The SIGMORPHON 2022 Shared Task on Morpheme Segmentation

Khuyagbaatar Batsuren[1] **Gábor Bella**[2] **Aryaman Arora**[3] **Viktor Martinovic**[4]
Kyle Gorman[5] **Zdeněk Žabokrtský**[6] **Amarsanaa Ganbold**[1] **Šárka Dohnalová**[6]
Magda Ševčíková[7] **Kateřina Pelegrinová**[6] **Fausto Giunchiglia**[2] **Ryan Cotterell**[8]
Ekaterina Vylomova[9]
[1]National University of Mongolia [2]University of Trento [3]Georgetown University
[4]University of Vienna [5]Graduate center, City University Of New York [6]Charles University
[7]University of Ostrava [8]ETH Zürich [9]University of Melbourne

Abstract

The SIGMORPHON 2022 shared task on morpheme segmentation challenged systems to decompose a word into a sequence of morphemes and covered most types of morphology: compounds, derivations, and inflections. Subtask 1, word-level morpheme segmentation, covered 5 million words in 9 languages (Czech, English, Spanish, Hungarian, French, Italian, Russian, Latin, Mongolian) and received 13 system submissions from 7 teams and the best system averaged 97.29% F1 score across all languages, ranging English (93.84%) to Latin (99.38%). Subtask 2, sentence-level morpheme segmentation, covered 18,735 sentences in 3 languages (Czech, English, Mongolian), received 10 system submissions from 3 teams, and the best systems outperformed all three state-of-the-art subword tokenization methods (BPE, ULM, Morfessor2) by 30.71% absolute. To facilitate error analysis and support any type of future studies, we released all system predictions, the evaluation script, and all gold standard datasets.[1]

1 Introduction

Many NLP applications, such as machine translation or question answering, require *subword tokenization*, i.e. splitting words into a sequence of substrings (Mielke et al., 2021). Such tokenizers are trained by an unsupervised algorithm, usually either Byte-Pair Encoding (BPE; Gage 1994; Sennrich et al. 2016) or Unigram Language Modeling (ULM; Kudo 2018). To give a few examples, contemporary language models RoBERTa (Liu et al., 2019) and GPT-3 (Brown et al., 2020) use a byte-level BPE (Radford et al., 2019) while XLNet (Yang et al., 2019) relies on ULM. These subword tokenization algorithms are not linguistically motivated but are rather based on statistical co-occurrences. Therefore, unsupervised and semi-supervised methods for morphological segmentation (Creutz and Lagus, 2005) have emerged in parallel, state-of-the-art methods of this kind being Morfessor variants (Grönroos et al., 2014, 2020). Ataman et al. (2017) and Schwartz et al. (2020) find that Morfessor-based language models can outperform BPE-based ones. Matthews et al. (2018); Nzeyimana and Rubungo (2022) show that enriching BPE with morphological analyzers can be beneficial for translation, while many others (Domingo et al., 2018; Macháček et al., 2018; Schwartz et al., 2020; Saleva and Lignos, 2021) find no conclusive improvements over BPE for machine translation.

System	type	motivation	segmentation
BPE	surface	sta.	in | val | uable
Morfessor2	surface	sta. & lin.	in | valuable
DeepSPIN-3	canonical	sta. & lin.	in | value | able

Table 1: Structural differences of subword tokenization (BPE), morphological segmentation (Morfessor2), and morpheme segmentation (DeepSPIN-3 – subtask 1 winning system); acronyms: sta. - statistics and lin. - linguistic

One of the core problems is that the state-of-the-art morphological segmentation and subword tokenization algorithms provide "surface-level" segmentation, which has several theoretical drawbacks with respect to "canonical" segmentation (e.g., segmented substrings are not considered as meaningful as morphemes). Cotterell et al. (2016) provided formal definitions for both: given a word w, its "surface" segmentation is a sequence of *surface substrings* the concatenation of which is w, e.g., *funniest* $\rightarrow$ *funn-i-est*. The purpose of canonical segmentation (Kann et al., 2016; Girrbach, 2022), on the other hand, is not only computing surface segmentation but also restoring standardized forms of morphemes, e.g., *funniest* $\rightarrow$ *fun-y-est*. More detailed structural distinctions between these segmentation types are shown in Table 1.

However, state-of-the-art studies in canonical segmentations have been limited to very low num-

[1] https://github.com/sigmorphon/2022SegmentationST

19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, pages 103 - 116

Lang	word	segmentation	category
eng	sheepiness	sheep @@y @@ness	010
	pokers	poke @@er @@s	110
hun	időpontod	idő @@pont @@od	101
	szőttetek	sző @@tt @@etek	100
mon	харах	харах	000
	гэмтлийг	гэмтэх @@л @@ийг	110

Table 2: Training samples for Subtask 1. Each sample consists of a word, its canonical segmentation, and a category encoding word formation processes.

bers of languages with sufficiently rich morphological resources (Kurimo et al., 2010a,b; Cotterell et al., 2016; Kann et al., 2018). With the goal of advancing research in this direction, we present a *morpheme segmentation shared task* and provide large-scale datasets over nine languages, evaluation metrics, and morphological annotations of five million word formations. In this, we rely on the latest release of UniMorph (Batsuren et al., 2022) which has introduced morpheme segmentations and derivational data from MorphyNet (Batsuren et al., 2021b). The resulting shared task is a follow-up to past morphological segmentation shared tasks such as "MorphoChallenge" (Kurimo et al., 2007, 2008, 2009) or "Multilingual parsing" (Zeman et al., 2017, where lemmatization as segmentation is a subtask).

2 Task and Evaluation Details

2.1 Subtask 1: Word-level Morpheme Segmentation

In subtask 1, participating systems were asked to segment a given word into a sequence of morphemes. The participants were initially provided with examples of segmentation to train and fine-tune their systems, as shown in Table 2. Each instance in the training set is a triplet consisting of a word, a sequence of morphemes, and a morphological category specifying the types of word formation (see Table 3). The morphological category is an optional feature that can only be used to oversample or undersample the training dataset (the word frequencies are imbalanced across the morphological categories, e.g., Italian has 431 compound words and 253K inflections). The test data only contained the initial word itself.

Key points of this subtask are:

- The task is focusing on canonical segmentation, i.e. given an input word, participants had to predict *a sequence of morphemes*. In canonical segmentation, the participating systems need to reconstruct internal morphophonological processes involved in word formation. For example, the word "intensive" will be decomposed into the base form "intens***e***" and the adjectival siffix '@@ive" (note that the ending '***e***' of the base word is inferred here);
- As shown in Table 4, the task is multilingual, with seven high-resource languages (English, Spanish, Hungarian, French, Italian, Russian, Latin) and two low-resource languages (Czech and Mongolian);
- The annotated corpus data represents a variety of morphological phenomena, including inflection, derivation, compounding (Table 4);
- A large-scale coverage as segmentations of five million words.

