Modeling Word-Internal Structures: Morphological Segmentation Across 58 Languages

Vojtěch John, Benjamin Reeves, Zdeněk Žabokrtský
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, Prague, Czech Republic
benjamin.reeves286@student.cuni.cz
{john, zabokrtsky}@ufal.mff.cuni.cz

Abstract

We present the largest multilingual experiment to date on word-to-morph segmentation, covering 58 typologically diverse languages. We describe a newly compiled collection of linguistically annotated resources for the task, providing broad coverage and enabling systematic cross-lingual evaluation. Second, we train two neural models on surface morphological segmentation, achieving 81% average word accuracy on the original datasets, slightly outperforming previous methods. Experiments on custom test sets reveal substantial variation in performance, highlighting the need for further harmonization and more robust multilingual approaches.

Keywords: morph, morphology, morphological segmentation, multilingual NLP

1. Introduction

Morph segmentation provides a practical intermediate layer between raw word forms and higher-level linguistic annotation. By dividing words into smaller meaning-bearing units, it can help reduce data sparsity, especially in morphologically rich languages (which can improve e.g. machine translation in low-resource settings (Mager et al., 2022)), and make patterns of inflection and derivation more transparent in structured datasets (as seen in e.g. (McCarthy et al., 2020)). Typically, morphological structure implicitly affects how lexical resources are organized, how annotations are aligned across tools and languages, and how systems are evaluated. Making word-internal structure explicit and cross-linguistically analyzable can therefore support clearer annotation guidelines, more controlled comparisons between approaches, and also provide valuable insights into language evolution, including phenomena arising from language contact. In some settings, morphological segmentation has proved to improve performance on downstream task, especially for low-resource languages.

Previous research, with the SIGMORPHON 2022 Shared Task on Morpheme Segmentation (Batsuren et al., 2022) being one of the most ambitious collective efforts, has still covered only a limited range of typologically diverse languages, which restricts the generalizability of its findings. Studying word-internal morphological segmentation across many languages is therefore important, as morphological systems differ substantially in structure and transparency. Truly polysynthetic or strongly introflexive languages remain largely beyond the current empirical scope and represent a longer-term goal, but even broadening coverage toward more agglutinative and fusional languages already should facilitate more robust evaluation, clearer identification of language-specific biases, and more broadly applicable modeling assumptions.

Although a multilingual perspective on segmentation is essential, advancement in the field is constrained by the fragmented and inconsistently annotated landscape of available morphological data. Data resources existing for individual languages vary widely in format, annotation guidelines, size, and quality, making cross-linguistic comparisons difficult and inconsistent. To enable robust evaluation, reproducible research, and meaningful benchmarking, there is a pressing need for standardized, harmonized, and high-quality morphologically segmented datasets that can be applied consistently across languages and systems. Our goal is to bring the world of morphological segmentation closer to the level of standardization and interoperability already achieved in the dependency syntax community through Universal Dependencies (de Marneffe et al., 2021).

Our long-term research questions focus on how to represent word-internal structure consistently across typologically diverse languages and how to standardize datasets and annotation practices to enable robust cross-linguistic evaluation, as well as how to develop systems that can perform this segmentation automatically. Looking further ahead, we aim to connect morph segmentation with other linguistic layers, explore typological effects on model performance, and establish benchmarks and metrics that support reproducible and comparable results.

The remainder of the paper is organized as follows. Section 2 summarized related work. Sec-

Proceedings of the Workshop on Structured Linguistic Data and Evaluation, pages 180–190

tion 3 describes the data collection that we use in our experiments. Section 4 details the evaluation setup. Section 5 presents the modeling approaches and experimental results, which are further interpreted in Section 6. Section 7 concludes and presents future work.

2. Basic notions and Related work

As noted by Haspelmath (2020), the notion of *morpheme* is subject to considerable terminological ambiguity in linguistics. In this work, we adopt the following distinction. A *morpheme* is an abstract linguistic unit defined as the smallest unit of language that carries semantic or grammatical function. A *morph*, by contrast, is the concrete realization of a morpheme in actual language use, i.e., a specific phonological or orthographic form that instantiates a given morpheme.

Morphs may constitute entire words (e.g. *house*, consisting of a single root morph), or they may form parts of more complex words (e.g. *un+kind+ness*, which contains three morphs). The central element of a word is the *root morphs*, conveying its core lexical meaning. Additional morphs, if present, are classified relative to the root. A root may be preceded by one or more *prefixes* (e.g. *un-* in *un+kind*), and followed by one or more *suffixes* (e.g. *-ness* in *kind+ness*). A suffix expressing inflectional categories such as number or tense (e.g. *-ed* in *walk+ed*) is often referred to as an *ending*. In compounds containing multiple roots, linking elements or *interfixes* may appear between the roots (e.g. *speed+o+meter*).

The term *morph* has been used in linguistics for less than a century. Although early attempts at a rigorous delimitation of the morpheme date back at least to the 1940s (Hockett, 1947), the notions of morph and morpheme are still frequently employed only implicitly and informally in linguistic practice. Nevertheless, there exist numerous attempts at lexicographic description of the morphemic structure of words in individual languages over the past decades, though these efforts remain largely isolated. Somewhat surprisingly, even though morphology is often considered a lower level of abstraction compared to syntax, there is still no *de facto* standard for morphological segmentation comparable, in terms of both community size and coverage of languages, to the Universal Dependencies project in syntax (de Marneffe et al., 2021). A few multilingual harmonization initiatives have emerged in the last decade, such as the UniMorph collection (McCarthy et al., 2020), the UniSegments collection (Žabokrtský et al., 2022), and the SIGMORPHON 2022 Shared Task on Morpheme Segmentation (Batsuren et al., 2022), but these resources remain considerably smaller, providing complete morphological segmentation for samples of words for 14, 32, and 9 languages, respectively.

Approaches to automatic morphological segmentation closely mirror the broader development of NLP. Focusing first on methods aimed at producing linguistically interpretable divisions of words into smaller units, the earliest approaches were rule-based (Klenk and Langer, 1989), followed by supervised machine learning methods (Na, 2015), and more recently by deep learning sequence labeling techniques (Girrbach, 2022) and large language models (Pranjić et al., 2024). In parallel, unsupervised segmentation has been an important research line, with Morfessor (Creutz and Lagus, 2005) as its most widely known representative; however, the resulting segmentations are typically of limited linguistic interpretability.

With the rise of deep learning as the dominant paradigm in NLP, the need to segment words into shorter units emerged rapidly. This was driven not by linguistic concerns, but by a practical technical constraint: the input layer of neural networks cannot feasibly accommodate the full vocabulary of a language. These shorter units, referred to as *subwords* (Sennrich et al., 2016), do not correspond to morphs or morphemes, yet subword segmentation methods are frequently used as baselines in morphologically informed NLP research.

Evaluating morphological segmentation is challenging due to the inherently ambiguous nature of morphemes and the variety of possible segmentation schemes. Commonly used metrics include precision, recall, and F1-score at the morph boundary level, which measure how well predicted boundaries match a gold-standard segmentation (Creutz and Lagus, 2005; Narasimhan et al., 2015). Another common approach is word-level accuracy, which requires the entire word to be segmented correctly, providing a stricter measure of performance (Batsuren et al., 2022). In multilingual settings, evaluation is often complicated by the heterogeneity of annotation schemes across languages, differing tokenization conventions, and the presence of allomorphy or non-concatenative morphology. Fully capturing linguistic validity remains an open challenge.

