala and Rey, 2008)), although morphological segmentation and word formation are closely related. We included only previously published data, even if created by semi-automatic or automatic segmentation methods; we did not attempt to create any new datasets ourselves, e.g. via application of automatic stemming algorithms such as Porter stemmer (Porter, 1980) or segmenters such as Morfessor (Smit et al., 2014). Other limiting factors included: non-existing or insufficient digitization of printed resources (e.g. in the case of Sokolová et al. (2005)), licenses disallowing redistribution, or actual inaccessibility of data.

[4] Search interface: `http://croderiv.ffzg.hr/`

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

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

Abbreviated name	Original name, version	Languages	License
CroDeriV	CroDeriV 1.0	Croatian	CC BY-SA-3.0
Démonette	Démonette-1.2	French	CC BY-NC-SA 3.0
DeriNet	DeriNet 2.1	Czech	CC BY-NC-SA 3.0
DerIvaTario	DerIvaTario	Italian	CC BY-SA 4.0
DerivBaseDE	DErivBase 2.0	German	CC BY-SA 3.0
DerivBaseRU	DerivBase.Ru 1.0	Russian	Apache-2.0
Échantinom	Échantinom	French	CC BY 4.0
KCIS	KCIS Resources	Marathi, Hindi, Malayalam, Kannada, Bangla	CC BY-NC 4.0
MorphoLex	MorphoLex, MorphoLex-FR	English and French	CC BY-NC-SA 4.0
MorphyNet	MorphyNet, v1	15 languages[a]	CC BY-SA 3.0
PerSegLex	Persian Morphologically Segmented Lexicon 0.5	Persian	CC BY-NC-SA 4.0
Uniparser	Uniparser morphological analyzer	7 languages[b]	MIT License
WordFormationLatin	Word Formation Latin 1.1	Latin	CC BY-NC-SA 4.0
CELEX	CELEX Lexical Database 2.0	Dutch, English, German	non-free[c]
KuznetsEfremDict	Dictionary of Morphemes of Russian	Russian	non-free[c]
MorphoChallenge	MorphoChallenge 2005, 2007-2010	English, Finnish, German, Turkish, (Arabic[d])	non-free[c]
TikhonovDict	Morphemic-spelling dictionary of the Russian language	Russian	non-free[c]

Table 1: Overview of segmentation resources harmonized in UniSegments 1.0.
[a] Catalan, Czech, English, Finnish, French, German, Hungarian, Italian, Mongolian, Polish, Portuguese, Russian, Serbo-Croatian, Spanish, and Swedish. [b] Eastern Armenian (Khurshudian and Daniel, 2009), Erzya, Komi-Zyrian, Meadow Mari, Moksha (all described in Arkhangelskiy (2019)), Tajik (Iskandarova, 2021) and Udmurt (Arkhangelskiy and Medvedeva, 2016). [c] Currently we are not aware of licenses that would allow us to distribute data derived from these resources publicly. [d]The data set contains Arabic, but we do not to include it, since we were unable to create the morph-morpheme alignment.

for French nouns, documenting nominal lemmas sampled from the Lexique (New et al., 2007) and flexique (Bonami et al., 2014) databases, based on frequency. It records affixation, conversion, compounding, or non-concatenative processes, as well as other features such as gender, and the derivational base, along with its POS category for each lemma. Each entry is also marked with a finer-grained label for this process from a set of 29 labels, including back-formation and reduplication.

KCIS. The KCIS datasets[7] (Rao et al., 2014; Bhat et al., 2017) contain treebanks (Tandon and Sharma, 2017); each word in a sentence is marked with an AnnCorra POS tag (Bharati et al., 2006), and a feature structure which includes a list of suffixes of the word form, such as case-markers, postpositions, or verbal inflections. Different language treebanks differ in certain aspects, including completeness, allomorphy, script-related issues (e.g. morph-initial vowels) as well as coverage.

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 morphemes. Words are taken from the English Lexicon Project (Balota et al., 2007), with Penn Treebank tags for English, (Santorini, 1990) and the French Lexicon Project (Ferrand et al., 2010) with added manual segmentation for French. Some inflectional morphemes are omitted from the segmentation, even when occurring inside the word stem (e.g. “ing” in “accordingly”).

MorphyNet. MorphyNet (Batsuren et al., 2021) is a multilingual database of derivational and inflectional morphology for 15 languages:[8] MorphyNet was extracted from Wiktionary using both hand-crafted and automated methods. Morphological information explicitly contained in Wiktionary was enriched by inferring more general (inflectional and derivational) patterns from the data. Each language has separate files in the MorphyNet resource for inflection (containing in-

[7] The treebanks were created by IIT-Bombay (Marathi), IIIT-Hyderabad (Hindi), CDIT, Trivandrum (Malayalam), Jadavpur University, Kolkata (Bengali), MIT-Manipal (Kannada), with contributions from (Angle et al., 2018; Todi et al., 2018; Redkar et al., 2016; Atmakuri et al., 2018). The annotation was funded by the Ministry of Electronics and Information Technology, Government of India.

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

flected forms for each lemma) and for derivation (marking derivational antecedent and last derivational affix for each lemma).

PerSegLex. Persian Morphologically Segmented Lexicon (Ansari et al., 2019) includes complete morphological segmentation of word forms that originate from Persian Wikipedia, popular Persian corpus BijanKhan, and Persian Named Entity corpus. Homonymy of word forms is handled by classifying them into disambiguating categories. The Hazm toolkit (Persian preprocessing and tokenisation tools) was used for segmentation; however, high-frequency words were segmented manually. The file format adheres to the Arabic ordering (right to left).