2.2 Subtask 2: Sentence-level Morpheme Segmentation

The second subtask is a context-dependent morpheme segmentation and focuses on resolving ambiguity in segmentations. Consider the following example containing a Mongolian homonym:

(1) Гэрт эмээ хоол хийв
Гэр @@т эмээ хоол хийх @@в
Home.DAT grandma meal cook.PRS.PRF
'Grandma just cooked a meal at home.'

(2) Би өдөр эмээ уусан
Би өдөр эм @@ээ уух @@сан
I afternoon medicine.PSSD take.PST
'Afternoon I took my medicine.'

where "эмээ" is a homonym of two different words; in the first sentence, it is "grandmother", and in the second sentence — an inflected form of "medicine". Thus, the form in the second case can be segmented. However, the modern subword segmentation tools consider no contextual differences in word forms.

Key points of this subtask are:

- Morpheme segmentation is context-dependent;
- We organize it for three languages: English, Czech, and Mongolian;
- For Czech and Mongolian we asked native speakers to manually annotate the data. The details of data collection are provided in Section 3.

Category	Infl.	Deri.	Comp.	Description	English example (input ==> output)
000	-	-	-	Root words (free morphemes)	progress ==> progress
100	✓	-	-	Inflection only	prepared ==> prepare @@ed
010	-	✓	-	Derivation only	intensive ==> intense @@ive
001	-	-	✓	Compound only	hotpot ==> hot @@pot
101	✓	-	✓	Inflection and Compound	wheelbands ==> wheel @@band @@s
011	-	✓	✓	Derivation and Compound	tankbuster ==> tank @@bust @@er
110	✓	✓	-	Inflection and Derivation	urbanizes ==> urban @@ize @@s
111	✓	✓	✓	Inflection, Derivation, Compound	trackworkers ==> track @@work @@er @@s

Table 3: Morphological categories and descriptions of segmented words in subtask 1

Category	English	Spanish	Hungarian	French	Italian	Russian	Czech	Latin	Mongolian
000	101938	15843	6952	13619	21037	2921	-	50338	1604
100	126544	502229	410662	105192	253455	221760	-	831991	7266
010	203102	18449	24923	67983	41092	72970	-	0	2201
001	16990	248	3320	1684	431	259	-	0	5
101	13790	458	101189	478	317	1909	-	0	35
011	5381	82	1654	506	140	328	-	0	0
110	106570	346862	323119	126196	237104	481409	-	0	7855
111	3059	343	54279	186	158	2658	-	0	0
total words	577374	884514	926098	382797	553734	784214	38682	882329	18966

Table 4: Word statistics across morphological categories on subtask 1

Language	train	dev	test
Czech	1,000	500	500
English	11,007	1,783	1,845
Mongolian	1,000	500	600

Table 5: The number of samples in each language in Subtask 2.

2.3 Evaluation

In order to evaluate and compare the systems, we used four metrics: (i) ***precision***, the ratio of correctly predicted morphemes over all predicted morphemes; (ii) ***recall***, the ratio of correctly predicted morphemes over all gold-label morphemes; (iii) ***f-measure***, the harmonic mean of the precision and recall; (iv) ***edit distance*** - average Levenshtein distance between the predicted output and the gold instance. For convenience, we provided the python tool[2] to evaluate these metrics on both subtasks. In addition, for subtask 1 this tool also provided detailed results across the morphological categories.

3 Data

We collected our morphological data from various sources to account for all types of morphology: derivational, inflectional, compounding. We also collected base forms. For derivational and inflectional morphology, we have used the segmentation data from UniMorph 4.0 (Batsuren et al., 2022) and MorphyNet (Batsuren et al., 2021b). UniMorph contains inflectional paradigms collected from linguistic sources as well as Wiktionary, while MorphyNet represents derivations scraped from various editions of Wiktionary. Compounds and base forms were also extracted from Wiktionary (see Section 3.2 for more details on the data extraction). We then used the data to produce morpheme segmentations for seven high-resource languages. For Czech and Mongolian, as low-resource languages, we asked native speakers and linguists to develop the resources (Section 3.3 provides more details). For English sentence data, we have used the universal dependency treebank of English (Silveira et al., 2014).

3.1 Data Statistics

The data for the shared task was moderately multilingual, containing nine unique languages of five genera including Germanic, Italic, Slavic, Mongolic, and Uralic. In subtask 1, we have over 5 million samples of morpheme segmentations that cover nine languages over nine morphological categories, as shown in Table 4. In subtask 2, Table 5 displays the data statistics of three languages.

3.2 Extraction from Wiktionary

Language-specific editions of Wiktionary contain a considerably large amount of derivations and compounds.

Compound extraction rules were applied to the

[2] https://github.com/sigmorphon/2022SegmentationST/tree/main/evaluation

etymology sections of Wiktionary entries to collect the Morphology template usages, such as for the English *newspaper*:

Equivalent to **news** + **paper**.

where we have a morphology entry from the Wiktionary XML dump as follows:

{{compound | en | news | paper}}

Most of compound entries use "compound" etymology template while some cases use "affix" templates, e.g., *basketball* and *volleyball*.

Root (and base) word extraction is a two-step procedure. In the first step we collected words, inherited from earlier phases of corresponding languages. For example, English 'book' is traced back to the Middle English 'bok', according to the etymology section of Wiktionary. We extracted 279,173 words from 6 languages from CogNet, a cognate database containing 8.1 million cognate pairs of 335 languages from Wiktionary (Batsuren et al., 2019a, 2021a). In the second step, we filtered out 116,863 words from the earlier extracted derivational and compound data, resulting in 162,310 root words in 6 languages. Similar Wiktionary data extraction procedures have been applied to a wide range of linguistic data, e.g., etymology (Fourrier and Sagot, 2020), multilingual lexicons - DBnary (Sérasset, 2015) and Yawipa (Wu and Yarowsky, 2020).

3.3 Collecting data for Czech and Mongolian

We had two languages with limited amount of data, Czech and Mongolian. For each language, we used a different development methodology than for the other seven languages (with larger amount of available data).

Mongolian: we asked two linguists (who are also native speakers of Mongolian) to annotate morpheme segmentations of 3,810 words from Mongolian WordNet (Batsuren et al., 2019b). After manual annotation, we received 1,604 base forms, 2201 derived forms, and 5 compounds. To account for inflectional morphology, we have used the Mongolian transducer tool (Munkhjargal et al., 2016) to generate inflected forms of the 3,810 annotated words. In total, we collected morpheme segmentations of 18,966 Mongolian words for subtask 1. For subtask 2, the same two linguists annotated 2,100 Mongolian sentences.