3. Data Collection

This section describes a newly compiled collection of diverse language resources relevant to morphological segmentation. The collection makes use of the harmonization annotation scheme of UniSegments 1.0 (Žabokrtský et al., 2022), and reuses some of its data components too (after a revision). Compared to UniSegments 1.0, the expanded collection covers nearly twice as many languages. A subset of the current collection – comprising only

converted resources whose original licenses permit modification and redistribution – is now publicly available as UniSegments 1.5 (John et al., 2026).

Table 1 summarizes all integrated resources (including those that we cannot distribute further) along with their main qualitative and quantitative properties. The original datasets differ along multiple dimensions, some of which are described in the following paragraphs.

The collected resources reflect the landscape of modern morphological research, bridging the gap between small-scale, linguistically grounded "gold" data and large-scale, automatically generated "silver" and "bronze" datasets. This collection encompasses scripts as diverse as Malayalam, Greek, and Kanji, with a primary objective of maximizing the volume of fully segmented data. Within this framework, resources providing segmentation for roots, inflections, and derivations are categorized as Complete. Conversely, those where the stem remains unsegmented are designated as Partial. These partial resources typically feature a single segmentation boundary per word, usually distinguishing between the stem and its inflectional or derivational affixes. While the granularity of segment labeling varies significantly across the collection, most labeled resources identify at least the prefix, suffix, and root, though a substantial portion remains unlabeled.

Data quality is largely a function of the segmentation methodology. Manual segmentation serves as the gold standard, whereas automatic methods, while more scalable, frequently introduce noise. To mitigate this, many resources in this set were either automatically generated and human-verified or developed through rule-based methods to ensure cross-linguistic consistency. Notably, 55 resources contain manual annotations, 15 of which originate from Metamorphosis, a newly developed manually annotated dataset. These expert-led approaches generally yield higher-fidelity data than unsupervised or semi-supervised machine learning models. Consequently, the collection as a whole represents a strategic trade-off between high-volume noisy data and high-quality, small-scale datasets, with resource sizes ranging from 200 to over 740,000 wordforms or lemmas.

The final corpus comprises 85 resources covering 58 languages from 14 language families, including Niger-Congo, Indo-European, Uralic, Turkic, and Austronesian. While the set spans four continents, there is a recognized geographic imbalance: North and South American languages are absent, and African and Asian languages are underrepresented relative to European ones. Furthermore, while 20 distinct scripts are represented, five resources have been transliterated into the Latin alphabet, a process that may obscure native orthographic cues essential for morphological analysis.

This diverse linguistic composition enables a granular analysis of morphological systems across the typological spectrum. This ranges from the strictly isolating profile of Mandarin Chinese to the extreme polysynthetic complexity of Adyghe. The dataset further represents fusional structures through Latin, Czech, and Greek, alongside the highly productive agglutinating systems of Turkish, Hungarian, and Korean. Beyond broad typological classification, the collection captures specialized morphological processes: Amharic provides a case study for non-concatenative (root-and-pattern) morphology, while Indonesian, Tagalog, Swahili, Zulu, Xhosa, and Hindi illustrate diverse reduplicative patterns. Furthermore, extensive compounding is evident in the Germanic and Uralic families, as well as Chinese, while the inclusion of the Romance group and Modern Greek facilitates the study of cliticization.

4. Evaluation strategies

Although superficially harmonized, the resources differ in several important respects. Firstly, some of the datasets cover only a highly specific part of lexicon (e.g. *SlavickovaDict* or *CroDeriV* covers only verb infinitives). Hence, it is unclear how models trained on such data would behave on a more representative sample. In the resources, that were not originally developed for morphological segmentation (e.g. *DerIvaTario* for word formation or *Uniparser* for morphological analysis), the quality of the segmentation varies both in and across the resources. For example, in DerIvaTario, the infinitives are usually not segmented, because they are often regarded as the unmotivated words. Furthermore, for some of the languages, we have only resources with partial segmentation. Most of the resources are segmented wordlists, often without any further annotation. As a result, we did not attempt to resolve differences in segmentation between homonyms. Since we aim for broad coverage, we have used all of the resources irrespective of their overall quality. In this section, we describe issues this causes for evaluation of the automatic segmenters, evaluation strategies we have adopted to mitigate them, and the limitations this imposes on the interpretation of the results.

As a result of the widely differing quality of the resources, evaluating the automatic segmentation on a test set sampled from the same resource as the train set is problematic. The metrics can devolve to mere consistency of the automatic segmentation with the resource, rather than actual quality of segmentation. It is unclear how well such results approximate the performance on complete segmentation of arbitrary words. For exam-