Uniparser. Uniparser is a finite-state-transducer-like morphological analyzer, optionally combined with constraint grammars (Arkhangelskiy et al., 2012), for 11 languages. The authors also publish lexicons of annotated words extracted from corpora. In addition to lemmatizing and tagging texts, the grammar description can be used to delimit boundaries between the inflectional morphemes of word forms, which is used in the grammars of 7 languages (see Table 1 for a list).

WordFormationLatin. The Word Formation Latin database (Litta et al., 2016) encompasses Latin derivation, compounding, and conversion, also marking POS tags and inflectional categories. The lemma list was compiled from three Classical and Late Latin dictionaries; most of the derivational relationships were either created automatically using a set of different rules or semi-automatically.

3.2. Data Resources with Non-free or Unspecified Licenses

CELEX. CELEX 2 (Baayen et al., 1995) is a phonological and morphological resource for German, Dutch and English. Lemmas are divided into both the constituent affixes and stems that can be used to infer derivational series, and hierarchically into morphemes. Morphemes are classified as free or bound.

KuznetsEfremDict. Dictionary of Morphemes of the Russian Language (Kuznetsova and Efremova, 1986) contains manually annotated morphological segmentations of lemmas. While homonymy of lemmas is partly resolved by assigning POS categories to the lemmas, allomorphy is not handled.

MorphoChallenge. This dataset comes from the MorphoChallenge shared tasks (2005–2010) for morphological segmentation (Kurimo et al., 2010). Its format depends on the year, e.g. whether morphs are labelled with syntactic function or usage of zero morphemes. The data encoding depends on the language, e.g. the Arabic is transliterated via the Buckwalter transliteration. The year 2007 also contains vowelized Arabic.

TikhonovDict. Russian Morphological dictionary (Tikhonov, 1996) contains lemmas segmented (completely) to morphs. The dataset is written in Cyrillic and some words contain hyphens or apostrophes as accent marks.

4. Diversity of Segmentation Annotations

Unsurprisingly, the resources vary widely along several factors. Some of these include: selection of segmented material, principle decisions of which morphological processes are handled (inflectional/derivational, or selected set of affixes as opposed to complete segmentation), as well as manner of treating specific phenomena (such as zero morphemes, compounds, allomorphy, homonymy), completeness of segmentation (single delimited affix as opposed to complete decomposition), manner and extent of annotation (providing information such as POS tags, lemmas, or the semantic nature of affix), presentation format (e.g. hierarchical vs. plain delimitation), and label conventions. See Table 2 for an overview of the key characteristics.

4.1. Segmented Units: Lexical Material

The resources differ widely in size and applied strategies for selecting lexical material. See the *Number of segmented units* in Table 2, showing whether the dataset segments word forms or lemmas, and *POS categories*, showing further constraints on selected material.

Typically, resources use either lists of lexemes from pre-existing lexical resources or frequency lists extracted from corpora, possibly further pruned by selection processes such as random sampling with respect to frequency distributions (one may have different priorities as to coverage of high-frequency or rare words), constraints on POS categories, or retention of only certain word-formation processes.[9]

4.2. Origin of Segments

The original segmentations in the surveyed resources were mostly annotated manually, see column *Segmentation origin* in Table 2, which is more costly but leads to considerably higher quality. However, there are still resources (namely DerivBaseDE, DerivBaseRU and resources in the Uniparser collection) that have been created entirely automatically, and they exploit sets of rules for inflectional and derivational morphology extracted e.g. from grammar books. Combined approaches are used too.[10]

4.3. Nature of Delimited Segments: Morphs, Morphemes, or Both

For simplicity, let us assume that any written word form can be fully decomposed into a sequence of

[9] Interestingly, Échantinom controls for homophony.

[10] A special case here is Démonette which has merged all the existing resources for French and then only manually resolved inconsistencies.

Resource	Number of segmented units:	POS categories:[d]	Segmentation origin:	Segment info:		Completeness of segmentation:	Classification of segments:	Zero morpheme allowed:	Hierarchical segm.:
	k = ×1,000, L = lemmas, W = word forms	N = noun, A = adjective, V = verb, D = adverb, O = other	M = manual, A = automatic	morphs or morpheme (or both)		C = complete, P = partial, S = single affix only	T = stem, R = root, P = prefix, I = interfix, S = suffix, E = ending		
CroDeriV	16 kL	V	M	✓	–	C	R, P, S, E	✓	–
Démonette	42 kL	N, V, A	M + A	✓	–	S	T, S	–	✓
DeriNet	1,039 kL	N, A, D, V, O	M + A	✓	✓	C	R, P, S	–	✓
DerIvaTario	11 kL	N, A, V, O	M	–	✓	C	R	✓	✓
DerivBaseDE	61 kL	N, A, V	A	✓	–	S	P, S	–	✓
DerivBaseRU	156kL	N, V, A, D, O	A	✓	–	S	P, S, E	–	✓
Échantinom	5 kL	N	M	✓	–	S	R, P, S	–	–
KCIS	avg. 26 kW	N, V, O, A, D	M + A	–	✓[a]	P	R, S	–	–
MorphoLex	avg. 43 kW	N, V, A, D, O[b]	M	–	✓	C	R, P, S	–	–
MorphyNet	362 kW+kL	N, A, V, D, O[c]	M + A	✓	–	S	R, P, S	–	–
PerSegLex	8 kW	–	M	✓	–	C	–	–	✓
Uniparser	avg. 277 kW	N, A, V, D, O	A	✓	–	P	T, P, S	✓	–
WordFormationLatin	36 kL	N, A, V, D, O	M + A	–	✓	P	R, P, S	–	✓
CELEX	avg. 77 kL	N, A, V, O, D	M	–	✓	C	R, P, I, S	✓	✓
KuznetsEfremDict	73 kL	N, V, A, D, O	M	✓	–	C	R	–	–
MorphoChallenge 2005	avg. 1 kL	–	M + A	✓	–	C	–	–	–
MorphoChallenge 2007-2010	avg. 2.5 kL	–	M + A	✓	✓	C	–	–	–
TikhonovDict	103 kL	–	M	✓	–	C	–	–	–