Czech: we merged hand-segmented word forms from four sources for the purpose of subtask 1: (a) segmentations previously created within DeriNet (Vidra et al., 2019), a project aimed at capturing derivational relations in Czech (9,508 word forms), (b) segmentations of Czech verb lemmas imported from a partially digitized version of a printed dictionary (Slavíčková et al. 2017; 13,162 word forms in addition, i.e. not counting overlaps), (c) segmentations available in the MorfCzech dataset (Pelegrinová et al., 2021), mostly extracted from dictionaries and grammar books existing for Czech (additional 11,137 word forms), and (d) word forms that we annotated newly in order to reach complete coverage of Czech subtask 2 sentences (see below; additional 4,887 word forms). In total, the subtask 1 dataset contains 38,694 unique Czech word forms segmented to morphs.

All annotations were performed by native speakers with linguistic education, and underwent careful harmonization if the input resources disagreed, as well as numerous consistency checks. However, because of rich allomorphy in Czech, we have not been able to merge allomorph sets under more abstract umbrella morphemes so far, and thus words are represented as sequences of morphs (whose concatenation perfectly matches the original word forms), not of morphemes.

The Czech subtask 2 dataset contains in total 2,000 sentences from the Czech subset of Universal Dependencies (de Marneffe et al. 2021; more specifically, 1000, 500, and 500 first sentences from the train, dev, and test sections, respectively, of the Prague Dependency Treebank subset of UD 2.9). Given that homonymy resulting in different morph boundaries is extremely rare in Czech, words are segmented basically regardless of their contexts.

3.4 Data Splits

From each language's collection of morpheme segmentations in subtask 1, we sampled 80% for the training, 10% for development, and 10% for test sets.[3] All splits of subtask 1 are balanced w.r.t. the nine morphological categories, described in Table 3. While sampling the training and development sets for the subtask 1, we excluded words that were present in the test sentences of subtask 2. This was done in order to avoid situations when the subtask 1 data could directly influence the results of subtask 2 (since we allowed the multi-task learnings between both subtasks).

[3] All the data splits can be obtained from https://github.com/sigmorphon/2022SegmentationST/tree/main/data

Team	Description	System	System features				
			Neural	Ensemble	Data+	Multilingual	Multi-task
Baseline	(Schuster and Nakajima, 2012)	WordPiece*	-	-	-	-	-
	(Kudo, 2018)	ULM*	-	-	-	-	-
	(Virpioja et al., 2013)	Morfessor2*	-	-	-	-	-
AUUH	(Rouhe et al., 2022)	AUUH_A*	✓	-	✓	✓	✓
		AUUH_B*	✓	-	-	✓	✓
		AUUH_C	✓	-	✓	-	✓
		AUUH_D	✓	-	-	-	✓
		AUUH_E*	✓	-	✓	-	-
		AUUH_F*	✓	-	-	-	-
CLUZH	(Wehrli et al., 2022)	CLUZH	✓	✓	-	-	-
		CLUZH-1	✓	✓	-	-	-
		CLUZH-2	✓	✓	-	-	-
		CLUZH-3	✓	✓	-	-	-
DeepSPIN	(Peters and Martins, 2022)	DeepSPIN-1	✓	-	-	-	-
		DeepSPIN-2	✓	-	-	-	-
		DeepSPIN-3	✓	-	-	-	-
GU	(Levine, 2022)	GU-1	✓	-	✓	-	-
		GU-2	✓	-	✓	-	-
NUM DI	(Zundui and Avaajargal, 2022)	NUM DI	✓	-	-	-	-
JB132	(Bodnár, 2022)	JB132	-	-	-	-	-
Tü Seg	(Girrbach, 2022)	Tü_Seg-1	✓	-	-	-	-
		Tü_Seg-2	✓	-	-	-	✓

Table 6: The list of participating systems submitted to the shared task and baseline systems; Systems marked with * are submitted to both subtasks

4 Baseline Systems

The shared task provided predictions and results of baseline systems to participants that covered all languages and both subtasks. We chose three baseline systems: First is WordPiece, one of the state-of-the-art subword tokenization algorithms used in BERT (Devlin et al., 2019), which is based on Schuster and Nakajima (2012) and somewhat resembles BPE (Sennrich et al., 2016). Second is ULM (Unigram Language Model Kudo (2018)), another popular subword tokenization, used in XLNet (Yang et al., 2019). Third is Morfessor2, one of the state-of-the-art unsupervised morphological segmentations (Virpioja et al., 2013).

In future shared tasks, we aim to include more state-of-the-art tokenization tools including other Morfessor variants (Grönroos et al., 2014; Ataman et al., 2017; Grönroos et al., 2020), BPE-dropout (Provilkov et al., 2019), dynamic programming encoding (DPE) (He et al., 2020) or its variant (Hiraoka et al., 2021; Song et al., 2022), multi-view subword regularization (Wang et al., 2021), Charformer (Tay et al., 2021), space-treatment variants of BPE and ULM (Gow-Smith et al., 2022).

5 System Descriptions

The SIGMORPHON 2022 Shared Task on Morpheme Segmentation received submissions from 7 teams with members from 10 universities and institutes. Many teams submitted more than one system while some focused on a specific set of languages like Romance. In total, we had 24 unique systems over two subtasks, including the baseline system. More system details can be seen in Table 6.

AUUH Researchers at the Aalto University and the University of Helsinki produced six submission systems: two were transformer models and four were bidirectional GRU models created with several innovations of Morfessor feature enrichment, multi-task learning, and multilingual learning. Morfessor (Creutz and Lagus, 2002, 2007) is the famous language-independent unsupervised and semi-supervised segmentation tool and has a big family of Morfessor variants (Virpioja et al., 2013; Grönroos et al., 2014; Ataman et al., 2017; Grönroos et al., 2020). They have used the first variant of Morfessor (Creutz and Lagus, 2005) for enriching input words along with their Morfessor subword segmentations. AUUH_A, AUUH_C, AAUH_E systems used this Morfessor-based feature enrichment. The key innovation of AUUH