ISO	Family	Resource Name	Script	Count	Meth.	C/P	Cat.	Avg M/W	Avg M.Len	Avg W.Len	Reference
ady	NWC	MorphAGram	Cyrillic	999	A+H	C	–	5.92	2.08	12.30	(Eskander et al., 2020)
aka	NC	LDC_RLP	Latin	2047	A+R	C	–	2.38	2.28	5.44	(Tracey and Strassel, 2020)
amh	AA	Amharic	Latin^T	135184	M	C	R	3.75	2.45	9.21	(Yeshambel et al., 2021)
bel	IE	Slounik	Cyrillic	35219	M	C	D, I, R, X	3.77	2.43	9.17	(Morozov et al., 2025)
ben	IE	kcis	Bengali	822	M+A	P	R, S	2.00	2.81	5.61	(LTRC, 2018)
cat	IE	Morphynet	Latin	19777	M+A	P	P, R, S	1.90	4.92	9.32	(Batsuren et al., 2021)
ces	IE	SlavickovaDict	Latin	14582	M+A	C	–	2.59	2.24	5.79	(Slavíčková, 1975)
ces	IE	Morphynet	Latin	351251	M+A	P	P, R, S	2.17	4.43	9.61	(Batsuren et al., 2021)
deu	IE	MorphoChallenge	Latin	2863	M+A	C	–	2.98	3.51	10.46	(Kurimo et al., 2010)
deu	IE	CELEX	Latin	47583	M	C	Int, P, R, S	2.37	2.72	6.44	(Baayen et al., 1995)
deu	IE	Morphynet	Latin	236544	M+A	P	P, R, S	2.10	5.30	11.16	(Batsuren et al., 2021)
ell	IE	GreekAnnot.Dict.	Greek	8419	A+S	C	D, I, R	2.40	3.12	7.49	(UI and NTUA, 2021)
eng	IE	MorphoChallenge	Latin	3355	M+A	C	–	2.30	3.67	8.44	(Kurimo et al., 2010)
eng	IE	CELEX	Latin	43957	M	C	Int, P, R, S	1.40	3.87	5.42	(Baayen et al., 1995)
eng	IE	MorphoLex	Latin	68549	M	C	P, R, S	2.21	3.77	8.34	(Sánchez-Gutiérrez et al., 2018)
eng	IE	Morphynet	Latin	387418	M+A	P	P, R, S	2.18	4.86	10.58	(Batsuren et al., 2021)
fas	IE	PerSegLex	Perso-Ar.	45369	M	C	–	2.12	3.20	6.80	(Ansari et al., 2019)
fin	Ur	MorphoChallenge	Latin	3859	M+A	C	–	3.43	3.81	13.04	(Kurimo et al., 2010)
fin	Ur	Morphynet	Latin	740711	M+A	P	P, R, S	2.31	5.65	13.07	(Batsuren et al., 2021)
fra	IE	MorphoLex	Latin	15954	M+A	C	P, R, S	2.03	3.15	6.40	(Mailhot et al., 2020)
fra	IE	demonette	Latin	81494	M+A	P	C, Int, P, R, S	1.72	4.70	8.20	(Hathout and Namer, 2014)
fra	IE	Morphynet	Latin	132094	M+A	P	P, R, S	1.84	5.66	10.42	(Batsuren et al., 2021)
hbs	IE	Morphynet	Latin, Cyrillic	4915	M+A	P	P, R, S	1.84	4.60	8.46	(Batsuren et al., 2021)
hin	IE	kcis	Devanagari	1603	M+A	P	R, S	1.89	3.60	6.78	(LTRC, 2018)
hin	IE	LDC_RLP	Latin^T	2029	A+H	C	R	1.72	2.33	4.40	(Tracey and Strassel, 2020)
hrv	IE	CroDeriV	Latin	15657	M	C	E, Int, P, R, S	4.08	2.30	9.66	(Šojat et al., 2014)
hun	Ur	LDC_RLP	Latin	2027	A+H	C	R	2.48	3.42	8.47	(Tracey and Strassel, 2020)
hun	Ur	Morphynet	Latin	28176	M+A	P	P, R, S	2.23	4.13	9.21	(Batsuren et al., 2021)
hye	IE	Uniparser	Armenian	593542	A+R	P	P, R, S	2.56	3.75	9.59	(Arkhangelskiy et al., 2012)
ind	AN	LDC_RLP	Latin	2035	A+H	C	–	1.79	4.23	7.58	(Tracey and Strassel, 2020)
ita	IE	DerIvaTario	Latin	11287	M	C	R	2.80	3.86	10.92	(Talamo et al., 2016)
ita	IE	Morphynet	Latin	80532	M+A	P	P, R, S	2.32	2.65	6.14	(Batsuren et al., 2021)
jpn	Ja	MorphAGram	Kanji, Hirigana	999	M	C	–	2.24	1.43	3.21	(Eskander et al., 2020)
kan	Dr	kcis	Kannada	25953	M+A	P	Int, P, R, S	4.36	2.46	9.45	(LTRC, 2018)
kat	Ka	MorphAGram	Mkhedruli	999	A+H	C	–	3.04	2.38	7.24	(Eskander et al., 2020)
kor	Ko	LDC2004T03	Latin^T	6455	A+H	C	–	1.77	5.47	9.66	(Han, 2004)
kpv	Ur	Uniparser	Cyrillic	216128	A+R	P	P, R, S	2.52	3.38	8.54	(Arkhangelskiy et al., 2012)
lat	IE	WFL	Latin	27998	M+A	P	P, R, S	2.39	3.53	8.48	(Litta Modignani et al., 2023)
mal	Dr	kcis	Malayalam	33657	M+A	P	P, R, S	1.99	4.70	12.40	(LTRC, 2018)
mar	IE	kcis	Devanagari	32523	M+A	P	Int, R, S	2.54	3.27	8.30	(LTRC, 2018)
mdf	Ur	Uniparser	Cyrillic	104672	A+R	P	R, S	2.28	3.91	8.90	(Arkhangelskiy et al., 2012)
mhr	Ur	Uniparser	Cyrillic	251373	A+R	P	P, R, S	2.44	3.52	8.59	(Arkhangelskiy et al., 2012)
mon	Mo	Morphynet	Cyrillic	11428	M+A	P	R, S	1.93	4.40	8.48	(Batsuren et al., 2021)
myv	Ur	Uniparser	Cyrillic	162564	A+R	P	R, S	2.40	3.76	9.03	(Arkhangelskiy et al., 2012)
nbl	NC	Sadilar	Latin	14753	A+H	C	–	3.70	2.63	9.72	(Gaustad and McKellar, 2024)
nld	IE	CELEX	Latin	100620	M+A	C	Int, P, R, S	2.44	4.34	10.81	(Baayen et al., 1995)
nso	NC	Sadilar	Latin	5338	A+H	C	–	2.52	3.04	7.68	(Gaustad and McKellar, 2024)
pol	IE	Morphynet	Latin	58711	M+A	P	P, R, S	1.48	6.66	9.83	(Batsuren et al., 2021)
por	IE	Morphynet	Latin	11774	M+A	P	P, R, S	1.49	7.18	10.67	(Batsuren et al., 2021)
rus	IE	Crosslexica	Cyrillic	27873	M	C	D, I, R	3.76	2.57	9.67	(Bolshakov, 2013)
rus	IE	KuznetsEfremDict	Cyrillic	73447	M	C	Int, P, R, S	4.34	2.27	9.85	(Kuznetsova and Efremova, 1986)
rus	IE	Morphynet	Cyrillic	93039	M+A	P	P, R, S	3.86	2.67	10.28	(Batsuren et al., 2021)
rus	IE	Tikhonov	Cyrillic	96046	M	C	D, I, R, X	2.87	2.11	6.05	(Alimardanova, 2024)
sot	NC	Sadilar	Latin	5667	A+H	C	–	2.46	3.00	7.38	(Gaustad and McKellar, 2024)
spa	IE	MATS	Latin	3541	A	C	–	2.13	3.07	6.55	(García et al., 2025)
spa	IE	Morphynet	Latin	214483	M+A	P	P, R, S	2.11	4.95	10.45	(Batsuren et al., 2021)
ssw	NC	Sadilar	Latin	14493	A+H	C	–	3.57	2.74	9.79	(Gaustad and McKellar, 2024)
swa	NC	LDC_RLP	Latin	2023	A+H	C	R	3.27	2.44	7.98	(Tracey and Strassel, 2020)
swe	IE	Morphynet	Latin	87024	M+A	P	P, R, S	2.52	3.89	9.79	(Batsuren et al., 2021)
tam	Dr	LDC_RLP	Latin^T	2029	A+H	C	–	1.99	4.92	9.82	(Tracey and Strassel, 2020)
tgk	IE	Uniparser	Cyrillic	230646	A+R	P	P, R, S	2.08	3.71	7.71	(Arkhangelskiy et al., 2012)
tgl	AN	LDC_RLP	Latin	2005	A+H	C	–	2.17	3.54	7.69	(Tracey and Strassel, 2020)
tsn	NC	Sadilar	Latin	5946	A+H	C	–	2.48	3.18	7.90	(Gaustad and McKellar, 2024)
tso	NC	Sadilar	Latin	5202	A+H	C	–	2.08	3.79	7.87	(Gaustad and McKellar, 2024)
tur	Tu	MorphoChallenge	Latin^T	6643	M+A	C	–	3.45	3.05	10.51	(Kurimo et al., 2010)
udm	Ur	Uniparser	Cyrillic	401382	A+R	P	Inf, R, S	2.59	3.47	8.97	(Arkhangelskiy et al., 2012)
uig	Tu	thuuymorph	Uyghur Ar.	20958	A+R	C	–	2.17	4.22	9.16	(Gulinigeer et al., 2021)
ven	NC	Sadilar	Latin	5273	A+H	C	–	1.93	3.84	7.42	(Gaustad and McKellar, 2024)
xho	NC	Sadilar	Latin	15480	A+H	C	–	3.96	2.35	9.32	(Gaustad and McKellar, 2024)
zul	NC	Sadilar	Latin	14798	A+H	C	–	3.81	2.45	9.33	(Gaustad and McKellar, 2024)