Table 2: Diversity of morphological information in the original resources.
[a] All KCIS datasets except Marathi mark morphemes rather than morphs. [b] Information for English; the French resource does not contain POS values. [c] The tag ordering has been obtained from the union of all MorphyNets. Note that the ordering may be different for the individual languages. [d] POS categories are ordered by the number of occurrences in their respective resource.

graphemic segments corresponding to individual morphemes. The resources under study approach this decomposition in three different ways: they may specify either a sequence of morphs, or of morphemes, or of morph+morpheme pairs (see *Segment info* in Table 2). Where morphs are used, they are specified straightforwardly as a sequence of characters and all morphs in a morph sequence are mutually non-overlapping. In most resources, all morphs are contiguous (the exception being Uniparser grammars where an infix may split a root into two non-contiguous parts), and if the segmentation is complete, they result in the whole word form when concatenated.

There is more variability when it comes to specification of morphemes, as morphemes require more abstraction. We observed three approaches to morpheme specification. In the first case, a morpheme is specified using one of its allomorphs, selected in some canonical way (see the morpheme *ad* corresponding to the contextually conditioned prefix morph *ab* in Example 4, or the root morpheme *frais* in Example 12),[11]

In the second case, a morpheme is specified by referring to (the citation form of) the base word; this difference is more obvious with lexical roots (see Example 5) rather than with affixes,[12]

In the third case, a morpheme is specified as a fully abstract unit, without mentioning any form (e.g. *PL* in Example 11 or *[VB]* in Example 12).

In all three cases, identifying exact boundaries between morphs (if only a sequence of morphemes is given in the original resource) is non-trivial and approximative solutions are necessary in some cases (see Section 5.3.1).

Ultimately we would like to have complete segmentation. However, some resources only delimit a single affix added during the last derivation/inflection step; the segmentation may also be incomplete in some other way (see *Completeness of segmentation* in Table 2);for some of the resources, a more complete segmentation

[11] Instead of choosing a single allomorph, regular-expression-like notation in Word Formation Latin is used to represent the full set of allomorphs: see Example 15 for a morpheme beginning with an optional *t*, followed by either *udo* or *udin*; this clearly comes with a risk of overgeneration.

[12] See also (Cotterell et al., 2016) for the notion of canonical segmentation, in which e.g. the German noun *Zulassung* is segmented to “canonical morphemes” *zu lassen ung*.

can be obtained by recursively accumulating boundaries from all derivational antecedents of the word (see Section 5.3.2).

The usual format of resources that store morphological segmentation is a list of input units segmented into desired segments, i.e., morph(eme)s. However, many of the described resources include also a derivational history of the input units, based on which the segments can be traced and identified if they are not in the original resource, cf. *Hierarchical segmentation* in Table 2. For example, Word Formation Latin includes such derivational history of words in a form of the rooted tree for each family of derivationally related words. In these trees, the roots are the shortest unmotivated words, while the leaves are the most complex words in terms of morphology (i.e., a number of segments). This allows us to induce a more detailed segmentation than the resource originally contained (note that WFL contains only partial segmentation of affixes). In addition to the rooted tree data structure of derivational history, the CELEX dataset contains annotations of hierarchy resembling phrase structure trees of segments.

4.4. Classification of Segments

Most resources also classify the split segments to specify whether they are either stems/roots, prefixes/suffixes, inflectional endings, or zero morph(em)s. Despite using the same labels (e.g., root, stem, prefix, interfix, suffix, ending), the resources follow different definitions of the classes. For example, while Démonette uses the label stem for segments containing derivational affixes (i.e. in line with the above mentioned delimitation), CroDeriV marks some segments as stems even though they do not include derivational affixes. Some resources distinguish derivational prefixes and suffixes from inflectional endings (e.g. DerivBaseRU).

CroDeriV, DeriNet, MorphoLex and Uniparser classify each segment into one of the above-mentioned classes; other resources classify only selected segments, such as root morphemes or affixes, or do not classify them at all. Different inventories of segment types are summarized in the column *Classification of segments* in Table 2.

5. Our Harmonized Scheme

The main goal of our work is to provide datasets for many languages in one format. Since the original datasets vary in not only their storage format, but also their annotation schemes and the information they contain (see Section 4), the conversion is necessarily more complex than simple re-interpretation of source data. A balance has to be found between forcing all datasets into one mould (either omitting annotations from datasets that are too rich, or manually adding missing information to sparser sources) and making the mould too loose (thus essentially failing at the stated goal of unification).

5.1. Basic Design Choices

After surveying the available resources, we decided to keep intact the parts which require deep in-language expertise (word forms and lemmatization, where present), unify the information which is available in most resources (POS categories and, on some level, the segmentation itself), and keep as much of the language- or resource-specific information as possible unchanged for users who need it. This ensures the existence of common ground between the diverse resources, while losslessly preserving the extra bits. We do not add missing POS categories or lemmas ourselves.

We decided to omit word-formation information from the converted resources. These annotations are better captured in the Universal Derivations project (Kyjánek et al., 2021) and are considered out of scope here.

We decided to make the notion of morph primary and to represent segmentation by grouping graphemes from the segmented word form (or lemma) and annotating the groups. This required inferring morphs from morphemes in resources that don't delimit morphs explicitly (See Section 5.3.1), but ensures uniform representation across languages, as the notion of morphs as grapheme strings is common to all included languages, while the annotation schemes for morphemes vary. Where applicable, the original morphemes are attached to the inferred morphs as extra annotation.