System	ces	eng	fra	ita	lat	rus	mon	hun	spa	macro avg.
WordPiece	20.42	23.06	12.66	9.08	8.84	13.81	14.58	24.00	16.57	15.89
ULM	23.71	32.32	16.08	10.65	10.42	15.67	25.82	31.27	19.58	20.61
Morfessor2	29.43	37.65	22.38	9.02	14.53	17.71	37.80	40.96	20.64	25.57
AUUH_A*	93.65	92.32	-	-	-	-	98.19	-	-	94.72
AUUH_B*	93.85	93.20	-	-	-	-	98.31	-	-	95.12
AUUH_E*	90.71	87.10	90.78	92.39	98.71	94.33	96.06	-	-	92.87
AUUH_F	90.28	86.40	90.81	92.56	98.85	93.68	95.32	98.34	97.25	**93.72**
CLUZH	93.81	92.70	94.80	96.93	99.37	98.62	98.12	98.54	98.74	**96.85**
DeepSPIN-1	93.42	92.29	91.66	96.01	99.37	98.75	98.03	98.56	98.79	96.32
DeepSPIN-2	**93.88**	93.39	95.29	**97.47**	99.36	99.30	98.00	98.68	99.02	97.15
DeepSPIN-3	93.84	**93.63**	**95.73**	97.43	**99.38**	**99.35**	**98.51**	**98.72**	**99.04**	**97.29**
GU-1*	-	-	83.44	88.69	-	-	-	-	-	86.07
GU-2*	-	-	83.38	87.49	-	-	-	-	95.95	88.94
JB132	64.65	65.43	46.20	33.44	91.39	50.55	57.82	72.64	43.39	**58.39**
NUM DI*	-	83.56	-	89.55	-	-	85.59	95.91	-	88.65
Tü_Seg-1	93.38	90.51	93.76	95.73	99.37	98.21	97.02	98.59	97.93	**96.06**

Table 7: Subtask 1 word-level results by system: The f-measure performance of systems by language; and macro average f-measure of all languages in the last column. Systems marked with * are partial submissions of a specific language set. The performances in bold are best performance of corresponding languages.

systems was multilingual and multi-task traning. They used a similar preprocessing technique (Johnson et al., 2017) to distinguish tasks and languages from one another, and then trained multilingual neural models which work on both subtasks. Their transformer-based multilingual and multi-task model, AUUH_B was the subtask 2 winning system (by its macro average f-measure) and also quite competitive with the subtask 1 winning systems on its partial three-language submissions.

CLUZH Researchers at the University of Zurich ensembled four submissions (Wehrli et al., 2022) by extending their previous neural hard-attention transducer models (Makarov and Clematide, 2018b,a, 2020). For subtask 1, they submit the following strong ensemble **CLUZH** composed of 3 models without encoder dropout and 2 models with encoder dropout of 0.15. In the sentence-level subtask 2, they submitted three ensembles, and treated this problem as the word-level problem by tokenizing sentences into words. They have also used POS tags as additional features to provide a light for the context of words. All individual models have an encoder dropout probability of 0.25 and vary only in their use of features: **CLUZH-1** with 3 models without POS features, **CLUZH-2** with 3 models with POS tag features, and **CLUZH-3** with combined all the models from CLUZH-1 and CLUZH-2. In overall, the **CLUZH-3** system was the subtask 2 winning system (by winning two out of three languages) and in subtask 1 **CLUZH** was the only system, outranked one (DeepSPIN-1) of three DeepSPIN systems.

DeepSPIN Researchers submitted three neural seq2seq models: (1) **DeepSPIN-1**, a character-level LSTM with soft attention (Bahdanau et al., 2014) with softmax trained with cross-entropy loss; (2) **DeepSPIN-2**, a character-level LSTM with soft attention in which softwax is replaced with its sparser version, 1.5-entmax (Peters and Martins, 2019); (3) **DeepSPIN-3**, a subword-level transformer (Vaswani et al., 2017) with the proposed 1.5-entmax, in which subword segments are modelled using ULM (Kudo, 2018). This design was one of most innovative architectures among all submitted systems. The authors previously experimented with the 1.5-entmax function on other tasks, demonstrating its utility, especially in the tasks with less uncertainty in the search space (e.g., compared to language modelling or machine translation) such as morphological and phonological modelling (Peters and Martins, 2020). The final results of this year's shared task confirm these observations: **DeepSPIN-2** and **DeepSPIN-3** achieve superior results and are the winner of the shared task.

GU One team from Georgetown University produced two submissions for three Romance languages of the word-level subtask, based on the GRU-based encoder-decoder model (Levine, 2022). In initial attempts, they tried to use additional features from the Wiktionary lists of prefixes and suf-

inf.	drv.	cmp.	eng	fra	ita	rus	mon	hun	spa	macro avg.
-	-	-	**83.80** CLUZH	84.08 DeepSPIN-3	82.69* DeepSPIN-3	82.56* DeepSPIN-1	93.37 JB132	**85.52** DeepSPIN-3	83.58 DeepSPIN-2	83.6 DeepSPIN-3
-	-	✓	93.23 AUUH_A	**81.80** CLUZH	**58.10*** CLUZH	**77.67** DeepSPIN-2	100.00 all systems	85.89 DeepSPIN-3	**57.89*** DeepSPIN-3	**78.60** DeepSPIN-3
-	✓	-	94.12 DeepSPIN-3	87.36* DeepSPIN-3	94.62 DeepSPIN-3	91.4 DeepSPIN-3	**92.41** DeepSPIN-3	94.96 DeepSPIN-3	92.47 DeepSPIN-3	92.48 DeepSPIN-3
✓	-	-	91.29* CLUZH	96.37 CLUZH	96.27 CLUZH	99.75 DeepSPIN-3	99.66 DeepSPIN-3	98.31 DeepSPIN-3	98.81 DeepSPIN-2	96.97 DeepSPIN-3
-	✓	✓	95.74 DeepSPIN-2	80.61 DeepSPIN-3	70.59* DeepSPIN-3	92.13 DeepSPIN-3	- -	89.82 DeepSPIN-3	97.3 DeepSPIN-3	87.65 DeepSPIN-3
✓	-	✓	96.89 DeepSPIN-3	96.60 DeepSPIN-2	94.97 DeepSPIN-3	100 DeepSPIN-3	100 all systems	98.71 DeepSPIN-3	96.15 DeepSPIN-1	97.45 DeepSPIN-3
✓	✓	-	97.54 DeepSPIN-3	99.03 DeepSPIN-3	99.23 DeepSPIN-3	99.97 DeepSPIN-3	99.74 DeepSPIN-3	99.41 DeepSPIN-2	99.75 DeepSPIN-3	99.24 DeepSPIN-3
✓	✓	✓	97.13 DeepSPIN-3	100 DeepSPIN-3	100 DeepSPIN-2	99.88 DeepSPIN-2	- -	99.28 DeepSPIN-2	97.04 DeepSPIN-2	98.23 DeepSPIN-2

Table 8: Subtask 1 word-level results by morphological category: f-measure performance of best performing system on a corresponding language and a category; Numbers in bold are worst performance of their corresponding language. Performances marked with * are worst performances of their morphological category.

fixes to train the model. However, such additional features decreased the main performances across morphological categories, so they excluded these features from the final submissions. Later on, they focus on data sharing between Romance languages. In French, the training data were augmented with four morphological category data from Italian and Spanish training and development datasets. These categories include non-inflection categories of `000`, `001`, `010`, `011`. With these experiments, they made minor improvements to these three languages. For these results, more research is needed to understand that transfer learning is useful.