Table 1: Resources used in the training of the models. Languages codes specified by ISO 639-3 are used in the ISO column. Abbreviations used in the Family column: NWC = Northwestern Caucasian, NC = Niger-Congo, AA = Afro-Asiatic, IE = Indo-European, Ur = Uralic, AN = Austronesian, Ja = Japonic, Dr = Dravidian, Ka = Kartvelian, Ko = Koreanic, Mo = Mongolic, ST = Sino-Tibetan, Tu = Turkic. ^{T} in the Script means transcription from the original writing system to the Latin script. Count = the number of segmented lemmas or wordforms available in the resource. Abbreviations used in the Meth. column: A = Automatic, Un = Unsupervised, R = Rule-Based, M = Manual, H = Human Reviewed. Abbreviations used in the C/P column: C = Complete (fully segmented), P = Partial (only some boundaries). Abbreviations for distinguished morph types in the Cat. column: R = Root, D = Derivational, I = Inflectional, X = Other, S = Suffix, P = Prefix, C = Connector, E = Ending, Int = Interfix, Inf = Infix. Avg. M/W = average number of morphs per lemma or word form. Avg M.Len = the average number of characters per morph. Avg W.Len = the average number of characters per lemma or word form.

ISO	Family	Script	Avg M/W	Avg M.Len	Avg W.Len
ces	IE	Latin	4.05	2.3	9.31
deu	IE	Latin	2.5	4.01	10.01
ell	IE	Greek	2.24	3.5	7.84
eng	IE	Latin	1.93	4.46	8.62
epo	Con.	Latin	2.41	2.59	6.26
fra	IE	Latin	1.66	5.86	9.75
hrv	IE	Latin	2.64	2.12	5.61
ita	IE	Latin	1.58	7.12	11.28
lat	IE	Latin	2.5	2.77	6.94
pol	IE	Latin	2.71	2.23	6.05
rus	IE	Cyrillic	1.65	6.32	10.45
spa	IE	Latin	2.25	3.48	7.81
tel	Dr	Telugu	1.49	4.82	7.16
ukr	IE	Cyrillic	2.64	2.15	5.67
zho	ST	Hanzi	1.93	1.05	2.12

Table 2: Metamorphosis resources. All resources are manually annotated, complete, and contain 200 wordforms. Languages identified by ISO 639-3 code. Avg M/W = average number of morphs per lemma or word form, Avg M.Len = the average number of characters per morph., Avg W.Len = the average number of characters per lemma or word form.

ple, we expect nominally good results for automatically generated datasets, (the model might reverse-engineer the original, often simple pipeline), and for datasets with partial segmentation. The comparability across resources is also limited.

However, the evaluation using manually segmented data, even if they are consistent across languages, is not necessarily more accurate. It might be strongly influenced by differences in annotation approaches between training and test data. Take as an illustration two pairs of highly accurate datasets: *MorphoLex* with *MorphoChallenge* for English, and *Tikhonov* with *KuznetsEfremDict* for Russian. From the 1941 words common to the training subsets of *eng-MorphoLex* and *eng-MorphoChallenge*, the same segmentation appears for only 72.7%. In the case of *Tikhonov* and *KuznetsEfremDict*, it is even less (namely 58.3%, out of 53,048 words). For example, the English word *stabilized* is segmented as *stabil+iz+ed* in *MorphoChallenge* but as *stabil+ize+d* in *MorphoLex*.

We have decided to evaluate our models in two ways. Firstly, we have split the original resources into two mutually exclusive subsets – training data and test data. The test data from the resources were sampled randomly according to the size of the resource (2000 words for resources with more than 5000 words, else 200 words), while the rest of the resource was used for training. Secondly, we have evaluated our models on a small, manually annotated dataset, where available (see Table 2). As evaluation metrics, we use whole-word accuracy, and precision, recall and F1 on boundaries between morphs.

5. Experiments and Results

Although transformers-based models have been used for segmentation in a high-resource setting (Batsuren et al., 2022), the performance gap between transformer-based models and simpler LSTM-based models appears to be quite small. Among other advantages of LSTM-based models is the comparatively low number of parameters, and therefore faster training time. Such an architecture also seems more appropriate for small-size datasets, where it achieved competitive results (Olbrich and Žabokrtský, 2025). We have thus opted for employing two LSTM-based architectures, trained on each resource separately.

Similarly to Girrbach (2022), we encode segmentation as character classification, using the BMES encoding (Ruokolainen et al., 2014). Each character is classified as beginning (B), middle (M), or end (E) of morpheme, or as single-character morpheme (S). For example, the word *be-ing-s* would be annotated as *b:B, e:E, i:B, n:M, g:E, s:S*.

The first architecture we have used is a two-layer LSTM, as presented by Girrbach (2022) and implemented in the publicly available Python package *morphseg* (Winkelman et al., 2026). Models using this architecture (TüSeg) have achieved competitive results in the SIGMORPHON 2022 Shared Task on Morpheme Segmentation (Batsuren et al., 2022). The implementation we use automatically removes non-alphabetical characters. We automatically restore most of these characters before evaluation. As we don't get any prediction for these characters, we treat them by default as morpheme boundaries. While not the best performing architecture in the original shared task, it might be more appropriate for our setting, which includes small datasets and concentrates of surface segmentation.

We have also created our custom architecture (CNN-LSTM), inspired by Olbrich and Žabokrtský (2025). Our models, after embedding the individual characters (with embedding size 128), apply convolutions with kernel sizes 1 to 8 and 64 channels each in order to capture local dependencies. Concatenated outputs of these convolutions are then fed into a bidirectional LSTM layer (of dimension 512, close to the optimum according to (Olbrich and Žabokrtský, 2025)), followed by a dense layer and a final dense classification layer. We train for 100 epochs with batch size 64 for large resources, 16 for small resources, with early stopping on word accuracy on validation set. It isn't strictly guaranteed that the output of the neural network will be a valid BMES-encoded segmentation. There are multiple possible strategies of extracting a valid segmentation from the output of the neural network, including conditional random fields or Viterbi decod-