Other examples of segment annotations available for selected resources include classification of morph types as per Section 4.4, information about the word-formation process that added the morph (in DerIvaTario), or the part-of-speech category for roots (in DerIvaTario and Échantinom). Where practical, the classification of types is added even though it is not present in the original resource (e.g. the annotation of free and bound morphemes from CELEX is converted to roots and affixes).

We allow for non-contiguous morphs, which are used to capture infixation, but zero morphs are not allowed. The segmentation need not be complete – in the extreme, it is possible to store unsegmented lexical material. Unsegmented items were kept in the converted resources, but not counted in Tables 2 and 4.

5.2. File Format

The file format is a combination of line-oriented tab-separated-values format with JSON. There are five columns: the word form, lemma and POS category of the segmented word, a simplified version of the segmentation (intended for viewing and easy browsing of the data) and a JSON map containing all other annotations, including the full segmentation.

The simplified segmentation is the word form with "+" signs inserted between morphs, with no other annotations (as in the last column of Table 3). It is meant as a guide for visual orientation in the file, since the representation of the full segmentation is geared towards programmatic use and too complex to scan for humans.

Ex.	Resource	Data samples in their original formats		Morph segmentation in UniSegments 1.0
1	CELEX	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	→	Leucht + bombe (*photoflash bomb*)
2	CELEX	5290\brinksmanship\0\C\\1\N\N\N\N\Y\brink+s+man+ship\NxNx\SASA\N\N\Y\###\N\N\SASA\ ((brink)[N],(s)[N|N.Nx],(man)[N],(ship)[N|NxN.])[N]\N\N\Y	→	brink + s + man + ship (*brinksmanship*)
3	Démonette	”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”	→	abaiss + e + ment (*lowering*)
4	DerIvaTario	3951;ABBATTIMENTO;BATTERE:vrb_th;ACons:ad:mt2:ms2b;MENTO:mento:mt4:ms1;;;;	→	ab + batt + i + mento (*breakdown*)
5	DerIvaTario	15744;CADENZAMENTO;CADERE:vrb_th;NZA:nza:mt1:ms2b;CONVERSION:N_V; MENTO:mento:mt1:ms1;;;	→	cade + nza + mento (*cadence*)
6	DerivBaseDE	Großstadt_Nf Großstädterin_Nf 2 Großstadt_Nf dNN05:(sfx ”er” & opt uml & try (rsfx ”er” ”r” .||. dsfx ”e” .||. opt (dsfx ”en” .|. rsfx ”en” ”n”) .||. try (dsfx ”ien” .|. rsfx ”ien” ”i”)) & try (rsfx ”ia” ”i”) & opt (rsfx ”a” ”i”)) nouns mNouns> Großstädter_Nm dNN02:(sfx ”in” & try (dsfx ”e”)) nouns nouns> Großstädterin_Nf	→	Großstädt + er + in (*female city dweller*)
7	DerivBaseRU	вымор noun повыморить verb rule887(по + noun + и1(ть) -> verb) PFX,SFX	→	по + вымори + ть (*become extinct*)
8	Échantinom	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	→	alpin + iste (*alpinist*)
9	KCIS (Marathi)	2.2 हवामानामुळे N_NN <fs af=’हवामान,n,n,sg,,o,मुळे,ा_मुळे’ name=’हवामानामुळे’>	→	हवामान + ा + मळे (*due to the weather*)
10	KuznetsEfrem- Dict	вязальщик,”[’вяз’, ’а’, ’льщик’]”,[’вяз’],S,”[0, 3, 4]”,”[[0, 2]]”	→	вяз + а + льщик (*knitter*)
11	MorphoChallenge	act:act_V ion:ion_s s:+PL	→	act + ion + s (*actions*)
12	MorphoLex	rafraîchissant <re«a<(frais)>[VB]»sant>	→	r + a + fraîchis + sant (*refreshing*)
13	MorphyNet	abbezahlen Abbezahlung V N ung suffix	→	Abbezahl + ung (*repayment*)
14	WordFormation- Latin	(15086,’expergefacio’,’V5’,”,’VmM’,’e1596’,’expergefacio’,’VERB’,NULL,’B’) (15092,’expergo’,’V3’,”,’VmH’,’e1601’,’expergo’,’VERB’,NULL,’B’) (15506,’facio’,’V5’,”,’VmM’,’f0048’,’facio’,’VERB’,NULL,’B’) (29306,’pergo’,’V3’,”,’VmH’,’p1180’,’pergo’,’VERB’,NULL,’B’) (15092,1,15086,’221’,’a’,’2017-08-01 08:42:36’) (1550,2,15086,’221’,’a’,’2017-08-01 08:42:36’) (29306,1,15092,’8’,’a’,’2015-11-17 15:06:00’) (’V+V=V’,’Compounding’,’221’,”,’v1*; v2*; v3*; v4*; v5*; v6* + v1*; v2*; v3*; v4*; v5*; v6*’,”,”,’v1*; v2*; v3*; v4*; v5*; v6*’,’assue-facere’) (’V-To-V’,’Derivation_Prefix’,’8’,’e(x)’,’v1*; v2*; v3*; v4*; v5*; v6*’,”,”,’v1*; v2*; v3*; v4*; v5*; v6*’,’e-duc-o’)	→	ex + perg + e + facio (*awaken*)
15	WordFormation- Latin	(32949,’pulchritudo’,’N3B’,’f’,’NcC’,’p4439’,’pulchritudo’,’NOUN’,NULL,’B’) (32945,’pulcher’,’N2/1’,’*’,’Af-’,’p4435’,’pulcher’,’ADJ’,NULL,’B’) (32945,1,32949,’62’,’m’,’0000-00-00 00:00:00’) (’A-To-N’,’Derivation_Suffix’,’62’,”,’n6; n7*’,’(i)’,’(t)udo/udin’,’n31’,’inquiet-udo, -udin-is’)	→	pulch + ri + tudo (*beauty*)

Table 3: Samples of segmentation data before and after harmonization (simplified). Full harmonized samples are shown in Table 5 in Appendix B.