NUM DI A single submission from the National University of Mongolia (Zundui and Avaajargal, 2022) is a transformer-based neural model. Their model architecture is simple as single-layered encoder-decoder classic architecture. All the hyperparameter settings are same as fairseq's standard tutorial tool. Their submission is also limited by four languages of subtask 1 due to human error.

JB132 The Charles University team (Bodnár, 2022) designed the Hidden Markov model, trained with the expectation-maximization algorithm. This model architecture has two sub-models. The first sub-model takes words as input and converts them into candidate morphemes. The second sub-model takes candidate morphemes and generates morphs as output. The first sub-model has three generators for accounting prefixes, root words, and suffixes. It is the only system not using neural methods among all submitted systems and the system's prediction is interpretable and can be useful for error analysis.

Tü Seg The University of Tübingen (Girrbach, 2022) team submitted two systems for each of subtasks. Both systems extend the sequence-labeling method proposed by (Hellwig and Nehrdich, 2018; Li and Girrbach, 2022). Their systems are very innovative and unique among all other neural models for considering the main segmentation task as a sequence-labeling task. All other neural systems used seq2seq architecture. Their neural model used a plain two-layer BiLSTM architecture. By its design, Tü Seg systems have at least two advantages over the main seq2seq alternative: (a) the number of parameters is much fewer, so the model can be trained fast and process quickly; (b) the system predictions are more interpretable compared to other neural systems and can help with the error analyses of high-resource datasets.

6 The System Results

All system results can be found and downloaded from the shared task GitHub page.[4]

6.1 Subtask 1 word-level results

Relative system performance of subtask 1 is provided in Table 7 which shows each system's f-measure by languages. The best performance of each language from submitted systems is in bold.

Two teams exploited external resources in some form: AUUH and GU. In general, any relative performance gained was minimal. AUUH submitted two systems that used additional resources, they received extra 1% compared to the team's other

[4] https://github.com/sigmorphon/2022SegmentationST/tree/main/results

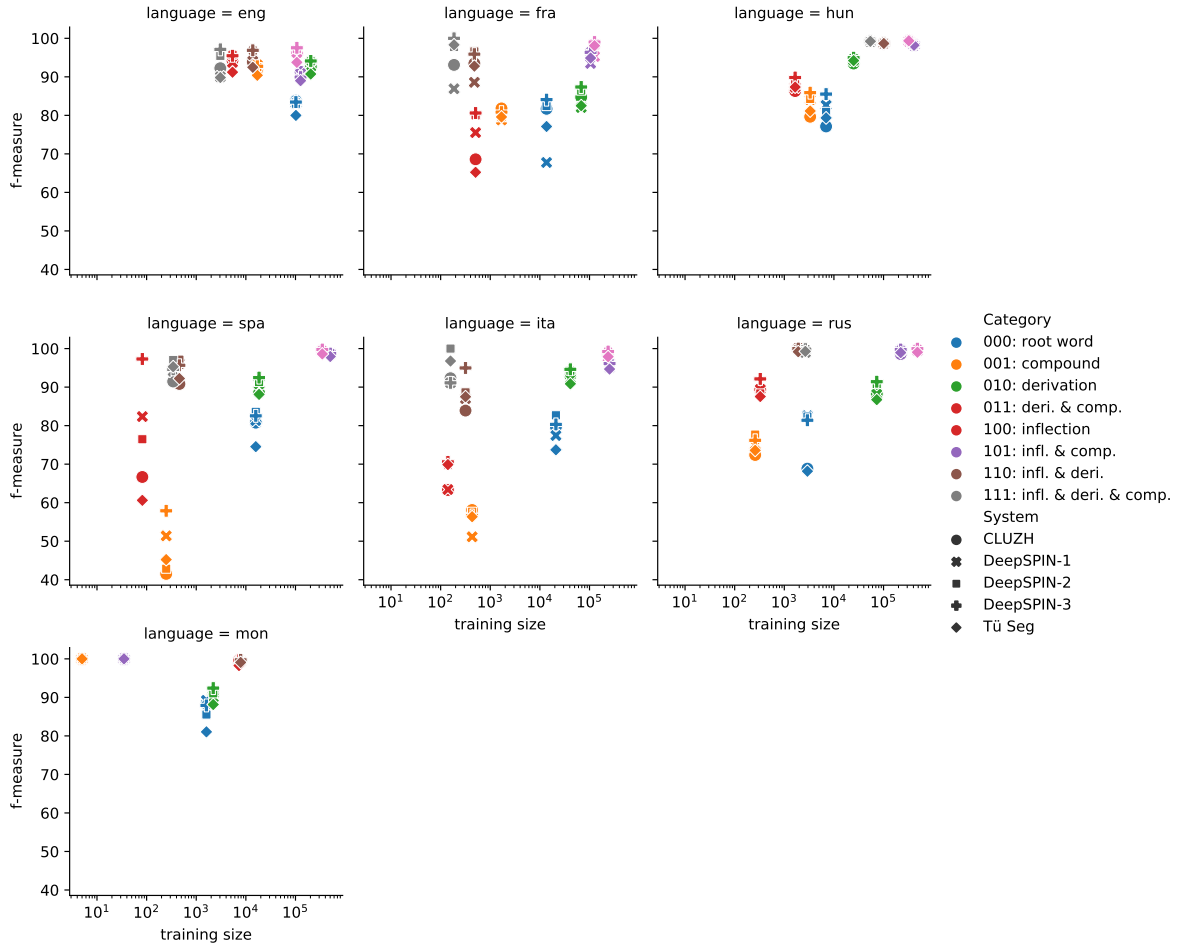

Figure 1: Impact of training sizes over languages and morphological categories: Results from top5-ranked systems of word-level subtask 1

systems. Similarly, GU and their submitted systems saw some minimal improvements over the performances. This details can be seen from their system description paper (Levine, 2022).

Only two of all the systems submitted to subtask 1 were multilingual and multi-task learning at same time. These two systems were proposed by AUUH team, but partial-language submissions were for English, Czech, and Mongolian. The important insight from this experiment is that the multi-task and multilingual learning approaches are quite beneficial for the task because their partial performances are quite competitive with the winning systems, DeepSPIN-3, DeepSPIN-2, and CLUZH.