	Morfessor				CNN-LSTM				TüSeg			
resource	WAcc	MBPrec	MBRec	MBF	WAcc	MBPrec	MBRec	MBF	WAcc	MBPrec	MBRec	MBF
ady-MorphAGram	0.10	0.76	0.64	0.69	**0.61**	0.96	0.95	**0.95**	0.58	0.94	0.96	**0.95**
aka-LDC_RLP	0.24	0.47	0.85	0.60	**0.80**	0.91	0.91	**0.91**	0.74	0.87	0.89	0.88
amh-Amharic	0.04	0.49	0.37	0.42	**0.76**	0.93	0.95	**0.94**	0.75	0.93	0.95	**0.94**
bel-Slounik	0.05	0.50	0.43	0.46	0.91	0.98	0.98	**0.98**	**0.92**	0.98	0.98	**0.98**
ben-kcis	0.14	0.32	0.74	0.44	**0.86**	0.87	0.87	**0.87**	0.74	0.87	0.76	0.81
cat-Morphynet	0.07	0.15	0.69	0.24	**0.87**	0.91	0.93	**0.92**	**0.87**	0.91	0.91	0.91
ces-Morphynet	0.24	0.32	0.67	0.44	**0.96**	0.97	0.98	**0.98**	0.93	0.95	0.96	0.96
ces-SlavickovaDict	0.00	0.59	0.44	0.51	0.94	0.99	0.99	**0.99**	**0.95**	0.99	0.99	**0.99**
deu-CELEX	0.15	0.35	0.72	0.47	**0.88**	0.95	0.96	**0.95**	0.83	0.91	0.94	0.93
deu-MorphoChallenge	0.01	0.27	0.80	0.41	**0.59**	0.83	0.86	**0.84**	0.53	0.84	0.81	0.82
deu-Morphynet	0.09	0.22	0.56	0.31	**0.93**	0.95	0.97	**0.96**	0.69	0.78	0.81	0.80
ell-GreekAnnotatedDict	0.09	0.26	0.71	0.39	**0.83**	0.92	0.92	**0.92**	0.60	0.86	0.76	0.80
eng-CELEX	0.12	0.20	0.71	0.32	0.86	0.92	0.92	**0.92**	**0.87**	0.91	0.94	**0.92**
eng-MorphoChallenge	0.07	0.22	0.88	0.36	**0.62**	0.81	0.79	**0.80**	0.59	0.79	0.80	0.79
eng-MorphoLex	0.17	0.31	0.76	0.44	0.92	0.96	0.97	0.96	**0.93**	0.96	0.97	**0.97**
eng-Morphynet	0.19	0.31	0.74	0.44	**0.90**	0.95	0.96	**0.95**	**0.90**	0.94	0.97	**0.95**
fas-PerSegLex	0.11	0.25	0.74	0.38	**0.85**	0.93	0.91	**0.92**	**0.85**	0.92	0.92	**0.92**
fin-MorphoChallenge	0.01	0.24	0.60	0.34	0.58	0.86	0.86	**0.86**	**0.59**	0.86	0.86	**0.86**
fin-Morphynet	0.05	0.24	0.40	0.30	**0.99**	1.00	1.00	**1.00**	0.97	0.98	1.00	0.99
fra-demonette	0.03	0.07	0.39	0.12	**0.89**	0.95	0.90	**0.92**	**0.89**	0.95	0.90	**0.92**
fra-MorphoLex	0.03	0.14	0.72	0.23	0.66	0.74	0.74	0.74	**0.67**	0.77	0.73	**0.75**
fra-Morphynet	0.08	0.17	0.75	0.28	**0.86**	0.92	0.91	**0.91**	**0.86**	0.89	0.92	0.90
hbs-Morphynet	0.07	0.18	0.84	0.30	**0.85**	0.89	0.90	**0.89**	0.74	0.83	0.80	0.81
hin-kcis	0.18	0.44	0.86	0.58	0.82	0.91	0.85	0.88	**0.83**	0.91	0.88	**0.89**
hin-LDC_RLP	0.04	0.19	0.89	0.32	**0.82**	0.88	0.88	**0.88**	0.75	0.90	0.77	0.83
hrv-CroDeriV	0.00	0.39	0.28	0.33	**0.93**	0.98	0.99	**0.98**	**0.93**	0.98	0.99	**0.98**
hun-LDC_RLP	0.03	0.23	0.88	0.36	**0.58**	0.80	0.77	**0.78**	0.54	0.79	0.71	0.75
hun-Morphynet	0.18	0.33	0.76	0.46	**0.89**	0.93	0.95	0.94	**0.89**	0.94	0.95	**0.95**
hye-Uniparser	0.18	0.47	0.61	0.53	**0.84**	0.95	0.94	**0.94**	0.83	0.91	0.97	**0.94**
ind-LDC_RLP	0.04	0.15	0.88	0.26	**0.89**	0.93	0.90	**0.91**	0.83	0.88	0.89	0.88
ita-DerIvaTario	0.05	0.26	0.53	0.35	**0.70**	0.90	0.89	**0.89**	0.67	0.89	0.87	0.88
ita-Morphynet	0.03	0.06	0.47	0.11	0.91	0.88	0.95	**0.92**	**0.92**	0.92	0.93	**0.92**
jpn-MorphAGram	0.41	0.60	0.69	0.65	0.67	0.76	0.92	0.83	**0.72**	0.81	0.91	**0.86**
kan-kcis	0.12	0.63	0.41	0.50	0.72	0.91	0.94	0.92	**0.74**	0.92	0.94	**0.93**
kat-MorphAGram	0.04	0.34	0.69	0.46	0.58	0.88	0.83	0.85	**0.59**	0.87	0.87	**0.87**
kor-LDC2004T03	0.05	0.12	0.74	0.21	**0.82**	0.86	0.84	**0.85**	0.66	0.64	0.78	0.70
kpv-Uniparser	0.18	0.44	0.61	0.51	**0.81**	0.91	0.94	**0.92**	0.80	0.90	0.94	**0.92**
lat-WFL	0.08	0.30	0.56	0.39	**0.71**	0.84	0.91	**0.88**	0.70	0.89	0.85	0.87
mal-kcis	0.04	0.31	0.42	0.36	**0.62**	0.86	0.83	**0.84**	0.60	0.83	0.82	0.83
mar-kcis	0.19	0.44	0.71	0.54	0.82	0.91	0.93	**0.92**	**0.83**	0.91	0.94	**0.92**
mdf-Uniparser	0.14	0.34	0.65	0.44	**0.89**	0.94	0.95	**0.95**	0.88	0.93	0.96	0.94
mhr-Uniparser	0.09	0.35	0.59	0.44	**0.71**	0.86	0.86	**0.86**	**0.71**	0.88	0.82	0.85
mon-Morphynet	0.33	0.32	0.92	0.48	**0.97**	0.98	0.99	**0.99**	**0.97**	0.98	0.99	**0.99**
myv-Uniparser	0.10	0.34	0.59	0.43	**0.85**	0.91	0.95	**0.93**	0.83	0.94	0.91	0.92
nbl-Sadilar	0.07	0.49	0.51	0.50	**0.64**	0.90	0.91	**0.91**	0.62	0.88	0.93	0.90
nld-CELEX	0.27	0.42	0.69	0.52	**0.93**	0.97	0.98	**0.97**	0.92	0.97	0.97	**0.97**
nso-Sadilar	0.04	0.20	0.56	0.30	**0.84**	0.92	0.93	**0.92**	0.82	0.90	0.93	**0.92**
pol-Morphynet	0.03	0.08	0.70	0.14	0.85	0.82	0.85	**0.84**	**0.86**	0.84	0.84	**0.84**
por-Morphynet	0.03	0.07	0.67	0.12	0.85	0.84	0.86	**0.85**	**0.86**	0.87	0.83	**0.85**
rus-Crosslexica	0.08	0.65	0.48	0.55	**0.96**	0.99	0.99	**0.99**	**0.96**	0.99	0.99	**0.99**
rus-KuznetsEfremDict	0.06	0.68	0.43	0.53	**0.89**	0.98	0.98	**0.98**	0.88	0.98	0.98	**0.98**
rus-Morphynet	0.03	0.09	0.55	0.15	0.86	0.89	0.87	0.88	**0.87**	0.89	0.89	**0.89**
rus-Tikhonov	0.05	0.55	0.43	0.48	**0.91**	0.98	0.98	**0.98**	0.89	0.98	0.97	0.97
sot-Sadilar	0.04	0.21	0.62	0.32	**0.82**	0.92	0.92	**0.92**	**0.82**	0.91	0.92	**0.92**
spa-MATS	0.01	0.22	0.82	0.35	**0.67**	0.81	0.84	**0.83**	0.66	0.81	0.84	0.82
spa-Morphynet	0.07	0.22	0.57	0.31	**0.95**	0.97	0.97	**0.97**	**0.95**	0.97	0.98	**0.97**
ssw-Sadilar	0.04	0.45	0.54	0.49	**0.77**	0.94	0.96	**0.95**	0.75	0.93	0.95	0.94
swa-LDC_RLP	0.01	0.35	0.59	0.44	**0.66**	0.91	0.90	**0.91**	**0.66**	0.91	0.91	**0.91**
swe-Morphynet	0.10	0.41	0.66	0.51	**0.94**	0.97	0.98	**0.98**	**0.94**	0.97	0.98	**0.98**
tam-LDC_RLP	0.02	0.15	0.72	0.25	**0.64**	0.73	0.75	**0.74**	0.58	0.75	0.65	0.70
tgk-Uniparser	0.24	0.44	0.73	0.55	**0.88**	0.92	0.98	**0.95**	0.87	0.92	0.98	**0.95**
tgl-LDC_RLP	0.01	0.18	0.88	0.30	**0.72**	0.86	0.79	**0.83**	0.68	0.81	0.80	0.81
tsn-Sadilar	0.04	0.20	0.58	0.30	**0.78**	0.87	0.91	**0.89**	0.76	0.89	0.86	0.88
tso-Sadilar	0.02	0.16	0.68	0.25	**0.78**	0.88	0.87	**0.88**	0.77	0.87	0.88	0.87
tur-MorphoChallenge	0.02	0.31	0.54	0.40	**0.44**	0.83	0.81	**0.82**	0.38	0.79	0.76	0.78
udm-Uniparser	0.14	0.40	0.59	0.48	**0.89**	0.97	0.93	**0.95**	**0.89**	0.94	0.97	**0.95**
uig-thuuymorph	0.12	0.27	0.64	0.38	**0.92**	0.96	0.97	**0.96**	**0.92**	0.96	0.97	**0.96**
ven-Sadilar	0.04	0.15	0.62	0.24	**0.74**	0.82	0.84	**0.83**	0.73	0.80	0.85	0.82
xho-Sadilar	0.10	0.54	0.55	0.55	**0.80**	0.95	0.95	**0.95**	0.79	0.95	0.95	**0.95**
zul-Sadilar	0.07	0.50	0.51	0.51	**0.75**	0.93	0.95	**0.94**	0.74	0.92	0.95	**0.94**
Macroaverage	0.09	0.32	0.64	0.39	**0.81**	**0.91**	**0.91**	**0.91**	0.78	0.89	0.90	0.90