The full segmentation is represented as a list of morphs, with each morph specified by a list of indices indicating which Unicode codepoints of the word form belong to this morph; see Table 5 in Appendix B for examples.

5.3. Examples of Resource Specific Conversion Issues

5.3.1. Aligning Morphemes to Morphs

One adopted algorithm for mapping morphemes (represented as canonical allomorphs) to morphs present in a word form/lemma is based on Levenshtein edit distance. By finding the lowest-cost mapping from the string obtained by concatenating all canonical allomorphs to the word form, we find which allomorph graphemes correspond to which form graphemes and which allomorph graphemes are deleted or added. In the CELEX and MorphoLex databases, edit distance functions as a good indicator of alignment between grapheme spans and morphemes.

For the KCIS Kannada and Malayalam datasets, morph boundaries for morphemes were inferred by greedily choosing the best boundary for each morpheme left to right using string matching between substrings of the word form and all generated transformations of the morpheme. The transformations accounted for the commonest observed types of allomorphy or boundary changes, such as elision of short vowels[13] and viramas[14], and switching between language-specific pairs

[13] We also preliminarily mapped both forms of vowels to a single form in the word form as well as morphemes; since the word form and listed morpheme may differ in this for the same vowel.

[14] A virama is a character suppressing the inherent vowel of a consonantal character.

of consonants; we also applied pairs of transformations. Candidate boundaries were ranked on length of the match as well as constraints on distance from the previous boundary (smaller is better). This approach is geared towards maintaining precision on these morphological rich data, that show high levels of allomorphy with specific patterns, while allowing e.g. overlapping or interfixes.

We apply a similar idea of generating candidate morphs per morpheme while processing words of the MorphoChallenge dataset segmented to abstract morphemes (e.g. "PAST"). In this case, we use exhaustive search to find a combination of candidate morphs[15] that would produce the original word, that is, we pick a candidate for the first morph, and if it is a prefix of the remaining part of the word, we go deeper into the recursion; else, we backtrack. If the process fails to find a combination, we broaden the set of candidate morphs and try again. This allows us to first test the more linguistically plausible results. This approach focuses on choosing allomorphs per morpheme that fit best given the rest of the segmentation, since we may not always have a good starting point for string comparison.

5.3.2. Partial to (More) Complete Segmentations

Word Formation Latin only captures the last word-formation step through which the lemma was coined (derivation, compounding, or conversion; except for base lemmas); in the case of derivation, it also contains the last-added affix, represented by a regular expression that matches every allomorph of that morpheme. However, complete segmentation is possible, because each such item additionally contains an reference to each of its ancestors – one if it is a product of derivation or conversion, or several it is a compound.

Starting with a given non-base item, we can take each morpheme that was last added, add it to the segmentation using regular expression matching [16], and jump to the associated ancestor reference(s). If we do this recursively, we eventually reach a base item as we branch off whenever a compound is encountered. The recursion is broken when the base item in the last branch is reached, completing the segmentation. The procedure always halts, assuming there is no circle of references present anywhere in the dataset. An example of a sequence of incomplete segmentation that has been recursively completed in this manner can be found in Table 3.

5.4. Resulting Collection: UniSegments 1.0

We converted data from all resources listed in Table 1 into the harmonized scheme by automatic converters implemented in Python. The resulting collection consists of 47 datasets for 32 languages stored in the same file format.

The collection was divided into the public edition and non-public edition. The public edition contains 38 harmonized segmentation datasets (30 languages) converted only from original resources with sufficiently free license policies that allow creation and distribution of derived data. The public edition of UniSegments 1.0 can be downloaded from the LINDAT/CLARIAH-CZ repository;[17] an original license is attached to each data file. The public edition is thus readily available e.g. to researchers interested in multilingual morphological segmentation. Future version of the public edition will be made available on the project web page.[18] Resources from the non-public edition (with restrictive or unclear licenses) can be built from source data using converters published in the UniSegments repository.[19]

Basic statistical properties of segmentation data contained both in the public and non-public edition are presented in Table 4 in the appendix.

6. Conclusions and Future Work

We reviewed a number of existing data resources for morphological segmentation, converted them into a unified scheme, and published the resulting data. To the best of our knowledge, no other comparably wide survey of such resources has been published (in terms of the number of resources), and the same holds for the scope of the harmonized data collection.

We believe that UniSegments 1.0 will be useful both for linguists and NLP researchers; among other goals, we would like to use our collection in a future shared on multilingual morphological segmentation.

Our work can be naturally extended to more resources. One of the biggest challenges will be inclusion of highly multilingual inflectional and derivational resources (especially UniMorph, McCarthy et al. (2020)) which, however, deal with segmentation only in a very indirect way.

7. Acknowledgments

We thank all authors of the original resources. This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, the Grant No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935), the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Rep. (project LM2018101), and by the SVV project No. 260 575.

[15] The candidate morphs are, for example, the morpheme itself, shortened versions, and representations of the abstract morpheme seen in parts of the dataset showing both morphemes and allomorphs. In the case of German, where we do not have such representations, we inferred allomorphs of abstract morphemes from word forms that were otherwise simple concatenations. E.g. "aufzustellen auf zu stell_V +INF" → +INF has an allomorph "en".

[16] Since stem allomorphy is not covered in Word Formation Latin, the longest common string was used with stems.

[17] `http://hdl.handle.net/11234/1-4629`

[18] `https://ufal.mff.cuni.cz/universal-segmentations`

[19] `https://github.com/ufal/universal-segmentations`

8. Bibliographical References

Angle, S., Rao, B. A., and Muralikrishna, S. (2018). Kannada morpheme segmentation using machine learning. *International Journal of Engineering and Technology (UAE)*, 7(2):45–49.