Impact of training size: In subtask 1, the training datasets' sizes vary across languages and morphological categories. It might have impacted the top-ranked systems. Therefore, we plotted the top5-ranked systems over training size and f-measure performance across morphological categories, as shown in Figure 1. Here, in high-resource setting (as greater than 10^5) in all morphological categories, any of the top5-ranked systems always achieves 80% f-measure greater than 80%.

The root words are present in all types of resources settings from high to low. All the systems in this category of root words achieved no more than 85.5% f-measure except for Mongolian.

The two inflectional categories `100` and `110` are always in high-resource setting, having more than 10^6 training instances (except for two low-resource languages Czech and Mongolian). All systems achieved their best system performance over these two categories, compared to other categories.

Impact of word length: In many NLP tasks, the length of the input sequence is strongly correlated with the difficulty of their tasks (Yin et al., 2017; Wu et al., 2018). So, we present how the

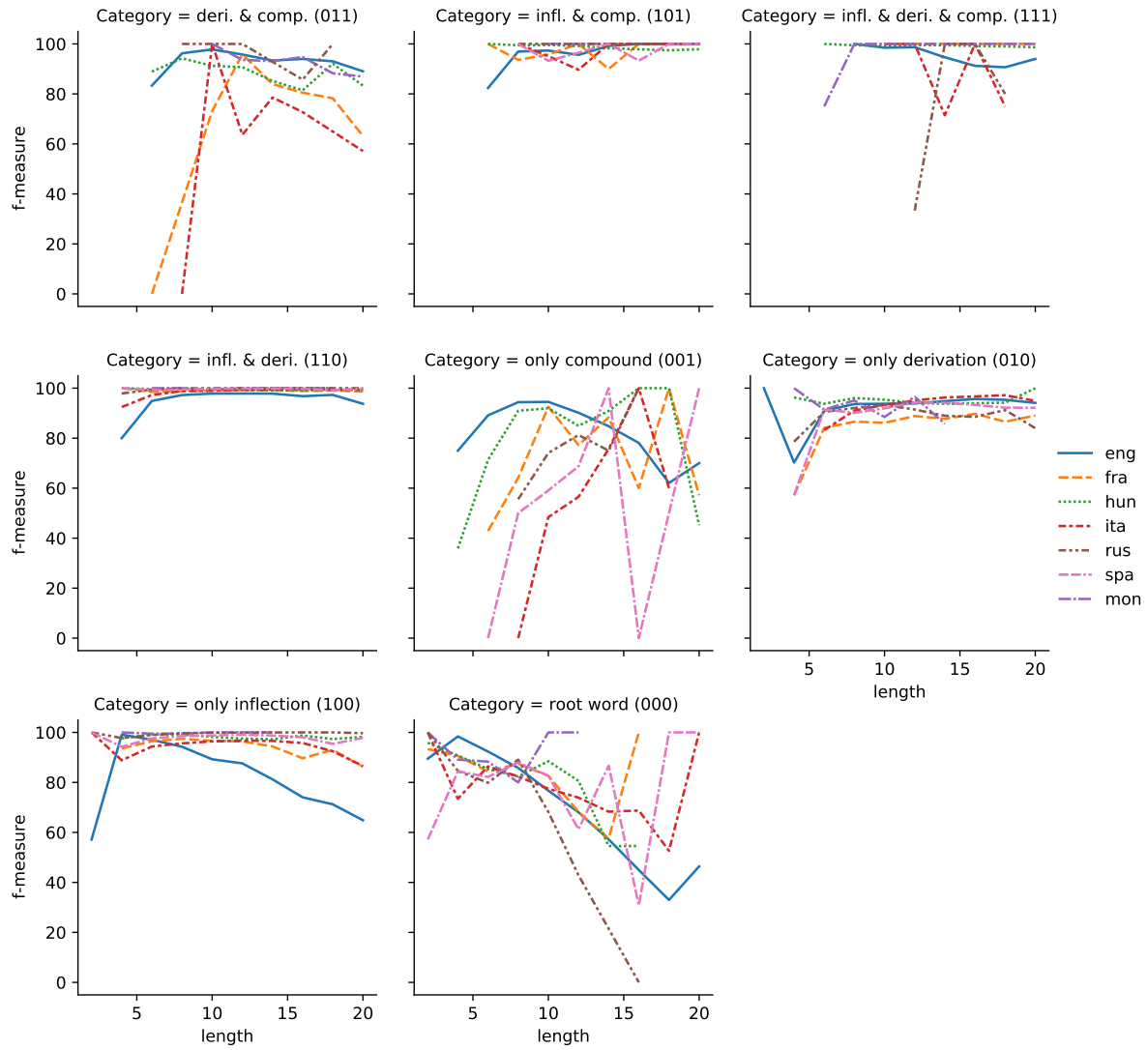

Figure 2: Impact of word length over languages and morphological categories: Results from DeepSPIN-3, the winning system of subtask 1, word-level morpheme segmentation

DeepSPIN-3's (subtask 1 winning system) performance relates to the word length across languages and morphological categories. Figure 2 shows various related facts: (i) for root words `000`, overall performance decreases across languages with increasing word length; (ii) inflectional morphology is systematically far more productive than other morphological categories, so this fact is reproduced here: the main inflectional category `100` has consistently high performance across languages and word lengths.

Difficulty of morphological categories: Even though the top-ranking systems perform very well on their own, other systems may have some complementary information across morphological categories. Therefore, we listed the best-performing systems for combinations of each language and each morphological category in Table 8. In the table, the lowest scores in corresponding languages are provided in bold. For instance, English root words (83.80 f-measure) are much harder to predict than other morphological categories in English. The hardest morphological categories are roots `000`, compounds `001`, and derivation and compound words `011`. The winning system, DeepSPIN-3 (marked with + in Figure 1), is consistently winning in these three categories across languages. Another observation from Figure 2 is that compound and root words are getting harder to predict across languages with the increase of word length. Also, identifying inflections from short