Table 3: Evaluation of the two model architectures - our CNN-LSTM architecture and the TüSeg architecture, trained on each resource, evaluated on data sampled from the corresponding resources. WAcc = word accuracy, MBPrec = morph boundary precision, MBRec = morph boundary recall, MBF = Morph boundary F1). The resource are listed in format [iso 639-3 language code]-[resource name/code].

ing. Since invalid outputs occur rarely, we simply split the word on characters tagged E or S.[1].

[1]The training and evaluation pipeline is available at

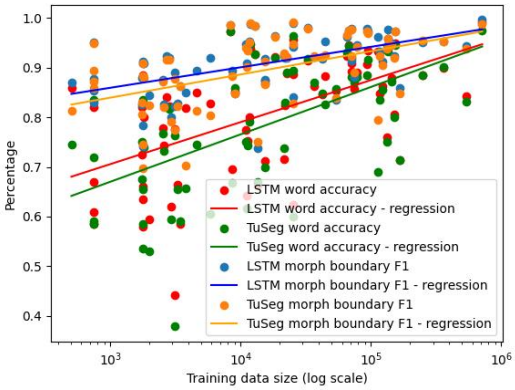

Figure 1: Effect of training data size on model performance. The training size is shown in logarithmic scale

As a simple baseline, we have trained a semi-supervised morphological segmentation model using Morfessor 2.0 (Smit et al., 2014). Some of our languages are very low-resource, and introducing corpora of varying sizes would confound comparability both across the baseline models and between baseline and neural models. Hence, the Morfessor Baseline models were trained exclusively on word types (lists of unique words without corpus frequencies) from the training sets of our resources, used both as the annotated and the unannotated data.

As mentioned in the previous section, we evaluate the models both on the original resources and on custom evaluation data. The results of the evaluation on the training resources are in Table 3. Both architectures achieved high word accuracy and morph-boundary F-measure. On (macro)average, our CNN-LSTM architecture performed slightly better than TüSeg in all the metrics, achieving 81 % word accuracy and 90.6 % morph boundary F1, as compared to 78 % and 89.5 % respectively for TüSeg. Both the architectures have vastly outperformed the Morfessor baseline on all the languages.

The performance gap between the two architectures in terms of word accuracy appears to diminish with growing size of training data, with the CNN-LSTM architecture being better for smaller datasets (see Figure 1). The observed effect of training data size, while consistent with observe linear improvement in performance with exponential increase of training data, as observed by (Olbrich and Žabokrtský, 2025), is overshadowed by the variability caused by the resource- and language-dependent factors.

While the evaluation on training resources yields promising numbers, there is a striking discrepancy between these and the values achieved while evaluating on the manually segmented data (see Table 4), where both the architectures only achieved (macro)average word accuracy of approximately 40 %.

6. Discussion

While it seems that the CNN-LSTM architecture outperforms the TüSeg architecture, some of the largest differences between the two are artifacts. For example, the German *MorphyNet* allows and consistently uses multi-word stems. The CNN-LSTM architecture allows multi-word stems, while TüSeg does not. Morphemes spanning multiple orthographical words are rare, if indeed they exist. In this particular case, the TüSeg is in fact the better segmenter, although nominally it is the worse one. Similarly, for the Korean data, many apparent errors are caused by treating the dashes by default as morpheme boundaries. While this effect is most visible for these two resources, it may on a lesser scale appear in others as well. In some cases, the apparent errors are in fact improvements. For example, *unfailingly* is segmented as *un+failing+ly* in the training data (*eng-MorphoChallenge*), but both the TüSeg and the CNN-LSTM model unfailingly segmented it as *un+fail+ing+ly*.

With respect to the large difference between evaluation measures on the original resources and evaluation on gold data, a large part of the difference is grounded in differences between gold segmentation and segmentation, as presented in the resources (for an overview, see Table 4). Here follows an overview of several types of data incompatibilities and corresponding errors.

- Unrepresentativeness of the training data, leading to false generalizations. For example, *SlavickovaDict* includes only Czech verbs in a slightly archaic infinitive form. As a consequence, in the training data, the last two characters always consist of the Czech inflective affix *-ti*. As a result, the models trained on this dataset tend to segment off the last two characters.

- Incomplete segmentation in the training data. It is noteworthy that the macroaverage precision is significantly higher than the macroaverage recall. For no model is recall more than 10 % higher than precision, while in several cases, the precision is more than 30 % higher than recall.