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

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, December. The COLING 2012 Organizing Committee.

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, January. Association for Computational Linguistics.

Aronoff, M. and Fudeman, K. (2011). *What is Morphology?* Wiley-Blackwell, Malden, MA, 2nd edition.

Aronoff, M. (1976). *Word Formation in Generative Grammar*. The MIT Press, Cambridge.

Atmakuri, S., Shahi, B., Rao, B. A., and Muralikrishna, S. (2018). A comparison of features for pos tagging in kannada. *International Journal of Engineering and Technology (UAE)*, 7(4):2418–2421.

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, Aug.

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, August. 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.

Bodnár, J., Žabokrtský, Z., and Ševčíková, M. (2020). Semi-supervised induction of morpheme boundaries in czech using a word-formation network. In *23rd International Conference on Text, Speech and Dialogue*, pages 189–196, Cham, Switzerland. Springer.

Bonami, O., Caron, G., and Plancq, C. (2014). Construction d'un lexique flexionnel phonétisé libre du français. In *SHS Web of Conferences*, volume 8, pages 2583–2596. EDP Sciences.

Cotterell, R., Vieira, T., and Schütze, H. (2016). A joint model of orthography and morphological segmentation. In *Proceedings of NAACL*. Association for Computational Linguistics.

de Marneffe, M.-C., Manning, C., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

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, May.

Gala, N. and Rey, V. (2008). POLYMOTS : une base de données de constructions dérivationnelles en français à partir de radicaux phonologiques. In *Actes de la 15ème conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, pages 91–100, Avignon, France, June. ATALA.

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

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 Paul Boucher, editor, *Many Morphologies*, pages 178–209. Cascadilla, Somerville, Mass.

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

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.

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, June.

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.

New, B., Brysbaert, M., Veronis, J., and Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied psycholinguistics*, 28(4):661–677.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*.

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.

Redkar, H., Singh, S., Ghag, N., Paranjape, J., Joshi, N., Kulkarni, M., and Bhattacharyya, P. (2016). Verbframator: Semi-automatic verb frame annotator tool with special reference to marathi. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 299–304.

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, jan.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Smit, P., Virpioja, S., Grönroos, S.-A., Kurimo, M., et al. (2014). Morfessor 2.0: Toolkit for statistical morphological segmentation. In *The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26-30, 2014*. Aalto University.

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.

Sokolová, M., Ološtiak, M., Ivanová, M., Šimon, F., Czéreová, B., Vužňáková, K., Benko, V., and Moško, G. (2005). *Slovník koreňových morfém slovenčiny*. Prešovská univerzita.

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, August.

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

Tandon, J. and Sharma, D. M. (2017). Unity in diversity: A unified parsing strategy for major indian languages. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 255–265.

Tanguy, L. and Hathout, N. (2002). Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web. In Jean-Marie Pierrel, 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.

Todi, K. K., Mishra, P., and Misra Sharma, D. (2018). Building a kannada pos tagger using machine learning and neural network models. *arXiv e-prints*, pages arXiv–1808.

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.

9. Language Resource References

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.

Baayen, R. Harald and Piepenbrock, Richard and Gulikers, Leon. (1995). *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, ISLRN 204-698-863-053-1.

Bhat, R. A., Bhatt, R., Farudi, A., Klassen, P., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. M., Vaidya, A., Vishnu, S. R., et al. (2017). The hindi/urdu treebank project. In *Handbook of Linguistic Annotation*, pages 659–697. Springer.

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

Hajič, Jan and Hlaváčová, Jaroslava and Mikulová, Marie and Straka, Milan and Štěpánková, Barbora. (2020). *MorfFlex CZ 2.0*. Institute of Formal and Applied Linguistics, LINDAT/CLARIN, Charles University.

Kyjánek, L., Žabokrtský, Z., Vidra, J., and Ševčíková, M. (2021). Universal derivations v1.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Rao, A. B., Murali Krishna, S., and Nayak, A. (2014). Developing a dependency treebank for kannada. *International Journal of Engineering Sciences and Research*, 15:5–7.

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.