System	Czech				English				Mongolian				Macro avg.	
	P	R	F_1	Lev.	P	R	F_1	Lev.	P	R	F_1	Lev.	F_1	Lev.
WordPiece	38.47	31.45	34.61	17.88	62.02	65.13	63.53	5.54	19.82	29.20	23.62	29.19	40.59	17.54
ULM	41.98	30.39	35.26	16.39	62.32	69.24	65.60	5.68	38.79	35.58	37.12	20.76	45.99	14.28
Morfessor2	49.89	36.95	42.45	13.09	54.61	69.75	61.25	6.00	50.88	45.91	48.26	17.16	50.65	12.08
AUUH_A	89.70	87.53	88.60	4.97	96.66	95.78	96.22	1.86	83.49	80.94	82.19	5.42	89.00	4.08
AUUH_B	91.89	89.00	90.42	3.96	**96.82**	**95.79**	**96.31**	**1.39**	83.74	81.46	82.59	5.16	**89.77**	**3.50**
AUUH_C	50.60	69.19	58.45	71.37	84.77	71.67	77.67	19.13	79.07	73.45	76.15	17.33	70.76	35.94
AUUH_D	45.07	67.82	54.15	80.67	93.29	83.41	88.07	10.58	77.99	74.15	76.02	17.88	72.75	36.38
AUUH_E	57.39	67.22	61.92	55.92	95.23	76.82	85.04	12.36	73.34	72.01	72.67	24.88	73.21	31.05
AUUH_F	62.36	43.82	51.47	61.84	91.50	74.84	82.34	13.30	75.50	59.22	66.38	33.91	66.73	36.35
CLUZH-1	92.03	90.69	91.35	1.93	89.74	89.20	89.47	9.86	82.98	81.48	82.22	5.28	87.68	5.69
CLUZH-2	92.41	91.13	91.76	1.87	89.71	89.22	89.47	9.79	83.29	81.83	82.55	5.19	87.93	5.62
CLUZH-3	**92.63**	**91.35**	**91.99**	**1.80**	89.83	89.25	89.54	9.84	**83.71**	**82.07**	**82.88**	**5.10**	88.14	5.58
Tü_Seg-2	89.52	88.42	88.97	2.50	87.83	89.58	88.69	1.78	69.59	67.55	68.55	9.85	82.07	4.71

Table 9: Subtask 2 sentence-level results: F-measure across 3 languages

words (word length < 5) is one of the unsolved challenges in all languages (except for English), as shown in Figure 2.

6.2 Subtask 2 sentence-level results

Relative system performance is described in Table 9, showing all four evaluation metrics by each combination of system and language. In the sentence-level subtask 2, we have two winners: CLUZH-3 (won two out of three languages) and AUUH_B (F1 89.77 as maximum macro- average among submissions).

The performance of systems in the sentence-level subtask significantly decreased by 15% in Mongolian compared to the results of the word-level subtask. One reason is that all submitted systems treated this problem as a zero-shot solution of word-level subtask 1, and mostly ignored its context by their design.

7 Future Directions

The submitted systems achieved unexpectedly high accuracy across nine languages. This result suggests that the neural systems may have more capabilities beyond segmenting morphemes. For the next year, we plan to modify the task design and enrich the dataset with more fine-grained analysis. For example, *truckdrivers* → *truck @@drive @@er @@s* → *truck $$drive @@er ##s* where $$ is compound, @@ is derivation, and ## is inflection. In another direction, we will explore possibilities of adapting other morphological resources including word-formation resources (Zeller et al., 2013; Talamo et al., 2016; Vidra et al., 2019; Vodolazsky, 2020) or segmentation resources, UniSegments (Žabokrtský et al., 2022; Žabokrtský et al., 2022). Our shared task team welcomes continued contributions from the community.

8 Conclusion

The SIGMORPHON 2022 Shared Task on Morpheme Segmentation significantly expanded the problem of morphological segmentation, making it more linguistically plausible. In this task, seven teams submitted 23 systems for two subtasks in total of nine languages, achieving at minimum F1 30.71 improvement over the three baselines of the state-of-the-art subword tokenization and morphological segmentation tools, being used to train large language models, e.g., XLNet (Yang et al., 2019). The results suggest many directions for improving morpheme segmentation shared task.

Acknowledgements

We thank Garrett Nicolai and Eleanor Chodroff for their advice and support. The authors also thank Ben Peters and Simon Clematide for their invaluable contributions and advice, including developing the evaluation tool and early detection of data errors.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2019a. Cognet: A large-scale cognate database. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3136–3145.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021a. A large and evolving cognate database. *Language Resources and Evaluation*.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2021b. 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.

Khuyagbaatar Batsuren, Amarsanaa Ganbold, Altangerel Chagnaa, and Fausto Giunchiglia. 2019b. Building the Mongolian WordNet. In *Proceedings of the 10th Global Wordnet Conference*, pages 238–244, Wroclaw, Poland. Global Wordnet Association.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina J. Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. Unimorph 4.0: Universal morphology.

Jan Bodnár. 2022. Jb132 submission to the sigmorphon 2022 shared task 3 on morphological segmentation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016. A joint model of orthography and morphological segmentation. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning*, pages 21–30.

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

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Miguel Domingo, Mercedes Garcıa-Martınez, Alexandre Helle, Francisco Casacuberta, and Manuel Herranz. 2018. How much does tokenization affect neural machine translation? *arXiv preprint arXiv:1812.08621*.

Clémentine Fourrier and Benoît Sagot. 2020. Methodological aspects of developing and managing an etymological lexical resource: Introducing etymdb 2.0. In *LREC 2020-12th Language Resources and Evaluation Conference*.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Leander Girrbach. 2022. Sigmorphon 2022 shared task on morpheme segmentation submission description: Sequence labelling for word-level morpheme segmentation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Edward Gow-Smith, Harish Tayyar Madabushi, Carolina Scarton, and Aline Villavicencio. 2022. Improving tokenisation by alternative treatment of spaces. *arXiv preprint arXiv:2204.04058*.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2020. Morfessor em+ prune: Improved subword segmentation with expectation maximization and pruning. *arXiv preprint arXiv:2003.03131*.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor flatcat: An hmm-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185.

Xuanli He, Gholamreza Haffari, and Mohammad Norouzi. 2020. Dynamic programming encoding for subword segmentation in neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3042–3051.

Oliver Hellwig and Sebastian Nehrdich. 2018. Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2754–2763.

Tatsuya Hiraoka, Sho Takase, Kei Uchiumi, Atsushi Keyaki, and Naoaki Okazaki. 2021. Joint optimization of tokenization and downstream model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 244–255.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2016. Neural morphological analysis: Encoding-decoding canonical segments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 961–967, Austin, Texas. Association for Computational Linguistics.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Mikko Kurimo, Mathias Creutz, and Matti Varjokallio. 2007. Unsupervised morpheme analysis evaluation by a comparison to a linguistic gold standard-morpho challenge 2007. In *CLEF (Working Notes)*.

Mikko Kurimo, Ville Turunen, and Matti Varjokallio. 2008. Overview of morpho challenge 2008. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 951–966. Springer.

Mikko Kurimo, Sami Virpioja, Ville Turunen, and Krista Lagus. 2010a. Morpho challenge 2005-2010: Evaluations and results. In *Proceedings of the 11th Meeting of the ACL Special Interest Group on Computational Morphology and Phonology*, pages 87–95.

Mikko Kurimo, Sami Virpioja, Ville T Turunen, Graeme W Blackwood, and William Byrne. 2009. Overview and results of morpho challenge 2009. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pages 578–597. Springer.