- Errors in the training data, especially in automatically generated data (e.g. *MorphyNet*) and/or resources not originally designed for morphological segmentation (e.g. *DerIvaTario*, *Uniparser*).

https://github.com/johnvojtech/unisegmenter

				CNN-LSTM				TüSeg			
resource	Both	Same	Same (%)	WAcc	MBPrec	MBRec	MBF	WAcc	MBPrec	MBRec	MBF
ces-Morphynet	39	126	31	0.30	0.71	0.31	**0.44**	**0.34**	0.75	0.30	0.43
ces-SlavickovaDict	0	0	–	**0.30**	0.62	0.65	**0.64**	0.21	0.63	0.53	0.58
deu-CELEX	64	40	63	**0.49**	0.92	0.48	**0.63**	0.46	0.91	0.46	0.61
deu-MorphoChallenge	6	6	100	0.41	0.76	0.61	**0.68**	**0.46**	0.84	0.56	0.67
deu-Morphynet	21	8	38	**0.38**	0.82	0.32	**0.46**	**0.38**	0.83	0.31	0.45
ell-GreekAnnotatedDict	28	12	43	**0.33**	0.55	0.46	**0.50**	0.22	0.53	0.40	0.46
eng-CELEX	121	103	85	**0.71**	0.61	0.36	**0.45**	0.70	0.73	0.32	0.44
eng-MorphoChallenge	4	4	100	**0.77**	0.64	0.63	0.64	0.74	0.64	0.65	**0.65**
eng-MorphoLex	165	143	87	**0.84**	0.74	0.76	0.75	0.82	0.78	0.75	**0.76**
eng-Morphynet	40	26	65	**0.76**	0.75	0.56	**0.64**	0.72	0.72	0.55	0.62
fra-demonette	33	7	21	**0.35**	0.61	0.11	**0.18**	0.34	0.67	0.09	0.16
fra-MorphoLex	69	39	57	**0.45**	0.78	0.38	**0.51**	**0.45**	0.88	0.36	**0.51**
fra-Morphynet	44	11	25	0.36	0.64	0.22	0.33	**0.37**	0.70	0.25	**0.37**
ita-DerIvaTario	3	0	0	0.14	0.37	0.21	**0.27**	**0.15**	0.37	0.17	0.24
ita-Morphynet	25	1	4	0.16	0.47	0.05	**0.10**	**0.17**	0.65	0.05	0.09
lat-WFL	73	44	60	**0.46**	0.83	0.56	**0.67**	0.42	0.86	0.53	0.66
pol-Morphynet	27	1	3	**0.16**	0.70	0.14	**0.23**	**0.16**	0.71	0.13	0.22
rus-Crosslexica	24	11	46	0.38	0.84	0.64	0.72	**0.40**	0.84	0.66	**0.74**
rus-KuznetsEfremDict	59	40	68	**0.49**	0.82	0.75	0.78	**0.49**	0.84	0.75	**0.79**
rus-Morphynet	26	1	4	**0.15**	0.74	0.09	0.15	0.14	0.70	0.10	**0.17**
rus-Tikhonov	51	33	65	**0.45**	0.88	0.59	0.70	**0.45**	0.90	0.59	**0.71**
spa-MATS	60	35	0.58	0.44	0.71	0.58	**0.64**	**0.45**	0.71	0.55	0.62
spa-Morphynet	48	10	0.21	**0.25**	0.34	0.10	**0.16**	0.23	0.33	0.11	**0.16**
Macroaverage				**0.42**	0.69	0.42	**0.49**	0.40	0.72	0.40	0.48

Table 4: Evaluation of the neural segmenters on gold data. Both = how many words are both in training data and in the golden test set. Same=number of words in both train and test set that are segmented in the same way. WAcc = word accuracy, MBPrec = morph boundary precision, MBRec = morph boundary recall, MBF = Morph boundary F1

- Incompatible annotation decisions. For example, in cases of character reduplication (like *Tomm-y* and *Tom-my*), it is unclear whether to increase allomorphy of stems or of suffixes. Our test data as a rule increase allomorphy of stems (i.e., the segmentation would be *Tomm-y*), which doesn't necessarily hold in other resources. Also, some morpheme boundaries are more debateable than others, and the authors of the original resources might have been generally less prone to segment (e.g. the *GreekAnnotatedDictionary*).

Morphological complexity does play a significant role in the performance, as witnessed by the high results for English (even in the case of the fairly inaccurate English MorphyNet). Nevertheless, it seems that typological features impose smaller limitations on the results than size and quality of the training data. For example, results for Russian are consistently high, even though Russian is very morphologically complex.

7. Conclusions

In this paper, we have presented the largest morphological segmentation experiment to date, covering 58 languages. We have trained two state-of-the-art neural models on surface morphological segmentation, achieving 81 % average word accuracy on the original resources. The models can be further used in automatically annotating text corpora on the morpheme level, which is so far very rare. However, the much worse performance on our custom test sets points to the need for further harmonization of morphological segmentation approaches. There is also a large scope for further research in the technical side of the experiment, as both the architecture and training can be potentially improved. For example, if there are multiple inconsistent resources for a language, it might be advantageous to (approximately) train a model succesively on all of these, in increasing order of segmentation quality, or use transfer learning for similar languages. This would avoid false generalizations when possible, while not significantly decrease the quality of the segmentation. If multiple models aren't available, it might still be useful to pretrain on a dataset segmented by an unsupervised method. Our preliminary results indicate improvement in precision, but large decrease in this scenario. Apart from experiments with different architecture (perhaps using another architecture from Batsuren et al. (2022), especially a Transformers-based architecture), we might also try to train a joint model for all the languages, to enable cross-lingual transfer across related languages (Olbrich and Žabokrtský, 2025). We would also like to make use of other morphological features present in the datasets (e.g. lemma or part of speech), as well as devise semi-supervised systems using such information for improved morphological segmentation.

Acknowledgements

This work was supported by the Charles University Research Centre program No. 24/SSH/009, by

the Charles University project GA UK No. 101924, by the Czech Science Foundation project No. 26-21822S, by the SVV project No. 260 821, and by the Ministry of Education, Youth and Sports of the Czech Republic, project No. LM2023062. We also thank anonymous reviewers for their useful remarks.

8. Bibliographical References

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 103–116, Seattle, Washington. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0. In *Unsupervised morpheme segmentation and morphology induction from text corpora using Morfessor 1.0*. Helsinki University of Technology.

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Leander Girrbach. 2022. SIGMORPHON 2022 shared task on morpheme segmentation submission description: Sequence labelling for word-level morpheme segmentation. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 124–130, Seattle, Washington. Association for Computational Linguistics.

Martin Haspelmath. 2020. The morph as a minimal linguistic form. *Morphology*, 30(2):117–134.

Charles F Hockett. 1947. Problems of morphemic analysis. *Language*, 23(4):321–343.

Ursula Klenk and Hagen Langer. 1989. Morphological segmentation without a lexicon. *Literary and Linguistic Computing*, 4(4):247–253.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Thang Vu. 2022. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Arya D McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2020. Unimorph 3.0: Universal morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Seung-Hoon Na. 2015. Conditional random fields for Korean morpheme segmentation and POS tagging. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 14(3):1–16.

Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. 2015. An unsupervised method for uncovering morphological chains. *Transactions of the Association for Computational Linguistics*, 3:157–167.