A. Selected Statistical Properties of UniSegments 1.0

		Distribution of				Mean	Mean unit	Mean
		morphs per unit [%]				morphs	length	morph
Resource name	Size	1	2	3	4+	per unit	[char]	len [char]
ben-KCIS	1 kW	0	100	0	0	2.0	5.6	2.8
cat-MorphyNet	516 kL	0	100	0	0	2.0	10.5	5.1
ces-DeriNet	1,039 kL	8	16	19	57	4.1	11.1	2.7
ces-MorphyNet	67 kL	0	100	0	0	2.0	9.4	4.6
deu-CELEX	48 kL	14	40	34	13	2.5	10.0	4.0
deu-DerivBaseDE	61 kL	36	59	4	0	1.7	11.2	6.6
deu-MorphoChallenge	3 kL	4	27	42	27	3.0	10.5	3.5
deu-MorphyNet	29 kL	0	100	0	0	2.0	10.6	5.1
eng-CELEX	44 kL	30	51	16	3	1.9	8.6	4.5
eng-MorphoChallenge	3 kL	16	49	27	9	2.3	8.4	3.7
eng-MorphoLex	69 kW	21	45	27	7	2.2	8.3	3.8
eng-MorphyNet	292 kL	0	100	0	0	2.0	10.7	5.1
fas-PerSegLex	45 kW	34	31	24	10	2.1	6.8	0.0
fin-MorphoChallenge	4 kL	3	18	35	44	3.4	13.0	3.8
fin-MorphyNet	400 kL	0	100	0	0	2.0	10.7	5.2
fra-Démonette	63 kL	46	80	3	0	1.7	9.9	5.9
fra-Échantinom	5 kL	53	40	6	1	1.5	7.8	5.1
fra-MorphoLex	16 kW	43	44	12	1	1.7	8.2	4.7
fra-MorphyNet	363 kL	0	100	0	0	2.0	10.7	5.1
hbs-MorphyNet	34 kL	0	100	0	0	2.0	10.3	4.9
hin-KCIS	2 kW	29	71	0	0	1.7	4.4	2.3
hrv-CroDeriV	16 kL	0	1	20	79	4.1	9.7	2.3
hun-MorphyNet	428 kL	0	100	0	0	2.0	10.6	5.1
hye-Uniparser	594 kW	9	41	37	13	2.6	9.6	3.7
ita-DerIvaTario	11 kL	1	46	31	21	2.8	10.9	3.9
ita-MorphyNet	599 kL	0	100	0	0	2.0	10.5	5.1
kan-KCIS	26 kW	0	11	25	64	4.4	9.4	2.5
kpv-Uniparser	205 kW	9	40	35	16	2.6	8.7	3.3
lat-WordFormationLatin	36 kL	16	52	27	5	2.2	9.0	3.2
mal-KCIS	33 kW	2	98	0	0	2.0	12.5	4.7
mar-KCIS	32 kW	0	51	43	6	2.5	8.3	3.3
mdf-Uniparser	105 kW	10	50	31	8	2.4	9.0	3.8
mhr-Uniparser	260 kW	9	38	36	17	2.7	9.1	3.4
mon-MorphyNet	35 kL	0	100	0	0	2.0	10.2	4.9
myv-Uniparser	164 kW	10	41	36	13	2.5	9.1	3.6
nld-CELEX	101 kL	11	52	25	12	2.4	10.8	4.3
pol-MorphyNet	508 kL	0	100	0	0	2.0	10.5	5.1
por-MorphyNet	449 kL	0	100	0	0	2.0	10.6	5.1
rus-DerivBaseRU	156 kL	31	35	23	10	2.1	10.3	4.8
rus-KuznetsEfremDict	73 kL	1	7	17	75	4.3	9.9	2.3
rus-MorphyNet	692 kL	0	100	0	0	2.0	10.5	5.1
rus-TikhonovDict	103 kL	6	11	22	61	3.8	10.2	2.7
spa-MorphyNet	541 kL	0	100	0	0	2.0	10.4	5.1
swe-MorphyNet	438 kL	0	100	0	0	2.0	10.6	5.1
tgk-Uniparser	232 kW	17	56	24	3	2.1	7.8	3.6
tur-MorphoChallenge	7 kL	3	19	34	45	3.4	10.5	3.0
udm-Uniparser	375 kW	8	35	36	21	2.8	9.1	3.3

Table 4: Basic statistics of all included datasets. The first three letters of each resource name are ISO 639-3 language codes, italicized name marks non-public datasets. The “Size” column lists the size in units of thousands of lemmas (L) or inflected word forms (W), depending on what information the dataset contains. In the other columns, “unit” means either lemma or form. All statistics only consider segmented units; the resources may contain additional unsegmented lexical material, which is ignored here.