Mikko Kurimo, Sami Virpioja, Ville T Turunen, et al. 2010b. Proceedings of the morpho challenge 2010 workshop. In *Morpho Challenge Workshop; 2010; Espoo*. Aalto University School of Science and Technology.

Lauren Levine. 2022. Sharing data by language family: Data augmentation for romance language morpheme segmentation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Jingwen Li and Leander Girrbach. 2022. Word segmentation and morphological parsing for sanskrit. *arXiv preprint arXiv:2201.12833*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Dominik Macháček, Jonáš Vidra, and Ondřej Bojar. 2018. Morphological and language-agnostic word segmentation for nmt. In *International Conference on Text, Speech, and Dialogue*, pages 277–284. Springer.

Peter Makarov and Simon Clematide. 2018a. Imitation learning for neural morphological string transduction. *arXiv preprint arXiv:1808.10701*.

Peter Makarov and Simon Clematide. 2018b. Uzh at conll-sigmorphon 2018 shared task on universal morphological reinflection. Association for Computational Linguistics.

Peter Makarov and Simon Clematide. 2020. CLUZH at SIGMORPHON 2020 shared task on multilingual grapheme-to-phoneme conversion. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 171–176, Online. Association for Computational Linguistics.

Austin Matthews, Graham Neubig, and Chris Dyer. 2018. Using morphological knowledge in open-vocabulary neural language models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1435–1445.

Sabrina J Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y Lee, Benoît Sagot, et al. 2021. Between words and characters: A brief history of open-vocabulary modeling and tokenization in nlp. *arXiv preprint arXiv:2112.10508*.

Zoljargal Munkhjargal, Altangerel Chagnaa, and Purev Jaimai. 2016. Morphological transducer for mongolian. In *International Conference on Computational Collective Intelligence*, pages 546–554. Springer.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. Kinyabert: a morphology-aware kinyarwanda language model. *arXiv preprint arXiv:2203.08459*.

Kateřina Pelegrinová, Viktor Elšík, Radek Čech, and Ján Mačutek. 2021. MorfoCzech. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Ben Peters and André F. T. Martins. 2020. One-size-fits-all multilingual models. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 63–69, Online. Association for Computational Linguistics.

Ben Peters and André F. T. Martins. 2022. Beyond characters: Subword-level morpheme segmentation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Ben Peters and André FT Martins. 2019. IT–IST at the sigmorphon 2019 shared task: Sparse two-headed models for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 50–56.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2019. Bpe-dropout: Simple and effective subword regularization. *arXiv preprint arXiv:1910.13267*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Aku Rouhe, Stig-Arne Grönroos, Sami Virpioja, Mathias Creutz, and Mikko Kurimo. 2022. Morfessor-enriched features and multilingual training for canonical morphological segmentation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Jonne Saleva and Constantine Lignos. 2021. The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 164–174, Online. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Lane Schwartz, Francis Tyers, Lori Levin, Christo Kirov, Patrick Littell, Chi-kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, et al. 2020. Neural polysynthetic language modelling. *arXiv preprint arXiv:2005.05477*.

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.

Gilles Sérasset. 2015. Dbnary: Wiktionary as a lemon-based multilingual lexical resource in rdf. *Semantic Web*, 6(4):355–361.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Eleonora Slavíčková, Jaroslava Hlaváčová, and Patrice Pognan. 2017. Retrograde morphemic dictionary of czech - verbs. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Haiyue Song, Raj Dabre, Chenhui Chu, Sadao Kurohashi, and Eiichiro Sumita. 2022. Self-supervised dynamic programming encoding for neural machine translation.

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

Yi Tay, Vinh Q Tran, Sebastian Ruder, Jai Gupta, Hyung Won Chung, Dara Bahri, Zhen Qin, Simon Baumgartner, Cong Yu, and Donald Metzler. 2021. Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jonáš Vidra, Zdeněk Žabokrtský, Magda Ševčíková, and Lukáš Kyjánek. 2019. Derinet 2.0: Towards an all-in-one word-formation resource. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*, pages 81–89, Praha, Czechia. ÚFAL MFF UK.

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, Mikko Kurimo, et al. 2013. Morfessor 2.0: Python implementation and extensions for morfessor baseline.

Daniil Vodolazsky. 2020. Derivbase. ru: A derivational morphology resource for russian. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 3937–3943.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2021. Multi-view subword regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482.

Silvan Wehrli, Simon Clematide, and Peter Makarov. 2022. Cluzh at sigmorphon 2022 shared tasks on morpheme segmentation and inflection generation. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

Wei Wu, Houfeng Wang, Tianyu Liu, and Shuming Ma. 2018. Phrase-level self-attention networks for universal sentence encoding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3729–3738.

Winston Wu and David Yarowsky. 2020. Computational etymology and word emergence. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3252–3259, Marseille, France. European Language Resources Association.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Wenpeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. 2017. Comparative study of cnn and rnn for natural language processing. *arXiv preprint arXiv:1702.01923*.

Zdeněk Žabokrtský, Nyati Bafna, Jan Bodnár, Lukáš Kyjánek, Emil Svoboda, Magda Ševčíková, Jonáš Vidra, Sachi Angle, Ebrahim Ansari, Timofey Arkhangelskiy, Khuyagbaatar Batsuren, Gábor Bella, Pier Marco Bertinetto, Olivier Bonami, Chiara Celata, Michael Daniel, Alexei Fedorenko, Matea Filko, Fausto Giunchiglia, Hamid Haghdoost, Nabil Hathout, Irina Khomchenkova, Victoria Khurshudyan, Dmitri Levonian, Eleonora Litta, Maria Medvedeva, S. N. Muralikrishna, Fiammetta Namer, Mahshid Nikravesh, Sebastian Padó, Marco Passarotti, Vladimir Plungian, Alexey Polyakov, Mihail Potapov, Mishra Pruthwik, Ashwath Rao B, Sergei Rubakov, Husain Samar, Dipti Misra Sharma, Jan Šnajder, Krešimir Šojat, Vanja Štefanec, Luigi Talamo, Delphine Tribout, Daniil Vodolazsky, Arseniy Vydrin, Aigul Zakirova, and Britta Zeller. 2022. Universal segmentations 1.0 (UniSegments 1.0). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Britta Zeller, Jan Šnajder, and Sebastian Padó. 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: Long Papers)*, pages 1201–1211.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajic, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, et al. 2017. Conll 2017 shared task: Multilingual parsing from raw text to universal dependencies. In *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19. Association for Computational Linguistics.

Tsolmon Zundui and Chinbat Avaajargal. 2022. Wordlive morpheme segmentation using transformer neural network. In *19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*.

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 13th International Conference on Language Resources and Evaluation (LREC 2018)*, Marseille, France. European Language Resources Association (ELRA).