Michal Olbrich and Zdeněk Žabokrtský. 2025. Morphological segmentation with neural networks: Performance effects of architecture, data size, and cross-lingual transfer in seven languages. In *Text, Speech, and Dialogue*, pages 275–286, Cham. Springer Nature Switzerland.

Marko Pranjić, Marko Robnik-Šikonja, and Senja Pollak. 2024. LLMSegm: Surface-level morphological segmentation using large language model. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10665–10674, Torino, Italia. ELRA and ICCL.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2014. Painless semi-supervised morphological segmentation using conditional random fields. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 84–89, Gothenburg, Sweden. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 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.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for

statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

9. Language Resource References

Shaxlo Ashurmamatovna Alimardanova. 2024. Representation of Linguistic Differentiation in the "Word-Formation Dictionary of the Russian Language" by A.N. Tikhonov. *American Journal of Language, Literacy and Learning in STEM Education*, 2(2):463–465.

Ebrahim Ansari, Zdeněk Žabokrtský, Hamid Haghdoost, and Mahshid Nikravesh. 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.

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

Baayen, R. Harald and Piepenbrock, Richard and Gulikers, Leon. 1995. *The CELEX Lexical Database (CD-ROM)*. University of Pennsylvania. Linguistic Data Consortium, ISLRN 204-698-863-053-1. Catalogue No. LDC96L14.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021. MorphyNet: a large multilingual database of derivational and inflectional morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online. Association for Computational Linguistics.

Igor Bolshakov. 2013. Crosslexica: a universe of links between Russian words. *Business Informatics*, 7(3):19–26.

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith Klavans, and Smaranda Muresan. 2020. MorphAGram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 7112–7122, Marseille, France. European Language Resources Association.

Alba Táboas García, Piotr Przybyła, and Leo Wanner. 2025. Exploring morphology-aware tokenization: A case study on Spanish language modeling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30505–30518, Suzhou, China. Association for Computational Linguistics.

Tanja Gaustad and Cindy A. McKellar. 2024. Updated Morphologically Annotated Corpora for 9 South African Languages. *Journal of Open Humanities Data*, 10(38):1–5.

Abudouwaili Gulinigeer, Abiderexiti Kahaerjiang, Wushouer Jiamila, Shen Yunfei, Maimaitimin Turenisha, and Yibulayin Tuergen. 2021. Morphological analysis corpus construction of Uyghur. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1076–1086, Huhhot, China. Chinese Information Processing Society of China.

Na-Rae Han. 2004. Morphologically Annotated Korean Text LDC2004T03. Web Download. ISBN 1-58563-284-8.

Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. *Linguistic Issues in Language Technology*, 11(5).

Vojtěch John, Zdeněk Žabokrtský, Benjamin Reeves, Magda Ševčíková, Haridanmu Abdukerim, Abduhalik Abduwaiti, Abdukerim Abliz, Ebrahim Ansari, Timofey Arkhangelskiy, Lizaveta Astapenka, Niyati Bafna, Khuyagbaatar Batsuren, Gábor Bella, Pier Marco Bertinetto, Chiara Celata, Hélène Deacon, Konstantinos Diamantopoulos, Šárka Dohnalová, Alexei Fedorenko, Matea Filko, Antonín Forst, Federica Gamba, Timur Garipov, Tanja Gaustad, Fausto Giunchiglia, Anna Glazkova, Hamid Haghdoost, Nabil Hathout, Irina Khomchenkova, Victoria Khurshudyan, Maria Klyucheva, Lukáš Kyjánek, Eleonora Maria Litta, Olga Lyashevskaya, Joël Macoir, Hugo Mailhot, Sun Maosong, Cindy McKellar, Maria Medvedeva, Dmitry Morozov, S.N. Muralikrishna, Fiametta Namer, Anna Nedoluzhko, Mahshid Nikravesh, Aleš Manuel Papáček, Marco Passarotti, Alexey Polyakov, Mihail Potapov, Mishra Pruthwik, Chrysanthi Raftopoulou, Ashwath Rao, Husain Samar, Claudia Sánchez-Gutiérrez, Iurii Savelev-Galiaminskii, Dipti Misra Sharma, Eleonora Slavíčková, Abishek Stephen, Emil Svoboda, Krešimir Šojat, Vanja Štefanec, Luigi Talamo, Jonáš Vidra, Arseniy Vydrin, Maximiliano A. Wilson, Liu Yang, and Aigul Zakirova. 2026. Universal Segmentations 1.5 (UniSegments 1.5). http://hdl.handle.net/11234/1-6130

LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, Prague.

Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. 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.

A.I. Kuznetsova and T.F. Efremova. 1986. *Dictionary of Morphemes of the Russian Language*. Firebird Publications, Incorporated.

Eleonora Litta Modignani, Greta Franzini, Rachele Sprugnoli, Giovanni Moretti, and Marco Passarotti. 2023. CIRCSE/WFL: WFL 1.0.1.

LTRC. 2018. KCIS dependency and coreference annotated corpora. Web Download. Annotation funded by KCIS, DeiTY, Govt. of India. Dependency annotation follows the Paninian Grammar Framework.

Hugo Mailhot, Maximiliano A. Wilson, Joël Macoir, S. Hélène Deacon, and Claudia Sánchez-Gutiérrez. 2020. MorphoLex-FR: A derivational morphological database for 38,840 French words. *Behavior Research Methods*, 52:1008–1025.

Dmitry Morozov, Lizaveta Astapenka, Anna Glazkova, Timur Garipov, and Olga Lyashevskaya. 2025. BERT-like models for Slavic morpheme segmentation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6795–6815, Vienna, Austria. Association for Computational Linguistics.

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

Eleonora Slavíčková. 1975. *Retrográdní morfematický slovník češtiny s připojenými inventárními slovníky českých morfémů kořenových, prefixálních a sufixálních*. Academia, Praha. Souběžné názvy v ruštině, angličtině a francouzštině.

Krešimir Šojat, Matea Srebačić, Tin Pavelić, and Marko Tadić. 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.

Luigi Talamo, Chiara Celata, and Pier Marco Bertinetto. 2016. DerIvaTario: An annotated lexicon of Italian derivatives. *Word Structure*, 9(1):72–102.

Jennifer Tracey and Stephanie Strassel. 2020. Basic Language Resources for 31 Languages (Plus English): The LORELEI Representative and Incident Language Packs. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6252–6259, Marseille, France. European Language Resources Association.

UI and NTUA. 2021. Greek annotated dictionary with morphological and linguistic features. Dataset distributed by European Language Grid (ELG), downloaded from https://live.european-language-grid.eu/catalogue/lcr/7360.

Donald Winkelman, Cynthia Wolf, Nathan abd Kong, Alexis Therrien, and Taoran Ye. 2026. Morphseg: An efficient and easy-to-use morpheme segmentation library. `https://github.com/TheWelcomer/MorphSeg`. GitHub repository, accessed 22 Feb 2026.

Tilahun Yeshambel, Josiane Mothe, and Yaregal Assabie. 2021. Morphologically Annotated Amharic Text Corpora. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2349–2355, New York, NY, USA. Association for Computing Machinery.

Zdeněk Žabokrtský, Niyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, and Jonáš Vidra. 2022. Towards universal segmentations: Unisegments 1.0. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1137–1149.