B. Samples of Converted Resources

Ex.	Word form	Lemma	POS	Simple segmentation	Full data
1	Leuchtbombe	Leuchtbombe	NOUN	Leucht + bombe	{”annot_name”: ”gCELEX”, ”morpheme_order”: ”N;N”, ”older_ortho”: ”Leuchtbombe”, ”segmentation”: [{”morpheme”: ”licht”, ”span”: [0, 1, 2, 3, 4, 5], ”type”: ”root”}, {”morpheme”: ”bombe”, ”span”: [6, 7, 8, 9, 10], ”type”: ”root”}], ”segmentation_hierarch”: ”(((licht)[A],(e)[N|A.])[N],(Bombe)[N])[N]”, ”segmentation_stem”: ”Leuchte;Bombe”}
2	brinksmanship	brinksmanship	NOUN	brink + s + man + ship	{”annot_name”: ”eCELEX”, ”morpheme_order”: ”S;A;S;A”, ”segmentation”: [{”morpheme”: ”brink”, ”span”: [0, 1, 2, 3, 4], ”type”: ”root”}, {”morpheme”: ”s”, ”span”: [5], ”type”: ”interfix”}, {”morpheme”: ”man”, ”span”: [6, 7, 8], ”type”: ”root”}, {”morpheme”: ”ship”, ”span”: [9, 10, 11, 12], ”type”: ”suffix”}], ”segmentation_hierarch”: ”((brink)[N],(s)[N|N.Nx],(man)[N],(ship)[N|NxN.])[N]”, ”segmentation_stem”: ”brink;s;man;ship”}
3	abaissement	abaissement	NOUN	abaiss + e + ment	{”annot_name”: ”derif”, ”gender”: ”masc”, ”number”: ”sg”, ”root”: ”abaiss”, ”segmentation”: [{”span”: [0, 1, 2, 3, 4, 5], ”type”: ”root”}, {”span”: [6], ”type”: ”interfix”}, {”span”: [7, 8, 9, 10], ”type”: ”suffix”}]}
4	abbattimento	abbattimento	NOUN	ab + batt + i + mento	{”annot_name”: ”DerIvaTario”, ”colfis_id”: ”3951”, ”root”: ”battere”, ”root_type”: ”vrb_th”, ”segmentation”: [{”doubling”: true, ”morpheme”: ”acons”, ”ms”: ”2b”, ”mt”: ”2”, ”ordering”: 1, ”span”: [0, 1], ”type”: ”prefix”}, {”morpheme”: ”battere”, ”ordering”: 0, ”root_type”: ”vrb_th”, ”span”: [2, 3, 4, 5], ”type”: ”root”}, {”ordering”: 0, ”span”: [6], ”type”: ”interfix”}, {”morpheme”: ”mento”, ”ms”: ”1”, ”mt”: ”4”, ”ordering”: 2, ”span”: [7, 8, 9, 10, 11], ”type”: ”suffix”}], ”upos”: ”NOUN”}
5	cadenzamento	cadenzamento	NOUN	cade + nza + mento	{”annot_name”: ”DerIvaTario”, ”colfis_id”: ”15744”, ”root”: ”cadere”, ”root_type”: ”vrb_th”, ”segmentation”: [{”conv_type”: ”n_v”, ”ordering”: 2, ”process_type”: ”conversion”, ”span”: [], ”type”: ”suffix”}, {”morpheme”: ”cadere”, ”ordering”: 0, ”root_type”: ”vrb_th”, ”span”: [0, 1, 2, 3], ”type”: ”root”}, {”morpheme”: ”nza”, ”ms”: ”2b”, ”mt”: ”1”, ”ordering”: 1, ”span”: [4, 5, 6], ”type”: ”suffix”}, {”morpheme”: ”mento”, ”ms”: ”1”, ”mt”: ”1”, ”ordering”: 3, ”span”: [7, 8, 9, 10, 11], ”type”: ”suffix”}], ”upos”: ”NOUN”}
6	Großstädterin	Großstädterin	NOUN	Großstädt + er + in	{”annot_name”: ”DErivBase-2.0”, ”segmentation”: [{”span”: [0, 1, 2, 3, 4, 5, 6, 7, 8], ”type”: ”unsegmented”}, {”span”: [9, 10], ”type”: ”suffix”}, {”span”: [11, 12], ”type”: ”suffix”}]}
7	повыморить	повыморить	VERB	по + вымори + ть	{”annot_name”: ”DerivBase.Ru-1.0”, ”segmentation”: [{”span”: [0, 1], ”type”: ”prefix”}, {”span”: [2, 3, 4, 5, 6, 7], ”type”: ”unsegmented”}, {”span”: [8, 9], ”type”: ”ending”}]}
8	alpiniste	alpiniste	NOUN	alpin + iste	{”annot_name”: ”echantinom”, ”base”: ”alpin”, ”base_pos”: ”ADJ”, ”gender”: ”masc”, ”last_morph_process”: ”suffix”, ”last_process_broad”: ”suffix”, ”segmentation”: [{”morpheme”: ”alpin”, ”span”: [0, 1, 2, 3, 4], ”type”: ”root”}, {”allomorph”: ”ist”, ”morpheme”: ”iste”, ”span”: [5, 6, 7, 8], ”type”: ”suffix”}]}
9	हवामानामुळे		NOUN	हवामान + ◌ा + मुळे	{”AnnCorra_tag”: ”N_NN”, ”annot_name”: ”kcis”, ”case”: ”o”, ”case_marker”: ”मुळे”, ”gender”: ”n”, ”lcat”: ”n”, ”number”: ”sg”, ”root”: ”हवामान”, ”segmentation”: [{”span”: [0, 1, 2, 3, 4, 5], ”type”: ”root”}, {”span”: [6], ”type”: ”interfix”}, {”span”: [7, 8, 9, 10], ”type”: ”suffix”}]}
10	вязальщик	вязальщик	NOUN	вяз + а + льщик	{”annot_name”: ”Dictionary of Russian Morphemes”, ”segmentation”: [{”span”: [0, 1, 2], ”type”: ”root”}, {”span”: [3], ”type”: ”suffix”}, {”span”: [4, 5, 6, 7, 8], ”type”: ”suffix”}]}
11	actions	actions	X	act + ion + s	{”annot_name”: ”MorphoChallenge2010”, ”segmentation”: [{”morph”: ”act”, ”morpheme”: ”act”, ”span”: [0, 1, 2], ”type”: ”X”}, {”morph”: ”ion”, ”morpheme”: ”ion”, ”span”: [3, 4, 5], ”type”: ”X”}, {”morph”: ”s”, ”morpheme”: ”+PL”, ”span”: [6], ”type”: ”X”}]}
12	rafraîchissant	rafraîchissant		r + a + fraîchis + sant	{”annot_name”: ”MorphoLex_fr”, ”segmentation”: [{”morpheme”: ”re”, ”span”: [0], ”type”: ”prefix”}, {”morpheme”: ”a”, ”span”: [1], ”type”: ”prefix”}, {”morpheme”: ”frais”, ”span”: [2, 3, 4, 5, 6, 7, 8, 9], ”type”: ”root”}, {”morpheme”: ”sant”, ”span”: [10, 11, 12, 13], ”type”: ”suffix”}]}
13	Abbezahlung	Abbezahlung	NOUN	Abbezahl + ung	{”annot_name”: ”MorphyNet deu”, ”segmentation”: [{”span”: [0, 1, 2, 3, 4, 5, 6, 7], ”type”: ”root”}, {”span”: [8, 9, 10], ”type”: ”suffix”}]}
14	expergefacio	expergefacio	VERB	ex + perg + e + facio	{”annot_name”: ”Word Formation Latin”, ”segmentation”: [{”morpheme”: ”e(x)”, ”span”: [0, 1], ”type”: ”prefix”}, {”morpheme”: ”pergo”, ”span”: [2, 3, 4, 5], ”type”: ”root”}, {”morpheme”: ”facio”, ”span”: [7, 8, 9, 10, 11], ”type”: ”root”}]}
15	pulchritudo	pulchritudo	NOUN	pulch + ri + tudo	{”Declension”: ”c”, ”Gender”: ”Fem”, ”annot_name”: ”Word Formation Latin”, ”segmentation”: [{”morpheme”: ”pulcher”, ”span”: [0, 1, 2, 3, 4], ”type”: ”root”}, {”morpheme”: ”(t)udo/udin”, ”span”: [7, 8, 9, 10], ”type”: ”suffix”}]}

Table 5: Examples from Table 3 converted into the UniSegments format.